

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Spanish and German Legal Translation Performance of Large Language Models
and Neural Machine Translation

verfasst von | submitted by

Lorenz Albert Alex Hochmair BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of
Master of Arts (MA)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt |
Degree programme code as it appears on the
student record sheet:

UA 070 331 351

Studienrichtung lt. Studienblatt | Degree pro-
gramme as it appears on the student record
sheet:

Masterstudium Translation Deutsch Spanisch

Betreut von | Supervisor:

Assoz. Prof. Mag. Dr. Dagmar Gromann BSc

Danksagung

Mein besonderer Dank gilt Assoz.-Prof. Mag. Dr. Dagmar Gromann, BSc, die mich zu diesem Thema inspiriert und bei der Ausarbeitung meiner Forschungsfrage sowie durch wertvolles Feedback während des gesamten Entstehungsprozesses dieser Arbeit unterstützt hat.

Ebenso möchte ich mich ganz herzlich bei Mag. Dr. Edelmann bedanken, der mir nicht nur die spanischen Rechtstextpassagen zur Verfügung stellte, sondern auch die für diese Arbeit benötigte Humanübersetzung anfertigte. Diese Unterstützung war für die Durchführung dieser Erhebung von großem Wert.

Ich danke auch allen teilnehmenden Übersetzern für ihre Zeit und Expertise, die sie für diese Erhebung mitgebracht haben.

Von Herzen danke ich auch all meinen Freunden, insbesondere Nina, Stefan und Jana, die mir in den unzähligen schwierigen Phasen während des Schreibprozesses für diese Masterarbeit mit viel Geduld immer beigestanden haben und mich in so vielen Momenten wieder aufgebaut und ermutigt haben, weiterzumachen und nicht aufzugeben.

Ich bedanke mich bei meiner ganzen Familie für den bedingungslosen Rückhalt und das Gefühl der Geborgenheit während dieser gesamten Zeit. Insbesondere möchte ich meinem Vater, Alex Villazón, danken, der mir trotz der großen räumlichen Distanz zwischen Österreich und Bolivien stets mit Rat und moralischer Unterstützung zur Seite stand und etliche Sitzungen und Stunden per Videochat mit mir verbracht hat, um meine Gedanken und Herangehensweise für diese Masterarbeit zu strukturieren. ¡Gracias, papá!

Abstract (English)

The translation profession has been marked by both limited societal visibility and limited acknowledgment in the labor market. With recent advances in artificial intelligence, the same question about the future role of translators has re-emerged. Against this background, there is a constant demand for quality assessment that compares human and machine translation, to make each strength and weakness more transparent. Building upon this demand, the present master's thesis investigates this question through a systematic quality assessment comparing human and machine translation in the legal translation domain. More specifically, five professional translators assessed and rated translations of Spanish criminal law excerpts into German, including those provided and translated by a legal expert, those generated by DeepL, and two translations generated by ChatGPT (GPT-5.1) using a naïve and a detailed prompt. The evaluation consisted of a two-part structure. The first part was a comparative quality assessment with Best-Worst Scaling, based on the methodology of Hiebl and Gromann (2024), following a criteria-based evaluation, in which the chosen best and worst translations were further assessed based on content, terminology, text function, and grammar and syntax. Based on the sum and average ratings, the survey revealed that human translation was rated best overall, while DeepL was rated worst and ranked lower than ChatGPT's translation with a naïve prompt; the long-prompt GPT translation surpassed its naïve-prompt translation. The results of the quality criteria-based evaluations put these rankings further into perspective: neither the human translation nor the machine translations ranked first in all four quality categories. Human translation surpassed all machine translations, prevailing in linguistic accuracy, more precisely in terminology, grammar, and syntax, while the long-prompt GPT translation was leading in the categories content and text function. Due to the low number of participants, the results shall be interpreted only as a possible trend and are to be understood as exploratory. Future research could aim for a larger sample size and improve the survey design, for instance by reducing cognitive load and conducting the survey in person rather than online. Comparative quality assessment using Best-Worst Scaling was initially intended as a time-efficient, cognitively undemanding method for translation assessment. The added criteria-based evaluation, however, significantly increased the survey's total effort. Therefore, future research needs to weigh up efficiency and the level of detail in the assessment.

Abstract (Deutsch)

Der Übersetzungsberuf war seit jeher sowohl von begrenzter gesellschaftlicher Sichtbarkeit als auch von mangelnder Anerkennung am Arbeitsmarkt geprägt. Durch die neuesten Fortschritte im Bereich der künstlichen Intelligenz stellt sich erneut die Frage nach der zukünftigen Relevanz des Berufsbildes von Übersetzer*innen. Vor diesem Hintergrund besteht ein stetiger Bedarf an Qualitätsbewertungen, die menschliche und maschinelle Übersetzung miteinander vergleichen, um die jeweiligen Stärken und Schwächen sichtbar zu machen. Die vorliegende Masterarbeit untersucht diese Frage anhand einer systematischen Qualitätsbewertung im Bereich der Rechtsübersetzung und vergleicht menschliche und maschinelle Übersetzung. Konkret bewerteten fünf professionelle Übersetzer*innen Übersetzungen spanischer Strafrechtstexte ins Deutsche. Die menschliche Übersetzung wurde von einem Rechtsexperten angefertigt, welcher auch die spanischen Rechtstextpassagen bereitstellte, während die maschinellen Übersetzungen von DeepL sowie von ChatGPT (GPT-5.1) mithilfe einer naiven Prompt-Strategie und eines ausführlichen Prompts generiert wurden. Die Bewertung erfolgte zweiteilig. Zuerst wurde eine vergleichende Qualitätsbewertung mittels Best-Worst Scaling, basierend auf der Methodik von Hiebl und Gromann (2024), durchgeführt, gefolgt von einer kriterienbasierten Bewertung, bei der die jeweils ausgewählten besten und schlechtesten Übersetzungen zusätzlich hinsichtlich Inhalt, Terminologie, Textfunktion sowie Grammatik und Syntax beurteilt wurden. Auf Basis der Summen- und Durchschnittswerte zeigte die Umfrage, dass die Humanübersetzung insgesamt am besten bewertet wurde, während DeepL am schlechtesten abscheidet und zudem niedriger eingestuft wurde als die ChatGPT-Übersetzung mit naivem Prompt; die Übersetzung mit dem ausführlichen Prompt übertraf wiederum jene mit dem naiven Prompt. Die Ergebnisse der kategoriebasierten Bewertung relativierten dieses Ranking jedoch. Weder die Humanübersetzung noch eine der maschinellen Übersetzungen belegten in allen vier Qualitätskategorien den ersten Platz. Die Humanübersetzung übertraf alle maschinellen Übersetzungen in sprachlicher Genauigkeit, insbesondere in Terminologie, Grammatik und Syntax, während die Übersetzung mit dem ausführlichen GPT-Prompt in den Kategorien Inhalt und Textfunktion führend war. Aufgrund der geringen Teilnehmerzahl sind die Ergebnisse lediglich als möglicher Trend zu interpretieren und als explorativ zu verstehen. Zukünftige Forschung sollte eine größere Stichprobe anstreben und das Design der Erhebung verbessern, etwas durch Reduzieren der kognitiven Belastung und die Durchführung der Befragung vor Ort statt online. Die vergleichende

Qualitätsbewertung mittels Best-Worst Scaling war ursprünglich als zeiteffiziente, kognitiv wenig belastende Methode zur Übersetzungsbewertung konzipiert. Die zusätzliche kategoriebasierte Bewertung erhöhte jedoch den Gesamtaufwand der Umfrage deutlich. Zukünftige Forschung muss daher Effizienz und Detailtiefe der Bewertung gegeneinander abwägen.

Table of Contents

1. Introduction	11
1.1. Research Questions and Assumptions	13
1.2. Motivation and Objective	15
2. Machine Translation.....	16
2.1. Historical Development of Machine Translation.....	16
2.2. Rule-based Machine Translation	19
2.2.1. Direct Machine Translation	20
2.2.2. Transfer-based Machine Translation	21
2.2.3. Interlingual Machine Translation	21
2.3. Statistical Machine Translation.....	22
2.4. Neural Machine Translation	24
2.5. Examples of Online Machine Translation Systems	26
2.5.1. Systran	26
2.5.2. Google Translate.....	27
2.5.3. DeepL	28
2.6. Machine Translation with Large Language Models	29
2.6.1. ChatGPT	30
2.7. Prompt Engineering	31
2.7.1. Prompt Engineering for Translation	33
3. Legal Language and Legal Translation	35
3.1. Characteristics of Legal Language.....	35
3.1.1. Characteristics of Spanish Legal Language.....	38
3.1.2. Characteristics of German Legal Language	41
3.2. Legal Translation	44
3.2.1. Machine Translation in the Legal Domain	46
4. Translation Quality.....	48
4.1. Perspective of Translation Studies.....	48
4.1.1. Equivalence.....	49
4.1.2. Adequacy and Fluency	51
4.1.3. Skopos-Theory.....	52
4.1.4. Translation Quality Assessment	53
4.1.5. Error Annotation and MQM	54
4.2. Perspective of Machine Translation.....	55

4.2.1. Reference-based Evaluation Metrics	56
4.2.2. Neural Language Model Based Metrics	60
4.2.3. Comparative Translation Quality Assessment with Best-Worst-Scaling	63
5. Related Work	65
6. Methodology	69
6.1. Text Selection	69
6.2. Translation Process for Source Text Samples	72
6.2.1. Translation process with DeepL	73
6.2.2. Translation Process with GPT-5.1	73
6.3. Participant Selection	75
6.4. Survey Structure and Design	76
6.4.1. Professional and Demographic Background	76
6.4.2. Best-Worst Scaling with Criteria-based Assessment	77
6.4.3. Post-Task Questionnaire	83
6.4.4. Methodological Evaluation and Closing	84
6.5. Data Analysis	84
7. Results	87
7.1. Participants	87
7.2. Best-Worst Scaling Ratings	88
7.3. Criteria-based Assessment	90
7.4. Set-specific Participant Comments	92
8. Discussion	94
9. Conclusion	100
10. Bibliography	102
Appendix 1	115
Appendix 2	115
Appendix 3	118
Appendix 4	121
Appendix 5	122
Appendix 6	123
Appendix 7	136

List of Figures

Figure 1: Vauquois triangle as depicted in Wang and Sawyer (2023: 77; based on Vauquois 1968)	20
Figure 2: Example screenshot of the first BWS task, showing the Spanish source text and the four German translations, with radio buttons to select the best and the worst translations	77
Figure 3: Example screenshot of the second BWS task for assigning points to the best and worst translations.....	79
Figure 4: Example screenshot of criteria-based task. Personal quality evaluation	80
Figure 5: Example screenshot of text field for adding comments.....	83

List of Tables

Table 1: Count of words (W), count of characters (C), source text (ST) in Spanish (ES).....	71
Table 2: Translation order within each set (HT = Human Translation; DL = DeepL; GPT-N = ChatGPT with a naïve prompt; GPT-N = ChatGPT with an extensive prompt).....	78
Table 3: BWS result overviews: average, sum, and total results	88
Table 4: BWS rating details across participants, strategy, and the selected best and worst translation modes (HT = Human Translation; DL = DeepL; GPT-N = ChatGPT with a naïve prompt; GPT-N = ChatGPT with an extensive prompt).....	89
Table 5: Aggregated sum and normalized average scores for quality categories	90
Table 6: Global overview of average points obtained through Best-Worst Scaling (BWS) and Criteria-based (CB) assessment, including the total number of ratings per translation mode	92

1. Introduction

The translation profession has long occupied an uncertain position between limited public visibility and the ongoing struggle for professional recognition in the labor market (Dam & Korning Zethsen 2008; Katan 2009). The invention of Machine Translation (MT) systems and their advancements towards Neural Machine Translation (NMT) have already cast doubt on the relevance of the profession. Free online machine translation systems such as DeepL, which can quickly produce translations of increasingly high quality, have intensified concerns about displacement and loss of professional standing (Pym & Torres-Simón 2021).

With the revolutionary technology breakthrough of Large Language Models (LLMs) and generative AI (GenAI), this relevance has had to be questioned again across other professional sectors as well as the language industry (Meer 2024; Yu et al. 2023). Unlike conventional translation programs, LLMs and GenAI allow users to refine translation output through prompt engineering (He 2024), enabling them to specify relevant parameters such as target audience, target culture, text function, text conventions, and language register, among others. These capabilities could lead to the assumption that LLMs can produce translations that meet professional standards without human intervention.

The translation of legal texts is a specialized area within translation, governed by strict norms established by the International Organization for Standardization (ISO), because small errors in translation can have significant economic or political consequences, endangering human rights or threatening human lives (Bajčić & Golenko 2024; Biel et al. 2024). For that reason, not every translator is allowed to work in this sector of the language industry, as it requires a certificate, e.g., for certified translations (Biel et al. 2024). This, again, raises the question of whether MT and LLMs could replace the field of legal translation through automation processes. A complete replacement, however, does not appear feasible at the time of writing, as certified translation requires that the authenticity and accuracy of the translated document be confirmed by a human expert, as prescribed by law (Ordóñez-López 2021). Moreover, legal translation is subject to strict data protection regulations, which raise significant concerns regarding the input of personal and confidential information into translation software and LLMs, as online services may expose sensitive data through third-party storage and transmission vulnerabilities. For this reason, the use of such tools in legal contexts

requires careful consideration and human oversight (Feretakis et al. 2025; Matvienko & Khomenko 2025).

Nevertheless, the translation capabilities of MT and LLMs remain an active area of research across multiple disciplines (Chatzikoumi 2019). Translation studies, (computer) linguistics, and informatics measure the quality of translations through translation quality assessment (TQA). Originally, TQA was performed only by human annotators, though the challenging aspects of this process are its time-consuming, expensive nature, the subjectivity of human quality perception, and the cognitive load. For this reason, there is a growing effort to use automated evaluation methods (AEMs) to address this issue and obtain evaluation data more efficiently and objectively (Gladkoff et al. 2022). However, the paradigm that humans are ultimately responsible for judging translation quality still holds (Chatzikoumi 2019).

In one study, Hiebl and Gromann (2024) developed a method that uses Best-Worst Scaling (BWS), a survey method derived from psychology, to determine which translation an annotator rates as best and which as worst, and to assign points accordingly. The goal of this method is to reduce the cognitive load on the evaluators and minimize the time required, ultimately enabling them to distinguish more efficiently between higher- and lower-quality translations. This master's thesis adopts and extends that methodology, complementing the holistic ranking approach with a criteria-based quality assessment as a second evaluation dimension. Unlike the original study, which used students from translation studies as evaluators and texts from various domains, this master's thesis focuses on the legal field and enlists professional legal translators as evaluators, with an emphasis on the Spanish-German language pair. Additionally, rather than comparing only human translations with NMT, this master's thesis also includes translations from LLMs for assessment.

This master's thesis is divided into three main theoretical chapters. The first chapter deals with machine translation, from its beginnings to the current state of the art. It also covers LLM translation capabilities and prompting characteristics. The second chapter addresses legal language and legal translation. It examines the general properties of legal language, the characteristics of Spanish and German legal language, and the challenges of machine translation in legal translation. The final theoretical chapter focuses on TQA. It assesses perspectives from the field of translation studies, including the concepts of equivalence, Skopos theory, and MQM, as well as the perspective of machine translation, with a focus on reference-based and neural-language-model-based evaluation metrics. Finally, the method relevant to the empirical

part of this master's thesis is explained. Subsequently, the methodology of the empirical part is presented, and the results and findings are discussed.

1.1. Research Questions and Assumptions

The increasing use of MT and LLMs in professional contexts has raised important questions about the quality of their outputs (Massey et al. 2024), particularly in specialized domains such as legal translation. This master's thesis investigates how professional translators evaluate translations produced by different translation systems relative to a human translation, using a two-part evaluation approach that combines Best-Worst Scaling with a criteria-based assessment. Against this background, the following two research questions guide this master's thesis:

1. Which translation do professional translators rate as best and worst when comparing a human translation, a neural machine translation, and two large language model translations generated with prompts of varying length and specificity of Spanish legal texts into German?
2. Which differences in translation quality do professional translators identify across the categories of content, terminology, grammar and syntax, and text function when evaluating the four translation modes?

Both research questions share the same framework: the four translation modes, the legal domain, and the Spanish-to-German language direction. While the first research question addresses the overall ranking of the four translation modes based on the holistic judgments of professional translators, the second research question examines the specific strengths and weaknesses of each mode across four quality categories.

With regard to the first question, it is assumed that human translation will be rated as the best translation overall. This assumption is based on findings from studies on NMT post-editing, which suggest that human translations tend to be more nuanced in terms of terminology and sentence structure (Moorkens & Guerberof Arenas 2024). Furthermore, Webster et al. (2020) observed that in literary translation, NMT aligns more closely with the source text's sentence structure, whereas human translations exhibit richer vocabulary and greater local cohesion. Although these findings stem from different translation domains and were obtained under different experimental conditions, they provide a reasonable basis for assuming that

human translation will also be perceived as superior in the legal domain. It is further assumed that the prompt's level of specificity will influence the quality of translations produced by the large language model, with longer, more detailed prompts yielding better results than shorter ones. This assumption is supported by prior studies investigating the translation output quality of LLMs, notably Gao et al. (2024) and He (2024), who independently concluded that persona prompts and prompts with supplementary contextual information yield higher translation quality than naïve prompts.

Regarding the second question, it is assumed that the criteria-based assessment results will not simply mirror the overall rating established in the first research question. While the translation rated best is expected to generally achieve higher scores across the quality categories, even the worst-rated translation may perform comparably well in individual categories. On the contrary, the translation rated as best may show relative weaknesses in specific categories. The criteria-based assessment is therefore expected to reveal a more nuanced picture of translation quality that goes beyond the holistic rating established in the first research question.

1.2. Motivation and Objective

One of the main motivations behind this master's thesis is to build upon the work of Hiebl and Gromann (2024) by extending their Best-Worst Scaling methodology for comparative translation quality assessment with a second criteria-based evaluation and by including LLM-based translation as an additional translation mode, as their original study was limited to human translation and NMT, surveyed translation students rather than professional translators, and did not focus on a specific subject domain. In doing so, this master's thesis contributes to an area of translation quality that remains comparatively underexplored, particularly in the context of legal translation. While prompt engineering for LLM-based translation has been investigated in general terms, its effect on translation quality specifically within the legal domain has received little attention to date. By involving professional translators and interpreters, most of whom hold certified or sworn translator/interpreter status, in the empirical evaluation, this master's thesis grounds its findings in the judgment of practitioners with direct professional experience in legal translation.

Apart from this gap in research, the findings could benefit different groups of interest active in translation and translation studies. Translators in general, which Moorkens and Guerberof Arenas (2024) describe as a highly exposed professional group to LLM technology, could gain a better understanding of the different translation modes in general as well as get some insight into the specific strengths and weaknesses of each translation mode across the predefined criteria assessed for this survey. This could help them to better put into perspective what the current translation capabilities of MT and LLMs are in relation to human translation. Especially legal translators, whether those working institutionally or as freelancers, who already face challenges linked to cost reduction through MT systems and Translation Memories (TMs) (Biel et al. 2024), could use the insights of the criteria-based assessment to see which categories are more or less reliably handled by machine-generated translations to identify where their professional expertise remains most indispensable. The findings could also be interesting for legal experts who often rely on translated legal texts without being trained translators themselves, as they could gain a better insight into judging how much a given translation can be trusted, particularly in cases involving machine-generated output.

2. Machine Translation

Machine Translation (MT) has developed considerably since its earliest research efforts in the postwar period, moving from rule-based systems grounded in explicit linguistic knowledge to statistical and, subsequently, neural approaches (Gaspari 2024). Closely tied to this development are the parallel advances in computing power and linguistic theory, as well as a steadily growing demand for fast and reliable translation. In this chapter, the historical development of MT is presented first, before turning to the different types of MT systems that emerged over the decades, namely rule-based Machine Translation, with its direct, transfer-based, and interlingual subtypes, as well as statistical and neural MT. It then turns to some of the most renowned commercially available MT systems to date, namely Systran, Google Translate, and DeepL, and provides an overview of the use of large Large Language Models (LLMs) for translation, including their general capabilities and the role of prompt engineering in shaping their output.

2.1. Historical Development of Machine Translation

Although the first MT systems were implemented on early computers shortly after World War II, the origins of MT do not, as it is often assumed, begin in the 1940s but can be traced back several centuries (Hutchins 2004; Stein 2018). In the 17th century, philosophers such as Gottfried Willhelm Leibnitz explored the idea of a language that could represent all human thought systematically, building upon the 13th-century concept of philosopher Ramon Llull, Leibnitz advanced the notion of *termini primi*, the fundamental units of meaning from which all conceivable thoughts could be constructed (Hutchins 2004; Stein 2018). Around the same time, in 1661, German physician and alchemist Johann Joachim Becher proposed a system, *Character pro notitia linguarum universali*, a system aimed at representing words across languages through multilingual dictionaries linked by numerical symbols. Although purely conceptual, both contributions are considered early attempts to formalize language relationships in ways that later proved relevant to MT because Leibnitz’s concept relates to universal language for borderless communication, and Becher’s coded notation for hidden communication (Hutchins 2004; Stein 2018).

The first legitimate precursors of MT emerged in the early 20th century, with two independent patents for an electromechanical multilingual dictionary device, filed simultaneously in 1933 by Georges B. Artsrouni and Petr Petrovič Trojanskij (Hutchins 2004).

Though similar in concept, Trojanskij's patent was more advanced than Artsrouni's general-purpose machine, featuring proposals for encoding and interpreting grammatical roles using universal symbols and outlining the basic design of a comprehensive translating machine. The translation process Trojanskij envisioned for his multilingual translation dictionary devices was intended to be carried out in three stages. First, a human operator analyzed the source text and assigned grammatical codes to each word. Second, the device identified the target language, and third, assembled a morphologically accurate translation (Hutchins 2004; Kenny 2018). Trojanskij's work can be viewed as an early forerunner of the MT models developed in the 20th century, as he envisioned a dedicated device solely for specialist translation purposes, whereas most MT systems came to be used for general applications, as the field eventually came to rely on digital computers that were still unknown to Trojanskij at the time of his death in 1950 (Kenny 2018).

The main driving force behind modern MT development goes back to a 1949 memorandum by Warren Weaver, then-director of the Division of Natural Sciences at the Rockefeller Foundation. Inspired by the advances made in cryptography during World War II, Weaver proposed a cryptographic approach to language, suggesting that translation could be understood as decoding an encrypted message (Gaspari 2024; Ramati & Pinchevski 2018). As ideas were drafted during the Cold War, the Rockefeller Foundation sought to align this vision with pro-American and anti-Soviet ideological interests, which increased strategic investments in MT research (Gaspari 2024; Ramati & Pinchevski 2018).

In 1954, initial research and extensive further development led to the first public demonstration of an operational MT system in New York, known as the Georgetown-IBM experiment. The IBM-Georgetown MT system prototype was capable to translate Russian into English using a limited set of vocabulary of approximately 250 words and six basic rules (Gaspari 2024). The system transferred meaning through word-for-word translation from the source to the target language (Garvin 1967; Gaspari 2024). While the system's output was limited, the demonstration generated significant interest and prompted the US government, military, and intelligence agencies to invest in MT as part of the Cold War arms race (Gaspari 2024; Ramati & Pinchevski 2018). In other words, it was the first Rule-based Machine Translation (RBMT) system of its kind, marking, as Ramati and Pinchevski (2018) define it, the beginning of the rule-based era of MT.

Despite early enthusiasm, the technological limitations of early digital computers and the adequacy of contemporary linguistic theories soon became apparent. Yehoshua Bar-Hillel, the first full-time MT researcher at the Massachusetts Institute of Technology (MIT), became an

influential skeptic, arguing that fully automatic high-quality translation was not achievable in the foreseeable future (Gaspari 2024; Hutchins 2000; Rehm et al. 2018). His pessimism was formally endorsed in 1966, when the Automatic Language Processing Advisory Committee (ALPAC) concluded that MT was slower, less accurate, and more costly than human translation, with no realistic short- or medium-term prospects for improvement (ALPAC 1966; Gaspari 2024). As a consequence, US federal funding for MT was largely withdrawn, and research came to a near standstill for more than a decade (Gaspari 2024; Hutchins 2001). During this period, development shifted to bilingual countries such as Canada and Western Europe, where practical translation needs, such as French-English translation in Canada and multilingual documentation within the European Economic Community, continued to thrive (Hutchins & Somers 1997, 1997; Poibeau 2017).

One prominent example of this post-ALPAC development is the TAUM-Météo system, developed by a research group at the Université de Montréal and operational from 1977 until 2002. Météo translated weather reports between English and French and, at its peak, processed 30 million words per year at an estimated accuracy of 90 percent (Poibeau 2017). Its success was largely due to the linguistic simplicity of weather reports, with their predictable vocabulary and limited grammatical structures. Not only did Météo represent a practical operational solution within a specialized domain, but it also contributed significantly to the rehabilitation of MT's reputation in the scientific community (Poibeau 2017; Rawling & Wilson 2018).

From the 1980s onward, commercial MT systems emerged in North America and Japan, some of which targeted the emerging personal computer market yet still produced low-quality output that required considerable post-editing. The conceptual shift toward statistical approaches was first proposed by Makoto Nagao in 1981, who argued that translation could be modeled as the retrieval and adaptation of previously translated examples rather than the application of predefined grammar rules (Ramati & Pinchevski 2018). This concept marked the beginning of the shift from rule-based to Statistical Machine Translation (SMT). However, this shift could not be realized until the 1990s and 2000s, when rising computational power and the rapid growth of the internet provided unprecedented amounts of parallel texts (Ramati & Pinchevski 2018; Rawling & Wilson 2018). By the late 1990s, MT was freely available online, and tools such as Babel Fish helped expand its worldwide visibility (Gaspari & Hutchins 2007). Although MT gained limited traction among professional translators due to concerns about output quality, the growing availability of Translation Memory (TM) systems played an important role in the further development of MT. TM tools were used by human translators to produce large amounts of bilingual, aligned text, or bi-text (Poibeau 2017) that provided

valuable training material for SMT while also offering an editing environment in which MT output could be reviewed alongside human translations. Through this dynamic, human translation contributed data to MT development while MT supported human translators, creating a mutually beneficial relationship that steadily improved SMT's capabilities and popularity (Rawling & Wilson 2018).

SMT dominated from the mid-2000s until the mid-2010s, consistently outperforming other approaches in annual MT evaluations. During this period, major technology companies, including IBM, Microsoft, and Google, integrated statistical methods into their pipelines (Koehn 2020). Following the terrorist attacks on the World Trade Center, MT research received a significant boost in funding from US defense agencies, particularly for Arabic and Mandarin, while the ongoing expansion of the European Union simultaneously increased translation demands (Rawling & Wilson 2018). Among these developments, Google Translate, launched in 2006, became the most widely adopted MT system, utilizing documents from the United Nations, international treaties, and multilingual corporate websites as training data, thus rapidly displacing commercial competitors (Ramati & Pinchevski 2018). However, performance improvements in SMT slowed over time, prompting research into new methods (Rawling & Wilson 2018). Neural approaches, though theorized since the 1950s and briefly explored in the late 1990s, had long been considered impractical due to limited computing power and insufficient training data (Mohamed et al. 2021; Poibeau 2017). By the 2010s, improved hardware, most notably Graphics Processing Units (GPUs), enabled a revival, and in 2016, Google's shift to NMT for Google Translate (Ramati & Pinchevski 2018; Wu et al. 2016) marked a turning point for the entire MT research sector (Koehn 2020; Wu et al. 2016).

2.2. Rule-based Machine Translation

Rule-Based Machine Translation (RBMT), also referred to as the classical approach to MT, is an umbrella term for machine translation systems that rely on explicit linguistic rules of the source and target languages, typically encoded in bilingual dictionaries and grammars capturing their core semantic, morphological, and syntactic regularities, which also explains the label rule-based. RBMT systems follow a methodology in which linguistic rules undergo a three-phase process: analysis, transfer, and generation (Okpor 2024; Sonbol et al. 2023). As discussed in Section 2.1, RBMT is the first type of MT, with its origins in the postwar period of the Second World War and during the Cold War. Its roots were, to some extent, influenced by ideas from cryptography and secret languages, which conceptualized translation primarily as decoding a

message in a foreign language (Poibeau 2017). Over the decades, various strategies, methods, and approaches have been developed and employed for MT. Although each method addresses the translation task differently, RBMT systems are generally classified into three main subtypes: Direct Machine Translation (DMT), Transfer-based Machine Translation (TBMT), and Interlingual Machine Translation (IMT) (Okpor 2024; Poibeau 2017). These three subtypes differ in how deeply the source text is analyzed linguistically before translating. To better illustrate this relationship, the Vauquois triangle can be used, which draws on Vauquois' (1968) work, in which he outlines the classical approaches to MT. Figure 1 shows one possible representation of the Vauquois triangle, adapted from Vauquois (1968), in which direct, transfer-based, and interlingual MT are arranged by increasing depth of linguistic analysis and increasing abstraction from the surface form of the source text (Wang & Sawyer 2023). The following sections describe these subtypes in greater detail from the top down.

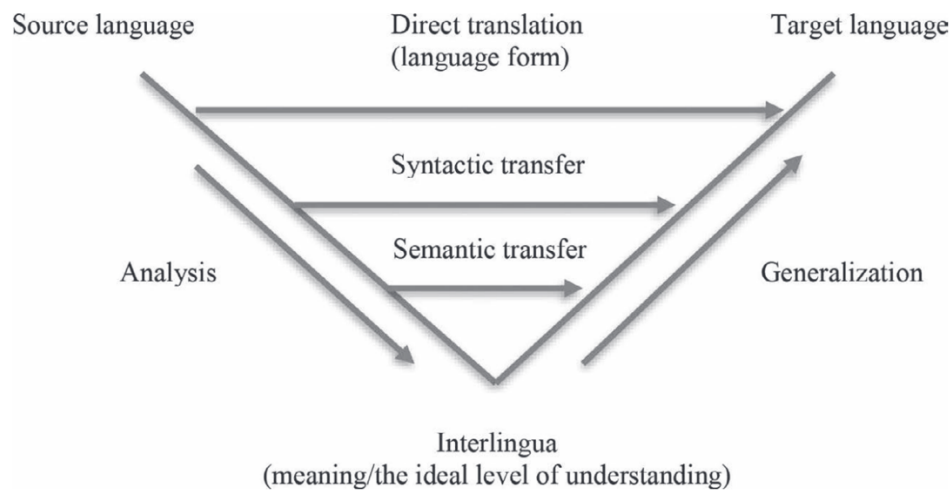


Figure 1: Vauquois triangle as depicted in Wang and Sawyer (2023: 77; based on Vauquois 1968)

2.2.1. Direct Machine Translation

The oldest and apparently least sophisticated approach in RBMT, DMT, performs a direct word-for-word translation with simple local grammar, such as morphological preprocessing for base-word extraction and word-order adjustments (Okpor 2024; Sonbol et al. 2023). Words from the source language are translated directly into a single target language. The analysis of source texts is specifically tailored to this target language, resulting in direct translation systems

that are fundamentally bilingual and unidirectional. As the earliest MT approach of its kind, DMT exhibits frequent lexical inaccuracies and syntactic structures that closely mirror those of the source language, which is largely due to the absence of systematic analysis of the internal structure of source texts, particularly syntactic relations, as well as to limited linguistic modeling and computational resources at the time (Okpor 2024).

2.2.2. Transfer-based Machine Translation

Transfer-based Machine Translation (TBMT) emerged within the second generation of rule-based MT research and generates translations based on intermediate representations of a source text's meaning rather than direct word-for-word translation (Hutchins 2001). Unlike DMT, which fully depends on the language pair, TBMT depends on the language pair used for the translation process only to a certain degree (Hutchins & Somers 1997). Based on the structural differences between the source and target languages, transfer-based systems are typically divided into three stages (Okpor 2024). First, a source-language parser analyzes the source text and generates an abstract representation of it (Hutchins & Somers 1997; Okpor 2024). Second, this representation is converted into one oriented towards the target language, which is where the language pairing is still occurring. Finally, the target-language text is produced by applying morphological analysis to the transferred representation. By using this approach, relatively high translation quality can be achieved, though the challenges of TBMT lie in the need to define and maintain rules at each stage of the translation process, namely analysis, transfer, and generation. Because transfer rules are language-pair dependent, analysis and generation modules are difficult to reuse across different language pairs. Moreover, transfer modules must account for numerous structural differences between source and target languages, thereby increasing their complexity and making them difficult to maintain (Okpor 2024).

2.2.3. Interlingual Machine Translation

Historically, IMT was the first approach to indirect MT systems (Hutchins & Somers 1997). Like TBMT, IMT first analyzes the source text and then converts it into an indirect, intermediate representation. The difference from TBMT is that, instead of converting the source text into a language-pair-dependent representation, IMT's intermediate stage is language-pair-independent, i.e., the source text is converted into a language-neutral form (Okpor 2024). This language-pair-independent stage is called the interlingua, meaning it is independent of both the source and target languages (Hutchins & Somers 1997; Okpor 2024). A generation module then

transforms this abstract representation into linguistically and grammatically correct sentences. One advantage in this regard is that the interlingua becomes increasingly valuable as the number of target languages it can be converted into grows (Hutchins & Somers 1997; Okpor 2024; Poibeau 2017). Interlingual systems require both a comprehensive analysis of the source text and language and accurate generation components that are capable of producing linguistically valid sentences in multiple target languages. Put differently, the main idea of the IMT approach is that these systems understand language at a semantic level (Sonbol et al. 2023). However, this poses several challenges to the approach. Text understanding remains a largely abstract and only partially formalizable process, requiring decisions about which types of information are relevant for translation. Due to subtle semantic distinctions and context-dependent variations in meaning, predicting all potential uses of a given word is not feasible (Okpor 2024). Put differently, a true interlingua would be independent of any particular natural language and would rely on a universal representation capable of expressing meanings across all source and target languages included in the system (Kenny 2018). Furthermore, extracting meaning from source texts to construct intermediate representations poses significant challenges, especially given the complexity of representing both explicit and inferred information in a formalized, language-independent manner (Poibeau 2017). Therefore, a truly universal language has never been achieved in practice (Giammarresi & Lapalme 2016).

2.3. Statistical Machine Translation

Compared to RBMT and its subtypes, Statistical Machine Translation (SMT) does not fall within the process defined by the Vauquois triangle, because the translated text is not generated on the basis of linguistic analysis. Rather than applying a set of rules for the source and target languages, SMT's transfer component is non-deterministic, meaning that translation candidates are ranked by probability rather than derived through fixed transfer rules (Wang & Sawyer 2023). From the input of a new source sentence to the generation of a target-language output, an SMT system typically proceeds through a set of steps. For a given source sentence, the system selects the most probable target-language translation using statistical models trained on data. These models include a Language Model (LM) trained on extensive monolingual target-language corpora that capture statistical regularities, such as target-language fluency, and a translation model derived from large parallel corpora to estimate the likelihood of correspondences between source- and target-language units (Kenny 2018). These probabilities are estimated during a training phase, in which the model's parameters are learned

automatically from the underlying corpora. During the decoding phase, when translating new input, the system generates multiple possible translations, referred to as hypotheses, and then selects the most probable hypothesis through search algorithms guided by the trained models (Liu & Zhang 2023; Wang & Sawyer 2023).

