MASTERARBEIT | MASTER'S THESIS

Titel | Title

Auswirkungen von Pre-Editing auf die neuronale maschinelle Übersetzungsqualität aus Sicht von Translationsexpert*innen und Lai*innen

verfasst von | submitted by

Karin Niederreiter BA MSc

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

Master of Arts (MA)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 070 331 342
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Translation Deutsch Englisch
Betreut von | Supervisor:	Assoz. Prof. Mag. Dr. Dagmar Gromann BSc

Danksagung

Mein ganz besonderer Dank gilt Assoz. Prof. Mag. Dr. Dagmar Gromann, BSc, die mich während des gesamten Masterarbeitsprozesses mit großer Geduld und beständiger Unterstützung begleitet hat. Ihre fachliche Expertise sowie motivierenden Worte, von der ersten Idee bis zur Fertigstellung dieser Masterarbeit, waren für mich von unschätzbarem Wert.
Vielen Dank für alles!

Ebenso möchte ich meiner Familie und meinen Freund*innen von Herzen danken. Eure aufmunternden Worte haben mir in herausfordernden Momenten Kraft gegeben und mich stets motiviert, mein Bestes zu geben. Ihr habt mir den Weg zu diesem Ziel wesentlich erleichtert. Dafür danke ich euch von Herzen.

Inhaltsverzeichnis

Abkürzungsverzeichnis

ALPAC = Automatic Language Processing Advisory Committee

AS = Ausgangssprache

ASR = Automatic Speech Recognition

AT = Ausgangstext

BERT = Bidirectional Encoder Representations from Transformers-Modell

BLEU = Bilingual Evaluation Understudy

BPE = Byte Pair Encoding

BWS = Best-Worst-Scaling

CAT = Computer-Assisted Translation

CEFR = Common European Framework of Reference for Languages

CL = Controlled Language

CNN = Convolutional Neural Network

DCE = Discrete Choice Experiments

DL = Deep Learning

ESA = European Space Agency

EST = English for Science and Technology

FAHQT = Fully Automatic High Quality Translation

GM = General Motors

GNMT = Google Neural Machine Translation

GPT = Generative Pre-Trained Transformer

HAMT = Human-Aided Machine Translation

KNN = Künstliches Neuronales Netz

KI = Künstliche Intelligenz

kP = Kein Pre-Editing

LISA = Localization Industry Standards Association

MAHT = Machine-Aided Human Translation

METEOR = Metric for Evaluation of Translation with Explicit Ordering

MNL = Multinomiales Logit-Modell

MQM = Multidimensional Quality Metrics

MTPE = Machine Translation Post Editing

MÜ = Maschinelle Übersetzung

MWER = Multi-Reference WER

NMÜ = Neuronale Maschinelle Übersetzung

NLP = Natural Language Processing

PE = Post-Editing

QE = Quality Estimation

RBMÜ = Regelbasierte Maschinelle Übersetzung

RNN = Recurrent Neural Network

RUT = Random Utility Theory

SAE = Society of Automotive Engineers

SMÜ = Statistische Maschinelle Übersetzung

TER = Translation Edit Rate

TQA = Translation Quality Assessment

UCT = User-Centered Translation

WER = Word Error Rate

ZS = Zielsprache

ZT = Zieltext

Abbildungsverzeichnis

Tabellenverzeichnis

Formelverzeichnis

1 Einleitung

Künstliche Intelligenz (KI) ist ein Begriff, der vor über 65 Jahren im Zuge einer wissenschaftlichen Konferenz entstanden ist (Görz et al. 2020: 4) und als Forschungsgegenstand heute mehr denn je den gesellschaftlich Diskurs prägt. Er beschreibt die Simulierung und Operationalisierung von menschlichen Denkweisen und Verhaltensmustern durch Computer, Roboter oder Netzwerke (2020: 1). Neben den Bereichen Telekommunikation, Gewerbe, Gesundheitswesen und Finanz spielt die Automatisierung menschlicher Tätigkeiten auch in der Sprachindustrie, wie etwa bei der Verarbeitung von sprachlichen Informationen und der Übersetzung eine wichtige Rolle (Gabriel 2019: 97).