Early statistical translation models were based on unigrams, i.e., single words. These early approaches are commonly referred to as word-based SMT, as translation probabilities were estimated at the word level (Liu & Zhang 2023). This line of research was originally developed in the context of early SMT systems at IBM Research and provided the theoretical basis for word alignment, which later approaches continue to rely on (Liu & Zhang 2023). However, this strictly word-based approach soon proved insufficient, as it fails to account for local contextual information and multi-word expressions. Subsequent developments extended this approach beyond single words to contiguous multi-word sequences of variable lengths (Koehn 2009). These multi-word sequences can be imagined as grouped surface forms of words also referred to as phrases, bringing the term phrase-based SMT into use (Maibam & Syam Purkayastha 2023). Translation knowledge in these systems is stored in so-called phrase tables, which pair source-language word strings with target-language word strings and assign probabilities indicating the likelihood that each is a translation of the other (Kenny 2018; Liu & Zhang 2023). Phrase-based SMT processes these multi-word sequences by often using a lexicalized reordering model that considers the orientation of a phrase relative to the previous phrase or block of phrases. By memorizing local dependencies such as short reordering, idiom translation, and context-sensitive insertion and deletion of words, phrase-based SMT can capture local context more effectively than word-based approaches (Durrani et al. 2015). The phrase-based SMT approach has shortcomings, though, including struggles with non-local dependencies, meaning that two words that belong together are not recognized when separated by intervening word sequences, e.g., German verbs that come at the end of a sentence (Durrani et al. 2015). A second constraint is the limit set for reorderings. Reordering is used to arrange translated phrases into their correct order in the target sentence. To prevent the phrase-based SMT system from testing each possible combination, a reordering limit is applied, which in turn prevents the system from producing correct translations in cases that require reorderings beyond that limit (Durrani et al. 2015). Another limitation is that because these phrases are learned from statistical information in the training data and do not necessarily correspond to linguistically motivated units, phrase-based SMT systems can struggle with grammatical agreement and highly inflected or agglutinative languages. Additional issues include omissions of words and inconsistent translations of the same source-language item (Kenny 2018; Liu & Zhang 2023).

2.4. Neural Machine Translation

Neural Machine Translation (NMT) marks a fundamental shift in MT research and has become one of the dominant paradigms since the mid-2010s. NMT systems are based on Artificial Neural Networks (ANNs), computational models composed of interconnected processing units (Stahlberg 2020; Wang et al. 2022). Neural language models, including NMT models, learn to identify and represent language patterns by iteratively processing large amounts of unannotated data during training. Earlier statistical approaches modeled translation through probabilities estimated over aligned words and phrases. By contrast, neural models learn internal representations directly from the data (Poibeau 2017). In NMT, the entire translation process is typically modeled using an encoder–decoder framework (Cho et al. 2014). The encoder processes the source sentence and transforms it into a comprised vector, which the decoder then uses to generate the corresponding sentence in the target language (Cho et al. 2014). In this framework, words are represented numerically and combined sequentially to form a representation of the entire sentence, giving the system the capability to handle translation within a unified model (Poibeau 2017). Assuming that words occurring in similar contexts tend to have related meanings, the system learns to represent these patterns directly from the translation data, i.e., parallel corpora. This makes the approach more robust to words unseen during training, as NMT can still handle them using learned patterns of subword units, i.e., units smaller than full words but larger than individual characters (Sennrich et al. 2016; Stahlberg 2020). By considering contextual, semantic, and relational information across the entire sentence altogether, NMT avoids decomposing the translation process into separate components and instead models multiple linguistic relations simultaneously. This processing of entire sentences makes the approach flexible and enables it to capture similarities and relations within the sentence (Poibeau 2017); however, it does not capture them within the entire document.

Despite its strong performance, NMT was associated with several conceptual and practical challenges. Early NMT systems, which are based on Recurrent Neural Networks (RNNs), performed well on short, structurally simple sentences but had difficulties on longer or more complex inputs (Poibeau 2017). Increasing the complexity led to efficiency issues and made it harder to ensure that all parts of the source sentence were translated reliably. A further challenge comes from the holistic nature of the RNN-based approach. Because sentences are processed globally rather than decomposed into smaller segments, some words or phrases may be omitted or receive insufficient attention, resulting in incomplete or distorted translations

(Poibeau 2017). To address these flaws, a new model called a transformer was introduced. A transformer applies a self-attention mechanism to capture dependencies among all words in a sentence, independent of where they occur in the sequence. By processing the entire input sequence in parallel rather than word-by-word, training efficiency and translation quality can be significantly improved (Liu & Li 2023), giving rise to what is commonly referred to as Transformer-based NMT. Transformers thus allow the MT system to dynamically focus on the most relevant parts of the source sentence at different stages of the translation process, particularly for long-distance dependencies or typologically diverse language pairs (Poibeau 2017; Vaswani et al. 2017).

From a structural perspective, NMT differs fundamentally from earlier statistical approaches in that it is described as having a more uniform, as Kenny (2018) describes it, monolithic architecture rather than a combination of specialized modules. While SMT systems could address specific problems by modifying individual modules, neural systems integrate analysis and generation into a single model, making targeted interventions, such as adding new vocabulary, more difficult (Kenny 2018; Stahlberg 2020). Translation decisions in NMT are guided by learned weights, which determine how strongly different elements of the source and target contexts influence each other, in a way comparable to the role of probabilities in statistical models. This architecture allows NMT to process full sentences rather than n -grams, but it also increases the computational cost of training and contributes to the opacity of the system's internal decision-making processes (Kenny 2018; Poibeau 2017).

NMT systems are therefore often described as powerful yet difficult to interpret. Because the internal representations are purely numerical, highly abstract, and distributed across many parameters, it is not straightforward to trace how specific translation decisions are produced or why certain errors occur (Poibeau 2017). This lack of transparency makes error analysis difficult and limits the extent to which neural models can be linguistically inspected or manually adjusted. Although current research increasingly aims to make internal representations more interpretable, the difficulty of explaining the hidden representations a neural network learns from remains a central challenge for the practical and theoretical understanding of NMT (Poibeau 2017). The discovery and development of Artificial Neural Networks (ANNs) and their application for NMT also laid the groundwork for later LLMs, discussed in Section 2.6.

2.5. Examples of Online Machine Translation Systems

Building on the previously discussed types of MT systems, this section examines some of the most widely used commercially available MT systems to date. Each system is described in terms of its historical and commercial trajectories, as well as the underlying technological shifts from rule-based and statistical methods to neural architectures that have shaped MT's evolution over time.

2.5.1. Systran

Systran is one of the oldest commercial MT systems, founded in 1968 by Peter Toma, who had previously worked on early MT research at Georgetown University. The system was initially developed for Russian-English translation (Koehn 2020; Liu & Zhang 2023; Poibeau 2017). From 1970 onward, Systran's Russian-English system was used by the US Air Force and later by NASA during the Apollo-Soyuz-Test-Project (Kenny 2018). A French-English version of Systran was purchased by the European Commission in 1976, after which additional European language pairs were developed to enable the system to translate between multiple language pairs, reflecting the expansion of the European Union (Kenny 2018; Koehn 2020; Liu & Zhang 2023). The system was then deployed at several major European institutions, including the European Commission, Euratom, and NATO, though its use at the Commission continued until 2010 and was discontinued thereafter (Kenny 2018). Systran is also known for providing the first free online translation tool, Babel Fish, for the then-web search engine AltaVista in 1997 (SYSTRAN Translate 2026). In 2003, Babel Fish was acquired by Yahoo and remained there until it was replaced by Bing Translator in 2012 (Poibeau 2017).

From a technological perspective, Systran's design must be understood in relation to the practical requirements of its main application domains over time. During the 1980s and 1990s, the system relied primarily on manually developed bilingual dictionaries and rule-based transfer approaches (SYSTRAN Translate 2026). Even as SMT became increasingly prominent in research, Systran continued to operate in contexts where access to large parallel corpora was limited, notably in the US defense industry and intelligence applications. Around the year 2000, a substantial share of the company's revenue was still generated in these domains, where training data are often highly confidential and difficult to share (Poibeau 2017). Still, military and intelligence institutions also require the ability to adapt translation systems locally without disclosing sensitive data. For these reasons, approaches based on dictionaries and transfer rules

remained relevant, since new lexical items can be added directly to existing resources without retraining the system on large bilingual datasets (Poibeau 2017).

Following the success of Google Translate, Systran introduced hybrid approaches that incorporated statistical information while keeping the core of the translation model relatively traditional (Poibeau 2017). More recently, and like other major MT providers, Systran has shifted its focus toward NMT based on deep learning techniques and, in collaboration with Harvard NLP, released the first open-source framework for NMT (Casacuberta Nolla & Abril Peris 2017; Poibeau 2017). In 2024, Systran was acquired by the French software company ChapsVision, which specializes in artificial intelligence and data intelligence. As of February 2026, Systran supports more than 60 languages, 150 language pairs, and domain-specific models (ChapsVision 2026; SYSTRAN Translate 2026).

2.5.2. Google Translate

Google Translate is widely recognized as the most popular free MT service. Google began researching MT in the early 2000s in order to develop its own solution. The technology initially supported Arabic, Chinese, and Russian in 2005, and, one year after its launch in 2006, the number of supported languages had increased to 25 (Kenny, 2018; Poibeau, 2017). In its early stages, Google relied on Systran, indicating that it was initially a rule-based system. However, as Google developed its own proprietary technology, this system was gradually replaced. Google's approach next followed a statistical model developed by IBM, with later advances in the underlying algorithms leading to increasingly sophisticated translation systems. From early on, Google also benefited from owning its own search engine and from having sufficient computing power to collect, align, and filter large-scale bilingual corpora available on the internet for training Google Translate (Poibeau, 2017). In October 2007, Google was among the first providers to introduce a self-learning MT system (Poibeau, 2017; Schmalz, 2019).

At the Workshop on Machine Translation (WMT) in 2014, Wu et al. (2016) showed that, compared with phrase-based SMT, RNN-based NMT models could reduce the error rate by an average of 60%. This demonstration of the new approach generated considerable interest in MT research, as, beyond reducing translation errors, NMT can be trained relatively efficiently, allowing new languages, dialects, and application domains, such as medical and investment-related contexts, to be added more easily (DeCamp 2023). When Google released its initial version of Google NMT (GNMT) in 2016 (Wu et al. 2016), it fundamentally transformed Google Translate and generated widespread media attention (Schmalz 2019). Unlike the

previous approach of translating from Language A to English and then into Language B, GNMT was capable of performing direct translations between Language A and Language B, known as zero-shot translation, without relying on English as an intermediary (DeCamp 2023). Prior to the release of GNMT, research on NMT showed promising results. Nevertheless, in practical evaluations, NMT systems still performed worse than mature, well-trained SMT systems. Google was therefore the first to successfully overcome several key weaknesses of early NMT, including slower training, difficulties with rare words, and incomplete translations (Schmalz 2019).

As of the time of writing, Google Translate supports a wide range of use cases that extend beyond text-based translation. Its integration into apps, websites, and third-party software enables translation to be embedded directly into everyday digital environments (Google Translate 2026; Kenny 2018). In addition to written input, users can interact with the system through speech and image-based text recognition, while features such as offline translation further enhance Google Translate's practical applicability (Kenny 2018; Schmalz 2019). Although speech-based translation remains technically challenging due to variability in spoken language and informal usage, existing Google Translate APIs perform reasonably well in many everyday situations (Schmalz 2019). Ongoing improvements are driven by large-scale usage, continuous data growth, and the option for users to provide feedback, enabling Google Translate to steadily expand its capabilities across languages, modalities, and domains (Kenny 2018; Schmalz 2019). At the time of writing, Google Translate supports 249 languages and, according to a 2023 Google blog post, has a user base of approximately 1 billion users (Gu 2023).

2.5.3. DeepL

DeepL originated from the startup Linguee, which was founded by ex-Google employee Gereon Frahling at the end of 2007 (DeepL 2026; R  th 2013). Linguee can be described as a translation search engine whose core concept is to provide an online dictionary that allows users to search for words in authentic usage rather than in isolation (Buyse & Verlinde 2013; Schmalz 2019). When a user searches for a translation in Linguee, a drop-down list appears displaying various word combinations, such as collocations and prepositional phrases. After one of these options is selected, all corresponding occurrences are shown within a large bilingual translation corpus, presented in a two-column format (Buyse & Verlinde 2013). Since its public release in 2009,

Linguee has processed over 10 billion requests from over 1 billion users, which helped the company to gather extensive, high-quality training data (DeepL 2022; Schmalz 2019).

Based on its large-scale data collection, Linguee launched the NMT system DeepL in 2017, led by then-CTO Jarosław Kutylowski (Neuhaus 2020; Schmalz 2019). The company reported that, in a blind evaluation, its translations were selected as the best three times more often than competing outputs (DeepL 2018; Schmalz 2019). According to DeepL, this performance is attributed to the use of Convolutional Neural Networks (CNNs), which were originally developed for image recognition, rather than RNNs, which were used in Google’s GNMT (Wu et al. 2016). In the field of machine translation, CNNs had previously been explored mainly in research settings, including work by Google and Facebook, but had not yet been implemented in commercial systems. One advantage of CNNs is that they enable parallel processing of all words, and optimized programming libraries for their computation were readily available (Schmalz 2019).

In 2022, DeepL announced a user base of 100 million, though this number is assumed to be higher at the time of writing (DeepL 2022), and expanded its range of services. In addition to its core translation service, DeepL offers a range of complementary services, including writing assistance, real-time speech translation, and integration options that embed its language AI directly into applications users already use. More recently, DeepL has expanded beyond traditional translation services by developing systems that support more autonomous, task-oriented workflows across multiple software environments, offering its services not only to individual users but also to businesses (DeepL 2026). In November 2025, DeepL announced the addition of around 70 languages in beta for enterprise customers, expanding its coverage to over 100 languages overall and making nearly all of the world’s twenty most widely spoken languages available on the platform (DeepL 2025).

2.6. Machine Translation with Large Language Models

Building on the earlier discussion of NMT, this section traces how developments in deep learning and language modeling have shifted MT from task-specific systems toward pre-trained, increasingly general-purpose models. Language modeling has evolved in parallel with MT. Statistical Language Models (SLMs) estimated word or phrase probabilities based on preceding words or phrases but struggled with data sparsity (see Section 2.3), while early Neural Language Models (NLMs) addressed this by representing words as dense vectors. Pre-trained Language Models (PLMs) introduced a two-stage setup of self-supervised pre-training on large

corpora followed by fine-tuning on task-specific datasets, enabling a single model to be adapted to a range of Natural Language Processing (NLP) tasks (Ali & Pyysalo 2024; Wang et al. 2025; Zhao et al. 2026). As PLMs were scaled up in both training data and parameter count, this led to the development of LLMs, which are mostly Transformer-based models with hundreds of billions of parameters or more, trained on massive datasets, and exhibit capabilities not observed in smaller models (Ali & Pyysalo 2024; Wang et al. 2025). This includes the ability to perform a wide range of NLP tasks, such as translation, without being specifically designed for them (Zhao et al. 2026). The current use of LLMs for translation, therefore, builds on broader progress in language modeling and deep learning, rather than representing a direct continuation of NMT.

One prominent family of Transformer-based LLMs is the Generative Pre-trained Transformer (GPT) family, a form of Generative Artificial Intelligence (GenAI) primarily designed for language generation. Unlike encoder-decoder models used in NMT and other LLMs, GPTs use a decoder-only design and generate text autoregressively, predicting the next word in a sequence based on previous words (Ali & Pyysalo 2024; Luo et al. 2022). The GPT series has undergone major development since the release of GPT-1 (Ali & Pyysalo 2024; Radford & Narasimhan 2018), the first decoder-only Transformer model, which demonstrated that unsupervised pre-training followed by task-specific fine-tuning could achieve strong results across a range of NLP tasks (Ali & Pyysalo 2024). With 175 billion parameters, GPT-3 marked a turning point by demonstrating that scaling up model size and training data enables LLMs to handle a wider range of complex tasks (Ali & Pyysalo 2024; Brown et al. 2020). GPT-4 further advanced this by improving factual accuracy and introducing multimodal input (Ali & Pyysalo 2024).

2.6.1. ChatGPT

A prominent application built on the GPT series is OpenAI's ChatGPT, which adapts these decoder-only LLMs for dialogue (Zhao et al. 2026) and, despite being based on a previously available underlying model, saw widespread adoption following the release of its accessible interface (Eloundou et al. 2024). Although primarily designed for chat, ChatGPT's instruction-following capabilities and broad NLP training have made it a subject of considerable interest for MT research (Jiao et al. 2023). Jiao et al. (2023) conducted one of the first evaluations of ChatGPT for MT, comparing it against commercial translation systems such as Google Translate and DeepL. They found that ChatGPT performs competitively on high-resource

European language pairs, but lags behind significantly on low-resource or distant languages, partly because low-resource translation tasks compete with other NLP tasks for model capacity.

Further research has explored ChatGPT’s capabilities as a translation tool across various language pairs and text types. Wang (2025), for instance, compared ChatGPT with NMT systems for translating literary text from a low-resource language. Their findings indicated that ChatGPT generally surpassed NMT systems in semantic accuracy, fluency, and cultural adequacy, especially when utilizing specifically designed prompts. By contrast, NMT systems exhibited more consistent but less flexible output. Since its initial release, the GPT series has continued to evolve significantly with each new version, reflecting advances in architecture, training data, and fine-tuning methodologies. Successive versions of ChatGPT, from ChatGPT-3.5 to ChatGPT-4o and ChatGPT-5, have demonstrated considerable performance gains (Sallam et al. 2025). These developments highlight the importance of both model advancement and prompt engineering in optimizing ChatGPT and other LLMs for translation tasks, with the latter addressed in the following section.

2.7. Prompt Engineering

Prompting is the primary approach to using LLMs to solve a wide range of tasks (Zhao et al. 2026). A prompt typically refers to text input that serves as the main interface for interaction with LLMs (Neumann et al. 2026). Given that the quality of a prompt substantially influences model performance on any given task, a considerable body of research has emerged around the systematic creation of effective prompts. This area of research broadly encompasses two approaches: manual prompt creation, commonly referred to as prompt engineering, and automatic prompt optimization (Zhao et al. 2026). In this section, the focus is on manual prompt creation.

Prompt engineering refers to the deliberate, manual process of designing prompts that are well-suited to obtain the desired capabilities of an LLM for a specific task (Zhao et al. 2026). Prompts can generally be delivered to an LLM through two distinct mechanisms: system prompts and user prompts. Neumann (2026) provides a concise explanation for both delivery modes:

System prompts, usually set by developers or deployers of an LLM, are guidelines that last throughout conversations. They take precedence over user prompts and create the top of an instruction hierarchy [...] and are used to set the general behaviour of the AI system. User prompts, by contrast, are issued by users and contain requests that apply mostly to immediate interactions. In a practical example, a system prompt might instruct a language model ‘Don’t

generate copyrighted content,’ while a user prompt could ask ‘Generate a picture of Mario and Luigi rescuing Princess Peach.’ The system prompt would override the user’s request, causing the model to refuse rather than comply with the potentially copyrightinfringing user instruction. (Neumann et al. 2026:2)

Zhao et al. (2026) define four core components for a well-constructed prompt: task description, input data, contextual information, and prompt style. The task description constitutes the foundational element of any prompt, providing the model with a specific instruction expressed in natural language. For tasks that involve non-standard input or output formats, detailed clarifications become particularly important, and the use of keywords to highlight special settings has been shown to improve task completion. Input data refers to the information that the model is asked to process or respond to. While straightforward for most natural language inputs, structured data types such as knowledge graphs or tables require additional handling. A common solution is linearization, which transforms structured records into sequential text, thereby rendering them readable for LLMs. Contextual information encompasses supplementary background material that may be necessary for certain tasks. Prompt style concerns the overall formulation and presentation of the prompt. Prompts should be expressed as clear questions or detailed instructions that the model can unambiguously understand. Certain prefixes or suffixes, e.g., “Let us think step by step” (Zhao et al. 2026:45) or “You are an expert on this task (or in this domain)” (Zhao et al. 2026:45), have been demonstrated to meaningfully enhance performance. Chain-of-Thought Prompting, i.e., step-by-step reasoning, and Persona Prompting, such as instructing the model to respond as a domain expert, have been shown to improve performance on specialized tasks by more effectively leveraging the model’s pretrained knowledge (Zhao et al. 2026). Complex tasks benefit from being decomposed into simpler, numbered subtasks, allowing the model to address each step individually (Zhao et al. 2026).

Beyond the structural ingredients of a prompt, Zhao et al. (2026) identify several design principles that contribute to more effective prompt construction. A central principle is the clear expression of the task goal. A prompt style with ambiguous or underspecified task descriptions is likely to produce inaccurate or inappropriate outputs. Effective task descriptions, therefore, explicitly formulate the task objective, specify the nature of the input and output, and include any relevant constraints on the response, such as length or format limitations. A second principle concerns the aforementioned decomposition of complex tasks into simpler sub-tasks. Rather than presenting a model with a single, extensive instruction, breaking the task into a sequence of more manageable steps allows the model to address each component in turn, which

has been shown to improve accuracy on complex tasks by reducing the computational load at each step. Third, the provision of few-shot demonstrations considerably benefits task performance. Including a small number of input-output examples in the prompt allows the model to learn the relationship between input and output without additional training, i.e., fine-tuning. Fourth, the format of the input data is also relevant, as structured data may require linearization or other formatting adaptations to ensure the model can process it correctly. According to OpenAI's documentation, including a model-format into the prompt, such as `###` or `"""`, which an LLM interprets as a stop symbol, can help it to distinguish the instruction from the context (OpenAI 2026a; Zhao et al. 2026). Also, since most LLMs show stronger task performance when prompted in English, using English-language instructions is generally recommended even for tasks involving other languages (Zhao et al. 2026). Finally, persona strategies have been shown to improve model performance. As LLMs are trained on large, varied text collections containing diverse materials, they possess an inherent ability to simulate different personas. Prompting a model to assume the role of a domain expert has been found to enable more accurate and contextually appropriate responses by more effectively activating domain-relevant knowledge (Zhao et al. 2026).

Combining these components and principles provides a practical framework for designing prompts that are more likely to produce higher quality LLM outputs. Prompt engineering is therefore particularly relevant in translation, where the design of the input prompt directly affects the quality of the generated translation. The following section examines how these principles apply to MT with LLMs.

2.7.1. Prompt Engineering for Translation

As established in previous sections, prompt engineering plays a central role in shaping LLM outputs across a wide range of tasks. In the context of translation, this role becomes particularly significant, given that LLMs, such as the underlying ChatGPT, process instructions differently from NMT systems. While conventional NMT tools produce relatively stable outputs, they offer little flexibility in response to user input. LLMs, by contrast, generate outputs based on patterns learned over large, general-purpose datasets, making their performance highly dependent on how a prompt is formulated (Wang 2025).

Research has demonstrated that the choice of prompt type has a considerable impact on translation quality. He (2024) compared several prompting approaches, ranging from simple task instructions to more elaborate professional briefs and persona-based formulations, and

found that assigning the model a translator persona produced the most natural results. This indicates that strategies developed for human translators do not necessarily apply to generative models in the same way, and that persona-based approaches may be better suited to how these systems operate.

Yamada (2023) took a different approach by drawing on professional translation standards, specifically ISO 17100 (2020), to inform prompt design. By specifying parameters such as the text's communicative purpose and intended audience, translation quality improved noticeably, particularly in handling idiomatic and culturally embedded expressions. Those findings suggest that prompt engineering can replicate aspects of the preparatory work that human translators typically carry out before beginning a translation.

Gao et al. (2024) examined the effect of prompt complexity on output quality and found that providing the model with additional contextual information, such as domain-specific indicators or grammatical annotations, consistently yielded better results than simple task commands, sometimes even outperforming established commercial translation systems. At the same time, excessively detailed prompts sometimes had the opposite effect, underscoring the importance of finding an appropriate level of specificity.

3. Legal Language and Legal Translation

Having developed a comprehensive understanding of the MT paradigm, this chapter explores legal language and legal translation as a specialized area within the translation field. Legal translation occupies a distinct position within the broader field of specialized translation (Biel et al. 2024). Unlike general-purpose translation, legal translation operates at the intersection of language and law, where terminological precision is crucial to ensure legal validity, and in some cases, fundamental rights (Biel et al. 2024; Scott & O'Shea 2021). Legal texts are deeply embedded in a country's legal system and culture, which requires not only translation competence but also a comparative understanding of the source and target legal cultures (Biel et al. 2024). These demands make legal translation a particularly challenging domain for MT systems, which, despite their growing fluency, have been shown to struggle with terminological consistency and accuracy in legal contexts. A study conducted by Stefaniak (2020) concluded that NMT between English and Polish generally produces well-formed sentences suitable for post-editing, making corrections relatively easy. Nevertheless, fewer than 20% of NMT segments were error-free, with most errors related to accuracy and terminology.

Because this study focuses on human and machine translation between legal Spanish and German, this chapter first outlines the fundamental characteristics of legal languages in general, including their inherent conservatism, syntactic complexity, and the use of Multi-word Terms (MWTs), before analyzing the linguistic properties of legal Spanish and German. It should be noted that many of the characteristics are not exclusive to legal language but also apply to technical languages in general and to their designated purposes (Li & Liu 2025; Pikor 2023). Features such as reduced lexical diversity and high-level syntactic complexity are recognized as common traits of technical and scientific registers (Li & Liu 2025). This overlap is amplified by the fact that legal language frequently borrows terminology from other specialized fields (Pikor 2023). This chapter does not aim to draw a strict line between features unique to legal language and those typical of specialized discourse, but rather to provide an overview of the most prominent characteristics of legal languages first, before turning to more language-specific properties of Spanish and German.

3.1. Characteristics of Legal Language

The features of written legal language are most evident in legal texts, such as statutes, contracts, and other legal documents. These types of text are traditionally composed in formal and

complex terms that are largely inaccessible to non-experts. Even though there are substantial differences among legal genres, written legal language is generally characterized by a conservative writing style, which legal professionals have developed over centuries to ensure predictability by relying on standardized formulas validated through long-term judicial interpretation (Biel et al. 2024). Legal-technical language seeks precision, effectiveness, formality, and impartiality, and makes use of its own terminology that is fixed by set formulas and phrases, which, in many cases, is not part of everyday language (Daum & Salán García 2016).

At the syntactic level, written legal language achieves precision through structural density, often at the expense of readability. This density results from long sentences with multiple levels of subordination and elaborate phrase structures (Biel et al. 2024). Beyond the frequent use of the passive voice, legal language also relies on impersonal constructions that avoid naming specific individuals and instead focus on abstract norms and general rules (Biel et al. 2024; Prieto Ramos 2015). Authors of legal texts regularly use conditional constructions, multiple negations, archaisms, and complex prepositions to ensure that a single legal rule covers all relevant situations (Biel et al. 2024).

In addition to the syntactic complexity, there is a high lexical density in legal texts, characterized by the extensive use of nominalizations and specialized terminology. Although the extent of these characteristics may vary from language to language, nominalization remains a basic tool in legal language in order to achieve abstraction and objectivity (Biel et al. 2024; Prieto Ramos 2015). As is the case in other specialized fields, legal texts often make use of MWTs to achieve a high level of legal precision. These terms compress complex legal concepts into compact units, and their meanings are frequently adjusted, leading to significant differences from their typical usage in everyday language. Moreover, this specialized vocabulary is typically tied to a restricted, standardized phraseology, often resulting in fixed formulas and structural redundancy (Biel et al. 2024).

Another pervasive aspect of legal language is the high degree of interdisciplinarity and thematic overlap between legal and non-legal domains, given that law regulates most aspects of social life and professional practice. Therefore, legal texts naturally incorporate the specialized terminology of the fields they govern, e.g., agriculture, energy, or finance. This leads to a high density of non-legal terms in legal documents, thereby increasing terminological and conceptual complexity (Biel et al. 2024). The variety of topics is best exemplified in institutional contexts such as the European Union, where legislation is categorized into 20 subject areas. Those range from fisheries and energy to industrial policy and the environment,

and it is legal language that needs to accommodate this wide range of specialized fields and professional practices (Biel et al. 2024). Furthermore, the system-bound nature of legal knowledge adds another layer of difficulty compared to other fields, such as medicine or science, where concepts often enjoy relative universal validity across cultures. Meanwhile, legal terms derive their meaning strictly from their specific legal system, making full equivalence between jurisdictions rare (Biel et al. 2024).

As briefly mentioned before, a feature closely related to legal language is the frequent use of archaisms. Instead of using contemporary vocabulary, legal drafters tend to use outdated, i.e., archaic expressions and fixed formulaic phrases that have been used for centuries (Biel et al. 2024). The practical use of archaisms is not merely a convention but serves to provide certainty about their meaning for legal practitioners, as established terms carry a history of interpretation, whereas newer, untested expressions do not. Legal contracts, for example, demand a high level of precision that leaves little room for interpretation, as every word is meant to be understood exactly as it is formulated (Biel et al. 2024). This feature increases the structural density of legal texts further, making them largely inaccessible to non-professionals (Biel et al. 2024).

It is worth noting that the features described above are not equally present across all legal text types. Their degree of formal complexity varies depending on their communicative purpose. Legal texts can be grouped into four genres: prescriptive, target, expository or persuasive, and enabling. Prescriptive genres, including legislation, together with target genres such as contracts and powers of attorney, exhibit a higher density of legal language because their main function is to establish legally binding rights and obligations (Biel et al. 2024). Expository or persuasive genres, by contrast, such as judgments, pleadings, and legal opinions, are usually less complex, as their purpose is to interpret or elaborate on legal positions rather than to prescribe them. Enabling genres like legal textbooks and similar educational materials contain less legal language as well to remain accessible to their intended learning audience (Bhatia 2006; Biel et al. 2024). While plain language movements have influenced the modernization of legal language, especially in the European Union, the shift has not been uniform. Legal contracts, in particular, remained resistant to simplification efforts, thus retaining much of their traditional formal complexity (Biel et al. 2024), which connects back to the role of archaisms discussed above.

3.1.1. Characteristics of Spanish Legal Language

The Spanish legal language, while sharing a range of the general characteristics outlined above, is further shaped by several distinct features. Those include the preference for nominal over verbal constructions, a high tolerance for lexical borrowing, an extensive use of archaisms, and a comparatively low degree of impersonality in judicial and administrative texts (Daum & Salán García 2016). Given the breadth of features that characterize Spanish legal language, the following section focuses on selected characteristics in greater depth, while others are addressed more briefly. It is important to note that the English translations provided in the following section should be regarded as approximations, as some examples lack direct grammatical equivalents in the English language. It should also be noted that the following section does not distinguish between Iberian and Latin American legal Spanish. While both varieties share a pan-Hispanic grammatical and syntactic core, differences exist primarily at the lexical and phraseological level, tied to the specific legal systems of individual jurisdictions (Cuadrado Rey 2022).

The use of archaisms in Spanish legal language manifests in the morphosyntactic structure, where the future subjunctive and the absolute participle are common. The future subjunctive is a tense that has become obsolete in modern-day Spanish but is still actively used in contracts and formal administrative documents, as well as in old sayings and literature. Its function is to express a hypothetical action, i.e., a legal condition whose implementation or fulfillment at the time of writing is still open, but which can have binding consequences if it occurs (Daum & Salán García 2016; Pérez Román 2020a). Examples are future imperfect subjunctive (future de subjuntivo imperfecto) *si procediere* (if applicable), *si hubiere lugar* (if the case arises), and the future perfect subjunctive (future de subjuntivo perfecto) *hubiere matado* (shall have killed), and conditional clauses like *si transcurrido el plazo no se hubieren abonado dichas sumas* (if, upon expiry of the deadline, the said sums have not been paid) (Daum & Salán García 2016; Pérez Román 2020a). The second feature within the morphosyntactic structure is the ablative absolute clause, a remnant of Latin, which serves as a compressed substitute for subordinate clauses to express the temporal and causal aspects of an action (Pérez Román 2020a). Instead of using a complete sentence, such as *una vez que se observaron las prescripciones legales* (once the legal requirements had been met), legal drafters prefer the participial clause *observadas las prescripciones legales* (having observed the legal requirements), which expresses the same meaning in a more concise and formal manner (Pérez Román 2020a). Other examples are *propados los hechos* (the facts having been established),

transcurrido el plazo (upon expiry of the deadline), and *instruido el expediente* (the file having been processed) (Daum & Salán García 2016; Pérez Román 2020a).

Another defining feature of Spanish legal language is its preference for nominal constructions, where verbs are systematically replaced with nominalized expressions (Daum & Salán García 2016). Instead of a direct verbal phrase such as *al tramitar el juicio* (when conducting the trial), the nominal phrase *en la tramitación del juicio* (in the conducting of the trial) removes any reference to a concrete actor and frames the action as an abstract legal process (Daum & Salán García 2016). One characteristic of nominalized expressions is the use of periphrases consisting of a meaningless verb and a meaningful noun, in which the verb serves solely as a grammatical frame while the noun carries the actual legal meaning (Daum & Salán García 2016). Typical examples are *acusar recibo* (to acknowledge receipt) instead of *recibir* (to receive), where *acusar* in modern Spanish means *to accuse*, but in this case refers to the archaic term *notifying* or *acknowledging*, or the expression *estar en posesión* (to be in the possession) instead of *poseer* (to possess) (Daum & Salán García 2016).

Closely related to periphrases for nominal constructions of Spanish legal language is the linguistic phenomenon of so-called Archiverbalisms (archiverbalismos) or verbal periphrases (Montolío Durán et al. 2011; Pérez Román 2020a, 2020b). Archiverbalisms are a key feature of legal-administrative Spanish, which generally prefers syntactically complex compound sentences (Pérez Román 2020a, 2020b). These structures can appear in the following patterns: “a verb + a preposition + participle/infinitive, a verb + a gerund, or a verb + a past participle” (Pérez Román 2020b: 6) and serve to transfer simple actions into a more formal expression. Examples are the use of the verb *venir* (to come), which replaces the verb *estar* (to be), in combination with a participle to create the expression *vienen recogidos en el artículo 5* (come contained in Article 5) instead of *están recogidos en* (are contained in Article 5). The construction *venir a* (come to) + an infinitive, such as in *vengo a formular* (I hereby submit), is considered more formal than a simple form of the main verb *formulo* (I submit). Furthermore, constructions with *dar por* and *tener por* (to regard/deem), are frequently used to describe something as established, given, or done, e.g., *dan por admitidos* (accept them as admitted) or *teniendo por presentando este escrito* (this complaint having been duly filed). Finally, the combination of a verb + a gerund is used to describe an action that extends from the past to the present, e.g., *prestando servicios* (providing/offering services) (Pérez Román 2020a, 2020b). Although these constructions are considered inconvenient and should be avoided in general language, they are deliberately used in a legal context to ensure the demanded formality and precision, even if it increases syntactic complexity (Daum & Salán García 2016).

Several other morphosyntactic features of Spanish legal language share the common purpose of conveying objectivity and impersonality, distancing texts from certain actors, and expressing legal norms in a neutral and authoritative way. Among those are the use of impersonal constructions with *se*, such as *se estima* (it is assumed), as well as an accumulation of prepositional and adverbial collocations (*locuciones*), like *en el supuesto de* (in the case that), *a instancia de parte* (at the request of), or *se rige según lo dispuesto* (is governed by the provisions of) (Daum & Salán García 2016). In terms of adjective placements, legal Spanish tends to place adjectives before the noun, in contrast to standard Spanish, which places adjectives after the noun. These adjectives are used in an attributive function, e.g., *correspondiente trámite* (the corresponding procedure) instead of *trámite correspondiente* (lit. the procedure corresponding) in standard Spanish (Daum & Salán García 2016). Other atypical verb tenses include the future subjunctive to indicate a hypothetical future, e.g., *si procediere* (wherever applicable) (Daum & Salán García 2016) and impersonal imperative (imperativo impersonal/impropio), combined with the reflexive pronoun *se* (oneself) to form expressions like *notifíquese* (notice is given) or *pronúnciese* (judgment is rendered). A final morphosyntactic feature worth noting is the accumulation of multiple gerunds within a single sentence (Pérez Román 2020a, 2020b). An excerpt from an example provided by Pérez Román (2020a) illustrates this: “*sentando los hechos y fundamentos de derecho que estimó pertinentes [...] terminando suplicando se dictara una sentencia [...] interesando se dictara una sentencia desestimatoria*” (Pérez Román 2020a: 28) (setting out the facts and legal grounds deemed relevant [...] concluding with a request that a judgment be delivered [...] requesting that a dismissal judgment be delivered).

At the lexical level, legal Spanish distinguishes itself by a highly conservative and technical vocabulary (Daum & Salán García 2016). A central phenomenon is legal polysemy, in which ordinary words adopt a technical meaning. Where the word *concurso* in common Spanish refers to *competition*, in legal Spanish, *concurso de acreedores* means *arrangement of creditors*, and *concurso de delitos* refers to *concurrence of multiple criminal offenses*. Similarly, the verb *allanar* means *to level* or *to flatten* in everyday Spanish, but its pronominal form *allanarse* (*a la pretension*) carries the specific legal meaning of *accepting* (*a claim*) (Pérez Román 2020a). In addition, numerous lexical archaisms that function as established technical terms have long since disappeared from general usage, such as *otrosí* (supplementary agreement), *incoar* (to initiate), *evicción* (dispossession), or *fehaciente* (credible). The conservative lexicon is complemented with Latin aphorisms, like *iura novit curia*, *ex tunc*, or *ipso iure*, helping to preserve the legal principles and the legacy of Roman Law (Daum & Salán

García 2016). To ensure clarity and avoid misinterpretation, Spanish legal language often employs redundancies as rhetorical devices, such as repetition, e.g., *uso y disfrute* (usufruct and enjoyment), *daños y perjuicios* (damages), and *libre de cargas y gravámenes* (free of charges and encumbrances) (Daum & Salán García 2016; Pérez Román 2020a).

3.1.2. Characteristics of German Legal Language

The German legal language distinguishes itself from other specialized registers through a combination of structural information density, expressed through a consistently nominal style, and the frequent use of extensive syntactic dependencies, as well as a morphological preference for complex compound words and strictly regulated technical vocabulary. (Li & Liu 2025; Pikor 2023).

First and foremost, it is important to note that in this context, German legal language does not refer exclusively to Germany but encompasses the entire German-speaking legal region, including Austria and Switzerland, which are identified as independent national varieties of the language (Wissik 2017). Since legal terminology is shaped autonomously by each national legal system, particularly reflected in the distinct legislative structures at national, federal, or cantonal levels, the terminological differences between these national varieties can be substantial (Biel et al. 2024; Wissik 2017). A comprehensive account of these variations would exceed the scope of this chapter. The following description, therefore, focuses on the legal language of Germany.

Like Spanish legal language, German legal language also employs a nominal style that replaces verbal constructions with nominalized phrases, although in German, the nominal style occurs even more frequently than in Spanish (Daum & Salán García 2016). The verb loses its central role and is frequently transformed into a prepositional phrase, e.g., *zwecks Identifizierung und Benachrichtigung* (for identification and notification purposes) instead of *um zu identifizieren und zu benachrichtigen* (to identify and notify), or into a light-verb construction, e.g., *zur Anwendung bringen* (lit. to bring to application) instead of *anwenden* (to apply) (Pikor 2023). Nominalization is also accompanied by genitive constructions, as in *die Nichtigkeitserklärung des Vertrags vom Gericht* (the declaration of nullity of the contract by the court), as well as agent derivation, in which verbs receive the suffix *-ung* to create a noun, like in *Abbedingung* (waiver) instead of *abbedingen* (waive), or *Vertagung* (adjournment) instead of *vertragen* (to adjourn) (Yuan Li & Liu 2025; Pikor 2023). Agent derivation is especially common in Spanish, English, and other legal languages. German legal language also relies heavily on this process, which can be seen in examples such as *der/die Beschuldigte* (the

defendant) from *beschuldigen* (to accuse) or *der/die Verdächtige* (the suspect) from *verdächtigen* (to suspect) (Pikor 2023). Furthermore, German legal language prefers the passive voice over the active voice to keep agents in the background and present norms as universally valid (Li & Liu 2025; Pikor 2023). The typical passive form used is the *Modalpassiv* (modal passive), which is formed with *sein* + *zu* + *Infinitive* (to be + to + infinitive), like in *die Entscheidung ist zu veröffentlichen* (the decision is to be published) (Pikor 2023). Further passive substitutes are *sich lassen* + *Infinitiv* (to let + infinitive), e.g., *Die Frist lässt sich nicht verlängern* (lit. The deadline cannot be let extended), or to create other passive substitutes with the suffix *-bar*; *-fähig*, or *-lich* (Li & Liu 2025; Pikor 2023), as in *anwendbar* (applicable) and *vollstreckbar* (executable), or *gesetzlich* (lawful) and *nichtehelich* (extramarital). German legal language also makes systematic use of impersonal pronouns such as *man* (one) and *es* (it), which frequently occur alongside passive constructions to further distance the text from any concrete subject, e.g., *Man ist zum Rettungseinsatz verpflichtet* (One is required to assist in a rescue operation) or *Es wird hiermit festgestellt* (It is hereby determined) (Li & Liu 2025). Also, the relative pronoun *Wer* (who) is frequently used to start a subject clause to define abstract natural or legal persons, e.g., *Wer eine Tat begeht* (Anyone who commits a crime) (Pikor 2023). In German legal language, a specific use of tenses and modes is observed, with the aim of assigning distinct functions to the present, past, and future tenses in legal texts. On that note, tenses that refer to the present or the future fulfill a descriptive and directive function (Pikor 2023). By contrast, tenses that refer to the past have an exclusively descriptive function. Legal sentences written in the indicative form can again have both a descriptive and a directive function, while the subjunctive form is solely descriptive. Imperative verb tenses are not used in legal texts to preserve their impartial character (Pikor 2023), since passive modes are used instead.

A further feature of German legal language syntax is the order of the subject, the object, and the verb. Main clauses are rigidly arranged in a subject-verb order, where the subject is placed before the verb, while subordinate clauses follow an object-verb order, with the verb placed at the end of a sentence. In combination with the frequent occurrence of multi-clause sentences (*Schachtelsätze*) that include conditional, relative, and temporal clauses, as well as participial and infinitive phrases, these structures make sentences longer, increasing the distance between the subject and predicate (Li & Liu 2025). Li and Liu (2025) conducted a corpus-based quantitative analysis of German legal language to evaluate its comprehensibility relative to everyday language. They found that this feature results in an elevated Mean Dependency Distance (MDD), as syntactically related elements are separated by a greater linear

distance. This requires readers to retain partial information in working memory until the verb at the end of the sentence resolves the sentence's meaning. Another of their findings was that German legal texts exhibit lower lexical richness than news texts, a result of the need for precision and clarity. By repeating the same standardized term, consistency can be ensured (Li & Liu 2025). Term precision is strengthened through legal definitions set by the legislature. German legal language, similar to other specialized languages, relies on these definitions to prevent ambiguity in interpretation (Pikor 2023).

The German language's capacity to form compound words is extensively utilized in legal texts, which not only increases word length but also morphological complexity (Li & Liu 2025). While precise for legal professionals, the language may pose a barrier to comprehension for non-professionals (Li & Liu 2025; Pikor 2023). Examples are: *Finanzdienstleistungsinstitut* (Financial Services Institution), *Grunderwerbsteuererklärung* (Real Estate Transfer Tax Return), *Pflanzengesundheitsverordnung* (Plant Health Regulation), *Konformitätsbewertungsverfahren* (Conformance Testing Procedure). This linguistic feature is also closely related to other lexical characteristics of German legal language, including polysemous words, as already discussed for Spanish legal language, where common words from daily life take on a distinct technical meaning, e.g., in legal contexts, the meaning of the word *Gewalt* (violence), which in everyday German usually refers to physical force, has since the 1970s been extended to also include psychological coercion (Li & Liu 2025). Finally, a common feature among legal languages in general is the frequent use of abbreviations. Pikor (2023) differentiates between three kinds of abbreviations commonly used in legal languages: Buchstabenwörter (acronyms), Silbenkurzwörter (syllabic abbreviations), and Mischwörter (mixed abbreviations), the latter of which are particularly typical for German legal language. Acronyms are created by taking the initial letters of words, e.g., *LKA* for *Landeskriminalamt* (State Criminal Police Office), or *DVD* for *Deutsche Vereinigung für Datenschutz* (German Privacy Association). Syllabic abbreviations are formed by shortening the determinative element and the primary word into one syllable, e.g., *Kripo* or *KriPo* for *Kriminalpolizei* (criminal police) or *BAföG* for *Bundesausbildungsförderungsgesetz*. (Federal Education Assistance Act). Lastly, the German legal language often makes use of mixed abbreviations, which are a mix of acronyms and syllabic abbreviations, as in *StPO* for *Strafprozessordnung* (Code of criminal procedure), *AbzG* for *Abzahlungsgesetz* (instalment act), or *InvZuIG* for *Investitionszulagengesetz* (capital investment grant's act).

3.2. Legal Translation

Legal translation, according to the ISO standard 20771 (2020) and discussed by Biel et al. (2024), covers the translation of law-related or specialist legal content in terms of both content and context, e.g., legal settings (Biel et al. 2024:322). Put differently, legal translation addresses translating legal texts of all genres, ranging from primary prescriptive (laws, regulations, and treaties), to judicial documents (court rulings and pleadings), to administrative and technical documents (patents, official certificates, and corporate documents), to name a few (Biel et al. 2024; Scott & O'Shea 2021). Legal translation involves numerous linguistic and procedural challenges throughout the process, which entail a degree of responsibility and risk for legal translators (Biel et al. 2024). With due regard to these elevated requirements, the field's professional standards are reflected in ISO standards. The ISO 20771 (2020) sets significantly higher qualification requirements for legal translators than the general translation standard ISO 17100 (2015). While ISO 17100 (2015) is the umbrella standard for translation services, primarily intended for Translation Service Providers (TSPs), and sets out the general competence and qualification requirements for translators, ISO 20771 (2020) was developed specifically for legal translation. Both standards differ most clearly in their qualification requirements. Whereas the ISO 17100 (2015) regards a translator as qualified on the basis of a single criterion, which is either a tertiary-education degree in translation, any tertiary degree combined with two years of full-time experience, or five years of full-time practice without a degree, ISO 20771 (2020) requires a tertiary-education degree, in translation or another field always combined with three to five years of full-time experience in the legal field (Biel et al. 2024; Popiołek 2023).

The strict qualification profiles form the foundation for the specific core competences that legal translators need to demonstrate in practice. ISO 20771 (2020), as discussed by Popiołek (2023), defines six complementary areas of competence that a legal translator must combine to deliver a translation appropriate for the purpose: translation, linguistic and textual competence, specialist legal field competence, research and information competence, legal culture, and technical competence. First, translation competence covers the actual transfer of legal content from one language into another. This includes the comprehension and reproduction of specialist material in line with a project's requirements and with appropriate register and terminology. Second, linguistic and textual competence, which refers to a thorough conception of the source language and fluent command of the target language. It also requires familiarity with how legal texts are conventionally structured and worded in both legal cultures, and the ability to put this

into practice when translating. Third, ISO 20771 (2020), specifies specialist legal field competence, which concerns the translator's understanding of the underlying subject matter so that it can be re-expressed accurately in the target language. When the translator is allowed to certify translations, it also involves understanding the proper certification process. Fourth, legal translators are required to demonstrate research and information competence, which describes how efficiently a translator can locate additional legal knowledge, documents, and terminology, as well as to judge how trustworthy those sources are and to work with research tools and self-built terminology resources. Fifth, legal-cultural competence involves awareness of the norms, values, and procedures specific to the legal cultures on both sides, together with an understanding of how differences between legal systems manifest in cultural and practical terms. Finally, technical competence relates to handling the technical side of the work, from translation software and terminology databases to electronic signature and data security systems.

The implementation of these competences varies depending on the translators' working environment. Biel et al. (2024) distinguish between two translator profiles: institutional and external translators. Institutional translators, who are employed directly by national authorities or supranational organizations such as the European Commission, the Court of Justice of the EU, or the UN, and external or outstitutional translators, as Biel et al. (2024) call them, who work in the private sector, either within law firms or as self-employed freelancers. Institutional recruitment requirements are often less strict than the ISO standard, as competencies are tested internally through competitive examinations, and new staff are extensively trained on the job. For external translators, by contrast, ISO 20771 (2020) serves as an objective benchmark for professionalization, helping them stand out in an otherwise largely unregulated market. Additionally, ISO 20771 (2020) prescribes procedural safeguards, since high individual qualifications alone are not sufficient to minimize risk in the legal field. The standard makes the four-eyes principle mandatory, meaning that every translation must undergo full revision by a second, equally competent legal translator (Biel et al. 2024). The ISO 20771:2020 standard also explicitly states that legal translators need to dedicate 5% of their working time, which is about 8-10 days per year, to Continuing Professional Development (CPD), to update their knowledge and improve their qualifications. For external translators, CPD is not a mandatory requirement, but rather a matter of personal commitment, except if a professional association membership obliges them to do so. In comparison, institutions tend to organize CPD for their translators to better identify knowledge gaps and training requirements (Biel et al. 2024).

The reason for a standard as demanding as ISO 20771:2020 lies in the exceptionally high-risk profile of legal translations. Errors in this domain can cause far-reaching consequences of legal, financial, and defamatory nature that affect not only individuals but also entire organizations. Legal translations have immediate repercussions on people's lives, companies' security, and the protection of fundamental rights and liberties, including endangerment of life in faulty medico-legal translations. A further risk is that even after delivery, a legal translation's status is not fixed and can change over time, e.g., when a translation that was initially created for informational purposes becomes legally binding due to a change in law or a lawsuit. Since liability claims are difficult to prove and pursue in court, the emphasis shifts toward preventing errors in the first place through such standards. The ISO 20771:2020 standard, therefore, responds to these risks by recommending professional liability insurance and requiring translators to inform clients if they are not insured (Biel et al. 2024; Popiolek 2023).

3.2.1. Machine Translation in the Legal Domain

Cost reduction through MT systems presents a challenge both for institutional and external legal translators. Institutions invest in these workflows not only to ease translators' work but also to reuse already translated content through automated processes (Biel et al. 2024). The effectiveness of MT for automation in the legal domain, however, remains uncertain, and the available evidence is mixed. Stefaniak (2020), for instance, found that NMT between English and Polish generally produced well-formed phrases that were relatively easy to post-edit, thereby facilitating the correction process. The fluency of this output is nonetheless deceptive, as fewer than 20% of the segments were entirely error-free, with most errors relating to accuracy and terminology. However, the system's performance differed noticeably by text type. Because the NMT system used, called eTranslation (European Commission 2026), was predominantly trained on a corpus of mostly legal texts, it handled the repetitive, standardized phrasing of legal texts comparatively well. By contrast, non-legal texts, with their rarer vocabulary and greater need for stylistic and register adaptation, produced more severe and less predictable errors (Stefaniak 2020). The findings of Bajčić and Golenko (2024) show that NMT has difficulty with Multi-word Terms (MWTs) and abbreviations, and that the quality highly depends on high-quality datasets and specialized corpora. The increasing fluency of NMT output can be deceptive, making error detection more difficult and thus requiring a high degree of expert knowledge for the revision process to achieve proficient post-editing. Further error patterns identified in NMT output included word repetitions, omissions, unnecessary additions, and

problems with upper- and lowercase. NMT also tends to translate legal terms too literally, which results in the translation not adhering to the established writing conventions. Their conclusion aligns with the risks discussed in Section 3.2, where erroneous use of terminology within institutions such as the EU and the faulty implementation of directives can lead to state liability as a legal consequence (Bajčić & Golenko 2024).

The emergence of LLMs, which can quickly adapt to new tasks and, in some cases, require minimal training data, depending on the use case, might prove advantageous for legal translations. Compared to specialized NMT systems that depend on large amounts of specialized training data, LLMs can adapt to new tasks through one- or few-shot prompting with minimal input. Specialized NMT systems trained on highly specialized domains still surpass LLMs because they can translate technical terminology more precisely. Though the main source of error remains terminology, which is why terminological knowledge of human translators remains vital for post-editing (Bajčić & Golenko 2024). Bajčić and Golenko (2024) note a lack of empirical research on the use of LLMs in the legal domain, which, at the time of writing, can still be claimed, as is discussed later in Chapter 5. This gap is particularly severe considering the increasing political and economic pressure to deploy LLMs before extensive testing of their capabilities is conducted (Bajčić & Golenko 2024).

As of now, these developments reshape the translator's role rather than eliminate it. Human translators increasingly act as professional post-editors, correcting machine output rather than producing it from scratch (Biel et al. 2024). This shift, though, carries the risk that reducing legal translation to an industrial production process disregards the high-stakes nature of the field, where errors can jeopardize rights and even lives (Bajčić & Golenko 2024; Biel et al. 2024).

4. Translation Quality

The chapters discussed so far have addressed machine translation, legal language, and legal translation. They form the foundation for the actual purpose of the empirical study of this master's thesis: the assessment of translation quality. How good or poor the quality of a translation is can be examined from different perspectives. Generally speaking, quality perception in Translation Quality Assessment (TQA) is largely subjective, and human translators or domain experts are widely regarded as the most appropriate judges of translation quality. This is also why no universally accepted and fully objective measure exists (Chatzikoumi 2019; Lommel et al. 2013; Martínez Mateo 2014; Nemati Parsa & Afrouz 2023). Both translation studies and other fields of research, such as computational linguistics and informatics, have developed approaches and methods for empirically assessing translation quality. However, methods that depend on human annotators are often time-consuming, expensive, and cognitively demanding, which can affect the consistency of results (Freitag et al. 2021). For this reason, Automated Evaluation Metrics (AEMs) have been developed to carry out this process automatically, with the goal of obtaining results that are uniform, consistent, and time-efficient. This chapter examines different perspectives on translation quality from the viewpoints of translation studies and machine translation, before concluding with the Best-Worst Scaling method for comparative TQA, which is introduced in the empirical part of this master's thesis.

4.1. Perspective of Translation Studies

Translation Studies have approached the question of translation quality from different theoretical standpoints, each foregrounding different criteria and relationships between source and target texts, which also partly contradict one another in their basic interpretations. While early approaches mainly defined quality in terms of the relationship between source and target texts, such as concepts of equivalence (Baker 1992; Koller 1979; Nida 1969), the focus shifted towards purpose and the communicative function of a translation (Vermeer & Reiß 1984). In parallel, methodological approaches were developed for the systematic assessment of translation quality, with the aim of enabling standardized, consistent evaluation. The following section sheds some light on these different perspectives and interpretations and shows how the understanding of translation quality has shifted from a source-oriented to a function-oriented

approach, and how there is an intent to standardize and objectivize the evaluation of translation quality, using the example of Multidimensional Quality Metrics (Lommel et al. 2013).

4.1.1. Equivalence

Equivalence has occupied a central position in translation studies since the discipline began to take shape in the second half of the twentieth century, though it has remained a disputed concept due to ongoing disagreements about its nature, definition and practical applicability. At its core, equivalence was introduced to indicate that a source text and a target text share some form of sameness, with different academics proposing varying definitions of the sameness this relationship entails (Panou 2013).

One of the earliest contributions came from Nida (1969, 2003), who in the 1960s distinguished between formal equivalence, where the target text closely mirrors the source text in both form and content, and dynamic equivalence, which prioritizes conveying the source message as naturally as possible into the target language. He developed a systematic and analytical framework that translators could apply across a wide range of text types, and, perhaps most importantly, he shifted attention toward the role of the reader by introducing the receptor and their cultural expectations as central factors in the translation process (Panou 2013).

Koller (1979) expanded this framework by identifying five distinct types of equivalence: denotative, connotative, text-normative, pragmatic, and formal aesthetic equivalence. Denotative equivalence refers to the correspondence between source and target text at the level of factual, lexical content, that is, whether the words used in the translation refer to the same extralinguistic reality as those in the original. Connotative equivalence refers to the equivalence of stylistic tone between source and target text, beyond mere factual meaning. Text-normative equivalence refers to the adherence to the linguistic and stylistic conventions of a given text type in the target language, such as those governing contracts, scientific texts, or business letters. Pragmatic equivalence refers to the adaptation of a text to the communicative needs and knowledge of the target audience, ensuring that the translation is appropriately understood by its intended readers. Formal-aesthetic equivalence refers to the reproduction of stylistic and aesthetic features of the source text in the target text, including elements such as rhyme, rhythm, metaphor, and individual authorial style (Koller 1979; Panou 2013). Koller (1979) argues that translators must establish a hierarchy among these equivalence types depending on the requirements of the target text (Panou 2013).

Baker (1992) approached equivalence from a more neutral standpoint, treating it as a relative notion shaped by linguistic and cultural factors, and structured her analysis around equivalence at the levels of word, grammar, text, and pragmatics. At the grammatical level, she highlighted that structural differences between languages may force translators to add or delete information in the target text. At the textual level, she emphasized the importance of cohesion and coherence, arguing that the translator's decision to maintain these features depends on the target audience, the purpose of the translation, and the text type. At the pragmatic level, she drew on the concept of implicature, arguing that translators must identify not only what is explicitly stated in the source text but also what is implied, and render this as accessibly as possible for the target audience (Baker 1992; Panou 2013).

Pym (2010) approaches equivalence as a relation of equal value between a source-text segment and a target-text segment, one that can be established at any linguistic level, ranging from form to function. He further distinguishes between two types of equivalence. Natural equivalence exists between languages independently of any translation act and is not affected by directionality. By contrast, directional equivalence gives translators the freedom to choose among several translation strategies that are not prescribed by the source text, with available options ranging from adherence to source-language norms on one end to adherence to target-language norms on the other. A key characteristic of directional equivalence is its asymmetry, meaning that producing an equivalent when translating in one direction does not imply that the same equivalent would result when translating in the opposite direction. Pym also notes that perfect equivalence between languages does not exist and that any equivalence established in translation is always assumed rather than absolute (Panou 2013).

Reviewing the different approaches to equivalence throughout the history of translation studies, Panou (2013) concludes that, despite the variety of proposed frameworks, a common thread across these theories is the recognition that perfect equivalence between languages is unachievable. Furthermore, equivalence has held an unstable position throughout the history of translation studies (Rędzioch-Korkuz 2023). Broeck (1978), for instance, doubted whether equivalent effect can be measured at all, given that no text produces the same response across different cultures and periods, while Lefevere (1993) argued that equivalence focuses too narrowly on single words and overlooks the broader textual and cultural context of a translation (Panou 2013). Having been continually reworked and, in some cases, abandoned altogether (Rędzioch-Korkuz 2023), the concept eventually gave way to other paradigms that shifted away from equivalence.

4.1.2. Adequacy and Fluency

Adequacy and fluency represent the two central and often competing goals of translation. Adequacy refers to the extent to which a translation retains the meaning of the source text, while fluency refers to the degree to which the target text reads naturally and correctly in the target language, as if it had originally been written there (Koehn 2020). The challenge of balancing these two goals becomes apparent when considering that even professional translators produce different translations of the same sentence, each reflecting different choices about how much weight to give to source fidelity versus natural expression in the target language. Translation is therefore always an approximation, and the decisions translators make in navigating this tension are shaped mainly by the text type they are working with (Koehn 2020).

Literary translation, for instance, prioritizes fluency and stylistic effect, sometimes at the expense of literal meaning, particularly evident in the translation of song lyrics, where emotional effect and rhythm may matter more than semantic precision (Koehn 2020). By contrast, legal and technical texts prioritize adequacy, as the precise and complete transmission of content takes precedence over stylistic elegance, even if this results in less natural phrasing in the target language (Koehn 2020). An overarching goal underlying both criteria is what Koehn (2020) calls the invisibility of translation, which means that a well-executed translation should not draw attention to itself, and the reader should have no reason to think the text was originally written in another language.

While Koehn (2020) frames invisibility as a mark of quality, Venuti (2017) approaches the concept critically. For Venuti (2017), the invisibility of a translation is not a sign of success but rather the result of a domesticating translation strategy that suppresses the foreign character of the source text in favor of fluent, familiar target-language expression. Domesticating translation draws on the dominant linguistic and cultural resources of the receiving situation and thus assimilates the source text's differences. Foreignizing translation, in contrast, draws on less dominant linguistic and cultural resources and can thereby make the translator's presence visible in the target text. However, Venuti (2017) notes that this distinction is not absolute, since even a foreignizing translation must be ultimately understandable to readers of the target language and is therefore domesticating to some degree. He argues that without a conscious reading practice that recognizes translation as an interpretation rather than as transparent communication, translation will remain invisible regardless of the strategies employed by the translator or the reader's knowledge. This stands in contrast to Koehn's

perspective, as Venuti does not regard invisibility as an ideal but as a problem that obscures the translator's role and the text's foreign origin (Venuti 2017).

4.1.3. Skopos Theory

The Skopos theory, first outlined by Vermeer (1978) and elaborated together with Reiß in their joint work (Vermeer & Reiß 1984) (Baer & Hofeneder 2025), as part of the functionalist tradition in German translation studies, represents a fundamental shift away from equivalence-based approaches to translation. The term Skopos derives from the Greek word for purpose and refers to the goal or intended function of translation. Rather than evaluating a translation by how closely it mirrors the source text, the Skopos theory holds that the purpose of the target text is the primary factor determining all translation decisions. In this framework, translation is understood as a purposeful action, and like any action, it is guided by the intentions of its agent and the circumstances in which it takes place (Du 2012; Vermeer & Reiß 1984). Reiß and Vermeer (1984) built on this foundation by placing the skopos at the center of translation theory and rejecting the idea that the source text alone determines the translation process. In their framework, the source text functions as an offer of information, and the translator's task is to produce a target text that serves a target purpose for a target audience in a target situation. This also implies that the same source text can legitimately be translated in different ways depending on the translation brief, i.e., the instructions specifying the purpose of the target text, and that no single translation can therefore be considered the only correct one (Du 2012; Vermeer & Reiß 1984).

Nord (2018) further developed the functionalist approach by introducing the loyalty principle, which addresses two limitations she identified in Reiß and Vermeer's (1984) original theory. First, she noted that the purpose of a translation cannot always satisfy all target readers, given the differences in their expectations. Second, she pointed out that if the translation brief requires a communicative purpose that conflicts with the intention of the source text's author, the Skopos theory, as formulated by Reiß and Vermeer (1984), provides no restriction on the range of possible translations. Nord's loyalty principle addresses this by emphasizing the translator's responsibility to all parties involved in the translational interaction, including the source-text author, the client, and the target audience (Du 2012; Nord 2018).

Within the framework of the Skopos theory, Reiß and Vermeer (1984) formulated six hierarchically ordered rules governing translation. The most important of these is the skopos rule itself, which states that a translation, or *translatum*, is determined by its purpose. This is

then followed by the coherence rule, which requires that the target text can be understood as coherent with the situation and knowledge of the target audience, and the fidelity rule, which requires that the target text maintain a recognizable relationship with the source text. The former is referred to as intratextual coherence, and the latter as intertextual coherence. Intratextual coherence, that is, the internal consistency of the target text and its alignment with the target audience's situation, takes superiority over intertextual coherence, which refers to the degree of fidelity maintained between the target text and the source text – both are subordinate to the skopos rule. This hierarchy effectively reduces the authority of the source text, since the purpose of the translation and the needs of the target audience take priority (Du 2012; Nord 2018). This stands in contrast to source-oriented approaches which treat the original as an absolute standard against which any translation is measured (Nord 2018).

4.1.4. Translation Quality Assessment

While the equivalence theories discussed in Section 4.1.1 approach translation quality primarily through the relationship between source and target text in various ways, House (2015) placed communicative function at the center of her framework, developing a systematic model of Translation Quality Assessment (TQA) that makes functional equivalence the primary criterion of evaluation.

House's (2015) model of translation quality assessment is grounded in the concept of functional equivalence. For House, translation aims to produce a target text that is semantically and pragmatically equivalent to the source text. This means that the target text must fulfill the same function as the source text in its original context. Function in this context is understood as the use or application a text has in a particular context of situation, and serves as the primary standard for evaluating translation quality (House 2015; Lauscher 2000). To make the functional equivalence methodically operational, House (2015) uses two parameters: Register and Genre. Register is further subdivided into Field, which refers to the subject matter and content of the text, Tenor, which captures the relationship between the author and the audience, and Mode, which describes the medium and channel of communication. The evaluation procedure consists of establishing a source text profile based on these parameters, determining the function of the source text, comparing the source text profile with the target text, and identifying matches and mismatches between the two (Lauscher 2000). Based on this framework, House further distinguishes between two translation strategies. Covert translation involves adapting the target text to the communicative preferences of the target culture, so that

the translation reads as if it were an original text in the target language. On the contrary, Overt translation maintains a close relationship with the source text and makes no attempt to disguise its origin in the source culture (Lauscher 2000).

In practice, House's (2015) model provides a systematic procedure for evaluating translation quality by analyzing both the source and target texts in terms of register and genre and assessing the extent to which the target text reproduces the original's communicative function. The evaluation thus focuses not on individual linguistic errors, but on the overall functional relationship between the two texts (House 2015). The result of such analysis is a qualitative statement that explains how and to what degree the communicative function of the source text has been reproduced in the target text, and where and why deviations from this function have occurred (House 2015).

4.1.5. Error Annotation and MQM

Having reflected on the theoretical foundations of TQA, the focus shifts to its methodological implementation through systematic error annotation. TQA has historically been characterized by a lack of standardized methods, relying instead on ad hoc procedures that have produced inconsistent and subjective results. The Multidimensional Quality Metrics (MQM) framework, developed by Lommel et al. (2013), addresses this by providing a standardized vocabulary for systematic description and categorization of translation errors. Complementing this theoretical framework from an applied research perspective, Popović (2018) argues that precise error analysis is essential to understanding the specific strengths and weaknesses of translation systems, particularly where overall quality scores alone fail to reveal where and why a system produces inadequate output (Lommel 2018; 2014; Popović 2018).

The MQM framework functions as a flexible master catalog of error types from which users select a relevant subset suited to their specific evaluation context. In its original form, Lommel et al. (2013) defined five primary dimensions: Accuracy, Fluency, Verity, Design, and Internationalization. Accuracy concerns the correct transfer of meaning from source to target text. Fluency refers to the linguistic correctness and naturalness of the target text. Verity captures the relationship between translated content and real-world facts. Design covers formatting and visual presentation. Internationalization addresses the adaptation of content to different cultural and linguistic contexts. Following the fusion with the Dynamic Quality Framework (DQF) model developed by the Translation Automation User Society (TAUS), a translation industry association, in 2014 and 2015, the framework was expanded to eight

dimensions by adding Locale convention, which concern region-specific formatting conventions, such as dates and currencies; Style, which refers to adherence to stylistic guidelines; and Terminology, which addresses the correct and consistent use of domain-specific terms. Error types are organized hierarchically in a tree structure with up to 4 levels of specificity, enabling analysis at varying levels of granularity. A key methodological principle is that quality assessment must always be grounded in the translation specifications established by the Standard Practice for Language Translation, ASTM F2575 (2014), a standard guide for quality assurance in translation. They define an error as only one that violates the documented requirements of a given project, ensuring that translators are not penalized for derivations irrelevant to the text's intended purpose (Lommel 2018; Lommel et al. 2014). Popović (2018) notes, however, that defining an appropriate error typology remains one of the greatest practical challenges, as more detailed taxonomies, while more informative, are harder to apply consistently.

The practical implementation of error annotation presents further challenges. Manual annotation is time-consuming, costly, and frequently produces low inter-annotator agreement, showing that categories such as mistranslation and terminology errors are particularly difficult to distinguish without intensive training (Lommel 2018; Popović 2018). To improve efficiency and consistency, Popović (2018) proposes using Hjerson (2011), an automated pre-annotation tool that identifies and tags potential errors in MT output by comparing it against a reference translation, after which human evaluators correct or extend the generated tags rather than annotating from scratch, significantly reducing the time and effort required for manual error classification. To quantify errors, MQM assigns penalty points based on severity levels, which in Lommel's framework from 2014 originally were Minor, Major, and Critical with multipliers of 1, 5, and 10, and were expanded to four levels in the 2014 revised framework, adding a Null category for changes without error status, with Critical errors weighted penalty points in relation to word count to produce a percentage-based quality score. The result of this process is an error profile that visualizes the distribution of error types across a translation, enabling targeting conclusions about system quality and guiding improvements (Lommel 2018; Lommel et al. 2014; Popović 2018).

4.2. Perspective of Machine Translation

The Machine Translation perspective on TQA focuses on developing Automatic Evaluation Metrics (AEMs). These aim to approximate human quality judgments through computational

means to address the efficiency and subjectivity limitations of human evaluation (HE) approaches for TQA. This section discusses reference-based metrics, examining Bilingual Evaluation Understudy (BLEU) (Papineni et al. 2002) and its n -gram approach to surface-level text comparison, before turning to Metric for Evaluation of Translation with Explicit ORdering (METEOR) (Banerjee & Lavie 2005), which was developed in direct response to BLEU’s limitations, and Translation Edit Rate (TER) (Snover et al. 2006), which quantifies translation quality in terms of how many edit operations MT output requires to produce an acceptable target text. It then addresses neural language model-based metrics, presenting Bidirectional Encoder Representations from Transformers Score (BERTScore) (Zhang et al. 2020), a pre-trained and fine-tuned LLM used for AEM, and describing Cross-lingual Optimized Metric for Evaluation of Translation (COMET) (Rei et al. 2020) and its enhanced version, eXplainable Crosslingual Optimized Metric for Evaluation of Translation (xCOMET) (Guerreiro et al. 2024), both of which are learned metrics trained directly on human quality judgments and capable of producing interpretable error-level output. It closes with Comparative Translation Quality Assessment and the Best-Worst-Scaling method as a more consistent alternative to absolute scoring approaches.