In der Translationswissenschaft stellt die Maschinelle Übersetzung (MÜ) den Brückenschlag zwischen KI und Sprachübersetzung dar. Im Gegensatz zur Humanübersetzung, bei der die Übersetzung vom Menschen, ggf. unter Zuhilfenahme von Computern oder digitalen Werkzeugen, angefertigt wird, ist die maschinelle Übersetzung ein rein von Maschinen automatisch durchgeführter sprachlicher Übertragungsprozess. Mit der Entwicklung von kostenlosen Online-Übersetzungsmaschinen Ende der 1990er Jahre und der Einbindung von MÜ in kommerzielle Übersetzungsworkflows in den 2000er und 2010er Jahren stieg das wissenschaftliche Interesse an maschineller Übersetzung und sie ist heutzutage als Forschungsgegenstand in der Translationswissenschaft nicht mehr wegzudenken (Kenny 2020: 305f.).

Grund für die Entwicklung des ersten maschinellen Online-Übersetzungswerkzeugs[1], das auf Kopfdruck automatisch Texte übersetzte, war die stetig wachsende Globalisierung, die ihrerseits ein Resultat des sich rasant weiterentwickelnden Internets ist (Schmalz 2019: 194). Neben der Globalisierung haben auch geopolitische Ereignisse wie Flüchtlingskrisen einen nicht unerheblichen Einfluss auf den Übersetzungsbedarf, so stieg die Anzahl der Übersetzungen in der Sprachkombination Arabisch-Deutsch bei dem Online-Übersetzungsdienst Google Translate in kurzer Zeit um ein Fünffaches (Lewis-Kraus 2016: 7). Die Kommunikationsbedürfnisse der globalisierten Welt führen also zu einem erhöhten Bedarf an Übersetzungen und stellen Übersetzer*innen vor immer größere Herausforderungen (Volkova 2018: 1). Für die Bewältigung dieser Herausforderungen stellt die maschinelle Übersetzung einen großen technischen Fortschritt dar.

Schnell stieß das maschinelle Übersetzungssystem jedoch an seine Grenzen und es zeigte sich, dass dem Computer das notwenige Textverständnis zur Identifizierung von Zusammenhängen und Kontexten fehlte, um Mehrdeutigkeiten korrekt auflösen zu können, was oft zu absurden Übersetzungen führte (Schmalz 2019: 195). Trotz einiger Fortschritte in der

[1] 1997, Babel Fish der Suchmaschine AltaVista, später Yahoo

Statistischen Maschinellen Übersetzung (SMÜ) war die Qualität weitestgehend unzufriedenstellend und die Komplexität von SMÜ-Systemen erschwerte weitere Fortschritte (Luong 2016: 7). Im Jahr 2016 wurde die SMÜ weitestgehend von der Neuronalen Maschinellen Übersetzung (NMÜ) abgelöst. Ziel der NMÜ ist es, mithilfe von Deep Learning (DL) und Künstlichen Neuronalen Netzen (KNN), die automatisch und eigenständig Wissen aus großen Datenmengen extrahieren, eine Annäherung an natürliche menschliche Sprache und damit einhergehend eine natürlicher klingende maschinelle Übersetzungen zu erreichen (Schmalz 2019: 194ff.).