4.2.1. Reference-based Evaluation Metrics

BLEU, introduced by Papineni et al. (2002), is an automated metric for MT evaluation that compares a machine translation against one or more human reference translations. It was designed to quickly and cost-effectively approximate human judgment to estimate its quality. The metric is based on the simple idea that the nearer a machine translation is to a human reference, the better its quality (Papineni et al. 2002). It is important to note that BLEU does not perform any semantic analysis and relies purely on surface-level text comparison.

The metric is based on two core mechanisms that rely on n -grams, defined as word sequences of length n . The first core mechanism is modified n -gram precision. An n -gram matching step measures how many sequences in the MT output also appear in the human reference translations. To prevent inflated scores through word repetition, a clipping step limits the count of each sequence to its maximum occurrence in any single reference translation. (Papineni et al. 2002). Papineni et al. (2002) argue that unigram matches ($n=1$) capture adequacy by measuring how many individual words in the MT output also appear in the reference translation, operating on the premise that the more words the system shares with the human reference, the more accurately it has transferred the content of the source text. Longer

sequences ($n > 1$) up to $n = 4$ consecutive words, according to Papineni et al. (2002), reflect fluency, indicating whether words have been arranged in a natural and grammatically coherent order. The second mechanism is the brevity penalty, which corrects for the fact that very short translations tend to achieve high precision scores but are entirely inadequate as translations. When the system output is shorter than the reference translation, the penalty reduces the score accordingly (Papineni et al. 2002).

As a result, BLEU outputs a score between 0 and 1, where 1 indicates a perfect match to the reference translation. Achieving a value of 1 is almost never possible, even by human translators, since there are many valid ways to translate a sentence (Papineni et al. 2002). The metric is most reliable at the corpus level, where individual scoring variations balance out across a large number of sentences. At the sentence level, however, BLEU can produce results that diverge from human judgments, primarily because it does not recognize synonyms. A correct and fluent translation that uses different but equally valid wording than the reference receives no credit, since BLEU requires exact word matches in the correct order. The authors argue that this rigidity remains acceptable because all systems are treated similarly. Since the metric applies the same strict rules to every translation, it effectively cancels out individual errors in comparisons. Therefore, BLEU's strongest assets lie not in providing an absolute quality grade for a single sentence, but in its ability to reliably rank different translation systems based on their overall closeness to human professional standards (Papineni et al. 2002).

Over the following years, BLEU became the industry standard for quick and cost-efficient evaluations of MT systems (Banerjee & Lavie 2005; Way 2018). However, it soon became apparent that certain downsides arose from the purely precision-based n -gram measure, especially since BLEU cannot recognize synonyms or lexical variants and, as a consequence, underrates the quality of otherwise fluent translations (Callison-Burch et al. 2006). Callison-Burch et al. (2006) therefore conclude that a higher BLEU score does not necessarily mean a better translation quality. They add that since BLEU allows n -gram matches to be drawn from any reference in the set simultaneously without checking whether the resulting translation is globally coherent, the metric is considered overly flexible. By that, they mean that an MT output can even be syntactically incoherent and semantically implausible and still receive a high score, since BLEU only checks whether individual word sequences appear in any of the provided reference translations, without evaluating whether the overall result created a meaningful sentence (Callison-Burch et al. 2006). While the use of multiple references is intended to capture natural variation in valid translations, BLEU often yields a crude statistical measure that lacks an understanding of grammatical or semantic plausibility (Way 2018), similar to how

direct machine translation functioned under rule-based approaches, as discussed in Section 2.2.1.

Banerjee and Lavie (2005) intended to counterweigh BLEU's constraints by introducing Metric for Evaluation of Translation with Explicit ORdering (METEOR). While both metrics share the general objective of measuring translation quality by comparing with human reference translations, METEOR was designed to address inherent limitations of BLEU, many of which were later criticized by Callison-Burch et al. (2006). The most significant difference from BLEU is METEOR's incorporation of recall alongside precision. Recall refers to how many words of the reference translation are covered by the MT output, weighted nine times more heavily than precision in the final score. This recall-driven approach was intended to replace BLEU's brevity penalty, which the authors argued does not adequately compensate for insufficient content coverage. The authors concluded that adding recall as an additional factor correlates more strongly with human judgments than precision alone (Banerjee & Lavie 2005).

METEOR also introduces greater linguistic flexibility through an incremental alignment process composed of three distinct mapping modules. The process begins with the exact module stage, which matches unigrams only if their surface forms are identical. This initial stage excludes a list of stop words to ensure that the alignment focuses on content-bearing lexical items (Banerjee & Lavie 2005), addressing BLEU's flaw in which content words and function words are weighted equally (Callison-Burch et al. 2006). For unigrams that remain unmapped, METEOR then applies the porter stem module via the Porter Stemmer, an algorithmic tool that reduces words to their stem to identify their morphological variants. As a final step, the WN synonymy module uses the WordNet lexical database (Fellbaum 1998; WordNet 2026) to identify semantic equivalences by mapping words to the same set of synonyms called Synset in WordNet (Banerjee & Lavie 2005). By prioritizing these modules in successive stages, METEOR ensures that the most precise matches are favored while still recognizing legitimate variations in word choice that BLEU's rigid n -gram identity requirement would ignore (Banerjee & Lavie 2005; Callison-Burch et al. 2006). Beyond flexible unigram matching, METEOR also differs from BLEU in how it evaluates translation structure. Rather than using higher-order n -grams as an indirect measure of grammaticality and word order, METEOR calculates an explicit fragmentation penalty based on the number of word blocks, referred to as chunks, which they define as blocks of matched words that appear consecutively in the output. If the matched words form a single uninterrupted sequence, the output is recognized as a single chunk and hence receives no penalty. Each interruption in the sequence creates an additional chunk, and the more chunks there are, the higher the fragmentation and the greater the penalty.

This allows METEOR to directly penalize deviations in word order rather than relying on the indirect signal provided by longer n -gram matches used in BLEU (Banerjee & Lavie 2005; Callison-Burch et al. 2006).

While BLEU and METEOR evaluate MT quality by measuring the similarity between the system output and the human reference translations (Papineni et al. 2002; Snover et al. 2006), Snover et al. (2006) follow the approach of quantifying the human effort necessary to correct a translation until it exactly matches a reference translation by introducing Translation Edit Rate (TER). TER takes a fundamentally different approach to measuring translation quality by quantifying the supposed human effort required to correct MT output to produce an acceptable target text. Rather than measuring surface similarity via n -gram overlap, TER computes an edit rate that reflects the number of edits required to transform the system output into a reference translation. This figure is divided by the average number of words in that reference, so that longer sentences do not automatically produce a higher score. In TER, a lower score indicates higher quality. Four types of edit operations can modify a system output to match a reference translation: inserting, deleting, substituting words, or shifting a contiguous sequence of words to another position. Each operation counts as one edit, no matter its category. Beyond these edit operations, TER also accounts for the formal surface-level features by treating punctuation marks as individual tokens and counting capitalization errors as substitution edits. This means that the final score reflects not only missing or incorrect words but also deviations in punctuation and capitalization that affect the accuracy of the target text. The shift operation is a central innovation of the metric, as it treats any phrasal reordering as a single edit step regardless of the distance over which the phrase is moved, directly capturing what BLEU can only approximate through longer n -gram matches. Because finding the optimal sequence of edits including shifts is computationally expensive, TER uses a greedy search algorithm that iteratively selects the shift that most reduces the remaining edit distance (Snover et al. 2006).

A notable finding from Snover et al. (2006) is that TER with a single reference already correlates as strongly with human quality judgments as BLEU does with four references. To push this further, the authors introduced Human-targeted TER (HTER), in which a human annotator produces a targeted reference by minimally editing the MT output until it is both fluent and meaning-equivalent to the original reference. This approach acknowledges that multiple valid expressions of the same meaning exist, as HTER does not penalize a machine translation that uses a correct synonym or an acceptable alternative word order not present in the original. A second annotation pass by independent reviewers then ensures the accuracy of

these targeted references. As a result, HTER produces a significantly lower error rate than standard TER, which is achieved through involving human annotators in the process. When a human annotator produces a targeted reference, valid synonyms and alternative phrasings used by the system are incorporated directly into that reference, so that they are not counted as errors in the final score. Snover et al. (2006) found that HTER aligns more closely with the average of human quality judgments than individual human raters do with one another, leading the authors to conclude that counting human-verified corrections provides a more consistent basis for TQA than subjective scalar ratings. Despite its high accuracy, the authors acknowledge that HTER is a resource-intensive method that requires an average of 3 to 7 minutes of human labor per sentence. Based on their findings, the authors suggest that HTER serves as a more consistent substitute for subjective human judgments. However, its high cost and time requirements typically limit its use to final evaluations rather than rapid, day-to-day system development (Snover et al. 2006).

4.2.2. Neural Language Model Based Metrics

Since reference-based metrics rely solely on surface-form similarity and string matching, their ability to recognize semantic equivalence of words in sentences or account for meaning-preserving lexical and compositional diversity is fundamentally limited (Zhang et al. 2020). To overcome the inherent shortcomings of string-based matching, the field of AEM has increasingly shifted toward neural metrics. Unlike reference-based approaches, which rely on surface-level statistics, these models estimate translation quality by leveraging learned semantic representations (Guerreiro et al. 2024; Zhang et al. 2020).

One prominent example of a neural metric is BERTScore, developed by Zhang et al. (2020). Rather than counting word-sequence matches, BERTScore computes translation quality by measuring the semantic similarity between individual tokens of a candidate translation and those of the reference translation. To do so, BERTScore relies on the contextual embeddings of a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model, which was trained on large text corpora. Because these models were trained to understand context rather than to evaluate translations specifically, BERTScore requires neither task-specific training nor human-annotated supervision (Zhang et al. 2020). For each token in the candidate, BERTScore identifies the most similar token in the reference by applying cosine similarity between their contextual representations, and vice versa. Precision, recall, and F1 scores are then derived from these pairwise matches. BERTScore-recall measures how well the meaning

of the reference is covered by the candidate translation, while BERTScore-precision captures how much of the candidate’s content is supported by the reference. The F1 score is the harmonic mean of precision and recall, and Zhang et al. (2020) recommend it as the most reliable measure because it performs consistently well across different languages and evaluation settings, whereas both precision and recall alternate in performance depending on the specific task. When conducting a translation quality assessment with BERTScore, it generates a score between 0 and 1. A score close to 1 indicates near-perfect semantic correspondence, and a score close to 0 indicates that the candidate and reference share little to no meaning (Zhang et al. 2020).

Cross-lingual Optimized Metric for Evaluation of Translation (COMET), introduced by Rei et al. (2020), marks a further advancement in neural language model based metrics. COMET is a learned neural metric that combines information from the source sentence and a human reference translation to predict translation quality. COMET bridges reference-based evaluation (see Section 4.2.1), which relies on the reference translation, and Quality Estimation (QE), a method that predicts quality using only the source text, without access to a human reference. The model is built on the pre-trained cross-lingual encoder XLM-RoBERTa (Conneau et al. 2020) and is trained to predict human quality judgments collected through HTER, MQM, and Direct Assessment (DA) (Rei et al. 2020). DA, in particular, is a method where human annotators rate how adequately a translation expresses the meaning of the reference or source text, using a scale, typically ranging from 0 to 100 (Guerreiro et al. 2024; Kocmi et al. 2022). By joining both processes, COMET can predict a single sentence-level score that reflects how well a translation aligns with its context. Although COMET achieves strong correlations with human judgments and represents a significant improvement over both lexical and unsupervised neural metrics, its output is limited to just that single numeric score, offering little insight into the nature or location of translation errors and makes it difficult to interpret the decisions that led to that score, as the black-box architecture of neural metrics makes the underlying decisions difficult to interpret (Guerreiro et al. 2024; Rei et al. 2023).

Building on the COMET framework, Guerreiro et al. (2024) introduce xCOMET, an improved learned metric that combines the strong correlation with human judgments of neural metrics such as COMET with the fine-grained error-detection capabilities of generative LLM-based approaches. With the emergence of GenAI models such as GPT-4, researchers began developing new AEM methods by prompting LLMs to identify and categorize translation errors, e.g., in line with the MQM framework, in order to achieve more interpretable and transparent output rather than single scores. Guerreiro et al. (2024) criticize the fact that, up to

that point, competitive performance with these approaches remained dependent on commercially operated models that were not publicly available. xCOMET was therefore developed as an open-source alternative (Guerreiro et al. 2024).

Like its predecessor, COMET, xCOMET builds on the pretrained cross-lingual encoder XLM-RoBERTa and extends it with two specialized prediction components. The first is a regression head that produces a global sentence-level quality score. The second is a sequence-tagging head that classifies each token in the candidate translation as Correct, Minor, Major, or Critical. From these token-level predictions, xCOMET identifies error spans and assigns each a severity level in line with the MQM framework, from which an overall MQM-based score is directly inferred, making the final quality judgment transparent and traceable (Guerreiro et al. 2024). This adaptation positions xCOMET as an example of Explainable QE that overcomes the limitations of earlier neural evaluation metrics, which offered no insight into how their single score was determined (Guerreiro et al. 2024; Unbabel 2024). The metric is available in xCOMET-XL (3.5 billion parameters) and xCOMET-XXL (10.7 billion parameters), as well as a lightweight version for users with limited computational resources (Guerreiro et al. 2024; Larionov et al. 2024). xCOMET supports three evaluation modes within a single model: it can assess a candidate translation using the source sentence alone, a human reference translation alone, or both together. Like COMET, xCOMET’s source-only mode enables reference-free quality estimation, which is useful in settings where no human reference translation is available. Guerreiro et al. (2024) demonstrate that xCOMET achieves state-of-the-art performance at both sentence and system level, outperforms GPT-3.5, and approaches GPT-4 in error span detection. It also shows particular robustness in identifying hallucinations and other critical translation pathologies that earlier metrics tend to miss.

Explainable QE approaches, as seen in xCOMET (Guerreiro et al. 2024; Unbabel 2024), have been used to address limitations in transparency in prior neural metric models. Nevertheless, like neural metrics in general, they still rely on scores from prior human annotations, posing further challenges. The established annotation methods, i.e., DA, HTER, and MQM (Rei et al. 2020), used for training, are widely criticized for being expensive, time-intensive, expert-dependent, and, in the case of DA, noisier and more variable in their results (Guerreiro et al. 2024; Hiebl & Gromann 2024; Kocmi et al. 2024; Zhao et al. 2024). To address these limitations and establish more reliable human baselines, various Comparative Quality Assessment (CQA) methodologies have been proposed. These approaches leverage the fact that humans find relative comparative judgments cognitively less demanding and more consistent than assigning absolute scores.

4.2.3. Comparative Translation Quality Assessment with Best-Worst-Scaling

CQA handles translation evaluation not by assigning absolute scores to individual translations in isolation, but by directly comparing multiple translations of the same source text. By doing so, CQA reflects a broader recognition within translation studies that quality assessment is inherently a relative and subjective process (Hiebl & Gromann 2024). The concepts of source-text-focused adequacy and target-text-focused fluency serve as a common denominator for evaluating both human translation (HT) and machine translation (MT), yet what constitutes a good translation ultimately depends on the evaluator and the context in which the translation is used (Hiebl & Gromann 2024). Approaches to TQA that are closest to CQA relate to ranking and rating. In the 2013 Workshop on Machine Translation (WMT) (Macháček & Bojar 2013), annotators ranked up to five different machine translations of the same source text relative to one another, allowing ties, and the collected rank labels were then aggregated into an overall score for each system. In 2017, the WMT evaluation shifted toward the already-mentioned DA, moving away from pairwise ranking toward absolute scoring (Bojar et al. 2017; Hiebl & Gromann 2024). Although these methods have each contributed to the field, they share a fundamental limitation, as they require annotators to evaluate translations either sequentially or in isolation, which increases cognitive load and makes it difficult to maintain consistent standards across annotators and evaluation sessions (Hiebl & Gromann 2024).

Best-Worst-Scaling (BWS) was originally developed by Louviere (2015) in the 1980s for the comparative evaluation of sets of items, in which participants identify only the best and worst options within each subset rather than scoring each item individually. The method is grounded in random utility theory, which holds that people choose the option that best satisfies their preferences, while acknowledging that individual decisions carry an element of unpredictability. BWS has since gone beyond its origins in psychology, being applied across diverse fields, including healthcare, linguistics, and consumer research (Hiebl & Gromann 2024). In translation research, it had previously been used to examine preferences for gender-fair language strategies (Paolucci et al. 2023), but Hiebl and Gromann (2024) note that its application to comparative translation quality assessment had not been proposed prior to their study. The key practical advantage of BWS over ranking and absolute rating methods lies in the simplicity and directness of its output. Annotators are required only to identify the best and worst translations in a given set. With four translations per set, this selection automatically implies a partial ordering in which one translation clearly emerges as best, another as worst,

and the two remaining options fall in between. Unlike ranking or absolute rating methods, BWS does not require a separate calculation of inter-annotator agreement, because the combination of selection choices and numerical ratings already enables a direct comparison of how different translation systems were evaluated across participants (Hiebl & Gromann 2024).

In their study, Hiebl and Gromann (2024) used BWS in order to capture not only which translation was perceived as best or worst, but also how strongly that preference was held. Annotators selected the best and worst translation from each set, then rated the best option on a scale from +4 to 0 and the worst from 0 to -4. Combining both approaches yields an overall score for each translation method that reflects both the frequency with which it was selected and the intensity of those evaluations. Five sets of four German translations of the same English source text were presented to ten master's students of translation studies. Each set contained one professionally published human translation and three machine translations produced by Google Translate, DeepL, and Microsoft Bing Translator. The source texts were retrieved from non-fiction books across five domains to ensure coverage of different registers while keeping the required level of domain expertise accessible to language learners rather than domain experts. The domains included in their survey were history, politics, finance, biology, and physics (Hiebl & Gromann 2024).

The results showed that BWS is an effective and consistent method for CQA in translation. Across the evaluated sets, participants most frequently selected the human translation as the best option, closely followed by DeepL, while Google Translate and Microsoft Bing Translator received predominantly negative ratings. Participant feedback on the method was mainly positive. Annotators acknowledged that having multiple translation options available at once made it easier to reach a decision than evaluating a single translation would have been. The survey was completed within a manageable timeframe of 20 to 35 minutes. Hiebl and Gromann (2024) conclude that BWS offers a practical and reliable way to capture subjective quality preferences in translation assessment without placing excessive demands on annotators.

5. Related Work

Of the works discussed in this chapter, the study by Hiebl and Gromann (2024) is most directly related to the methodology of the present master's thesis, as it serves as the basis for the experiment conducted in the present work. Since their approach has already been presented in detail in the previous section, the studies below are selected for how they address related aspects of the research questions explored here.

Otal and Limbach (2026) compare DeepL, Google Translate, and GPT-4o for Spanish-to-German translation. Their contrastive analysis examines the strengths and weaknesses of the three systems at the lexical, syntactic, and textual levels, assessing terminology selection and consistency, sentence structure and grammatical accuracy, and coherence and register appropriateness. Their central conclusion is that while automatically generated translations offer substantial productivity advantages, none of the systems yet ensure accurate translation, and post-editing by a human translator remains necessary. Although their study does not focus specifically on legal texts and uses Google Translate rather than a second LLM translation with a longer persona prompt, it provides the closest comparable study for the Spanish-German language pair using contemporary MT and LLM systems, as far as could be determined from the literature reviewed.

Wiesmann (2019) evaluated DeepL and MateCat on Italian-to-German legal texts across several genres, including contracts, court decisions, and legislative texts, assessing translation quality according to the criteria of comprehensibility, meaningfulness, and correspondence between source and target text. Her central finding was that correspondence between source and target text was rated substantially worse than meaningfulness across all genres and both systems, indicating that the systems produced fluent output that nonetheless failed to render the source's legal content accurately. Wiesmann concluded that the overall results were insufficient for legal translation, as the systems disregarded the translation context and lacked the legal expertise needed to produce an adequate target text.

Fernández Torné and Matamala (2021) compared SMT and NMT systems for Spanish-to-German industrial documentation, aiming to gather quantitative and qualitative data through a quasi-experimental field study. At the segment level, they collected judgments on fluency, adequacy, and ranking. At the document level, they measured subjective perceptions of quality and mental effort from 22 professional participants, as well as their general attitudes toward machine translation. Both levels assessed participants through point scales, with the document-

level assessment including the option for participants to add comments. The study involved 22 professional translators who are native speakers of German with at least two years of experience in this language pair.

Castro-Lesching (2025) conducted a contrastive analysis of English-to-German criminal law translations produced by DeepL and GPT-4o, identifying errors and inaccuracies in their performance at the lexical, syntactic, and textual levels, with particular emphasis on the challenges of legal comparison between the US Common Law and German Civil Law systems. The results show that both systems display considerable flaws at the lexical level, frequently mistranslating specific legal concepts and disregarding necessary umbrella terms. ChatGPT showed a greater tendency toward terminological consistency than DeepL through targeted prompting. It is concluded that MT in the legal domain continues to require critical post-editing by specialists, as current systems are unable to adequately capture the functional requirements of the translation brief or the fundamental differences between the legal systems involved. Similarly, Feng (2025) compared six MT systems, including DeepL and GPT-4, on Spanish legal texts translated into Chinese across three legal genres, combining the automated metrics BLEU and METEOR with HE following the DQF-MQM framework. The conducted error analysis revealed that across all systems, accuracy-related issues, such as mistranslation and terminology errors, were more frequent than fluency problems.

Briva-Iglesias et al. (2024) evaluated GPT-4 and other LLMs against Google Translate across four language pairs, Spanish, Catalan, Turkish, and Brazilian Portuguese, in the legal domain, combining AEMs (BLEU, TER, chrF3, COMET) with HE by professional translators. The language selection was intended to assess how the different systems handled the high-resource (Spanish), middle-resource (Brazilian Portuguese), and low-resource (Catalan, Turkish) languages. Their study revealed that Google Translate generally outperformed LLMs on AEMs, whereas professional translators rated GPT-4 comparably or slightly better in contextual adequacy and fluency. While Google Translate tended to translate certain legal terms differently across the text, GPT-4 demonstrated greater overall consistency. This finding shows a certain discrepancy between AEMs and human-based assessments. At the time of writing, the authors suggest that traditional AEMs appear better tailored for NMTs and are unable to fully capture the nuances of LLM translations.

Sanz-Valdivieso and López-Arroyo (2023) had two annotators compare Google Translate and GPT-3.5 on specialized Spanish-to-English terminology through TQA based on the MQM framework. To generate a translation with GPT-3.5, they used a simple Spanish prompt that requests a translation from Spanish to English. Their results show that GPT-3.5 made fewer

terminological errors, fewer mistranslations, and omitted fewer elements compared to Google Translate. In over one-third of the translations, the annotators rated them as error-free, while Google Translate was rated as error-free in fewer than one-fourth.

Block et al. (2025) compared DeepL, GPT-4, and Google Translate on German-language medical texts from a web portal translated into English, French, and Polish, using BLEU, METEOR, and COMET for automatic evaluation. The authors found that, according to METEOR and COMET, the translations showed good to very good quality, depending on the language. They criticize BLEU and METEOR for their limited ability to capture semantic accuracy and terminological correctness in medical translation, while COMET is considered more suitable for this context due to its neural architecture and integration of the source text. Since none of the metrics revealed statistically significant differences among the systems, the study raises the question of whether AEMs are sufficiently sensitive to differentiate among the three translation systems in specialized domains.

Zhang and Fan (2025) evaluated how specific prompt variables influence the translation quality of specialized Chinese texts in the news, medicine, and legal domains into English using GPT-4. They modified the prompts by adjusting the register, genre, and audience, and then compared them to a baseline prompt. The aim was to assess which prompting strategy best activates the model's domain-specific knowledge. Using BLEU, ChrF++, and COMET, the study finds that the genre-oriented prompts were universally effective, particularly for news texts, where BLEU scores showed the highest sensitivity and performance gains, whereas specialized domains like law required semantically aware metrics such as COMET for accurate quality assessment.

Gao et al. (2024) systematically evaluated the effect of different prompting strategies on ChatGPT's translation quality across high-resource (English, Spanish, French) and low-resource language pairs (Romanian, Turkish), comparing the generated output against DeepL and Google Translate using BLEU, ChrF++, and TER, as well as a multi-reference test. They demonstrated that ChatGPT's translation potential, without specifying the GPT version used, is significantly underestimated when evaluated with standard prompts and AEMs that rely on HT as the reference. Domain-specific and direction-specific prompts yield substantial quality improvements, and few-shot examples further enhance performance consistently across text types. Under the multi-reference evaluation, which uses 10 human translations rather than just one to better capture the natural diversity of HT, ChatGPT outperforms DeepL and Google Translate on high-resource language pairs, whereas single-reference metrics tend to unfairly penalize the model's linguistic flexibility. Performance in low-resource languages remains a

weakness, and the integration of part-of-speech tags as additional linguistic guidance proves unstable across translation directions.

Upon reviewing these studies, it becomes apparent that machine translation quality depends heavily on the language pair, the domain, and the evaluation method applied. LLMs, such as different versions of ChatGPT, tend to perform comparably to or better than NMT systems in high-resource language pairs when evaluated by human judges, while AEMs often fail to adequately capture the actual quality of LLM output. The legal domain in particular emerges as a persistent challenge across all systems, with terminological accuracy and legal-systemic differences posing the greatest difficulties. Prompt engineering further emerges as a relevant variable, with domain-specific and context-rich prompts yielding better results than standard instructions. The present thesis addresses a gap in this body of research: the systematic evaluation of LLM- and NMT-based translation quality in the Spanish-to-German legal domain using human evaluation, a language pair and domain combination that remains underrepresented in the literature at the time of writing. The following chapters outline the methodological framework and the results adopted to address this gap.

6. Methodology

This chapter details the methodology used in the empirical part of this master’s thesis. The overall aim is to assess how the four translation variants introduced in Section 1.1, that is, human translation, NMT, and two LLM-based outputs, are evaluated by professional translators, both in terms of their relative best and worst evaluations and in terms of their performance across specific quality categories. For this purpose, this chapter outlines the selection of Spanish legal source passages with particular attention to training data contamination, the production of the four translation variants, the screening and recruitment of participants for the required professional expertise, and the design of the survey used to obtain both best-worst judgments and quality ratings.

6.1. Text Selection

The text selection required careful consideration, particularly regarding the datasets on which contemporary NMT and LLM systems are trained. These systems are developed using vast monolingual, parallel, and multilingual corpora extracted from the internet, such as data from forums, digitized books, and news sources, to name but a few. This raised the methodological concern that NMT and LLM systems may recognize a source text and reproduce an existing human translation rather than generate an independent translation based on their inherent capacities (Kocyigit et al. 2025). A small preliminary trial using Google Translate confirmed this concern when translating passages from EU legislation. Google Translate reproduced the exact wording of the official human translation published on EUR-Lex. To reduce the risk of including source texts already present in large-scale multilingual training corpora, the text selection process therefore prioritized legal sources that are less likely to be represented in widely used datasets. The text selection strategy is intended to maximize the likelihood that the translation outputs genuinely reflect the systems’ internal translation capabilities. This was achieved by receiving Spanish source texts from a legal expert.

Another central methodological consideration concerns prompting. LLMs have been shown to be highly sensitive to variations in prompting, regarding phrasing, structure, and contextual framing (Zhuo et al. 2024). Minor changes in wording or punctuation can lead to substantial differences in an LLM’s output (Salinas & Morstatter 2024). Although prompt engineering has become an integral part of applied LLM research, there is no consensus on best

practices for the translation task. Prompting, therefore, remains a significant variable that can influence results. LLMs also exhibit inherent non-determinism, often generating different outputs for the same prompt due to their probabilistic nature and sampling methods (Ouyang et al. 2025). This needs to be considered for the translations generated by ChatGPT in this survey, as the same prompt can produce different outputs each time it is reused.

The first step involved selecting appropriate legal text passages to later serve as the basis for the comparative assessment. This process involved several stages. The initial aim was to compile Spanish legal texts translated into German, with a focus on authentic, certified documents. Accordingly, the first materials considered for inclusion were officially certified Spanish-German translations, as well as translation examples from a Spanish-German business correspondence textbook. During a consultation with a certified translator and lecturer, who also reviewed a preliminary version of the survey, concerns were raised regarding the professional authenticity of these text passages. The lecturer recommended narrowing the scope of the study to a clearly defined legal subdomain and suggested either focusing on document translations or on criminal law, as this would yield a more coherent and professionally relevant set of texts. Following the lecturer's advice, a legal expert with extensive experience across several legal domains and languages was contacted. Based on the criteria discussed with the lecturer, this expert compiled a selection of authentic Spanish criminal law texts, including statutory excerpts and legally relevant passages commonly encountered in professional translation contexts. In addition to the original text, a German translation by a legal expert was provided, which served as the human translation (HT) for the survey.

The following Spanish legal text passages, arranged from shortest to longest, were provided without source references. Their exact textual origin could therefore not be independently verified. The only exception is Passage 4, which closely follows the wording of Art. 237 of the Spanish Criminal Code (Código Penal 1996), albeit in a shortened form.

1. Que mediante el presente escrito vengo a formular demanda de juicio declarativo ordinario.
2. La parte demandada en el acto de Audiencia Previa se allanó totalmente a las pretensiones del actor, quedando las actuaciones vistas para sentencia.
3. No toda culpa da lugar a responsabilidad; en la culpa contractual, a diferencia de la extracontractual no rige el principio de incurrir en ella por culpa levísima.
4. Son reos del delito de robo los que, con ánimo de lucro, se apoderaren de las cosas muebles ajenas empleando fuerza en las cosas para acceder o abandonar el lugar donde éstas se encuentran o violencia o intimidación en las personas.
5. La tipicidad, la antijuridicidad y la culpabilidad son las características comunes a todo hecho delictivo. El punto de partida es siempre la tipicidad. Solo en algunos casos concretos las condiciones objetivas de penalidad o perseguibilidad también configuran un elemento autónomo de la teoría general del delito.

In order, the five passages address a statement announcing the filing of a legal claim (1), a description of a defendant's agreement with the claimant's demands during a court hearing (2), a statement distinguishing two types of legal fault and their consequences for liability (3), a definition of the criminal offense of robbery (4), and a description of the elements that must be present for an act to be considered a criminal offense (5). All passages fall within the domain of Spanish civil and criminal law. According to the expert, the texts represent practice-based legal texts of low to moderate difficulty, translatable by professional legal translators without extensive research.

Set	W ST (ES)	C ST (ES)
1	13	90
2	23	148
3	27	163
4	41	232
5	46	313

Table 1: Count of words (W), count of characters (C), source text (ST) in Spanish (ES)

Table 1 lists the word and character counts for the Spanish legal source text. The length of the Spanish source texts (ST-ES) ranges from 13 to 46 words and 90 to 313 characters, giving a total of 150 words. This text selection closely follows the methodological framework established by Hiebl and Gromann (2024) by keeping the text passages relatively short to prevent cognitive and temporal overload for participants. Given the high complexity of the legal domain, keeping participants' cognitive load low is also particularly important to consider in this survey. Although the excerpts remain rather short, averaging in 30 words, the word count of the Spanish source texts does increase gradually across the five sets. This aspect contrasts with Hiebl and Gromann's (2024) methodology, where the source texts were kept at a similar length. The length variation was not initially planned but resulted from practical constraints in the text selection process. The legal expert who provided the source texts and prepared the human translations was not informed in advance that the passages should not only be short but also have a similar length. Therefore, the final set of texts varied in word count. After considering different approaches, from arranging the texts from longest to shortest to potentially decrease cognitive load as participants progress in the survey, to arranging them from shortest to longest with the intention to provide participants with a gradual introduction to the task rather than confronting them with the longest and most demanding text at the start. It was then decided that the latter arrangement was considered to be the more desirable approach and was therefore chosen.

6.2. Translation Process for Source Text Samples

In this section, the translation process into German for each one of the Spanish source text samples is elaborated. The German translation, prepared by the same legal expert who provided the original Spanish texts, served as the reference for the human translation. The expert was explicitly asked to produce these translations for the purpose of this master's thesis. Since the translations were made available in their final form, no information on translation strategies has been provided. The expert, a former lecturer in legal and business translation for Spanish at the University of Vienna, was contacted on the recommendation of professional translators familiar with his expertise in this field. Two systems were employed for the MT output component: DeepL for the NMT translation output and OpenAI's ChatGPT 5.1 for the LLM-based translation output. Here it should be noted that both tools reflect the versions available in December 2025 when the translations for this study were generated.

6.2.1. Translation process with DeepL

Machine translation output from DeepL was generated using the free web-based version of the system. Currently, DeepL can be run on the Classic Language Model and the Next-generation Language Model. Since this master's thesis focuses on the latest generations of LLMs and NMTs available at the time of writing, the Next-generation Language Model was selected. For each translation, the language direction was manually set from Spanish to German beforehand, and the corresponding source text was then entered into the input field. DeepL's translation output was then copied directly from the interface in its unedited form and transferred into the survey. Each text was processed independently, and no glossaries or additional settings were applied.

6.2.2. Translation Process with GPT-5.1

Translations from LLMs were generated using OpenAI's GPT-5.1 model via a paid ChatGPT account. Two interaction approaches with the LLM were used for the translation process. For the naïve prompting strategy, each source text was submitted with a minimal, single-instruction prompt in German, requesting a translation from Spanish into German, with no additional instructions (see Appendix 1 for the exact prompt wording). This approach is intended to reflect casual use of an LLM for translation purposes. Each translation was carried out in a separate chat to ensure that no contextual information from previous interactions influenced the translation process or output. Within each chat, the translation request and the Spanish source text were submitted together in a single prompt. Additionally, it was ensured that the Memory settings section disabled the options Reference saved memories and Reference chat history to avoid contamination from previous chats. Specifically, Reference saved memories uses explicitly stored user information, while Reference chat history allows the model to draw on prior conversations to tailor responses; both were disabled to ensure that outputs reflect translation performance under controlled conditions (OpenAI 2026c). The procedure followed a typical interaction pattern with ChatGPT, i.e., entering the prompt and receiving the translation as an answer.

For the extended-prompt strategy, a GPT was created using ChatGPT's GPT Builder, which is a feature available in the paid Pro version. This feature can be used to customize ChatGPT versions for specific tasks or topics. Within the Configure tab of the GPT Builder, the following options are available to customize a GPT: Description, Instruction, and Knowledge.