Mit der Entwicklung der maschinellen Übersetzung rückte auch die Frage nach der Qualität maschineller Übersetzungen und ihrem generellen Nutzen für Translator*innen in den Fokus. Die Messbarkeit der Qualität von Übersetzungen ist seit Langem ein zentrales, kontroverses und vieldiskutiertes Thema in der Translationswissenschaft. Wissenschaftler*innen sind sich einig, dass es keine einheitlichen Standards und keine einzig richtige und objektive Methode zur Messung von Translationsqualität gibt (Drugan 2013: 35; House 2014; Koby et al. 2014). Darüber hinaus kann Translationsqualität unterschiedliche Bedeutungen haben, die personen-, gruppen- und kontextabhängig sind. Der Zweck der Übersetzung spielt ebenso eine wichtige Rolle, so unterscheiden sich die Qualitätsanforderung an Übersetzungen, die im Zuge eines Produktionsprozesses entstehen von Übersetzungen, die für wissenschaftliche Zwecke durchgeführt werden. In der Wirtschaft liegt der Fokus auf einer benutzer*innenfreundlichen Translationsqualität, die zur Kund*innenzufriedenheit beiträgt, wohingegen die wissenschaftliche Forschung vornehmlich auf Weiterentwicklung und Qualitätsverbesserung abzielt (Castilho et al. 2018: 11).

Im Bereich der Messbarkeit von Qualität und dem Nutzen von MÜ zählen die Qualität des Endprodukts, die Produktivität von Übersetzer*innen bzw. Post-Editierer*innen, die kognitiven Prozessen und die Wirtschaftlichkeit von MÜ zu den wichtigsten Bereichen, die mittels Analyse von unterschiedliche Kombinationen aus Humanübersetzungen und maschinellen Übersetzungen, mit und ohne Post-Editing, erforscht wurden (Kenny 2020: 308). Die alltägliche Sprache ist mehrdeutig, Formulierungen sind oft vage und die Bedeutung geht oft erst aus dem Kontext hervor. Im Gegensatz zu Computern bzw. Maschinen sind Menschen dank ihres Weltwissens und ihrer Intelligenz in der Lage Kontexte zu identifizieren oder zumindest Plausibilitätsannahmen zu treffen. Auf sprachlicher Ebene fehlt der MÜ zusätzlich unter anderem die Fähigkeit zu paraphrasieren, interpretieren oder unübersetzbare Inhalte richtig zu verarbeiten (Burchardt & Porsiel 2017: 15f.). Hinzu kommt, dass unterschiedliche Ansätze der MÜ zu unterschiedlichen Ergebnissen führen, die sich in der Qualität unterscheiden. Neben Sprachenpaar und Sachgebiet zählen die Qualität und die Art des Ausgangstexts zu den entscheidenden

Faktoren (2017: 11). Trotz der rasanten technologischen Fortschritte im Bereich der MÜ bleibt der Faktor Mensch also ein wichtiger Bestandteil der MÜ (Chan 2018: 1).

1.1 Forschungsfrage und Hypothese

Maschinell übersetzte Texte müssen häufig nachbearbeitet, also post-editiert werden, damit ein qualitatives Ergebnis erzielt wird. Bernth und Gdaniec (2001: 175) stellten fest, dass einige einfache Korrekturen des Ausgangstexts dazu beitragen können, die MTranslatability eines Quelltexts zu erhöhen und dadurch die Qualität des Resultats der MÜ verbessert werden kann. Der Begriff MTranslatability bezeichnet die maschinelle Übersetzbarkeit eines Textes. Guerberof Arenas (2019: 334) definiert diese Korrekturen des Quelltexts, das so genannte Pre-Editing als „the use of a set of terminological and stylistic guidelines or rules to prepare the original text before translation automation to improve the raw output quality."

Basierend auf der Feststellung von Burchardt und Porsiel (2017: 11), dass neben Sprachenpaar und Sachgebiet die Qualität des Ausgangstexts einen Einfluss auf das Ergebnis der MÜ hat und dementsprechend beeinflusst, wie brauchbar bzw. unbrauchbar die resultierende Übersetzung ist, soll die geplante Masterarbeit eruieren, ob bestimmte, auf Basis einer Fehlertypologie ausgearbeitete Pre-Editing-Regeln die Resultate von MÜ im Fachbereich Astronautik und der Sprachkombination Englisch-Deutsch qualitativ verbessern.

Die entsprechende Forschungsfrage lautet daher: Welche Pre-Editing-Regeln verbessern das Resultat neuronaler maschineller Übersetzung vom Englischen ins Deutsche im Fachgebiet Astronautik aus der Perspektive von Translationsexpert*innen und Lai*innen? Die Forschungsarbeit beruht einerseits auf der Annahme, dass MÜ-Systeme keine Übersetzungen produzieren können, die qualitativ mit Humanübersetzungen vergleichbar sind (Miyata & Fujita 2017: 54) und daher nicht als endproduktreif betrachtet werden können. Andererseits geht die Arbeit davon aus, dass Pre-Editing zu einer besseren MTranslatability führt und dadurch eine Qualitätsverbesserung der maschinellen Übersetzung erreicht werden kann.

Angesichts der vielfältigen Anwendungszwecke von MÜ und ihrer mittlerweile weit verbreiteten Nutzung durch Lai*innen (Krenz & Ramlow 2008: 26f.) geht die Arbeit darüber hinaus davon aus, dass die Perspektive dieser Zielgruppe in der Beurteilung von Übersetzungsqualität besondere Relevanz besitzt. Dies begründet sich vor allem in potenziell divergierenden Qualitätsauffassungen und -erwartungen zwischen Translationsexpert*innen und Lai*innen (Forcada 2017: 217). Da die Qualitätswahrnehmung darüber hinaus sehr stark von den subjektiven Erwartungen und dem Verwendungszweck der Endnutzer*innen geprägt ist (Bruhn 2020:

140), ermöglicht erst ein Vergleich beider Perspektiven eine umfassende und differenzierte Einschätzung der Übersetzungsqualität, die den unterschiedlichen Anwendungszwecken und Nutzer*innenbedürfnissen gerecht wird.

1.2 Motivation und Zielsetzung

Sowohl im internationalen professionellen Umfeld als auch im privaten Umfeld ist die MÜ heutzutage ein integraler Bestandteil geworden. Ihr Einsatzgebiet reicht von Lokalisierungen von Websites und Produkten über die Nutzung durch Lai*innen im touristischen Bereich wie in Restaurants und im Straßenverkehr bis hin zu virtuellen Assistent*innen, die Kund*innen mittels maschineller Übersetzung Kundenservice in ihrer Erstsprache anbieten (Mathangi 2020: 42). Maschinell erstellte Übersetzungen werden heutzutage oft nicht mehr explizit als MÜ gekennzeichnet, z. B. bei Facebook oder Amazon, und vielen Menschen ist daher gar nicht bewusst, wie sehr MÜ schon Bestandteil des täglichen Lebens geworden ist und wie viele solcher Übersetzungen tagtäglich wissentlich und/oder unwissentlich konsumiert werden. Aufgrund der steigenden Nutzung von MÜ von, für und durch Lai*innen ist diese Masterarbeit daher nicht nur für Translator*innen relevant, sondern ebenso für Lai*innen, die MÜ privat und/oder beruflich nutzen. Für Übersetzer*innen bringt eine Qualitätsverbesserung der MÜ eine Verringerung des Arbeitsaufwands und eine Produktivitätssteigerung durch automatisierte Übersetzungsprozesse mit verringertem Nachbearbeitungsaufwand und ohne relevanter Qualitätsverluste. Neben der praktischen Anwendung von erforschten Pre-Editing-Regeln würde eine Implementierung von Pre-Editing-Regeln, die zu einer nachweislichen Verbesserung der Übersetzungsqualität führen, als Plug-In in bestehende MÜ-Systeme zu kostengünstigen, aber auch qualitativen Übersetzungen führen, von denen Lai*innen, Translator*innen und Entwickler*innen solcher Systeme gleichermaßen profitieren.