The Description text field is used to add a description of what the GPT does. The Instructions text field provides the option to give the GPT instructions on what it should do, how it should behave, and what it should avoid doing. Knowledge allows the user to give the GPT more context to reference by uploading files (OpenAI 2026b). Of these fields, only the Instructions option was used for this experiment, as it functions as a persistent system-level directive that defines the model's behavior and guidelines across all interactions within the customized GPT. Hence, inserting the extended prompt into the Instruction text field caused that prompt to function as a system prompt. This approach to the translation process was chosen to reflect a plausible real-world use case in which a user defines the desired instructions once as a persistent prompt, rather than re-entering a lengthy prompt for each translation request, particularly for professional use. Professional use settings with LLMs further require reliable and consistent output, yet some studies (Jaffe et al. 2026; Li et al. 2024; Wu et al. 2025) mention that the consistency with which LLMs follow instructions deteriorates as input length increases. Considering these constraints for a translation task, this poses a risk of faulty or inconsistent translations over the course of a long conversation. The main reason for choosing a persistent system prompt over repeatedly entering the same lengthy prompt in an ordinary chat User Interface (UI) was therefore to keep the instructions consistently active in the background, independent of the model's ability to maintain instruction consistency within any given conversation.

The extended prompt was written in the translation's target language, German, as it is assumed that this could influence the translation quality. The prompt assigns the model the role of a specialized legal translator for German (Austria) and Spanish (Spain and Latin America), emphasizing awareness of jurisdictional and terminological variation across legal systems and regional language varieties. It further embeds key professional requirements derived from ISO 20771:2020, Legal Translation – Requirements, as discussed in Section 3.2, which include responsibility for accuracy, adherence to translation specifications, terminological consistency, and confidentiality, to name but a few (ISO 20771 2020). Furthermore, the prompt provides a structured, step-by-step instruction of tasks for the model to follow, requiring it to first analyze the entire source text and identify domain-specific terminology, then proceed with the translation while respecting the conventions of Austrian and Castilian Spanish, ensure terminological consistency, and to verify that the translation preserves the source text's legal intent. Cultural and legal differences between Spanish- and German-speaking jurisdictions must also be taken into account. In a final step, the system prompt also defines strict output restrictions for the LLM. Translations must be delivered exclusively as plain text, without

commentary, linguistic explanations, or interactive elements. Additionally, each input must be treated independently, ensuring that no contextual information is transferred between the source text and the translation output. The translation should be suitable for use in a certified translation context. See Appendix 2 for the exact prompt wording.

Once the extended prompt was entered into the Instruction text field, the customized GPT was tested with two trial inputs to verify that it correctly followed the requested output format, i.e., returning only plain text without commentary, explanations, or interactive elements. The prompt itself was not modified based on this trial run, as it only served to confirm the compliance with the output restrictions rather than to evaluate or refine translation quality. After confirming that the model behaved as instructed, the customized GPT was set for the actual translation process. The actual Spanish source text samples were then entered into GPT one by one, and the generated German translations were retrieved and added to the survey.

It is important to note that for both prompting strategies, no independent qualitative assessment of the translation outputs was conducted beyond the initial trial run for the extensive prompt, as this would have run counter to the purpose of the assessment, in which professional translators themselves evaluated the quality of the raw translation output. Instead, the output was only checked to confirm that the content had been rendered correctly from Spanish into German, and that no information from the source text had been omitted, which was confirmed to be the case for all translation outputs.

6.3. Participant Selection

Selecting participants was determined by the specialized and legal expertise required for this survey. Since interpreting the survey materials involves domain-specific terminology and discourse conventions in both German and Spanish law, only individuals with advanced professional competence in legal translation or interpreting were considered suitable. Students, bilingual laypersons, and translators without a legal specialization were therefore excluded, as their level of expertise does not fully guarantee the accuracy of their assessments of the source text and its translations.

Participants were recruited from a professional mailing list, including sworn translators and interpreters working between German and Spanish. Access to this list was granted through an established professional contact. All mailing list members received an invitation with a link to the online survey and were informed that participation was voluntary and anonymous.

To ensure the participant pool met the intended selection criteria, the survey began with a set of demographic and professional background questions. These items include questions about participants' native language(s), highest educational degree, years of professional experience, and self-evaluation of proficiency in German and Spanish. These data serve both to describe the participants' characteristics and abilities and to verify their alignment with the eligibility criteria. Only responses from participants whose profiles match the required expertise will be retained for survey analysis.

6.4. Survey Structure and Design

The survey structure is also based on the framework proposed by Hiebl and Gromann (2024). Several elements of their questionnaire structure were adopted and adapted for the present study. In particular, because this study includes not only NMT-generated translations but also LLM outputs, the original questions were revised to cover a broader range of translation types. Furthermore, the entire survey was conducted in German. The survey was created in the paid version of LimeSurvey and used a two-part structure, combining Best-Worst Scaling with a criteria-based assessment, along with optional qualitative comments. Particular attention was paid to interface design to minimize participant fatigue, further outlined in the upcoming chapters. As explained in Section 6.1, the source texts varied in length due to constraints in the text selection process. For the survey procedure, the five units were therefore arranged from the shortest to the longest passage.

6.4.1. Professional and Demographic Background

The first section of the survey was designed to validate participants' expertise, as only individuals with advanced professional competence in translation and/or interpreting (preferably in the legal domain) were considered eligible for the study. To establish this, participants were first asked to provide key background information, mostly through closed-ended items with predefined response options; where necessary, and the option Other allowed them to add a free-text response.

First, participants reported basic demographic data, including age, gender, and their highest educational degree. They were then asked about their professional qualifications, specifying their field of study. If they had a background in translation studies, they were asked to specify their language combinations (A, B, and C). To capture language competence more systematically, participants should provide a self-assessment of their Spanish and German

proficiency using the Common European Framework of Reference for Languages (CEFR) scale (e.g., C1 or C2). Professional status was further verified by asking whether participants were sworn and certified court interpreters and translators. In addition, the survey collected information on professional experience, including the number of years participants had worked in the field, with options ranging from None” to More than 8 years. Finally, the section included questions on technology use, asking how frequently participants use DeepL and LLMs such as ChatGPT, for which purposes they use these tools (professional, other business-related, or private), and how satisfied they are with their performance in practice.

6.4.2. Best-Worst Scaling with Criteria-based Assessment

Participants then proceeded to the core task, which was organized into two parts (see Figure 2) and consisted of an adapted version of the survey design proposed by Hiebl and Gromann (2024) and included four tasks. Participants were presented with a table containing five rows and three columns. The second row displayed the Spanish source text, while the following rows contained the four translation variants, which are human translation, NMT output, LLM output generated with a simple prompt, and LLM output generated with a system prompt.

*
Wählen Sie die für Sie beste und schlechteste Übersetzung in der unten stehenden Tabelle

	Beste	Schlechteste
Que mediante el presente escrito vengo a formular demanda de juicio declarativo ordinario.		
Mit diesem Schriftsatz bringe ich eine Klage im ordentlichen Erkenntnisverfahren ein.	☒	☐
Mit dem vorliegenden Schriftsatz bringe ich Klage auf Einleitung eines ordentlichen Feststellungsverfahrens ein.	☐	☐
Hiermit erhebe ich Klage im ordentlichen Feststellungsverfahren.	☐	☒
Mit diesem Schreiben reiche ich eine Klage auf Feststellung ein.	☐	☐

Figure 2: Example screenshot of the first BWS task, showing the Spanish source text and the four German translations, with radio buttons to select the best and the worst translations

To reduce cognitive load and participant fatigue, the source text and translation variants were presented in a single aligned table. This layout allowed participants to compare the Spanish source text directly with each translation and to identify differences more easily. In addition, the source-text row was visually emphasized with thicker upper and lower borders and black

text, whereas the translation variants were displayed in dark grey. This design choice was made to avoid a less efficient format in which the source text would appear separately in the question title and the translations would be shown in fixed-size answer boxes, which would compress the translation texts and make comparison more difficult due to line breaks and inconsistent formatting.

The second and third columns contained radio buttons that allowed participants to select the best translation (column 2) and the worst translation (column 3). After the best and worst options had been selected, the text of the chosen best translation was highlighted in green, while the chosen worst translation was highlighted in red.

The order of translations in each of the five sets was randomized in advance using the free randomization website RANDOM (Random 2026). Each translation mode corresponded to a specific number, used exclusively for internal randomization purposes and not visible to participants, in order to preserve the blind evaluation of the translation modes:

- Human translation = 1
- DeepL = 2
- ChatGPT with a naïve prompt = 3
- ChatGPT with an extensive prompt = 4

The numbers 1 to 4 were set as the sequence boundaries, and a randomized sequence was generated separately for each of the five sets. The resulting order was then applied consistently across all participants. Participants were not informed about the origin of the translations beforehand, that is, human translation vs. machine- or LLM-generated output. Table 2 shows the final randomized translation orders for each set:

Set	Option 1	Option 2	Option 3	Option 4
1	HT	GPT-E	GPT-N	DL
2	DL	GPT-N	HT	GPT-E
3	DL	HT	GPT-E	GPT-N
4	GPT-N	HT	GPT-E	DL
5	DL	GPT-E	GPT-N	HT

Table 2: Translation order within each set (HT = Human Translation; DL = DeepL; GPT-N = ChatGPT with a naïve prompt; GPT-E = ChatGPT with an extensive prompt)

*
Wie viele Punkte würden Sie der für Sie besten und der schlechtesten Übersetzung vergeben? (Es gibt keine richtigen oder falschen Antworten; es geht um Ihre individuelle Einschätzung)

	Beste Übersetzung ⓘ						Schlechteste Übersetzung ⓘ				
	0	1	2	3	4		0	-1	-2	-3	-4
Punktevergabe	☐	☐	☐	☒	☐		☐	☒	☐	☐	☐

Figure 3: Example screenshot of the second BWS task for assigning points to the best and worst translations

Figure 3 shows the second task where participants were required to assign points to the translation they selected as best and to the translation they selected as worst. In contrast to the survey design by Hiebl and Gromann (2024), this scoring step was implemented as a single combined question rather than two separate questions. The scoring options were presented in a table-based Likert-type rating format with two separate scales: one for the best translation (scores from 0 to 4) and one for the worst translation (scores from 0 to -4). Each scale used a single-choice response format with radio buttons, allowing participants to select exactly one score for the best and worst translations. To support consistent interpretation of the scale, additional info icons were placed next to the labels “best translation” and “worst translation”. These icons explained that, for the best translation, 0 represented the lowest score and 4 the highest, whereas for the worst translation, 0 represented the highest score and -4 the lowest.

*

Wie bewerten Sie die einzelnen Qualitätsmerkmale (Inhalt, Terminologie, Grammatik, Funktion) der **besten und schlechtesten** Übersetzung auf einer Skala von 1 bis 5?

Ausgangstext	Que mediante el presente escrito vengo a formular demanda de juicio declarativo ordinario.
Ihre gewählte beste Übersetzung	Mit diesem Schriftsatz bringe ich eine Klage im ordentlichen Erkenntnisverfahren ein.
Ihre gewählte schlechteste Übersetzung	Hiermit erhebe ich Klage im ordentlichen Feststellungsverfahren.

	Beste Übersetzung						Schlechteste Übersetzung				
	1	2	3	4	5		1	2	3	4	5
Inhalt ⓘ	☐	☐	☐	☐	☐		☐	☐	☐	☐	☐
Terminologie ⓘ	☐	☐	☐	☐	☐		☐	☐	☐	☐	☐
Grammatik und Satzbau ⓘ	☐	☐	☐	☐	☐		☐	☐	☐	☐	☐
Textfunktion ⓘ	☐	☐	☐	☐	☐		☐	☐	☐	☐	☐

Figure 4: Example screenshot of criteria-based task. Personal quality evaluation

For the criteria-based task (see Figure 4), participants were asked to provide a personal evaluation of the quality of the translations they had previously selected as the best and worst. This additional task was included to obtain more detailed insight into participants' individual perceptions of translation quality beyond the Best-Worst selection itself. As in the previous task, a table-based Likert-type matrix with two parallel rating sections, one for the selected best translation and one for the worst translation. For each of the four quality categories, which are content, terminology, grammar and syntax, and text function, participants assigned one rating on a 1-5 scale for the best translation and one rating on a 1-5 Likert scale for the worst translation. To support consistent interpretation of the quality dimensions, an additional info icon was placed next to each category label, providing a short explanation of the respective criterion.

For each of the four quality categories, participants were shown a short explanation via an info icon (ⓘ) placed next to the category label, displayed as a tooltip upon clicking or hovering over the icon. These explanations provided participants with guidance on what to consider for each category. For Content, the guidance addressed the extent to which the translation conveys the meaning correctly and completely. For Terminology, the guidance addressed how appropriate, consistent, and domain-specific the terminology used is. For Grammar and Syntax, the guidance addressed how correct, idiomatic, and readable the

translation is with regard to grammar, syntax, and linguistic fluency. For Text function, the guidance addressed the extent to which the translation fulfills the same communicative function as the source text, e.g., legally binding, informative, or appellative.

The criteria categories were partly inspired by Gutiérrez Rubio (2024), who used three quality dimensions in an MT evaluation study involving Czech-Spanish outputs from Google Translate, DeepL, and ChatGPT: fluency, grammatical correctness, and usability. Fluency refers to the naturalness of the translation in the target language. Grammatical correctness includes the coherence of the target text and is often treated in the literature as a subcomponent of fluency (Castilho et al. 2018; Gutiérrez Rubio 2024). Usability refers to the adequacy of vocabulary and style in relation to the communicative purpose of the translation (Suojanen et al. 2014). Gutiérrez Rubio (2024) intentionally excluded the dimension of accuracy, also referred to as adequacy in his study, in order to focus on the target texts' quality as perceived by non-professional end users, that is, native speakers of Spanish evaluating the translation as an independent text without reference to the source text. The present master's thesis, by contrast, reintroduces content and terminology as distinct categories to address the high requirements for factual accuracy in specialized translation, such as within the legal field.

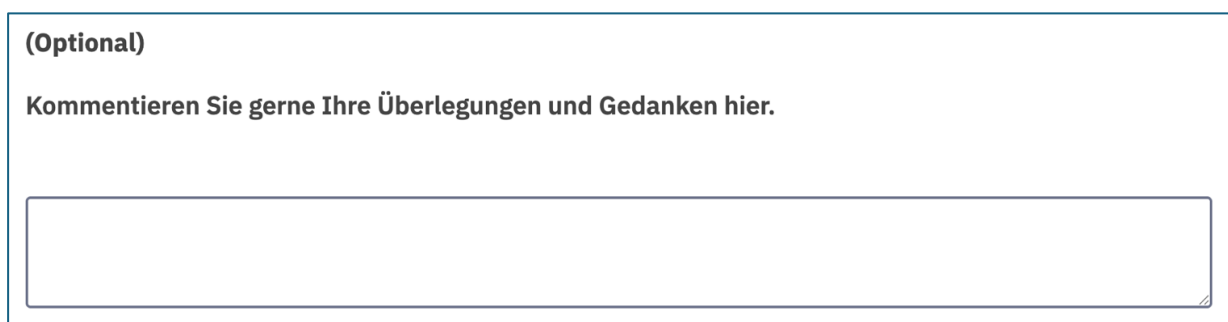
In this framework, Content and Terminology align with Gouadec's (2007) professional quality management model. Specifically, Content aligns with Gouadec's Material quality checks, which verify that the translation is complete, ensuring all relevant source material has been transferred, and that it strictly adheres to the initial project specifications. Complementary to this, the terminology category corresponds to Gouadec's Technical-factual-semantic quality checks as well as his Homogeneity and Consistency checks. Combining these checks ensures the adequacy of factual information and the precision of data while requiring that the specialized vocabulary and phraseology remain perfectly homogeneous throughout the entire translation. By including these dimensions, this survey aims to ensure high Transfer quality by accurately reflecting all key elements of the source texts in the target text and meeting professional standards, ideally as reflected in the participants' responses.

Grammar and syntax correspond directly to the dimensions of readability and linguistic fluency. These factors, including word choice and sentence structure, have a direct impact on Cognitive ergonomics, as they influence the mental workload and cognitive effort required by the reader to process information (Gouadec 2007; Suojanen et al. 2014). Within this context, grammatical and syntactic correctness serve to minimize Semantic noise, such as errors in punctuation or structure, which can disrupt the message and lead the user into unnecessary hesitation (Suojanen et al. 2014). Furthermore, as noted by Gutiérrez Rubio (2024), maintaining

grammatical correctness ensures that the text adheres to the rules and norms of the target language, thereby supporting the overall naturalness and fluency perceived by the reader.

Finally, the category Text function serves as the practical operationalization of the skopos. Grounded in the principles of User-Centered Translation (UCT), as proposed by Suojanen et al. (2014), this dimension evaluates whether the translation is fit for purpose by measuring the extent to which it enables users to achieve specified goals effectively, efficiently, and with satisfaction. Within the ISO 9241-11 (1998) framework, effectiveness is considered fundamental because it measures whether the translation enables specified users to achieve their intended goals. It is an approach that is based on the premise that reading a specialized text is an instrumental act of using. Thus, the success of the translation is determined by how well it functions as a tool for the recipient within its specific legal context. In this study, professional translators act as expert evaluators, adopting a method known in UCT as heuristic evaluation, which is conceptually equivalent to the professional practice of virtual qualification. The term heuristic refers to the use of established usability guidelines or rules of thumb that experts use to systematically identify potential obstacles in a product. By performing such a virtual qualification, the evaluator simulates the target user's perspective – metaphorically standing “in the end user's shoes” (Gouadec 2007: 82) – to assess the text's ergonomic conformity. This ensures that the translation is designed to minimize the reader's cognitive workload, allowing for a smooth interaction in which the user can achieve their purpose without unnecessary hesitation, questions, or mental effort (Gouadec 2007; Gutiérrez Rubio 2024; Suojanen et al. 2014).

An additional reference table was added above the quality criteria matrix to display the Spanish source text alongside the participant's previously selected best and worst translations. This table served as a visual reference during the assessment of quality criteria. Its inclusion was intended, again, to reduce cognitive load and participant fatigue by keeping the relevant comparison material visible at the point of rating. Without this embedded reference, participants would have had to repeatedly scroll back to earlier parts of the survey to review the source text and selected translations, likely increasing effort and disrupting the flow, focus, and thus the quality of the evaluation. More generally, the core task's interface was designed to keep as much relevant information as possible visible at a glance and minimize scrolling.



(Optional)

Kommentieren Sie gerne Ihre Überlegungen und Gedanken hier.

Figure 5: Example screenshot of text field for adding comments

At the end of each assessment task, an optional text field was provided for participants to submit comments if they desired (see Figure 5).

6.4.3. Post-Task Questionnaire

The post-task questionnaire served as a comprehensive meta-evaluation of the assessment process, designed to reflect on the participants' experience and validate the research methodology. This phase begins with four quantitative questions assessing the procedural difficulty of the task: Participants were asked to rate, on a five-point Likert scale (from 5 = very easy to 1 = very difficult), how easy or difficult it was to select the best translation, evaluate the best translation, select the worst translation, and evaluate the worst translation. To ensure that the study's framework aligns with professional reality, experts were then asked to indicate the degree to which the provided quality criteria, that is, Content, Terminology, Grammar and Syntax, and Text function, correspond to their own professional concept of translation quality. Regarding the complexity of the subject matter, the questionnaire includes an evaluation of how demanding the Spanish source texts were perceived to be in both technical and linguistic terms. Crucially, participants must explicitly state whether they consider subject-matter expertise, specifically, legal or translational knowledge, mandatory for a meaningful evaluation of the provided texts. To measure engagement and intellectual depth, participants were asked, using single-choice questions with four graded response categories, whether they felt appropriately challenged in their professional knowledge, ranging from To a high degree to Hardly or not at all, with an additional open-ended Other option. Additionally, they were asked how cognitively demanding they perceived the comparative evaluation tasks overall, on a scale from Very high to Very low. The survey assesses whether the experts felt appropriately challenged in their professional knowledge and how cognitively demanding they perceived the comparative

evaluation tasks to be overall. Furthermore, a specific question provides a subjective measure of self-reported consistency, requiring participants to reflect on how stable their judgment remained across the five individual Best-Worst sets.

6.4.4. Methodological Evaluation and Closing

The last part of the questionnaire asks participants to provide methodological and qualitative reflections and feedback. It includes open, voluntary questions about participants' prior experience with TQA for HT and/or MT, as well as their views on whether the Best-Worst scaling method is suitable for identifying quality differences. Participants were also asked which criteria or text features, e.g., authenticity or terminology, were most important for their decisions when selecting the best and the worst translations. Finally, the questionnaire concludes with feedback on specific likes or dislikes about the survey design, as well as an invitation to provide general suggestions for improvement or additional remarks.

6.5. Data Analysis

This section outlines the analytical procedures applied for BWS and criteria-based data, providing the methodological basis for interpreting the results presented in Chapter 7. For the BWS task, the analysis follows the evaluation guidelines proposed by Hiebl and Gromann (2024). The BWS survey output consisted of ratings for the translations selected as best and worst. By calculating and comparing these values across sets, it is possible to identify which translation variants were preferred overall.

First, the point ratings for the best and worst selections (positive and negative points) are summed across the five sets. These summed ratings produce a total score for each translation mode. An average score is then calculated by dividing the total score by the number of times the respective translation mode is selected by participants, denoted by n . Importantly, selections with a rating of 0 points (whether 0 points for the best or for the worst translation) are still counted as selections for this calculation. With best selections scored between 0 and 4 and worst selections scored between 0 and -4, the combined scale ranges from -4 to 4. The average score for each translation mode, therefore, also lies between -4 and 4, indicating its position on this scale.

In addition to analyzing summed and average points, the analysis examines how often each translation mode is selected as the best or worst. This is achieved by first counting the number of best and worst selections for each translation mode in the dataset. Relative

frequencies are then calculated separately for the best and the worst selections. For the best selections, the relative frequency of each translation mode is calculated by dividing the number of times that mode was chosen by the participants as best by the total number of responses, i.e., the product of the number of participants and the number of distinct text sets. Since exactly one translation mode is selected as best per set, the relative frequencies of the four translation modes sum up to 100% for the best selections. The same procedure and logic apply to the worst selections, whose relative frequencies also sum up to 100%. Converting the raw selection counts to relative frequencies allows the results to be compared across the four translation modes.

While BWS provides a holistic ranking of the translation variants, criteria-based assessment offers a more detailed analysis of the underlying quality markers. This second part is designed to identify specific strengths and weaknesses within the selected best and worst translations. As illustrated in the point distribution in Appendix 4, only the variants chosen in BWS were subjected to this detailed evaluation across the four quality categories. First, for each translation mode, the points in each category are summed across all five sets, regardless of whether the translation was selected as the best or the worst. In contrast to the BWS task, the criteria-based assessment task uses a Likert scale with only positive values, with 1 being the lowest and 5 the highest. The distinction between best and worst translations is therefore not considered at this stage, since, e.g., a translation not selected as best may nevertheless achieve higher scores in a specific category than the translation selected as best. Hence, all four category points assigned to each translation mode are summed together. To obtain the average score for each criterion category in each translation mode, the total points are divided by the number of times the translation mode was selected across all sets and participants. This results in an average score between 1 and 5, making it possible to compare the translation modes with one another and to see where each mode falls on the 1-5 scale for each criterion category. Additionally, a total score is calculated for each translation mode by summing the points across all four quality categories. Dividing by the number of evaluations yields an average total score on a scale of 4 (lowest) to 20 (highest), since each evaluation can award up to 5 points per category. Finally, set-specific totals and averages are calculated for each quality category and translation mode by summing the assigned category points and dividing by the number of selections in the respective set.

Together, these calculations standardize evaluation results, making them comparable across translation modes, sets, and participants for both BWS and criteria-based tasks. By converting raw selection counts and points totals into averages and relative frequencies, the

influence of unequal selection rates is accounted for, ensuring that no translation mode is disadvantaged solely because it was selected as best or worst less frequently than others.

7. Results

This chapter presents the results obtained from the survey conducted among professional translators and interpreters, following the two-part evaluation structure outlined in Chapter 6. It begins with an overview of the participant sample, before turning to the BWS results and, subsequently, the criteria-based assessment. At the end, an overview of the optional comments participants made throughout the evaluation process is provided.

7.1. Participants

The survey was available from December 30, 2025, to January 22, 2026. Of the 46 professionals on the mailing list to whom the survey was sent, 5 completed it and were therefore included in the analysis. It took the participants an average of 41 minutes to complete the survey. Of the five participants, two were aged between 55 and 65, two were older than 65, and one was between 35 and 44. Four participants were female, and one was male. All five participants hold a Magister's degree and have academic backgrounds in translation and/or interpreting, with specializations in conference interpreting. Based on the Common European Framework of Reference for Languages (CEFR), four participants reported C2 proficiency in Spanish, and one reported C1 proficiency. All five participants rated their German proficiency as C2. All five participants reported more than eight years of professional experience in translation and/or interpreting. Four participants were sworn and certified court interpreters and/or translators, and one holds an American Translation Association (ATA) certification. Regarding technology use, two participants reported using MT tools several times a week, while three reported never using them. Regarding LLMs, two participants used them several times a week, one used them daily, and two never used them. Participants were also asked for what purpose they used machine translation programs and/or large language models, if applicable, with three options: For professional translation, For other professional purposes, or For private use. Multiple answers were allowed. Of the five participants, one selected all three purposes, one selected both, For other professional purposes and Private use, one selected For other professional purposes only, and two selected none of the three purposes.

When asked which type of machine translation tool they prefer to use, one participant mentioned DeepL. For satisfaction with MT tool output, two reported being very unsatisfied, one reported being not very satisfied, one reported being neither satisfied nor unsatisfied, and one reported being quite satisfied.

7.2. Best-Worst Scaling Ratings

This section presents the results of the Best-Worst Scaling (BWS) task, in which participants selected and rated the best and worst translation from each of the five sets. Table 3 provides an overview of the BWS results, including summed and averaged ratings, as well as the total number of best and worst selections per translation mode.

	HT	DL	GPT-N	GPT-E
Sum Rating	32	-9	3	7
Avg. Rating	2	-0,6	0,3	0,7
Times Selected Best	12 (48%)	4 (16%)	4 (16%)	5 (20%)
Times Selected Worst	4 (16%)	11 (44%)	5 (20%)	5 (20%)
Total Selected	16	15	9	10

Table 3: BWS result overviews: average, sum, and total results

Each translation mode was selected at least once as best or worst. Regarding best selections, Human Translation (HT) was chosen most frequently, with 12 out of 25 selections (48%), followed by ChatGPT with an extensive prompt (GPT-E) with 5 selections (20%), while DeepL (DL) and ChatGPT with a naïve prompt (GPT-N) were selected equally infrequently, with 4 selections each (16%).

The pattern inverts for the worst selections, where DL was chosen most frequently, with 11 out of 25 selections (44%), followed by GPT-N and GPT-E, each with 5 selections (20%), and HT least frequently, with 4 selections (16%). Notably, GPT-N and GPT-E showed a relatively balanced distribution between best and worst selections: GPT-N was selected 4 times as best and 5 times as worst, and GPT-E was selected 5 times as both best and worst.

In terms of ratings, HT achieved the highest sum rating of 32 across $n=16$ selections, with an average of 2 points on a scale from -4 to 4, while DL was the only mode to receive a negative sum rating of -9 across $n=15$ selections, corresponding to an average of -0.6 points. Although HT and DL show comparable total selection counts of 16 and 15, their average ratings diverge considerably, indicating that DL was selected predominantly as worst rather than best. GPT-N and GPT-E ranked between these two extremes, with sum ratings of 3 and 7 across $n=9$ and $n=10$ selections, respectively, and average ratings of 0.3 (rounded) and 0.7, respectively, both remaining slightly above the neutral midpoint of the scale. Despite being selected equally often as best and worst, GPT-E achieved a positive average rating of 0.7, suggesting that participants

rated it more highly when selected as best than they rated it poorly when selected as worst. Based on both the total and average ratings, which yield identical rankings, HT ranked first, followed by GPT-E in second, GPT-N in third, and DL in last.

Set	Set 1				Set 2				Set 3				Set 4				Set 5			
Mode	HT	GPT-E	GPT-N	DL	DL	GPT-N	HT	GPT-E	DL	HT	GPT-E	GPT-N	GPT-N	HT	GPT-E	DL	DL	GPT-E	GPT-N	HT
ID 10	-2			2	-3		3			4		-2	2		-3		-2			3
ID 12		4		-4	0		4			4		0		4	0		0		4	
ID 13	-3	4			-2		4			3		-3		3	-4		-3	4		
ID 21		4		-4		3		-3	4	-3				-3		3	0		4	
ID 23		4		-3		-2	4			3		-3			-3	4	-1			4
Sum	-5	16	0	-9	-5	1	15	-3	4	11	0	-8	2	4	-10	7	-6	4	8	7
Best	0	16	0	2	0	3	15	0	4	14	0	0	2	7	0	7	0	4	8	7
Worst	-5	0	0	-11	-5	-2	0	-3	0	-3	0	-8	0	-3	-10	0	-6	0	0	0

Table 4: BWS rating details across participants, strategy, and the selected best and worst translation modes (HT = Human Translation; DL = DeepL; GPT-N = ChatGPT with a naïve prompt; GPT-E = ChatGPT with an extensive prompt)

Table 4 provides the detailed ratings of each participant across all five sets. HT and DL were selected most frequently overall, while GPT-N was not chosen as best or worst by any participant in Set 1, and GPT-E received no selections in Set 3. No translation mode consistently dominated across all five sets, though distinct patterns are identifiable for each mode. HT was the only mode to show predominantly consistent pointwise performance, receiving a total sum of -5 only in Set 1 and achieving positive results in all remaining sets. HT achieved the highest scores in Sets 2 and 3, with total sums of 15 and 11, and considerably outperformed all other translation modes in both sets. In Sets 4 and 5, HT achieved lower total sums of 4 and 7 yet ranked second among all translation modes in both sets. DL received negative sums in three of five sets, with total sums of -9 in Set 1, -5 in Set 2, and -6 in Set 5, with all five participants selecting it as the worst. In contrast, DL achieved positive sums in Sets 3 and 4, with total sums of 4 and 7, the latter exceeding HT’s total sum of 4 in Set 4. GPT-N achieved its highest total of 8 in Set 5, closely followed by HT with a total of 7 in the same set. In return, GPT-N received a total sum of -8 in Set 3, where 4 out of 5 participants selected it as worst, a score comparable to DL’s total sum of -9 in Set 1. GPT-E showed the greatest variation across all modes, with the highest set-specific total sum of 16 in Set 1, where 4 out of 5 participants selected it as best, and the lowest set-specific total sum of -10 in Set 4, where 4 out of 5 participants selected it as worst.

As these results further illustrate, the number of times a translation mode was selected does not necessarily reflect the quality of its ratings. GPT-N translations, for instance, were

selected the least frequently, with 9 selections, yet had a total sum of 3, while DL was selected 15 times but had a negative total sum of -9. The same pattern is observable between HT and GPT-E, where HT was selected 16 times and achieved a total sum of 32, while GPT-E was selected 10 times and achieved a total sum of 7.

7.3. Criteria-based Assessment

This section presents the results of the criteria-based assessment, in which participants evaluated their selected best and worst translations across four quality categories: Content (C), Terminology (T), Grammar and Syntax (G), and Text Function (TF), rated on a scale from 1 to 5 points.

Total Points	Category	Mode				Avg. Points	Category	Mode			
		HT	DL	GPT-N	GPT-E			HT	DL	GPT-N	GPT-E
	C	68	55	34	43		C	4,25	3,67	3,78	4,3
	T	63	45	29	34		T	3,94	3	3,22	3,4
	G	71	60	30	40		G	4,44	4	3,33	4
	TF	64	58	33	45		TF	4	3,87	3,67	4,5
	Total Points	266	218	126	162		Combined Avg. Score	16,63	14,54	14	16,2
							Avg. Score	4,16	3,64	3,50	4,05

Table 5: Aggregated sum and normalized average scores for quality categories

Table 5 provides an overview of total points (left) and average points (right) by criteria category for each translation mode, across all sets and participants. The data on the left side of the table for total points shows that HT achieved the highest total points across all four quality categories combined (266 points), followed by DL (218 points), GPT-E (162 points), and GPT-N (126 points); the total points only represent an intermediate result, as they are directly influenced by the number n of times each mode was selected and evaluated (HT: $n=16$, DL: $n=15$, GPT-E: $n=10$, GPT-N: $n=9$), and therefore do not yet allow for meaningful comparison across modes.

The average points presented on the right side of the table, calculated by dividing the total points by n for each criterion category, thus serve as the primary basis for the following results. Regarding the combined average scores, i.e., the sum of the four category-level averages (maximum 20 points), HT achieved the highest score of 16.63 out of 20 possible points, followed by GPT-E with 16.20 points, DL with 14.53 points, and GPT-N with 14.00 points. At the category level, the average scores reveal a more differentiated picture than the combined average scores alone suggest. In the Content category, GPT-E achieved the highest average

score of 4.30 points, followed by HT with 4.25 points, while GPT-N scored 3.78 points and DL 3.67 points. In the Terminology category, HT achieved the highest average score of 3.94 points, followed by GPT-E at 3.40 points, GPT-N at 3.22 points, and DL at 3.00 points. In the Grammar and Syntax category, HT achieved the highest average score of 4.44 points, followed by DL and GPT-E, both at 4.00 points, and GPT-N at 3.33 points. Finally, in the Text Function category, GPT-E achieved the highest average score of 4.50 points, followed by HT at 4.00 points, DL at 3.87 points, and GPT-N at 3.67 points. Across all four translation modes, Terminology received the lowest average scores among the four quality categories.

The criteria-based results reveal that no single translation mode ranked first across all four quality categories. HT achieved the highest average scores in Terminology and Grammar and Syntax, while GPT-E ranked first in Content and Text Function. Neither DL nor GPT-N achieved the highest average score in any individual category. GPT-N ranked last or second-to-last across all four categories, while DL ranked last in both Content and Terminology. In Appendix 4, the table of individual criterion-based scores shows that the upper section displays the category scores assigned by each participant to the translation modes they evaluated as best and worst within each set. The lower section provides the set-specific totals and average scores per quality category and translation mode, which serve as the basis for the following analysis. As shown in Table 5, the lower section of Table 6 first presents the total points as intermediate results, followed by the average scores below. As already described in Section 7.2, GPT-N was not evaluated in Set 1, and GPT-E was not evaluated in Set 3, as neither mode was selected as best or worst by any participant in those respective sets, which is also reflected in Table 6. Regarding the set-specific average scores, the results reveal considerable variation both across sets and across translation modes.

HT shows the greatest variation, with its lowest scores in Set 1, especially in Content at 1.50 and Text Function at 2.50, scoring a category average of 2.63 in that set, while achieving consistently high scores across all categories in Sets 2 to 5, with a category average of 4.35 across those four sets. GPT-E achieved the maximum score of 5.00 across all four categories in both Set 1 and Set 5 but scored a category average of 2.81 in Set 4, with scores of 1.75 in Terminology and 2.50 in Grammar and Syntax. DL achieved a perfect score of 5.00 in all four categories in Set 3, and an almost perfect score in Set 4, with Grammar and Syntax receiving 4.00 points, while scoring lower in Terminology in Sets 1, 2, and 5, with values of 2.50, 2.30, and 2.60, respectively. GPT-N recorded its lowest scores in Set 3, particularly in Terminology (1.75) and Content (2.75), whereas in Set 5 it achieved the maximum score of 5.00 in Content, Terminology, and Grammar and Syntax.

Mode	Avg. Points (BWS)	Avg. Points (CB)	Total ratings
HT	2	4,16	16
DL	-0,6	3,64	15
GPT (s)	0,33	3,5	9
GPT (l)	0,7	4,05	10

Table 6: Global overview of average points obtained through Best-Worst Scaling (BWS) and Criteria-based (CB) assessment, including the total number of ratings per translation mode

7.4. Participant Comments

This section presents the optional qualitative comments provided by participants after completing the questionnaire for each set. Not all participants submitted comments for every set, and in some cases, no comments were provided. The comments are intended to complement the quantitative results presented in Sections 7.2 and 7.3 by providing insight into participants' reasoning and observations. The full set-specific comments can be found in Appendix 5. Another important factor to consider is that the English translations of specific terms, whether Spanish or German, should be understood as approximations, since these terms can have different meanings in each language and legal system. In this section, participants are listed by number corresponding to their survey ID, which does not reflect the number of participants who completed this survey.

Several comments concerned terminological precision. Participant 10 questioned whether *Fahrlässigkeit* (negligence) is an adequate German equivalent for the Spanish legal term *culpa*, noting that this would require further research (Set 3). Participant 23 similarly criticized GPT-N's rendering of *quedando* as *entscheidungsreif* (to be ripe for adjudication) as a decisive factor in selecting it as the worst translation (Set 2). Participant 21 pointed out that all four translations in Set 4 mistakenly used *Diebstahl* (theft) without capturing the aspect of breaking and entering, even though the original sentence encompassed theft, robbery, and burglary at once.

A second recurring concern addressed by some participants related to syntactic and stylistic naturalness. Participant 21 criticized HT for introducing a semicolon not present in the original and for altering the sentence order (Set 2) and also described an HT translation as excessively convoluted for giving too much prominence to the concept of negligence (Set 3). The same participant characterized a GPT-N construction *dass man ... in sie verfällt* (that one lapses into

it) as entirely fabricated (Set 3). They also criticized DL for the stylistically weak use of *anwenden* (to apply) that appeared twice in the same sentence (Set 4), and considered GPT-N's use of parentheses inappropriate, as none appeared in the original, suggesting a footnote as a preferable alternative (Set 4).

In return, Participant 23 highlighted translations perceived as particularly natural and convincing. They noted that GPT-E's translation sounded natural and legally precise and could realistically appear in the corresponding legal documents, whereas DL was considered unsuitable in its word choice and phrasing, though not outright incorrect (Set 1). They also found HT the most convincing and realistic of the four options in Set 3.

A few comments addressed the assessment process itself rather than individual translations. Participant 23 noted that the translations were evidently intended for different countries and legal systems, which made the evaluation considerably more difficult (Set 2). Participant 10 reported being unable to identify any satisfactory translation variant (Set 4), and Participant 21 observed that the example sentences in Set 5 differed only marginally from one another.

8. Discussion

This chapter interprets and contextualizes the results presented in Chapter 7 in relation to the two research questions and the assumptions established in Section 1.1. The first research question asked which translation modes translation professionals rated as best and worst, while the second examined the differences in translation quality they identified across the four quality categories. With regard to the first research question, it was assumed that HT would be rated the best translation overall and that longer, more detailed prompts would yield higher LLM translation quality than shorter ones. Regarding the second research question, it was assumed that the criteria-based assessment results would not simply mirror the overall rating but would reveal a more nuanced picture of strengths and weaknesses across the four categories. The results are also analyzed in the context of the related work reviewed in Chapter 5, with the aim of situating the findings within the broader research landscape of MT- and LLM-based translation criteria-based assessment in specialized domains.

The survey results from the BWS task show a tendency towards the first assumption that HT was rated higher than automatically produced translations. HT was rated the best translation overall, receiving the highest sum and average ratings across all five sets and being selected as best most frequently. The finding is consistent with studies in the legal domain that have identified persistent weaknesses in MT output at the terminological and contextual levels, suggesting that MT systems continue to fall short of the professional standards in specialized text types (Castro-Lesching 2025; Otal & Limbach 2026; Wiesmann 2019). The qualitative comments provided by participants offered further insight into this result, as one participant explicitly noted that the selected translation sounded more natural, was more legally precise, and could realistically appear in the corresponding legal documents. Although it is only one comment, it could still suggest that professional translators evaluated the translations not only on linguistic grounds but also on their perceived authenticity within the legal register. This criterion extends beyond the capabilities of automatic metrics, which are able to capture surface-level linguistic accuracy but cannot assess whether a translation conforms to the conventions and register expected of authentic legal documents.

For DeepL, the results in the BWS task also present a noteworthy pattern. Despite being selected almost as frequently as HT, with 16 selections, DeepL, with 15 selections, was the only translation mode to receive a negative overall total rating of -9, indicating that it was selected predominantly as the worst translation. One participant comment supports this interpretation,

noting that the translation (referring to DeepL's output), while not incorrect per se, was considered unsuitable in terms of word choice and phrasing for the given legal context, pointing to a recurring challenge for NMT systems in specialized domains, as Wiesmann (2019), Castro-Lesching (2025), and San-Valdivieso and López-Arroyo (2023) have also concluded – namely, producing output that is linguistically acceptable but falls short of the register and precision requirements of professional legal translation. The criteria-based results, however, add nuance to this picture. DeepL achieved positive scores in Sets 3 and 4 and even outperformed HT in Set 4 on the total sum rating, suggesting that the overall ranking was not entirely stable across all five sets. Whether this variation is related to text length, which increases gradually across the sets, or to other textual factors cannot be determined from the present data alone.

The results for GPT-based translations show that the elaborateness of the prompt affected translation quality. GPT-E consistently outperformed GPT-N across BWS, with a higher sum rating of 7 compared to GPT-N's sum rating of 3. Since both translations were generated using the same underlying model, GPT-5.1, the observed difference can be partially attributed to differences in the prompting design. Though this and the further mentioned findings related to GPT-N and GPT-E translation performance need to be interpreted with certain caution, as is addressed in the limitations found later in this section. Nevertheless, this finding does partially confirm the second assumption of the present master's thesis and is consistent with prior research showing that domain-specific and contextually richer prompts yield better translation output than naïve prompts, as demonstrated by Gao et al. (2024) and Zhang and Fan (2025).

At the same time, both GPT outputs ranked between HT and DeepL, suggesting that while prompt engineering can improve LLM-based translation quality, it does not yet consistently match HT in the legal domain. Notably, GPT-N outperformed DeepL despite receiving only a naïve prompt, suggesting that even minimally prompted LLM-based translation surpassed the NMT system in the professional translators' judgment. GPT-E's strongest performance occurred in Set 1, where four out of five participants selected it as the best translation. It is worth noting, though, that Set 1 contains the shortest source text of all five sets, which may have made the task more manageable for the model. In Set 4, by contrast, four out of five participants selected GPT-E as the worst translation, which in addition received no selections at all in Set 3. GPT-N showed a comparable pattern of instability, achieving its highest score in Set 5 and its lowest in Set 3, where four out of five participants selected it as the worst, while receiving no selections in Set 1. Taken together, both GPT outputs were absent from participant selections in one set each, and both were selected as the worst translation in at least one set, pointing to a degree of inconsistency across the five texts that neither prompting strategy fully

overcame. This degree of variability may partly reflect the inherent non-determinism of LLMs, which can produce markedly different outputs even under controlled conditions, as noted in Chapter 6.

The criteria-based assessment results confirm the second assumption for this master's thesis, revealing a more nuanced picture of translation quality than the holistic BWS rankings alone suggest. The most prominent finding in the criteria-based assessment is that GPT-E outperformed HT in the categories Content and Text Function, despite ranking second overall in BWS. One possible explanation for this pattern could lie in the design of the extensive prompt itself, which explicitly instructed the model to analyze domain-specific terminology, preserve the legal intent of the source text, and ensure that the translation fulfills the same communicative function as the original. These specifications in the prompt might have had a measurable effect on the dimensions most directly related to functional adequacy, namely Content and Text Function, while HT retained its average in the more linguistically oriented categories of Terminology, and Grammar and Syntax. This interpretation, though, is rather speculative, as the present study did not test individual prompt elements in isolation, and the observed differences may also partly reflect the fact that GPT-E was selected considerably less often than HT, with 10 selections compared to 16. With only five participants, individual high ratings carry considerable weight, which limits the generalizability of this finding.

Across all four translation modes, Terminology received the lowest average scores of all quality categories. Even HT, which achieved its highest scores in Grammar and Syntax, recorded its lowest category average in Terminology. This finding aligns with Wiesmann (2019), Castro-Lesching (2025), and Feng (2025), who similarly identified terminological accuracy as the most persistent challenge in legal translation. If terminology already poses difficulties for human translators, this challenge is likely even greater for general-purpose MT systems and LLMs, where domain-specific and jurisdiction-specific knowledge cannot be assumed.

A further observation concerns the divergence between BWS and criteria-based rankings. While DeepL ranked last in BWS, it ranked third in the combined criteria-based average, ahead of GPT-N. One possible interpretation is that DeepL produced translations that were perceived as less convincing globally in direct comparison but nonetheless performed adequately on specific quality dimensions. In particular, DeepL outperformed GPT-N in Grammar and Syntax and Text Function, while GPT-N achieved higher scores in Content and Terminology. However, the differences between the two translation modes across all four categories were minimal, and the comparison is further limited by the fact that DeepL was chosen much more

often than GPT-N, with 15 selections compared to 9. The limitation is comparable to the frequency with which HT and GPT-E were chosen, making a direct comparison of scores less robust. This divergence between BWS and criteria-based rankings nonetheless illustrates that BWS and criteria-based assessment capture different aspects of translation quality and should be understood as complementary rather than interchangeable evaluation tools. By introducing criteria-based assessment as a complementary layer, the present study aimed to capture not only overall preferences but also the specific strengths and weaknesses of each translation mode across defined quality dimensions.

The post-survey responses offer additional insight into participants' experience of the evaluation tasks. The difficulty of selecting the best and worst translations varied considerably among participants, with some finding the task very difficult and others finding it relatively straightforward. This variability may partly reflect differences in participants' prior experience with translation quality assessment, as responses ranged from very extensive experience with both HT and MT evaluation to little or no prior experience with MT assessment. Selecting the worst translation was generally perceived as somewhat easier than selecting the best, which may reflect that clear deficiencies are more readily identifiable than subtle differences in overall quality. The difficulty of assigning point ratings to the selected translations followed a similar pattern, and one participant noted that they had consistently assigned 0 points to the worst translation because the meaning of the negative point scale had been unclear to them.

The quality categories and their definitions for criteria-based assessment were well received, with all five participants rating them as either very accurate or accurate in reflecting their professional understanding of translation quality. Participant 23 explicitly noted that the precision of the criteria was an improvement over previous evaluation tasks, where criteria had often been vaguely defined, and Participant 10 commented that the overall structure of the survey was well designed. All five participants considered domain-specific expertise strictly necessary for a meaningful assessment, and four of them reported feeling highly challenged in their professional knowledge. The cognitive demand of the task was rated as high or very high by all participants, consistent with their perception of the source texts as either rather demanding or very demanding in both technical and linguistic terms.

Four participants described their judgments as fairly consistent across the five sets, while one reported feeling rather inconsistent. To a certain degree, this may reflect similarities among some translation variants, as Participant 21 explicitly noted in Set 5 that the translations differed only marginally from one another, making it difficult to draw clear distinctions. A related observation was made by Participant 23, who noted that the translations were clearly aimed at

different legal systems and regional varieties of German, making the assessment more difficult. This observation is methodologically relevant, as the extensive prompt and the human translation explicitly targeted Austrian German and the Austrian legal context, whereas the naïve prompt and DeepL did not include such specifications. The possible terminological and stylistic variations across the four translation modes may therefore have introduced an additional source of difficulty for participants. Also, the instructions for this survey did not specify the variety of the Spanish source text or the German translations for the participants to consider during the assessment.

In retrospect, the high cognitive demand reported by participants may also partly be attributable to a specific aspect of the survey design. In the criteria-based assessment, the quality criteria ratings for the best and worst translations were presented side by side rather than sequentially, making more response fields visible at once. While this layout was intended to minimize scrolling and reduce cognitive load, it may have had the opposite effect for some participants, potentially contributing to the high dropout rate observed among respondents. Of the 23 who started the survey, 18 submitted incomplete responses, and only 5 completed the entire survey.

Further limitations need to be explicitly addressed. First, both GPT-based prompt conditions did not, as initially intended, differ only in length and specificity but also in terms of how the prompts were delivered to the model. The naïve prompt was entered to simulate an ad hoc written user prompt for each translation within the standard ChatGPT chat UI, whereas the extensive prompt was entered as a permanent system-level instruction (see Section 6.2.2). Thus, a discrepancy with the initially formulated research question should be pointed out, as prompt content elaborateness alone was not the only factor differentiating the two conditions. Since LLMs most certainly process prompt content differently at the system level rather than at the user level, the observed discrepancy in translation quality could, if at all, be attributed only in part to prompt extensiveness. At the same time, as discussed in Section 6.2.2, this approach can be understood as investigating two different types of interactions with LLMs, that is, prompts freshly entered for each request, as opposed to a single, one-time instruction applied automatically to every subsequent input without the need to be restated. This structural difference should be taken into account when interpreting the results for the GPT-based translations. Secondly, the criteria-based assessment offered additional nuance to the overall ratings from BWS but did not clarify which factors in the source texts and their translations influenced participants' ratings. Such reasoning was only accessible when participants voluntarily provided qualitative comments, though these comments did not consistently address

the individual quality categories. Some instead reflected on the BWS selection itself rather than on the criteria-based ratings. However, adding such a task would increase the length and cognitive demand of the survey further, and adding the criteria-based assessment alone had already taken it considerably beyond BWS's original aim of enabling quick, intuitive quality judgments, as it required more time and consideration. Third, the survey was administered online rather than in a controlled, in-person setting, meaning participation depended entirely on whether a sufficient number of professionals chose to participate. Fourth, the source texts across the five sets varied in length. This is likely because the request to the legal expert did not explicitly specify a need for comparable length across all texts. For this reason, text length was arranged to increase progressively across the five sets, as discussed in Section 6.1. Varying text length, though, is a considerable limitation, since a higher word count increases cognitive demand placed on participants. Moreover, a certain degree of fatigue over the course of the survey is to be expected regardless of text length, simply due to the repeated nature of the assessment task across five sets. A progressively increasing text length may have amplified this effect further.

Altogether, these findings address both research questions of the present master's thesis and offer an exploratory insight into the relative strengths and weaknesses of human translation, NMT, and LLM-based translation in the Spanish-to-German legal domain. Given the small sample size of only five participants and the methodological constraints discussed, the results need to be interpreted with caution. Nonetheless, they contribute to the growing body of research on evaluating MT and LLM outputs in specialized domains.

9. Conclusion

This master's thesis investigated the translation quality of human translation, NMT, and two LLM-based translation outputs in the Spanish-to-German criminal-law domain. A two-part evaluation approach was used, combining Best-Worst Scaling (BWS) with a criteria-based assessment. With respect to the first research question, the results show that human translation, overall, was selected as the best among the four translation modes, while DeepL, selected almost as often as human translation, received the lowest overall rating. The assumption that human translation would be perceived as the best overall translation can therefore be confirmed. Notably, the ChatGPT output with the naïve prompt surpassed DeepL's translation in BWS ratings. This could indicate that even without a domain-specific prompting strategy, LLM-based translations are perceived as more adequate by professional translators than those generated by general-purpose NMT systems such as DeepL. Furthermore, the extensive prompt, which combined a persona, contextual background information, and a structured, step-by-step task description, yielded results that slightly exceeded those of the naïve prompt. However, because the extensive prompt was implemented as a persistent system-level instruction rather than a user prompt, this advantage cannot be attributed solely to prompt specificity.

Regarding the second research question, the criteria-based assessment revealed that none of the four translation modes achieved first place in all four quality categories. Human translation achieved the highest average score in the categories Terminology and Grammar and Syntax, while the ChatGPT output with the extensive prompt achieved first place in the categories Content and Text Function. Neither DeepL nor ChatGPT, with the naïve prompt, achieved a leading position. As assumed, the results confirm that the ranking from criteria-based assessment does not reflect the holistic ranking from BWS. This also becomes evident in the case of DeepL, which ranked last in BWS but moved up to third place when considering the combined average score in the criteria-based assessment. The results show that BWS and criteria-based assessment are complementary dimensions of translation quality assessment and cannot be used interchangeably. Another overarching insight was that the Terminology category was rated the lowest across all translation modes, including human translation, indicating that terminological precision keeps being an ongoing challenge in legal translation.

The proceedings and findings of this master's thesis' survey provide new possibilities for future research. Future studies could replicate the design with two prompts delivered through

the same mechanism, either both as user prompts or system-level prompt instructions within a custom GPT, to isolate the effect of prompt content from that of where the prompt was entered. Related to this, given the pace at which LLMs continue to be developed, findings based on a specific model and a specific point in time may not hold for later model versions, thus making periodic replication a worthwhile pursuit. The present study also relied on ChatGPT as one of the most widely used LLMs. Future work could extend the comparison to other LLMs. It could also compare general-purpose systems, such as those used in the present study, to models specifically fine-tuned for legal translation, whether for NMT, LLM-based translation or with each other. At a broader level, this area of research offers a wide range of possible comparisons. These include NMT versus LLM-based translation, as well as LLM-based translations generated with different prompting strategies. Moreover, other language combinations, including combinations involving lower-resource languages and other legal or non-legal specialized domains, can be assessed. Finally, a related consideration for future research concerns the number of sets included in the survey. While a larger number of sets would allow for more robust conclusions, it would also further increase the cognitive load on participants, particularly given the combined BWS and criteria-based approach used in this study, which is inherently more demanding than BWS alone. A greater number of sets may therefore be more feasible if BWS is used on its own, without the additional criteria-based layer, though this would come at the cost of the insight into participants' reasoning that the criteria-based assessment provides. Alternatively, conducting the survey in a controlled, in-person setting with scheduled breaks could help offset the added cognitive fatigue of a larger number of sets. An in-person setting would also make it easier to explain the survey directly to participants and answer any questions on the spot, rather than leaving participants to interpret unclear task elements on their own. A factor that future research could consider is developing and optimizing the survey design itself; for instance, testing an analog rather than a digital format could make the combined BWS and criteria-based assessment more manageable for participants.

10. Bibliography

- Ali, Wazir & Pyysalo, Sampo (2024). A Survey of Large Language Models for European Languages. CoRR/abs/2408.15040.
- ALPAC (1966). *Language and Machines: Computers in Translation and Linguistics*. Washington, DC: The National Academies Press. DOI: 10.17226/9547.
- Baer, Brian James & Hofeneder, Philipp (2025). A tale of two Skopos theories: (Re-)siting translation theory. *Target. International Journal of Translation Studies*, 37 (3), 309–332. DOI: 10.1075/target.23140.hof.
- Bajčić, Martina & Golenko, Dejana (2024). Applying Large Language Models in Legal Translation: The State-of-the-Art. *International Journal of Language & Law (JLL)*, 13, 171–196. DOI: 10.14762/jll.2024.171.
- Baker, Mona (1992). *In other words: a coursebook on translation*. London: Routledge.
- Banerjee, Satanjeev & Lavie, Alon (2005). *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*. In: Goldstein, Jade, Lavie, Alon, Lin, Chin-Yew, & Voss, Clare (Eds.) *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics. 65–72.
- Bhatia, V. K. (2006). Legal Genres. In: *Encyclopedia of Language & Linguistics*. Elsevier. 1–7. DOI: 10.1016/B0-08-044854-2/04505-3.
- Biel, Łucja; Scott, Juliette & O’Shea, John (2024). Chapter 14 Legal translator profiles. In: Massey, Gary, Ehrensberger-Dow, Maureen, & Angelone, Erik (Eds.) *Handbook of the Language Industry*. De Gruyter. 321–348. DOI: 10.1515/9783110716047-015.
- Block, Pascal; Schaefer, Johanna; Maurer, Felix & Storf, Holger (2025). Quality of Machine Translations in Medical Texts: An Analysis Based on Standardised Evaluation Metrics. In: Röhrig, Rainer, Ganslandt, Thomas, Jung, Klaus, Kock-Schoppenhauer, Ann-Kristin, König, Jochem, Sax, Ulrich, Sedlmayr, Martin, Spreckelsen, Cord, & Zapf, Antonia (Eds.) *Studies in Health Technology and Informatics*. IOS Press. DOI: 10.3233/SHTI251380.
- Bojar, Ondřej; Chatterjee, Rajen; Federmann, Christian; Graham, Yvette; Haddow, Barry; Huang, Shujian; Huck, Matthias; Koehn, Philipp; Liu, Qun; Logacheva, Varvara; Monz, Christof; Negri, Matteo; Post, Matt; Rubino, Raphael; Specia, Lucia & Turchi, Marco (2017). *Findings of the 2017 Conference on Machine Translation (WMT17)*. In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics. 169–214. DOI: 10.18653/v1/W17-4717.
- Briva-Iglesias, Vicent; Camargo, Joao Lucas Cavaleiro & Dogru, Gokhan (2024). Large Language Models “Ad Referendum”: How Good Are They at Machine Translation in the Legal Domain? (Version 1). CoRR/abs/2402.07681.
- Broeck, R. (1978). The concept of equivalence in translation theory. Some critical reflections. *Literature and Translation, Leuven: Academic*, 29–47.

- Brown, Tom; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared D.; Dhariwal, Prafulla; Neelakantan, Arvind; Shyam, Pranav; Sastry, Girish; Askell, Amanda; Agarwal, Sandhini; Herbert-Voss, Ariel; Krueger, Gretchen; Henighan, Tom; Child, Rewon; Ramesh, Aditya; Ziegler, Daniel; Wu, Jeffrey; Winter, Clemens; Hesse, Chris; Chen, Mark; Sigler, Eric; Litwin, Mateusz; Gray, Scott; Chess, Benjamin; Clark, Jack; Berner, Christopher; McCandlish, Sam; Radford, Alec; Sutskever, Ilya & Amodei, Dario (2020). *Language Models are Few-Shot Learners*. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., & Lin, H. (Eds.) *Advances in Neural Information Processing Systems*. Curran Associates, Inc. Vol. 33. 1877–1901.
- Buyse, Kris & Verlinde, Serge (2013). Possible Effects of Free on Line Data Driven Lexicographic Instruments on Foreign Language Learning: The Case of Linguee and the Interactive Language Toolbox. *Procedia - Social and Behavioral Sciences*, 95, 507–512. DOI: 10.1016/j.sbspro.2013.10.675.
- Callison-Burch, Chris; Osborne, Miles & Koehn, Philipp (2006). *Re-evaluating the Role of Bleu in Machine Translation Research*. In: McCarthy, Diana, & Wintner, Shuly (Eds.) *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy: Association for Computational Linguistics. 249–256.
- Casacuberta Nolla, Francisco & Abril Peris, Peris (2017). Neural Machine Translation. *Tradumàtica Tecnologies de La Traducció*, (15), 66–74. DOI: 10.5565/rev/tradumatica.203.
- Castilho, Sheila; Doherty, Stephen; Gaspari, Federico & Moorkens, Joss (2018). Approaches to Human and Machine Translation Quality Assessment. In: Castilho, Sheila, Gaspari, Federico, & Doherty, Stephen (Eds.) *Translation Quality Assessment*. Cham: Springer International Publishing. Vol. 1. 9–38. DOI: 10.1007/978-3-319-91241-7_2.
- Castro-Lesching, Wilma Eleonore (2025). Kontrastive Analyse maschinell generierter Fachtranslate (EN-DE) im Strafrecht: DeepL und ChatGPT im Vergleich. *Lebende Sprachen*, 70 (1), 92–127. DOI: 10.1515/les-2025-0024.
- ChapsVision (2026). <https://www.chapsvision.com/about-us/> (Stand: 4.2.2026).
- Chatzikoumi, Eirini (2019). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26 (2), 137–161. DOI: 10.1017/S1351324919000469.
- Cho, Kyunghyun; Van Merriënboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger & Bengio, Yoshua (2014). *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation*. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics. 1724–1734. DOI: 10.3115/v1/D14-1179.
- Código Penal (1996). *BOE-A-1995-25444 Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal*. (Stand: 17.7.2026).
- Conneau, Alexis; Khandelwal, Kartikay; Goyal, Naman; Chaudhary, Vishrav; Wenzek, Guillaume; Guzmán, Francisco; Grave, Edouard; Ott, Myle; Zettlemoyer, Luke & Stoyanov, Veselin (2020). *Unsupervised Cross-lingual Representation Learning at Scale*. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

- Cuadrado Rey, Analía (2022). Estudio de la variación fraseológica diatópica del español de España y de Argentina basado en corpus de sentencias: propuestas de traducción al alemán. *AUC PHILOLOGICA*, 2021 (2), 173–187. DOI: 10.14712/24646830.2021.30.
- Dam, Helle V. & Korning Zethsen, Karen (2008). Translator Status: A Study of Danish Company Translators. *The Translator*, 14 (1), 71–96. DOI: 10.1080/13556509.2008.10799250.
- Daum, Ulrich & Salán García, María Engracia (2016). *Spanische Rechtssprache : ein Studien- und Arbeitsbuch*. München: Verlag C.H. Beck. 3. Auflage.
- DeCamp, Jennifer (2023). A History of Translation Technology in the United States. In: Sinwai, Chan *Routledge Encyclopedia of Translation Technology*. London: Routledge. 2nd ed. 421–449. DOI: 10.4324/9781003168348-27.
- DeepL (2018). *Press Information*.
<https://web.archive.org/web/20180906073819/https://www.deepl.com/press.html>
(Stand: 15.7.2026).
- DeepL (2022). *DeepL Is Hiring | Our Current Job Offers*.
<https://web.archive.org/web/20220310163154/https://www.deepl.com/jobs.html>
(Stand: 6.2.2026).
- DeepL (2025). *DeepL Unveils Next-Gen Language AI tools and Game-Changing Agentic Productivity Solution*. https://www.deepl.com/en/press-release/deepl_unveils_next-gen_language_ai_tools_and_game-changing_agentic_productivity_solution (Stand: 7.2.2026).
- DeepL (2026). *DeepL AI Platform: Translation, Voice & Agentic Workflows*.
<https://www.deepl.com> (Stand: 6.2.2026).
- Du, Xiaoyan (2012). A Brief Introduction of Skopos Theory. *Theory and Practice in Language Studies*, 2 (10), 2189–2193. DOI: 10.4304/tpls.2.10.2189-2193.
- Durrani, Nadir; Schmid, Helmut; Fraser, Alexander; Koehn, Philipp & Schütze, Hinrich (2015). The Operation Sequence Model—Combining N-Gram-Based and Phrase-Based Statistical Machine Translation. *Computational Linguistics*, 41 (2), 185–214. DOI: 10.1162/COLI_a_00218.
- Eloundou, Tyna; Manning, Sam; Mishkin, Pamela & Rock, Daniel (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384 (6702), 1306–1308. DOI: 10.1126/science.adj0998.
- European Commission (2026). *AI translation and language tools - Multilingualism, translation and language-based AI services*. https://translation.ec.europa.eu/tools-and-resources/ai-translation-and-language-tools_en (Stand: 15.7.2026).
- F43 Committee (2014). F2575–14 standard guide for quality assurance in translation. West Conshohocken: ASTM International. DOI: 10.1520/F2575-14.
- Fellbaum, Christiane (1998). *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.
- Feng, Hongxia (2025). Machine translation outputs in Spanish–Chinese legal translation: comparisons of different legal genres and domestic vs. non-domestic systems, error analysis and implications for post-editing. *Revista de Llengua i Dret*, 114–142. DOI: 10.58992/rld.i83.2025.4368.

- Feretzakis, Georgios; Vagena, Evangelia; Kalodanis, Konstantinos; Peristera, Paraskevi; Kalles, Dimitris & Anastasiou, Athanasios (2025). GDPR and Large Language Models: Technical and Legal Obstacles. *Future Internet*, 17 (4), 151. DOI: 10.3390/fi17040151.
- Fernández Torné, Anna & Matamala, Anna (2021). Human evaluation of three machine translation systems: from quality to attitudes by professional translators. *Vigo International Journal of Applied Linguistics*, (18), 123–148. DOI: 10.35869/vial.v0i18.3366.
- Freitag, Markus; Foster, George; Grangier, David; Ratnakar, Viresh; Tan, Qijun & Macherey, Wolfgang (2021). Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9, 1460–1474. DOI: 10.1162/tacl_a_00437.
- Gao, Yuan; Wang, Ruili & Hou, Feng (2024). *How to Design Translation Prompts for ChatGPT: An Empirical Study*. In: *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*. Auckland New Zealand: ACM. 1–7. DOI: 10.1145/3700410.3702123.
- Garvin, Paul L. (1967). The Georgetown-IBM Experiment of 1954: An Evaluation in Retrospect. In: Austin, William Mandeville (Ed.) *Papers in linguistics in honor of Léon Dostert*. De Gruyter. 46–56. DOI: 10.1515/9783111675886-006.
- Gaspari, Federico (2024). The History of Translation Technologies. In: Lange, Anne, Monticelli, Daniele, & Rundle, Christopher *The Routledge Handbook of the History of Translation Studies*. London: Routledge. 1st ed. 324–338. DOI: 10.4324/9781032690056-22.
- Gaspari, Federico & Hutchins, John (2007). *Online and free! Ten years of online machine translation: origins, developments, current use and future prospects*. In: Maegaard, Bente (Ed.) *Proceedings of Machine Translation Summit XI: Papers*. Copenhagen, Denmark.
- Giammarresi, Salvatore & Lapalme, Guy (2016). Computer science and translation: Natural languages and machine translation. In: *Benjamins Translation Library*. The Netherlands: John Benjamins Publishing Company. Vol. 126. 205–224.
- Gladkoff, Serge; Sorokina, Irina; Han, Lifeng & Alekseeva, Alexandra (2022). *Measuring Uncertainty in Translation Quality Evaluation (TQE)*. Paper presented at the “Thirteenth Language Resources and Evaluation Conference,” Marseille, France. 1454–1461. DOI: 10.63317/5oye6bgfxmey.
- Google Translate (2026). *Google Translate*. <https://translate.google.com/?sl=it&tl=en&op=translate> (Stand: 4.2.2026).
- Gouadec, Daniel (2007). *Translation as a Profession*. Amsterdam: John Benjamins Publishing Co. null Vol. 00073.
- Gu, Xinxing (2023). *New features make Translate more accessible for its 1 billion users*. <https://blog.google/products-and-platforms/products/translate/new-features-make-translate-more-accessible-for-its-1-billion-users/> (Stand: 5.2.2026).
- Guerreiro, Nuno M.; Rei, Ricardo; Stigt, Daan van; Coheur, Luisa; Colombo, Pierre & Martins, André F. T. (2024). xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Transactions of the Association for Computational Linguistics*, 12, 979–995. DOI: 10.1162/tacl_a_00683.

- Gutiérrez Rubio, Enrique (2024). Aproximación al análisis de la calidad de la traducción automática desde la perspectiva de la evaluación basada en el usuario : el caso de DeepL, Google Translate y ChatGPT en la combinación checo-español. *Études romanes de Brno*,(4), 65–86. DOI: 10.5817/ERB2024-4-4.
- He, Sui (2024). *Prompting ChatGPT for Translation: A Comparative Analysis of Translation Brief and Persona Prompts*. In: Scarton, Carolina, Prescott, Charlotte, Bayliss, Chris, Oakley, Chris, Wright, Joanna, Wrigley, Stuart, Song, Xingyi, Gow-Smith, Edward, Bawden, Rachel, Sánchez-Cartagena, Víctor M., Cadwell, Patrick, Lapshinova-Koltunski, Ekaterina, Cabarrão, Vera, Chatzitheodorou, Konstantinos, Nurminen, Mary, Kanojia, Diptesh, & Moniz, Helena (Eds.) *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*. Sheffield, UK: European Association for Machine Translation (EAMT). 316–326.
- Hiebl, Bettina & Gromann, Dagmar (2024). *Comparative Quality Assessment of Human and Machine Translation with Best-Worst Scaling*. In: Scarton, Carolina, Prescott, Charlotte, Bayliss, Chris, Oakley, Chris, Wright, Joanna, Wrigley, Stuart, Song, Xingyi, Gow-Smith, Edward, Bawden, Rachel, Sánchez-Cartagena, Víctor M., Cadwell, Patrick, Lapshinova-Koltunski, Ekaterina, Cabarrão, Vera, Chatzitheodorou, Konstantinos, Nurminen, Mary, Kanojia, Diptesh, & Moniz, Helena (Eds.) *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*. Sheffield, UK: European Association for Machine Translation (EAMT). 507–536.
- House, Juliane (2015). *Translation Quality Assessment*. Routledge. 1st ed. DOI: 10.4324/9781315752839.
- Hutchins, W. John (2000). *Early Years in Machine Translation. Memoirs and biographies of pioneers*. Amsterdam: John Benjamins Publishing Co. null Vol. 00097.
- Hutchins, W. John (2001). Machine Translation over fifty years. *Histoire Épistémologie Langage*, 23 (1), 7–31. DOI: 10.3406/hel.2001.2815.
- Hutchins, W. John (2004). Two precursors of machine translation: Artsrouni and Trojanskij. *International Journal of Translation*, 16 (1), 11–34.
- Hutchins, W. John & Somers, Harold L. (1997). *An introduction to machine translation*. London: Academic Press. 2. printing.
- ISO 9241-11 (1998). Ergonomic requirements for office work with visual display terminals (VDTs) — Part 11: Guidance on usability. Geneva: International Organization for Standardization.
- ISO 17100 (2015). ISO 17100:2015 Translation services — Requirements for translation services. International Organization for Standardization.
- ISO 20771 (2020). ISO 20771:2020 Legal translation — Requirements. International Organization for Standardization.
- Jaffe, Andrew; Reicin, Noah & Choi, Jinho D. (2026). RIFT: Reordered Instruction Following Testbed To Evaluate Instruction Following in Singular Multistep Prompt Structures. CoRR/abs/2601.18924.
- Jiao, Wenxiang; Wang, Wenxuan; Huang, Jen-Tse; Wang, Xing; Shi, Shuming & Tu, Zhaopeng (2023). *Is ChatGPT A Good Translator? A Preliminary Study*. CoRR/abs/2301.08745.

- Katan, David (2009). Occupation or profession: A survey of the translators' world. *Translation and Interpreting Studies*, 4 (2), 187–209. DOI: 10.1075/tis.4.2.04kat.
- Kenny, Dorothy (2018). Machine Translation. In: *The Routledge Handbook of Translation and Philosophy*. Routledge. 1st ed. 428–445. DOI: 10.4324/9781315678481.
- Kocmi, Tom; Bawden, Rachel; Bojar, Ondřej; Dvorkovich, Anton; Federmann, Christian; Fishel, Mark; Gowda, Thamme; Graham, Yvette; Grundkiewicz, Roman; Haddow, Barry; Knowles, Rebecca; Koehn, Philipp; Monz, Christof; Morishita, Makoto; Nagata, Masaaki; Nakazawa, Toshiaki; Novák, Michal; Popel, Martin & Popović, Maja (2022). *Findings of the 2022 Conference on Machine Translation (WMT22)*. In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. 1–45. DOI: 10.18653/v1/2022.wmt-1.1.
- Kocmi, Tom; Zouhar, Vilém; Avramidis, Eleftherios; Grundkiewicz, Roman; Karpinska, Marzena; Popović, Maja; Sachan, Mrinmaya & Shmatova, Mariya (2024). *Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation*. In: *Proceedings of the Ninth Conference on Machine Translation*. Miami, Florida, USA: Association for Computational Linguistics. 1440–1453. DOI: 10.18653/v1/2024.wmt-1.131.
- Kocyigit, Muhammed Yusuf; Briakou, Eleftheria; Deutsch, Daniel; Luo, Jiaming; Cherry, Colin & Freitag, Markus (2025). Overestimation in LLM Evaluation: A Controlled Large-Scale Study on Data Contamination's Impact on Machine Translation.
- Koehn, Philipp (2009). *Statistical Machine Translation*. Cambridge University Press. 1st ed. DOI: 10.1017/CBO9780511815829.
- Koehn, Philipp (2020). *Neural Machine Translation*. Cambridge University Press. 1st ed. DOI: 10.1017/9781108608480.
- Koller, Werner (1979). *Einführung in die Übersetzungswissenschaft*. Heidelberg Wiesbaden: Quelle & Meyer.
- Larionov, Daniil; Seleznyov, Mikhail; Viskov, Vasiliy; Panchenko, Alexander & Eger, Steffen (2024). *xCOMET-lite: Bridging the Gap Between Efficiency and Quality in Learned MT Evaluation Metrics*. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics. 21934–21949. DOI: 10.18653/v1/2024.emnlp-main.1223.
- Lauscher, Susanne (2000). Translation Quality Assessment: Where Can Theory and Practice Meet? *The Translator*, 6 (2), 149–168. DOI: 10.1080/13556509.2000.10799063.
- Lefevere, André (ed.) (1993). *Translating literature: practice and theory in a comparative literature context*. New York: The Modern Language Association of America.
- Li, Yuan & Liu, Qi (2025). A corpus-based quantitative analysis of German legal language: linguistic features and implications for accessibility. *International Journal of Legal Discourse*, 10 (2), 361–383. DOI: doi:10.1515/ijld-2025-2019.
- Li, Zekun; Peng, Baolin; He, Pengcheng & Yan, Xifeng (2024). *Evaluating the Instruction-Following Robustness of Large Language Models to Prompt Injection*. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics. 557–568. DOI: 10.18653/v1/2024.emnlp-main.33.

- Liu, Qun & Zhang, Xiaojun (2023). Machine Translation. In: Sin-wai, Chan *Routledge Encyclopedia of Translation Technology*. London: Routledge. 2nd ed. 112–127. DOI: 10.4324/9781003168348-6.
- Liu, Xueting & Li, Chengze (2023). Artificial Intelligence and Translation. In: Sin-wai, Chan *Routledge Encyclopedia of Translation Technology*. London: Routledge. 2nd ed. 280–302. DOI: 10.4324/9781003168348-16.
- Lommel, Arle (2018). Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In: Moorkens, Joss, Castilho, Sheila, Gaspari, Federico, & Doherty, Stephen (Eds.) *Translation Quality Assessment*. Cham: Springer International Publishing. Vol. 1. 109–127. DOI: 10.1007/978-3-319-91241-7_6.
- Lommel, Arle; Burchardt, Aljoscha & Uszkoreit, Hans (2013). *Multidimensional quality metrics: a flexible system for assessing translation quality*. In: *Proceedings of Translating and the Computer 35*. London, UK: Aslib.
- Lommel, Arle; Uszkoreit, Hans & Burchardt, Aljoscha (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista Tradumàtica: Tecnologies de La Traducció*, (12), 455–463. <https://revistes.uab.cat/tradumatica/article/view/n12-lommel-uzskoreit-burchardt>.
- Louviere, Jordan J.; Flynn, Terry N. & Marley, A. A. J. (2015). *Best-worst scaling: theory, methods and applications*. Cambridge: Cambridge University Press. First published. DOI: 10.1017/CBO9781107337855.
- Luo, Renqian; Sun, Liai; Xia, Yingce; Qin, Tao; Zhang, Sheng; Poon, Hoifung & Liu, Tie-Yan (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23 (6), bbac409. DOI: 10.1093/bib/bbac409.
- Macháček, Matouš & Bojar, Ondřej (2013). *Results of the WMT13 Metrics Shared Task*. In: Bojar, Ondrej, Buck, Christian, Callison-Burch, Chris, Haddow, Barry, Koehn, Philipp, Monz, Christof, Post, Matt, Saint-Amand, Herve, Soricut, Radu, & Specia, Lucia (Eds.) *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics. 45–51.
- Maibam, Indika & Syam Purkayastha, Bipul (2023). Reordering of Source Side for a Factored English to Manipuri SMT System. *International Journal of Electrical and Computer Engineering Systems*, 14 (3), 285–292. DOI: 10.32985/ijeces.14.3.6.
- Martínez Mateo, Roberto (2014). A Deeper Look into Metrics for Translation Quality Assessment (TQA): A Case Study. *Miscelánea: A Journal of English and American Studies*, 49, 73–93. DOI: 10.26754/ojs_misc/mj.20148792.
- Matvienko, Lesya & Khomenko, Liubov (2025). RISKS OF INFORMATION LEAKAGE USING ONLINE MACHINE TRANSLATION SERVICES. *Cybersecurity: Education, Science, Technique*, 3 (27), 284–293. DOI: 10.28925/2663-4023.2025.27.730.
- Meer, Anne-Maj Van Der (2024). Chapter 2 Evolution of the language industry. In: Massey, Gary, Ehrensberger-Dow, Maureen, & Angelone, Erik (Eds.) *Handbook of the Language Industry*. De Gruyter. 33–48. DOI: 10.1515/9783110716047-003.
- Mohamed, Shereen A.; Elsayed, Ashraf A.; Hassan, Y. F. & Abdou, Mohamed A. (2021). Neural machine translation: past, present, and future. *Neural Computing and Applications*, 33 (23), 15919–15931. DOI: 10.1007/s00521-021-06268-0.

- Montolío Durán, Estrella; García Asensio, Ma. Ángeles (María Ángeles); Gras Manzano, Pedro; López Samaniego, Anna; Polanco Martínez, Fernando; Taranilla, 1981-, Raquel & Yúfera Gómez, Irene (2011). Estudio de campo: lenguaje escrito. Comisión para la modernización del lenguaje jurídico.
- Moorkens, Joss & Guerberof Arenas, Ana (2024). Chapter 4 Artificial intelligence, automation and the language industry. In: Massey, Gary, Ehrensberger-Dow, Maureen, & Angelone, Erik (Eds.) *Handbook of the Language Industry*. De Gruyter. 71–98. DOI: 10.1515/9783110716047-005.
- Nemati Parsa, Rouhullah & Afrouz, Mahmoud (2023). Evaluating the evaluator: A novel perspective on translation quality. *The International Journal of Translation and Interpreting Research*, 15 (2), 184–188. DOI: 10.12807/ti.115202.2023.r03.
- Neuhaus, Elisabeth (2020). *Spekuliert Deepl auf den Google-Exit, Jaroslaw Kutylowski?* <https://www.businessinsider.de/gruenderszene/technologie/deepl-ceo-jaroslaw-kutylowski/> (Stand: 6.2.2026).
- Neumann, Anna; Pi, Yulu & Singh, Jatinder (2026). *Who Controls the Conversation? User Perspectives on Generative AI (LLM) System Prompts*. In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Barcelona Spain: ACM. 1–37. DOI: 10.1145/3772318.3791726.
- Nida, Eugene (1969). Science of Translation. *Language*, 45 (3), 483–498. DOI: 10.2307/411434.
- Nida, Eugene (2003). *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating (Second edition)*. Leiden Boston: BRILL. DOI: 10.1163/9789004495746.
- Nord, Christiane (2018). *Translating as a Purposeful Activity: Functionalist Approaches Explained*. Routledge. 2nd ed. DOI: 10.4324/9781315760506.
- Okpor, Margaret Dumebi (2024). Machine Translation Approaches: Issues and Challenges. DOI: 10.13140/RG.2.2.12055.38561.
- OpenAI (2026).a *Best practices for prompt engineering with the OpenAI API*. <https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api> (Stand: 27.3.2026).
- OpenAI (2026).b *Creating and editing GPTs*. <https://help.openai.com/en/articles/8554397-creating-and-editing-gpts> (Stand: 17.7.2026).
- OpenAI (2026).c *Memory FAQ*. <https://help.openai.com/en/articles/8590148-memory-faq> (Stand: 20.2.2026).
- Ordóñez-López, Pilar (2021). An examination of the accreditation process for sworn translators–interpreters in Spain. *Perspectives*, 29 (6), 900–916. DOI: 10.1080/0907676X.2020.1849323.
- Otal, Juan Cuartero & Limbach, Christiane (2026). Neuronale maschinelle und KI-generierte Übersetzungen: Versuch einer Kategorisierung von Stärken und Schwächen in der Sprachkombination Spanisch-Deutsch. *Lebende Sprachen*,. DOI: 10.1515/les-2025-0055.
- Ouyang, Shuyin; Zhang, Jie M.; Harman, Mark & Wang, Meng (2025). An Empirical Study of the Non-Determinism of ChatGPT in Code Generation. *ACM Trans. Softw. Eng. Methodol.*, 34 (2), 42:1-42:28. DOI: 10.1145/3697010.

- Panou, Despoina (2013). Equivalence in Translation Theories: A Critical Evaluation. *Theory and Practice in Language Studies*, 3 (1), 1–6. DOI: 10.4304/tpls.3.1.1-6.
- Paolucci, Angela Balducci; Lardelli, Manuel & Gromann, Dagmar (2023). *Gender-Fair Language in Translation: A Case Study*. In: Vanmassenhove, Eva, Savoldi, Beatrice, Bentivogli, Luisa, Daems, Joke, & Hackenbuchner, Janiça (Eds.) *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*. Tampere, Finland: European Association for Machine Translation. 13–23.
- Papineni, Kishore; Roukos, Salim; Ward, Todd & Zhu, Wei-Jing (2002). *Bleu: a Method for Automatic Evaluation of Machine Translation*. In: Isabelle, Pierre, Charniak, Eugene, & Lin, Dekang (Eds.) *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. 311–318. DOI: 10.3115/1073083.1073135.
- Pérez Román, Virginia (2020).a *Linguistic Features of Legal Spanish (I)* -. <https://vprspanishtranslations.com/linguistic-features-of-legal-spanish-i/> (Stand: 6.6.2026).
- Pérez Román, Virginia (2020).b *Linguistic Features of Legal Spanish (II)*. <https://vprspanishtranslations.com/linguistic-features-of-legal-spanish-ii/> (Stand: 6.6.2026).
- Pikor, Gabriela Anna (2023). Rechtssprache im deutsch-polnischen Kontrast. *Ślowo. Studia Językoznawcze*, (14), 173–189. DOI: 10.15584/slowo.2023.14.12.
- Poibeau, Thierry (2017). *Machine translation*. Cambridge, Massachusetts: The MIT Press.
- Popiołek, Monika (2023). An in-context overview of legal translator competences, qualifications and other professional requirements set in the ISO 20771:2020 *Legal translation – Requirements* standard. *Digital Translation*, 10 (2), 200–232. DOI: 10.1075/dt.00008.pop.
- Popović, Maja (2011). Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96 (1). DOI: 10.2478/v10108-011-0011-4.
- Popović, Maja (2018). Error Classification and Analysis for Machine Translation Quality Assessment. In: Moorkens, Joss, Castilho, Sheila, Gaspari, Federico, & Doherty, Stephen (Eds.) *Translation Quality Assessment*. Cham: Springer International Publishing. Vol. 1. 129–158. DOI: 10.1007/978-3-319-91241-7_7.
- Prieto Ramos, Fernando (2015). Quality Assurance in Legal Translation: Evaluating Process, Competence and Product in the Pursuit of Adequacy. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, 28 (1), 11–30. DOI: 10.1007/s11196-014-9390-9.
- Pym, Anthony (2010). *Exploring translation theories*. London New York: Routledge, Taylor & Francis Group.
- Pym, Anthony & Torres-Simón, Ester (2021). Is automation changing the translation profession? *International Journal of the Sociology of Language*, 2021 (270), 39–57. DOI: 10.1515/ijsl-2020-0015.
- Radford, Alec & Narasimhan, Karthik (2018). *Improving Language Understanding by Generative Pre-Training*. <https://api.semanticscholar.org/CorpusID:49313245>.

- Ramati, Ido & Pinchevski, Amit (2018). Uniform multilingualism: A media genealogy of Google Translate. *New Media & Society*, 20 (7), 2550–2565. DOI: 10.1177/1461444817726951.
- Random (2026). *RANDOM.ORG - True Random Number Service*. <https://www.random.org/> (Stand: 17.7.2026).
- Räth, Magdalena (2013). *Gereon Frahling (Linguee): „Wir haben uns 18 Monate vergraben“ | Gründerszene*. https://web.archive.org/web/20190301125429/https://www.gruenderszene.de/allgemein/linguee-gereon-frahling-startup-helden?interstitial_click (Stand: 6.2.2026).
- Rawling, Piers & Wilson, Philip (ed.) (2018). *The Routledge Handbook of Translation and Philosophy*. Routledge. 1st ed. DOI: 10.4324/9781315678481.
- Rędzioch-Korkuz, Anna (2023). Revisiting the concepts of translation studies: equivalence in linguistic translation from the point of view of Peircean universal categories. *Language and Semiotic Studies*, 9 (1), 33–53. DOI: 10.1515/lass-2022-0008.
- Rehm, Georg; Sasaki, Felix; Stein, Daniel & Witt, Andreas (2018). *Language Technologies For A Multilingual Europe: Tc3 Iii*. Zenodo. DOI: 10.5281/ZENODO.1291947.
- Rei, Ricardo; Guerreiro, Nuno M.; Treviso, Marcos; Coheur, Luisa; Lavie, Alon & Martins, André (2023). *The Inside Story: Towards Better Understanding of Machine Translation Neural Evaluation Metrics*. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Toronto, Canada: Association for Computational Linguistics. 1089–1105. DOI: 10.18653/v1/2023.acl-short.94.
- Rei, Ricardo; Stewart, Craig; Farinha, Ana C. & Lavie, Alon (2020). *COMET: A Neural Framework for MT Evaluation*. In: Webber, Bonnie, Cohn, Trevor, He, Yulan, & Liu, Yang (Eds.) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics. 2685–2702. DOI: 10.18653/v1/2020.emnlp-main.213.
- Salinas, Abel & Morstatter, Fred (2024). *The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance*. In: Ku, Lun-Wei, Martins, Andre, & Srikumar, Vivek (Eds.) *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics. 4629–4651. DOI: 10.18653/v1/2024.findings-acl.275.
- Sallam, Malik; Irshaid, Amal; Department of Clinical Laboratories and Forensic Medicine, Jordan University Hospital, Jordan; Snygg, Johan; Department of Management, Mediclinic City Hospital, Mediclinic Middle East, Dubai, United Arab Emirates; Department of Anesthesia and Intensive Care, University of Gothenburg, Sahlgrenska Academy, Sweden; Albadri, Rula; Department of Family Medicine, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai, United Arab Emirates; Sallam, Mohammed; Department of Pharmacy, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai, United Arab Emirates; Department of Management, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai, United Arab Emirates; Department of Management, School of Business, International American University, United States; & College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU), Dubai, United Arab Emirates (2025). Rapid Evolution of Large Language Models in Medical Education: Comparative Performance of ChatGPT-3.5, ChatGPT-5, and DeepSeek on Medical Microbiology MCQs.

- Contemporary Education and Teaching Research*, 6 (8), 295–309. DOI: 10.61360/BoniCETR252018770801.
- Sanz-Valdivieso, Lucía & López-Arroyo, Belén (2023). *Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation?* In: *Proceedings of the International Conference on Human-informed Translation and Interpreting Technology 2023*. INCOMA Ltd., Shoumen, Bulgaria. 97–107. DOI: 10.26615/issn.2683-0078.2023_008.
- Schmalz, Antonia (2019). Maschinelle Übersetzung. In: Wittpahl, Volker (Ed.) *Künstliche Intelligenz*. Berlin, Heidelberg: Springer Berlin Heidelberg. 194–211. DOI: 10.1007/978-3-662-58042-4_12.
- Scott, Juliette & O'Shea, John (2021). How Legal Documents Translated Outside Institutions Affect Lives, Businesses and the Economy. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, 34 (5), 1331–1373. DOI: 10.1007/s11196-020-09815-5.
- Sennrich, Rico; Haddow, Barry & Birch, Alexandra (2016). *Neural Machine Translation of Rare Words with Subword Units*. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics. 1715–1725. DOI: 10.18653/v1/P16-1162.
- Snover, Matthew; Dorr, Bonnie; Schwartz, Rich; Micciulla, Linnea & Makhoul, John (2006). *A Study of Translation Edit Rate with Targeted Human Annotation*. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas. 223–231.
- Sonbol, Riad; Rebdawi, Ghaida & Ghneim, Nada (2023). A Machine Translation Like Approach to Generate Business Process Model from Textual Description. *SN Computer Science*, 4 (3), 291. DOI: 10.1007/s42979-023-01742-z.
- Stahlberg, Felix (2020). Neural Machine Translation: A Review. *Journal of Artificial Intelligence Research*, 69, 343–418. DOI: 10.1613/jair.1.12007.
- Stefaniak, Karolina (2020). *Evaluating the usefulness of neural machine translation for the Polish translators in the European Commission*. In: Martins, André, Moniz, Helena, Fumega, Sara, Martins, Bruno, Batista, Fernando, Coheur, Luisa, Parra, Carla, Trancoso, Isabel, Turchi, Marco, Bisazza, Arianna, Moorkens, Joss, Guerberof, Ana, Nurminen, Mary, Marg, Lena, & Forcada, Mikel L. (Eds.) *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal: European Association for Machine Translation. 263–269.
- Stein, Daniel (2018). Machine Translation: Past, Present And Future. Zenodo. DOI: 10.5281/ZENODO.1291924.
- Suojanen, Tytti; Koskinen, Kaisa & Tuominen, Tiina (2014). *User-Centered Translation*. Routledge. 1st ed. DOI: 10.4324/9781315753508.
- SYSTRAN Translate (2026). *Over 50 Years of Developing Machine Translation Technology*. <https://www.systransoft.com/systran/> (Stand: 4.2.2026).
- Unbabel (2024). *Unbabel/XCOMET-XL*. <https://huggingface.co/Unbabel/XCOMET-XL> (Stand: 16.7.2026).
- Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukasz & Polosukhin, Illia (2017). *Attention is all you need*. In:

- Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. 6000–6010.
- Vauquois, Bernard (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. <https://hal.science/hal-04701802v1>.
- Venuti, Lawrence (2017). *The Translator's Invisibility: A History of Translation*. Third edition. | New York: Routledge, [2018] | Reissue of: Routledge. 1st ed. DOI: 10.4324/9781315098746.
- Vermeer, Hans J. (1978). Ein Rahmen für eine allgemeine Translationstheorie. *Lebende Sprachen*, 23 (3), 99–102. DOI: doi:10.1515/les.1978.23.3.99.
- Vermeer, Hans J. & Reiß, Katharina (1984). *Grundlegung einer allgemeinen Translationstheorie*. Berlin: De Gruyter. DOI: 10.1515/9783111351919.
- Wang; Chu, Zhibo; Doan, Thang Viet; Ni, Shiwen; Yang, Min & Zhang, Wenbin (2025). History, development, and principles of large language models: an introductory survey. *AI and Ethics*, 5 (3), 1955–1971. DOI: 10.1007/s43681-024-00583-7.
- Wang, Haifeng; Wu, Hua; He, Zhongjun; Huang, Liang & Church, Kenneth Ward (2022). Progress in Machine Translation. *Engineering*, 18, 143–153. DOI: 10.1016/j.eng.2021.03.023.
- Wang, Peng & Sawyer, David B. (2023). *Machine learning in translation*. London New York: Routledge. DOI: 10.4324/9781003321538.
- Wang, Qiufen (2025). Evaluating Uighur literary translation: A comparative study of ChatGPT, Google Translate, and Bing Translator. *PLOS One*, 20 (10), e0335261. DOI: 10.1371/journal.pone.0335261.
- Way, Andy (2018). Quality Expectations of Machine Translation. In: Moorkens, Joss, Castilho, Sheila, Gaspari, Federico, & Doherty, Stephen (Eds.) *Translation Quality Assessment: From Principles to Practice*. Cham : Springer International Publishing. 159–178. <https://doi.org/10.1007/978-3-319-91241-7>.
- Wiesmann, Eva (2019). Machine Translation in the Field of Law: A Study of the Translation of Italian Legal Texts into German. *Comparative Legilinguistics*, 37 (1), 117–153. DOI: 10.14746/cl.2019.37.4.
- Wissik, Tanja (2017). Niemiecka terminologia prawnicza w dziedzinie szkolnictwa wyższego – warianty narodowe a język angielski. *Comparative Legilinguistics*, 2, 127. DOI: 10.14746/cl.2010.2.11.
- WordNet (2026). *What is WordNet?* <https://wordnet.princeton.edu/homepage> (Stand: 16.7.2026).
- Wu, Xiaodong; Wang, Minhao; Liu, Yichen; Shi, Xiaoming; Yan, He; Xiangju, Lu; Zhu, Junmin & Zhang, Wei (2025). *LIFBench: Evaluating the Instruction Following Performance and Stability of Large Language Models in Long-Context Scenarios*. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics. 16445–16468. DOI: 10.18653/v1/2025.acl-long.803.
- 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. CoRR/abs/1609.08144.
- Yamada, Masaru (2023). *Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability*. In: Yamada, Masaru, & do Carmo, Felix (Eds.) *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*. Macau SAR, China: Asia-Pacific Association for Machine Translation. 195–204.
- Yu, Ping; Xu, Hua; Hu, Xia & Deng, Chao (2023). Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration. *Healthcare*, 11 (20), 2776. DOI: 10.3390/healthcare11202776.
- Zhang, Tianyi; Kishore, Varsha; Wu, Felix; Weinberger, Kilian Q. & Artzi, Yoav (2020). BERTScore: Evaluating Text Generation with BERT.
- Zhang, Yamin & Fan, Chenyang (2025). *The Impact of Prompt Strategies on Specialized Text Translation by Large Language Models: An Empirical Study Using ChatGPT*. In: *Proceedings of the 2025 International Conference on Educational Technology and Artificial Intelligence*. Wuhan China: ACM. 218–222. DOI: 10.1145/3766557.3766594.
- Zhao; Liu, Yilun; Tao, Shimin; Meng, Weibin; Chen, Yimeng; Geng, Xiang; Su, Chang; Zhang, Min & Yang, Hao (2024). *From Handcrafted Features to LLMs: A Brief Survey for Machine Translation Quality Estimation*. In: *2024 International Joint Conference on Neural Networks (IJCNN)*. Yokohama, Japan: IEEE. 1–10. DOI: 10.1109/IJCNN60899.2024.10650457.
- Zhao, Wayne Xin; Zhou, Kun; Li, Junyi; Tang, Tianyi; Dong, Zican; Hou, Yupeng; Zhang, Beichen; Min, Yingqian; Zhang, Junjie; Liu, Peiyu; Wang, Xiaolei; Du, Yifan; Yang, Chen; Chen, Yushuo; Chen, Zhipeng; Jiang, Jinhao; Ren, Ruiyang; Li, Yifan; Tang, Xinyu; Liu, Zikang; Hu, Yiwen; Nie, Jian-Yun & Wen, Ji-Rong (2026). A Survey of Large Language Models. *Frontiers of Computer Science*, 20 (12), 2012627. DOI: 10.1007/s11704-026-60308-3.
- Zhuo, Jingming; Zhang, Songyang; Fang, Xinyu; Duan, Haodong; Lin, Dahua & Chen, Kai (2024). *ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs*. In: Al-Onaizan, Yaser, Bansal, Mohit, & Chen, Yun-Nung (Eds.) *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics. 1950–1976. DOI: 10.18653/v1/2024.findings-emnlp.108.

Appendix 1

Übersetze folgende Textpassage vom Spanischen ins Deutsche:

[Source Text]

Appendix 2

Kontext:

Du bist ein Fachübersetzer für Deutsch (Österreich) und Spanisch (Spanien, Lateinamerika). Du bringst ein breites Spektrum an Fachwissen und Terminologie mit. Dir ist bewusst, dass sich deutsches Recht von österreichischem Recht unterscheidet und teils unterschiedliche Begriffe für dieselbe Sache verwendet werden können. Dir ist bewusst, dass die spanische Sprache vor allem in Spanien und in Lateinamerika gesprochen wird und jedes Land möglicherweise unterschiedliche Begriffe für denselben rechtlichen Sachverhalt verwenden kann. Deine Aufgabe ist es, eine Übersetzung zu erstellen, die für eine Beglaubigung geeignet ist.

Deine Rolle als Übersetzer:

Gemäß „ISO 20771 Legal Translation - Requirements“ musst du als Rechtsübersetzer bestimmte Anforderungen berücksichtigen, die deine primäre Aufgabe (Übersetzen) nicht unmittelbar betreffen, dir jedoch bewusst machen und verdeutlichen sollen, welche Verantwortung und welche fachlichen Anforderungen mit dieser Rolle verbunden sind:

- 1. Verpflichtung zur vollständigen Revision (obligatorisch) und Durchsicht (empfohlen) aller Übersetzungen*
- 2. Verantwortung des juristischen Übersetzers gegenüber dem Auftraggeber für den gesamten Prozess, einschließlich der Erfüllung sämtlicher Spezifikationen und der Sicherstellung, dass eine unabhängige Revision durchgeführt wird.*

3. *Erfordernis, eine Art Dienstleistungsvereinbarung und Leistungsbeschreibung für den Übersetzungsdienst zu haben.*
4. *Erfordernis, einen standardisierten Prozess für die Durchführung eines Übersetzungsprojekts einzuhalten*
5. *Empfehlung, die Übersetzung zu unterzeichnen (falls möglich) und diesen Schritt zu dokumentieren, um vollständige Transparenz und Nachverfolgbarkeit sicherzustellen*
6. *Anforderung für den Umgang mit Feedback, Beschwerden und der Umsetzung von Korrekturmaßnahmen*
7. *Anforderung zur Wahrung von Vertraulichkeit und Sicherheit*
8. *Empfehlung zum Abschluss einer Berufshaftpflichtversicherung sowie die Anforderung, dass der Auftraggeber informiert wird, falls der Übersetzer nicht versichert ist.*

Durchzuführende Schritte für die Übersetzung:

1. *Lies die gesamte, in Spanisch oder Deutsch verfasste Textpassage aus einem Dokument stammend ausführlich durch und erfasse den Kontext und die Kernbegriffe*
2. *Identifiziere alle möglichen komplexen Fachbegriffe und Formulierungen und stelle sicher, dass du klare Definitionen und den Kontext für eine genaue Übersetzung bereitstellst*
3. *Beginne mit dem Übersetzen und achte dabei besonders auf den Satzbau, die Formalitäten und Floskeln der deutschen Rechtssprache in Österreich oder spanischen Rechtssprache in Spanien*
4. *Stelle sicher, dass für sich wiederholende Begriffe dieselbe Übersetzung verwendet wird, um für eine einheitliche Terminologie zu sorgen*
5. *Überprüfe deine Übersetzung auf Kohärenz und Genauigkeit und stelle sicher, dass sie die juristische Absicht des Originaltextes beibehält*

Hinweise:

Beachte die kulturellen und rechtlichen Auswirkungen, die in verschiedenen Rechtssystemen unterschiedliche Bedeutungsnuancen [Spanisch (Spanien, Lateinamerika), Deutsch (Österreich)] haben können.

Aufgabe:

Übersetze die Textpassage aus einem Dokument vom Spanischen ins österreichische Deutsche und achte dabei auf die für eine Beglaubigung essenzielle Wahrung der Genauigkeit und die sinngemäße Übertragung der spanischen Rechtssprache in die deutsche Rechtssprache. Verwende einheitliche Terminologie und stelle sicher, dass die Übersetzung klar und kohärent ist.

Übersetzungsergebnis:

Die Übersetzung soll als reiner Text erscheinen. Deine einzige Aufgabe ist, die Texte nach bestem Wissen und Gewissen zu übersetzen. Beifügungen, Kommentare, Anmerkungen oder Interaktionen sind nicht erlaubt. Jeder Input und Output ist für sich alleine stehend und darf nicht vermischt werden.

Appendix 3

Spanish legal text excerpts and translation output per mode

Source Text	Human Translation	DeepL	ChatGPT with naïve prompt	ChatGPT with extensive prompt
Que mediante el presente escrito vengo a formular demanda de juicio declarativo ordinario.	Mit diesem Schriftsatz bringe ich eine Klage im ordentlichen Erkenntnisverfahren ein.	Mit diesem Schreiben reiche ich eine Klage auf Feststellung ein.	Hiermit erhebe ich Klage im ordentlichen Feststellungsverfahren.	Mit dem vorliegenden Schriftsatz bringe ich Klage auf Einleitung eines ordentlichen Feststellungsverfahrens ein.
La parte demandada en el acto de Audiencia Previa se allanó totalmente a las pretensiones del actor, quedando las actuaciones vistas para sentencia.	Die beklagte Partei hat in der vorbereitenden Verhandlung die Ansprüche des Klägers zur Gänze anerkannt; das Verfahren ist somit spruchreif.	Die beklagte Partei hat sich in der Vorverhandlung vollständig den Forderungen des Klägers unterworfen, sodass das Verfahren zur Urteilsverkündung bereit ist.	Die beklagte Partei erkannte in der Vorverhandlung die Ansprüche des Klägers vollständig an, sodass das Verfahren entscheidungsreif war.	Die beklagte Partei hat sich in der vorbereitenden Tagsatzung vollständig den Begehren des Klägers unterworfen, sodass die Rechtssache zur Urteilsfällung vorlag.
No toda culpa da lugar a responsabilidad; en la culpa contractual, a diferencia de la extracontractual no rige el principio de incurrir en ella por culpa levisima.	Nicht jede Fahrlässigkeit zieht eine Haftung nach sich; bei Fahrlässigkeit im Rahmen der Erfüllung einer vertraglichen Pflicht gilt im Gegensatz zur Fahrlässigkeit außerhalb eines Vertrags nicht der Grundsatz, dass für eine geringfügige Sorgfaltspflichtverletzung gehaftet wird.	Nicht jede Schuld führt zu einer Haftung; bei vertraglicher Schuld gilt im Gegensatz zu außervertraglicher Schuld nicht der Grundsatz, dass bereits geringfügige Schuld eine Haftung begründet.	Nicht jede Schuld führt zu einer Haftung; bei der vertraglichen Schuld gilt – anders als bei der außervertraglichen – nicht der Grundsatz, dass man bereits durch geringste Fahrlässigkeit in sie verfällt.	Nicht jedes Verschulden führt zu einer Haftung; im Vertragsverschulden gilt – anders als im Ausservertraglichen – nicht der Grundsatz, dass bereits leichtestes Verschulden haftungsbegründend ist.

Son reos del delito de robo los que, con ánimo de lucro, se apoderaren de las cosas muebles ajenas empleando fuerza en las cosas para acceder o abandonar el lugar donde éstas se encuentran o violencia o intimidación en las personas.	Es begeht die Straftat des Raubes, wer in Gewinnabsicht fremde bewegliche Sachen an sich nimmt und dabei Gewalt gegen Sachen anwendet, um Zugang zu dem Ort zu erlangen, an dem sich diese befinden, oder diesen zu verlassen, oder mit Gewalt oder Drohungen gegenüber Personen vorgeht.	Des Diebstahls schuldig sind diejenigen, die in Gewinnabsicht fremdes bewegliches Vermögen an sich nehmen, indem sie Gewalt an Gegenständen anwenden, um Zugang zu dem Ort zu erlangen, an dem sich diese befinden, oder diesen zu verlassen, oder indem sie Gewalt oder Einschüchterung gegen Personen anwenden.	Wegen des Delikts des Diebstahls (Raubes) sind diejenigen strafbar, die mit Bereicherungsabsicht fremde bewegliche Sachen an sich nehmen, indem sie Gewalt gegen Sachen anwenden, um den Ort, an dem sich diese befinden, zu betreten oder zu verlassen, oder indem sie Gewalt oder Einschüchterung gegen Personen einsetzen.	Täter des Delikts des Diebstahls sind jene, die sich in Bereicherungsabsicht fremde bewegliche Sachen aneignen, indem sie zur Erlangung des Zugangs zu dem Ort, an dem sich diese befinden, oder zu dessen Verlassung Gewalt gegen Sachen anwenden oder Gewalt oder Drohung gegen Personen einsetzen.
La tipicidad, la antijuridicidad y la culpabilidad son las características comunes a todo hecho delictivo. El punto de partida es siempre la tipicidad. Solo en algunos casos concretos las condiciones objetivas de penalidad o perseguibilidad también configuran un elemento autónomo de la teoría general del delito.	Die Tatbestandsmäßigkeit, die Rechtswidrigkeit und die Schuld bilden die gemeinsamen Merkmale einer jeder Straftat. Ausgangspunkt ist immer die Tatbestandsmäßigkeit. Nur in einigen bestimmten Fällen bilden die objektiven Voraussetzungen der Strafbarkeit oder Verfolgbarkeit auch ein eigenständiges Element der	Die Strafbarkeit, die Rechtswidrigkeit und die Schuld sind die gemeinsamen Merkmale jeder Straftat. Ausgangspunkt ist immer die Strafbarkeit. Nur in einigen konkreten Fällen bilden die objektiven Voraussetzungen für die Strafbarkeit oder Verfolgbarkeit auch ein eigenständiges Element der Theorie der Straftat.	Die Tatbestandsmäßigkeit, die Rechtswidrigkeit und die Schuld sind die gemeinsamen Merkmale jedes strafbaren Handelns. Ausgangspunkt ist stets die Tatbestandsmäßigkeit. Nur in einigen konkreten Fällen bilden die objektiven Bedingungen der Strafbarkeit oder Verfolgbarkeit ebenfalls ein eigenständiges Element der	Die Tatbestandsmäßigkeit, die Rechtswidrigkeit und die Schuld sind die allen Straftaten gemeinsamen Merkmale. Ausgangspunkt ist stets die Tatbestandsmäßigkeit. Nur in einigen konkreten Fällen bilden die objektiven Strafbarkeits- oder Verfolgungsvoraussetzungen ebenfalls ein eigenständiges Element der allgemeinen Verbrechenslehre.

	allgemeinen Theorie der Straftat.		allgemeinen Verbrechenslehre.	
--	--------------------------------------	--	----------------------------------	--

Appendix 4

Criteria-based Assessment: Point distribution and Set-specific totals and averages per quality category

ID	Set 1					Set 2					Set 3					Set 4					Set 5						
	Mode (Best & Worst)	C	T	G	TF	Mode (Best & Worst)	C	T	G	TF	Mode (Best & Worst)	C	T	G	TF	Mode (Best & Worst)	C	T	G	TF	Mode (Best & Worst)	C	T	G	TF		
10	B	DL	3	3	3	B	HT	5	3	4	5	B	HT	5	5	5	B	GPT-N	4	3	2	4	B	HT	4	4	4
10	W	HT	2	4	4	W	DL	4	2	4	4	W	GPT-N	3	1	3	W	GPT-E	4	2	2	4	W	DL	3	3	3
12	B	GPT-E	5	5	5	B	HT	5	5	5	5	B	HT	5	5	5	B	GPT-E	5	5	5	3	B	GPT-N	5	5	5
12	W	DL	1	1	5	W	DL	3	3	5	3	W	GPT-N	2	2	5	W	GPT-E	2	1	4	5	W	DL	2	1	4
13	B	GPT-E	5	5	5	B	HT	5	4	5	5	B	HT	4	4	5	B	HT	5	3	4	4	B	GPT-E	5	5	5
13	W	HT	1	1	4	W	DL	5	2	5	5	W	GPT-N	2	1	3	W	GPT-E	3	1	2	3	W	DL	2	2	3
21	B	GPT-E	5	5	5	B	GPT-N	5	5	2	5	B	DL	5	5	5	B	DL	5	4	5	4	B	GPT-N	5	5	5
21	W	DL	4	4	3	W	GPT-E	5	2	5	5	W	HT	3	3	3	W	HT	4	4	5	4	W	DL	5	4	5
23	B	GPT-E	5	5	5	B	HT	5	4	5	5	B	HT	5	5	4	B	DL	5	5	4	5	B	HT	5	4	4
23	W	DL	3	2	4	W	GPT-N	4	4	3	3	W	GPT-N	4	3	2	W	GPT-E	4	3	2	3	W	DL	5	3	3
Set 1					Set 2					Set 3					Set 4					Set 5							
Total:					C	T	G	TF	Total:					C	T	G	TF	Total:					C	T	G	TF	
HT					3	5	8	5	HT					22	22	22	20	HT					14	12	14	11	
DL					11	10	15	12	DL					12	7	14	12	DL					10	10	8	10	
GPT-N					-	-	-	-	GPT-N					9	9	5	8	GPT-N					4	3	2	4	
GPT-E					20	20	20	20	GPT-E					5	2	5	5	GPT-E					13	7	10	15	
Average:					C	T	G	TF	Average:					C	T	G	TF	Average:					C	T	G	TF	
HT					1.5	2.5	4	2.5	HT					5	4	4.75	5	HT					4.7	4	4.7	3.7	
DL					2.75	2.5	3.75	3	DL					4	2.3	4.7	4	DL					5	5	4	5	
GPT-N					-	-	-	-	GPT-N					4.5	4.5	2.5	4	GPT-N					2.75	1.75	3.25	3.5	
GPT-E					5	5	5	5	GPT-E					5	2	5	5	GPT-E					2.9	1.75	2.5	2.75	
Average:					2.6	2.6	2.6	2.6	Average:					4.7	4.7	4.7	4.7	Average:					4.7	4.7	4.7	4.7	
HT					2.5	2.5	2.5	2.5	HT					5	4	4.75	5	HT					4.7	4	4.7	3.7	
DL					3.75	3.75	3.75	3	DL					4	2.3	4.7	4	DL					5	5	4	5	
GPT-N					-	-	-	-	GPT-N					4.5	4.5	2.5	4	GPT-N					2.75	1.75	3.25	3.5	
GPT-E					5	5	5	5	GPT-E					5	2	5	5	GPT-E					2.9	1.75	2.5	2.75	
Average:					4.1	4.1	4.1	4.1	Average:					4.7	4.7	4.7	4.7	Average:					4.7	4.7	4.7	4.7	
HT					2.5	2.5	2.5	2.5	HT					5	4	4.75	5	HT					4.7	4	4.7	3.7	
DL					3.75	3.75	3.75	3	DL					4	2.3	4.7	4	DL					5	5	4	5	
GPT-N					-	-	-	-	GPT-N					4.5	4.5	2.5	4	GPT-N					2.75	1.75	3.25	3.5	
GPT-E					5	5	5	5	GPT-E					5	2	5	5	GPT-E					2.9	1.75	2.5	2.75	
Average:					4.1	4.1	4.1	4.1	Average:					4.7	4.7	4.7	4.7	Average:					4.7	4.7	4.7	4.7	
HT					2.5	2.5	2.5	2.5	HT					5	4	4.75	5	HT					4.7	4	4.7	3.7	
DL					3.75	3.75	3.75	3	DL					4	2.3	4.7	4	DL					5	5	4	5	
GPT-N					-	-	-	-	GPT-N					4.5	4.5	2.5	4	GPT-N					2.75	1.75	3.25	3.5	
GPT-E					5	5	5	5	GPT-E					5	2	5	5	GPT-E					2.9	1.75	2.5	2.75	
Average:					4.1	4.1	4.1	4.1	Average:					4.7	4.7	4.7	4.7	Average:					4.7	4.7	4.7	4.7	
HT					2.5	2.5	2.5	2.5	HT					5	4	4.75	5	HT					4.7	4	4.7	3.7	
DL					3.75	3.75	3.75	3	DL					4	2.3	4.7	4	DL					5	5	4	5	
GPT-N					-	-	-	-	GPT-N					4.5	4.5	2.5	4	GPT-N					2.75	1.75	3.25	3.5	
GPT-E					5	5	5	5	GPT-E					5	2	5	5	GPT-E					2.9	1.75	2.5	2.75	
Average:					4.1	4.1	4.1	4.1	Average:					4.7	4.7	4.7	4.7	Average:					4.7	4.7	4.7	4.7	
HT					2.5	2.5	2.5	2.5	HT					5	4	4.75	5	HT					4.7	4	4.7	3.7	
DL					3.75	3.75	3.75	3	DL					4	2.3	4.7	4	DL					5	5	4	5	
GPT-N					-	-	-	-	GPT-N					4.5	4.5	2.5	4	GPT-N					2.75	1.75	3.25	3.5	
GPT-E					5	5	5	5	GPT-E					5	2	5	5	GPT-E					2.9	1.75	2.5	2.75	
Average:					4.1	4.1	4.1	4.1	Average:					4.7	4.7	4.7	4.7	Average:					4.7	4.7	4.7	4.7	

Appendix 5

Set specific comments

ID	Set 1	Set 2	Set 3	Set 4	Set 5
10			Ob es in diesem Satz tatsächlich um Fahrlässigkeit bzw. fahrlässiges Verschulden geht, würde ich im Falle einer von mir angefertigten Übersetzung noch eingehender recherchieren.	Erneut kann ich mich für keine der Übersetzungsvarianten begeistern	
12					
13					
21		Satz Drei geht überhaupt nicht; die Übersetzung darf keine selbstgewählten Satzzeichen (hier: Strichpunkt) verwenden. Die Satzabfolge des Originals muss erhalten bleiben.	Satz Zwei ist viel zu schwurbelig. Die Vorstellung der Fahrlässigkeit wird viel zu stark betont, sie ist hier nur ein Nebengedanke von "culpa". Satz Vier "... dass man ... in sie verfällt" ist überhaupt erdichtet.	Bei Satz Vier stört mich der zweimalige Einsatz von "... anwenden". Bei Satz Eins geht die Klammer nicht - da sie auch im Original nicht vorkommt. Hier müsste evtl. eine Fußnote gemacht werden, weil manche Adressaten glauben, "robo" heißt immer nur Raub. Bei allen vier Sätzen liegt der Irrtum vor, dass sie "Diebstahl" sagen, während der Aspekt des "Einbruchs" in keinem erwähnt wird. Der Originaltext fasst "Diebstahl", "Raub(-überfall)" und "Einbruch in einem Satz zusammen.	Die Beispielsätze differieren recht wenig voneinander.
23	Die von mir ausgewählte beste Übersetzung klingt „natürlicher“, in diesem Fall juristisch präziser und könnte so tatsächlich in den dazugehörigen Unterlagen auftauchen. Die schlechteste Übersetzung ist zwar nicht an sich falsch, aber in Wortwahl und Formulierung eher unpassend für den Zweck/Kontext.	Die Übersetzungen sind offensichtlich für unterschiedliche Länder/Rechtssysteme gedacht, das macht die Beurteilung sehr schwer. Ausschlaggebend für die Auswahl der schlechtesten Übersetzung war zum einen die Übersetzung von "quedando" mit "(entscheidungsreif) war" sowie die gesamte Formulierung des Satzes.	Die als beste ausgewählte Übersetzung klingt für mich am überzeugendsten und realistischsten.		

Appendix 6

Entire survey

Herzlich Willkommen!

Sehr geehrte Teilnehmerin, sehr geehrter Teilnehmer,

vielen Dank für Ihre Teilnahme an dieser Studie zur Bewertung von humanen und maschinellen Übersetzungen im Bereich des Strafrechts.

Ziel der Erhebung ist es, zu untersuchen, wie Expertinnen und Experten unterschiedliche Übersetzungsvarianten auf Basis ihres fachlichen Wissens und ihrer beruflichen Erfahrung beurteilen.

Eine **genaue Erklärung des Ablaufs** der Umfrage folgt im nächsten Schritt. Sie haben jederzeit die Möglichkeit, über die Schaltfläche „Zurück“ zu vorherigen Seiten zurückzukehren, um Informationen nochmals nachzulesen.

Die Befragung dauert etwa **20–25 Minuten**. Alle erhobenen Daten werden ausschließlich zu wissenschaftlichen Zwecken im Rahmen einer **Masterarbeit** verwendet und anonym ausgewertet.

Vor Beginn der Erhebung werden einige wenige Angaben zu Ihrem **beruflichen und sprachlichen Hintergrund** erhoben.

Vielen herzlichen Dank für Ihre Unterstützung!

Datenschutzerklärung

Made in LimeSurvey 

Angaben zur Person

*

Ihr Alter:

- ☐ 18-24 Jahre
- ☐ 25-34 Jahre
- ☐ 35-44 Jahre
- ☐ 45-54 Jahre
- ☐ 55-65 Jahre
- ☐ Älter als 65 Jahre

*

Geschlecht

- ☐ Männlich
- ☐ Weiblich
- ☐ Divers/Nicht-binär
- ☐ Keine Angabe

*

Welcher ist Ihr höchster Bildungsabschluss?

- ☐ Bachelor (BA, BSc,...)
- ☐ Master (MA, MSc,...)
- ☐ Diplom (Mag.)
- ☐ Doktorat (Dr., PhD)
- ☐ Sonstiges:

*

Bitte geben Sie das/die Fachgebiet(e) an, in dem Sie Ihren Abschluss erworben haben.

Falls Sie einen translationswissenschaftlichen Abschluss haben, geben Sie bitte Ihre Sprachkombination an.

Sprache A

Sprache B

Sprache(n) C

*

Bitte geben Sie Ihre Spanischkenntnisse gemäß dem Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER) an.

- ☐ C2 - (Annähernd) muttersprachliche Kenntnisse
- ☐ C1 - Fachkundige Sprachkenntnisse
- ☐ B2 - Selbständige Sprachverwendung
- ☐ B1 - Fortgeschrittene Sprachverwendung
- ☐ A2 - Grundlegende Kenntnisse
- ☐ A1 - Anfänger

*

Bitte geben Sie Ihre Deutschkenntnisse gemäß dem Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER) an.

- ☐ C2 - (Annähernd) muttersprachliche Kenntnisse
- ☐ C1 - Fachkundige Sprachkenntnisse
- ☐ B2 - Selbständige Sprachverwendung
- ☐ B1 - Fortgeschrittene Sprachverwendung
- ☐ A2 - Grundlegende Kenntnisse
- ☐ A1 - Anfänger

*

Wie beurteilen Sie Ihre bisherige Übersetzungserfahrung?

- ☐ Ich habe sehr viel Übersetzungserfahrung
- ☐ Ich habe etwas Übersetzungserfahrung
- ☐ Ich habe keine Übersetzungserfahrung

*

Was ist Ihr derzeitiger Beruf?

*

Sind Sie allgemein beeidigte*r und zertifizierte*r Gerichtsdolmetscher*in (Österreich) bzw. öffentlich bestellte*r oder beeidigte*r Übersetzer*in und/oder Dolmetscher*in in einem anderen Land?

Bitte wählen Sie eine der folgenden Antworten:

- ☐ Ja
- ☐ Nein

☐ Sonstiges:

*

Wie viele Jahre Berufserfahrung haben Sie im Übersetzungsbereich?

- ☐ Keine
- ☐ Bis zu 1 Jahr
- ☐ 1 bis 3 Jahre
- ☐ 4 bis 8 Jahre
- ☐ Mehr als 8 Jahre

*

Wie oft verwenden Sie maschinelle Übersetzungsprogramme (z. B. DeepL, Google Translate, Bing)?

- ☐ Nie
- ☐ Bis zu 5 Mal im Jahr
- ☐ Einmal im Monat
- ☐ Einmal pro Woche
- ☐ Mehrmals die Woche
- ☐ Täglich

*

Wie oft verwenden Sie Large Language Models (LLMs - z. B. ChatGPT, Google Gemini)?

- ☐ Nie
- ☐ Bis zu 5 Mal im Jahr
- ☐ Einmal im Monat
- ☐ Einmal pro Woche
- ☐ Mehrmals die Woche
- ☐ Täglich

Falls Sie maschinelle Übersetzungsprogramme und/oder Large Language Models verwenden: Wofür verwenden Sie diese? (mehrere Optionen auswählbar)

- ☐ Für professionelle Übersetzung
- ☐ Für andere berufliche Zwecke
- ☐ Privat

Falls Sie maschinelle Übersetzungsprogramme verwenden: Gibt es eines, das Sie bevorzugen? Falls ja, geben Sie bitte unten an, welches bzw. welche.

*

Wie zufrieden sind Sie mit den Ergebnissen von maschinellen Übersetzungsprogrammen allgemein?

- ☐ Sehr zufrieden
- ☐ Eher zufrieden
- ☐ Weder zufrieden noch unzufrieden
- ☐ Nicht sehr zufrieden
- ☐ Sehr unzufrieden

In den folgenden Abschnitten sehen Sie jeweils einen spanischen Ausgangstext sowie vier deutsche Übersetzungen.

1. Wählen Sie in der Tabelle eine Übersetzung als beste und eine als schlechteste.
2. Bewerten Sie Ihre Auswahl mit Punkten: beste Übersetzung 0 bis +4, schlechteste Übersetzung 0 bis -4. Eine erste Einschätzung genügt.
3. Bewerten Sie anschließend die Qualität der beiden ausgewählten Übersetzungen. Hinweise zu den Kategorien finden Sie bei Bedarf über die ⓘ-Symbole.

Optional können Sie Ihre Einschätzung in einem Kommentarfeld näher erläutern.

Es gibt keine richtigen oder falschen Antworten – entscheidend ist Ihre fachliche Beurteilung.

[Datenschutzerklärung](#)

Made in LimeSurvey 

Bewerten: 1 von 5

*

Wählen Sie die für Sie beste und schlechteste Übersetzung in der unten stehenden Tabelle

	Beste	Schlechteste
Que mediante el presente escrito vengo a formular demanda de juicio declarativo ordinario.		
Mit diesem Schriftsatz bringe ich eine Klage im ordentlichen Erkenntnisverfahren ein.	☐	☐
Mit dem vorliegenden Schriftsatz bringe ich Klage auf Einleitung eines ordentlichen Feststellungsverfahrens ein.	☐	☐
Hiermit erhebe ich Klage im ordentlichen Feststellungsverfahren.	☐	☐
Mit diesem Schreiben reiche ich eine Klage auf Feststellung ein.	☐	☐

*

Wie viele Punkte würden Sie der für Sie besten und der schlechtesten Übersetzung vergeben? (Es gibt keine richtigen oder falschen Antworten; es geht um Ihre individuelle Einschätzung)

	Beste Übersetzung ①					Schlechteste Übersetzung ①				
	0	1	2	3	4	0	-1	-2	-3	-4
Punktevergabe	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

*

Wie bewerten Sie die einzelnen Qualitätsmerkmale (Inhalt, Terminologie, Grammatik, Funktion) der besten und schlechtesten Übersetzung auf einer Skala von 1 bis 5?

- 5 = Sehr gut
- 4 = Gut
- 3 = Befriedigend
- 2 = Schlecht
- 1 = Sehr schlecht

Ausgangstext	Que mediante el presente escrito vengo a formular demanda de juicio declarativo ordinario.
Ihre gewählte beste Übersetzung	—
Ihre gewählte schlechteste Übersetzung	—

	Beste Übersetzung					Schlechteste Übersetzung				
	1	2	3	4	5	1	2	3	4	5
Inhalt ①	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Terminologie ①	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Grammatik und Satzbau ①	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Textfunktion ①	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

(Optional)

Kommentieren Sie gerne Ihre Überlegungen und Gedanken hier.

Bewerten: 2 von 5

*

Wählen Sie die für Sie beste und schlechteste Übersetzung in der unten stehenden Tabelle

	Beste	Schlechteste
La parte demandada en el acto de Audiencia Previa se allanó totalmente a las pretensiones del actor, quedando las actuaciones vistas para sentencia.		
Die beklagte Partei hat sich in der Vorverhandlung vollständig den Forderungen des Klägers unterworfen, sodass das Verfahren zur Urteilsverkündung bereit ist.	☐	☐
Die beklagte Partei erkannte in der Vorverhandlung die Ansprüche des Klägers vollständig an, sodass das Verfahren entscheidungsreif war.	☐	☐
Die beklagte Partei hat in der vorbereitenden Verhandlung die Ansprüche des Klägers zur Gänze anerkannt; das Verfahren ist somit spruchreif.	☐	☐
Die beklagte Partei hat sich in der vorbereitenden Tagsatzung vollständig den Begehren des Klägers unterworfen, sodass die Rechtssache zur Urteilsfällung vorlag.	☐	☐

*

Wie viele Punkte würden Sie der für Sie besten und der schlechtesten Übersetzung vergeben? (Es gibt keine richtigen oder falschen Antworten; es geht um Ihre individuelle Einschätzung)

	Beste Übersetzung ①						Schlechteste Übersetzung ①				
	0	1	2	3	4		0	-1	-2	-3	-4
Punktevergabe	☐	☐	☐	☐	☐		☐	☐	☐	☐	☐

*

Wie bewerten Sie die einzelnen Qualitätsmerkmale (Inhalt, Terminologie, Grammatik, Funktion) der besten und schlechtesten Übersetzung auf einer Skala von 1 bis 5?

- 5 = Sehr gut
- 4 = Gut
- 3 = Befriedigend
- 2 = Schlecht
- 1 = Sehr schlecht

Ausgangstext	La parte demandada en el acto de Audiencia Previa se allanó totalmente a las pretensiones del actor, quedando las actuaciones vistas para sentencia.
Ihre gewählte beste Übersetzung	–
Ihre gewählte schlechteste Übersetzung	–

	Beste Übersetzung						Schlechteste Übersetzung				
	1	2	3	4	5		1	2	3	4	5
Inhalt ①	☐	☐	☐	☐	☐		☐	☐	☐	☐	☐
Terminologie ①	☐	☐	☐	☐	☐		☐	☐	☐	☐	☐
Grammatik und Satzbau ①	☐	☐	☐	☐	☐		☐	☐	☐	☐	☐
Textfunktion ①	☐	☐	☐	☐	☐		☐	☐	☐	☐	☐

(Optional)

Kommentieren Sie gerne Ihre Überlegungen und Gedanken hier.

Bewerten: 3 von 5

*

Wählen Sie die für Sie beste und schlechteste Übersetzung in der unten stehenden Tabelle

	Beste	Schlechteste
No toda culpa da lugar a responsabilidad; en la culpa contractual, a diferencia de la extracontractual no rige el principio de incurrir en ella por culpa levisima.		
Nicht jede Schuld führt zu einer Haftung; bei vertraglicher Schuld gilt im Gegensatz zu außervertraglicher Schuld nicht der Grundsatz, dass bereits geringfügige Schuld eine Haftung begründet.	☐	☐
Nicht jede Fahrlässigkeit zieht eine Haftung nach sich; bei Fahrlässigkeit im Rahmen der Erfüllung einer vertraglichen Pflicht gilt im Gegensatz zur Fahrlässigkeit außerhalb eines Vertrags nicht der Grundsatz, dass für eine geringfügige Sorgfaltspflichtverletzung gehaftet wird.	☐	☐
Nicht jedes Verschulden führt zu einer Haftung; im Vertragsverschulden gilt – anders als im Ausservertraglichen – nicht der Grundsatz, dass bereits leichtestes Verschulden haftungsbegründend ist.	☐	☐
Nicht jede Schuld führt zu einer Haftung; bei der vertraglichen Schuld gilt – anders als bei der außervertraglichen – nicht der Grundsatz, dass man bereits durch geringste Fahrlässigkeit in sie verfällt.	☐	☐

*

Wie viele Punkte würden Sie der für Sie besten und der schlechtesten Übersetzung vergeben? (Es gibt keine richtigen oder falschen Antworten; es geht um Ihre individuelle Einschätzung)

	Beste Übersetzung ①						Schlechteste Übersetzung ①				
	0	1	2	3	4		0	-1	-2	-3	-4
Punktevergabe	☐	☐	☐	☐	☐		☐	☐	☐	☐	☐

*

Wie bewerten Sie die einzelnen Qualitätsmerkmale (Inhalt, Terminologie, Grammatik, Funktion) der besten und schlechtesten Übersetzung auf einer Skala von 1 bis 5?

- 5 = Sehr gut
- 4 = Gut
- 3 = Befriedigend
- 2 = Schlecht
- 1 = Sehr schlecht

Ausgangstext	No toda culpa da lugar a responsabilidad; en la culpa contractual, a diferencia de la extracontractual no rige el principio de incurrir en ella por culpa levisima.
Ihre gewählte beste Übersetzung	–
Ihre gewählte schlechteste Übersetzung	–

	Beste Übersetzung						Schlechteste Übersetzung				
	1	2	3	4	5		1	2	3	4	5
Inhalt ①	☐	☐	☐	☐	☐		☐	☐	☐	☐	☐
Terminologie ①	☐	☐	☐	☐	☐		☐	☐	☐	☐	☐
Grammatik und Satzbau ①	☐	☐	☐	☐	☐		☐	☐	☐	☐	☐
Textfunktion ①	☐	☐	☐	☐	☐		☐	☐	☐	☐	☐

(Optional)

Kommentieren Sie gerne Ihre Überlegungen und Gedanken hier.

Bewerten: 4 von 5

*

Wählen Sie die für Sie beste und schlechteste Übersetzung in der unten stehenden Tabelle

	Beste	Schlechteste
Son reos del delito de robo los que, con ánimo de lucro, se apoderaren de las cosas muebles ajenas empleando fuerza en las cosas para acceder o abandonar el lugar donde éstas se encuentran o violencia o intimidación en las personas.		
Wegen des Delikts des Diebstahls (Raubes) sind diejenigen strafbar, die mit Bereicherungsabsicht fremde bewegliche Sachen an sich nehmen, indem sie Gewalt gegen Sachen anwenden, um den Ort, an dem sich diese befinden, zu betreten oder zu verlassen, oder indem sie Gewalt oder Einschüchterung gegen Personen einsetzen.	☐	☐
Es begeht die Straftat des Raubes, wer in Gewinnabsicht fremde bewegliche Sachen an sich nimmt und dabei Gewalt gegen Sachen anwendet, um Zugang zu dem Ort zu erlangen, an dem sich diese befinden, oder diesen zu verlassen, oder mit Gewalt oder Drohungen gegenüber Personen vorgeht.	☐	☐
Täter des Delikts des Diebstahls sind jene, die sich in Bereicherungsabsicht fremde bewegliche Sachen aneignen, indem sie zur Erlangung des Zugangs zu dem Ort, an dem sich diese befinden, oder zu dessen Verlassung Gewalt gegen Sachen anwenden oder Gewalt oder Drohung gegen Personen einsetzen.	☐	☐
Des Diebstahls schuldig sind diejenigen, die in Gewinnabsicht fremdes bewegliches Vermögen an sich nehmen, indem sie Gewalt an Gegenständen anwenden, um Zugang zu dem Ort zu erlangen, an dem sich diese befinden, oder diesen zu verlassen, oder indem sie Gewalt oder Einschüchterung gegen Personen anwenden.	☐	☐

*

Wie viele Punkte würden Sie der für Sie besten und der schlechtesten Übersetzung vergeben? (Es gibt keine richtigen oder falschen Antworten; es geht um Ihre individuelle Einschätzung)

	Beste Übersetzung ①					Schlechteste Übersetzung ①				
	0	1	2	3	4	0	-1	-2	-3	-4
Punktevergabe	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

*

Wie bewerten Sie die einzelnen Qualitätsmerkmale (Inhalt, Terminologie, Grammatik, Funktion) der besten und schlechtesten Übersetzung auf einer Skala von 1 bis 5?

- 5 = Sehr gut
- 4 = Gut
- 3 = Befriedigend
- 2 = Schlecht
- 1 = Sehr schlecht

Ausgangstext	Son reos del delito de robo los que, con ánimo de lucro, se apoderaren de las cosas muebles ajenas empleando fuerza en las cosas para acceder o abandonar el lugar donde éstas se encuentran o violencia o intimidación en las personas.
Ihre gewählte beste Übersetzung	–
Ihre gewählte schlechteste Übersetzung	–

	Beste Übersetzung					Schlechteste Übersetzung				
	1	2	3	4	5	1	2	3	4	5
Inhalt ①	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Terminologie ①	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Grammatik und Satzbau ①	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Textfunktion ①	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

(Optional)

Kommentieren Sie gerne Ihre Überlegungen und Gedanken hier.

Bewerten: 5 von 5

*

Wählen Sie die für Sie beste und schlechteste Übersetzung in der unten stehenden Tabelle

	Beste	Schlechteste
La tipicidad, la antijuridicidad y la culpabilidad son las características comunes a todo hecho delictivo. El punto de partida es siempre la tipicidad. Solo en algunos casos concretos las condiciones objetivas de penalidad o perseguibilidad también configuran un elemento autónomo de la teoría general del delito.		
Die Strafbarkeit, die Rechtswidrigkeit und die Schuld sind die gemeinsamen Merkmale jeder Straftat. Ausgangspunkt ist immer die Strafbarkeit. Nur in einigen konkreten Fällen bilden die objektiven Voraussetzungen für die Strafbarkeit oder Verfolgbarkeit auch ein eigenständiges Element der allgemeinen Theorie der Straftat.	☐	☐
Die Tatbestandsmäßigkeit, die Rechtswidrigkeit und die Schuld sind die allen Straftaten gemeinsamen Merkmale. Ausgangspunkt ist stets die Tatbestandsmäßigkeit. Nur in einigen konkreten Fällen bilden die objektiven Strafbarkheits- oder Verfolgungsvoraussetzungen ebenfalls ein eigenständiges Element der allgemeinen Verbrechenslehre.	☐	☐
Die Tatbestandsmäßigkeit, die Rechtswidrigkeit und die Schuld sind die gemeinsamen Merkmale jedes strafbaren Handelns. Ausgangspunkt ist stets die Tatbestandsmäßigkeit. Nur in einigen konkreten Fällen bilden die objektiven Bedingungen der Strafbarkeit oder Verfolgbarkeit ebenfalls ein eigenständiges Element der allgemeinen Verbrechenslehre.	☐	☐
Die Tatbestandsmäßigkeit, die Rechtswidrigkeit und die Schuld bilden die gemeinsamen Merkmale einer jeder Straftat. Ausgangspunkt ist immer die Tatbestandsmäßigkeit. Nur in einigen bestimmten Fällen bilden die objektiven Voraussetzungen der Strafbarkeit oder Verfolgbarkeit auch ein eigenständiges Element der allgemeinen Theorie der Straftat.	☐	☐

*

Wie viele Punkte würden Sie der für Sie besten und der schlechtesten Übersetzung vergeben? (Es gibt keine richtigen oder falschen Antworten; es geht um Ihre individuelle Einschätzung)

	Beste Übersetzung ①					Schlechteste Übersetzung ①				
	0	1	2	3	4	0	-1	-2	-3	-4
Punktevergabe	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

*

Wie bewerten Sie die einzelnen Qualitätsmerkmale (Inhalt, Terminologie, Grammatik, Funktion) der besten und schlechtesten Übersetzung auf einer Skala von 1 bis 5?

- 5 = Sehr gut
- 4 = Gut
- 3 = Befriedigend
- 2 = Schlecht
- 1 = Sehr schlecht

Ausgangstext	La tipicidad, la antijuridicidad y la culpabilidad son las características comunes a todo hecho delictivo. El punto de partida es siempre la tipicidad. Solo en algunos casos concretos las condiciones objetivas de penalidad o perseguibilidad también configuran un elemento autónomo de la teoría general del delito.
Ihre gewählte beste Übersetzung	—
Ihre gewählte schlechteste Übersetzung	—

	Beste Übersetzung					Schlechteste Übersetzung				
	1	2	3	4	5	1	2	3	4	5
Inhalt ①	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Terminologie ①	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Grammatik und Satzbau ①	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Textfunktion ①	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

(Optional)

Kommentieren Sie gerne Ihre Überlegungen und Gedanken hier.

Abschließende Fragen

*

Fragen zur vergleichenden Qualitätsbewertung mit Best-Worst-Scaling:

- 5 = Sehr leicht
- 4 = Leicht
- 3 = Mittelmäßig
- 2 = Schwierig
- 1 = Sehr schwierig

	1	2	3	4	5
Wie einfach/schwierig war es für Sie die <u>beste</u> Übersetzung auszuwählen ?	☐	☐	☐	☐	☐
Wie einfach/schwierig war es für Sie die <u>beste</u> Übersetzung zu bewerten ?	☐	☐	☐	☐	☐
Wie einfach/schwierig war es für Sie die <u>schlechteste</u> Übersetzung auszuwählen ?	☐	☐	☐	☐	☐
Wie einfach/schwierig war es für Sie die <u>schlechteste</u> Übersetzung zu bewerten ?	☐	☐	☐	☐	☐

*

Inwieweit entsprachen die verwendeten Qualitätskriterien (Inhalt, Terminologie, Grammatik und Satzbau, Textfunktion) Ihrer eigenen Vorstellung von Übersetzungsqualität?

- ☐ Sehr genau
- ☐ Genau
- ☐ Mäßig genau
- ☐ Wenig präzise
- ☐ Überhaupt nicht präzise

*

Wie anspruchsvoll empfanden Sie die ausgewählten spanischen Ausgangstexte in fachlicher und sprachlicher Hinsicht?

- ☐ Sehr anspruchsvoll
- ☐ Eher anspruchsvoll
- ☐ Eher wenig anspruchsvoll
- ☐ Gar nicht anspruchsvoll

☐ Sonstiges:

*

Halten Sie es für notwendig, über fachliche Expertise (z. B. juristisches oder translatorisches Fachwissen) zu verfügen, um die vorliegenden Texte und Übersetzungen sinnvoll bewerten zu können?

Bitte wählen Sie eine der folgenden Antworten:

- ☐ Ja, fachliche Expertise ist zwingend erforderlich
- ☐ Fachliche Expertise ist hilfreich, aber nicht zwingend erforderlich
- ☐ Nein, sehr gute Sprachkenntnisse reichen aus
- ☐ Kann ich nicht beurteilen
- ☐ Sonstiges:

*

Haben Sie sich bei der Bearbeitung der Umfrage in Ihrem fachlichen Wissen angemessen gefordert gefühlt?

- ☐ Ja, in hohem Maße
- ☐ Ja, in angemessenem Maße
- ☐ Eher nicht, die Anforderungen waren gering
- ☐ Nein, kaum oder gar nicht
- ☐ Sonstiges:

*

Wie kognitiv belastend empfanden Sie die vergleichenden Bewertungsaufgaben insgesamt?

- ☐ Sehr hoch
- ☐ Hoch
- ☐ Gering
- ☐ Sehr gering

Wie konsistent empfanden Sie Ihre eigenen Bewertungen über die einzelnen Best-Worst-Aufgaben hinweg?

Bitte wählen Sie eine der folgenden Antworten:

- ☐ Sehr konsistent
- ☐ Eher konsistent
- ☐ Eher inkonsistent
- ☐ Sehr inkonsistent

Offene Fragen (optional)

Welche Erfahrung haben Sie mit der Beurteilung der Übersetzungsqualität (menschlich oder maschinell)?

Inwiefern halten Sie die Auswahl der besten und schlechtesten Übersetzungen für geeignet, um Qualitätsunterschiede zwischen den Übersetzungen abzubilden?

Gab es allgemeine Kriterien oder Merkmale, die für Sie ausschlaggebend bei der Entscheidung für die beste bzw. schlechteste Übersetzung waren?

Gab es etwas an dieser Umfrage, das Ihnen besonders gefallen oder nicht gefallen hat?

Haben Sie sonstige Verbesserungsvorschläge oder Anmerkungen?

[Datenschutzerklärung](#)

Made in LimeSurvey 

Appendix 7

Declaration of the Use of Generative AI Tools¹

I am informed about the possibilities and limitations of gen-AI used for this thesis. I have verified and checked the results provided by the AI regarding correctness and corrected/edited information, if needed. I acknowledge that the responsibility for this thesis lies with me, the author (Lorenz Hochmair), and not with the tools.

Activity	Description	Tools used	Comments
Translation EN-ES-DE	Spelling, grammar, and stylistic/phrasing adjustments, explaining of source text passages	Grammarly Pro Claude	To improve grammar, spelling, and style, including phrasing and sentence structure I frequently revised and used stylistic suggestions provided by Grammarly at times. In addition, I used Claude Sonnet to have individual passages from English or Spanish source texts I had difficulties to understand explained to me and, with my new understanding, checked independently whether these explanations coincide with the

¹ The declaration follows the template of:

<https://portal.fh-campuswien.ac.at/webservices/portaltoenabler/enablergetdocument.aspx?id=1982>

			original source text and made sense in the context. If I was not happy with my own phrasing, I used the rephrase function of Grammarly.
Conception and structure	Brainstorming ideas for concepts and structure	Claude ChatGPT	I used Claude and ChatGPT to support my decision-making regarding chapter conception and thesis structure, for example, by discussing possible outlines and weighing different ways of organizing my arguments.
Survey interface programing	Technical implementation and front-end programming of the survey interface, including interactive UI elements	ChatGPT	I used ChatGPT in an iterative, exploratory process, to generate and refine the HTML, CSS, and JavaScript code needed to implement my intended survey design in LimeSurvey. This involved formulating various prompts and testing the resulting code repeatedly, adjusting my prompts based on the output until the desired functionality and appearance were achieved.

Literature	Finding literature for my topics	Consensus.ai Elicit	At moments where I could not find any more literature, I used Consesus.ai or Elicit to ask for a quick overview and a first impression how good a topic has been covered by research.
------------	----------------------------------	------------------------	---