Das Ziel dieser Masterarbeit ist es durch Anwendung bestimmter Pre-Editing-Regeln eine Qualitätsverbesserung des maschinellen Übersetzungsresultats zu erreichen. Auf Basis einer Fehlertypologie sollen für den Fachbereich Astronautik in der Sprachkombination Englisch-Deutsch Pre-Editing-Regeln definiert werden, die zu einer Verringerung bzw. Eliminierung der Fehler im Resultat der maschinellen Übersetzung führen und dadurch die Qualität der Übersetzung verbessern sollen. Damit soll letztlich auch der Post-Editing-Aufwand reduziert werden. Diese Erkenntnisse könnten nicht nur zur Verbesserung der maschinellen Übersetzungsprozesse beitragen, sondern auch wertvolle Impulse für die Ausbildung von Übersetzer*innen im Umgang mit MÜ-Systemen liefern.

2 Grundlagen

Die Kernelemente dieser Masterarbeit umfassen die Themen NMÜ und Qualität, insbesondere die Evaluierung ebenjener, Pre-Editing, MTranslatability (Bernth & Gdaniec 2001) und Best-Worst-Scaling (BWS) (Louviere et al. 2015; Louviere & Woodworth 1990). Wichtige Grundlagen sind daher die Abgrenzung von MÜ zur Computer-Assisted Translation (CAT), die Entwicklung und die verschiedenen Formen der maschinellen Übersetzung inkl. Kurzüberblick über ihre verschiedenen Anwendungsbereiche; von regelbasierten Systemen über statistische Systeme bis hin zur neuronalen maschinellen Übersetzung.

Die Erforschung der MÜ geht mit der Frage nach der Qualität und besonders nach der Messbarkeit bzw. Evaluierung der Qualität ebenjener einher. Anhand von verschiedenen relevanten Qualitätsmetriken folgt die Erörterung des Begriffs der Qualität sowie ein Überblick über die wichtigsten Qualitätsstandards in Bezug auf MÜ. Der Faktor Mensch spielt nach wie vor eine wichtige Rolle für die Qualität von MÜ (Chan 2018: 1). Daher wird im weiteren Verlauf näher auf die menschliche Vorbearbeitung des Ausgangstexts, das Pre-Editing, eingegangen und der Begriff der MTranslatability näher definiert.

Darüber hinaus werden die sprachlichen Charakteristika des Fachgebiets Astronautik sowie die vielfältigen automatischen und von Menschen durchgeführten Evaluierungsmethoden von maschineller Übersetzung vorgestellt und das Best-Worst-Scaling als Evaluierungsmethode näher diskutiert, wobei ein besonderer Fokus auf den relativ neuen Anwendungsbereich dieser Methode für die Evaluierung der Qualität von MÜ liegt.

2.1 Maschinelle Übersetzung

MÜ bezeichnet die „automatic conversion of text from one natural language to another" (Kenny 2020: 305) und ist ein Teilgebiet der Computerlinguistik, das sich mit Ideen und Methoden verschiedener Disziplinen, darunter Linguistik, Informatik, Statistik, Künstliche Intelligenz und Informationstheorie, befasst (Maučec & Donaj 2020: 143).

Im Gegensatz zur Humanübersetzung, bei welcher Übersetzer*innen individuell oder kollaborativ und gegebenenfalls unter Zuhilfenahme von computergestützten Übersetzungstechnologien für die Übersetzung von einer Sprache in eine andere verantwortlich sind, ist MÜ die automatische, von Maschinen durchgeführte Übersetzung, gegebenenfalls erweitert um menschliche Arbeitskraft vor und/oder nach dem eigentlichen Übersetzungsprozess, um die ordnungsgemäße Funktionsfähigkeit des Übersetzungssystems sicherzustellen (Kenny 2020: 305f.).

2.1.1 Grade der Maschinengestützen Übersetzung

Die verschiedenen Abstufungen der Arbeitsteilung zwischen Mensch, also die klassische vollständige Humanübersetzung, und Maschine, die vollautomatisierte Übersetzung ohne menschliches Zutun, beschreiben Hutchins und Somers (1992: 148) als vierstufige Skala.

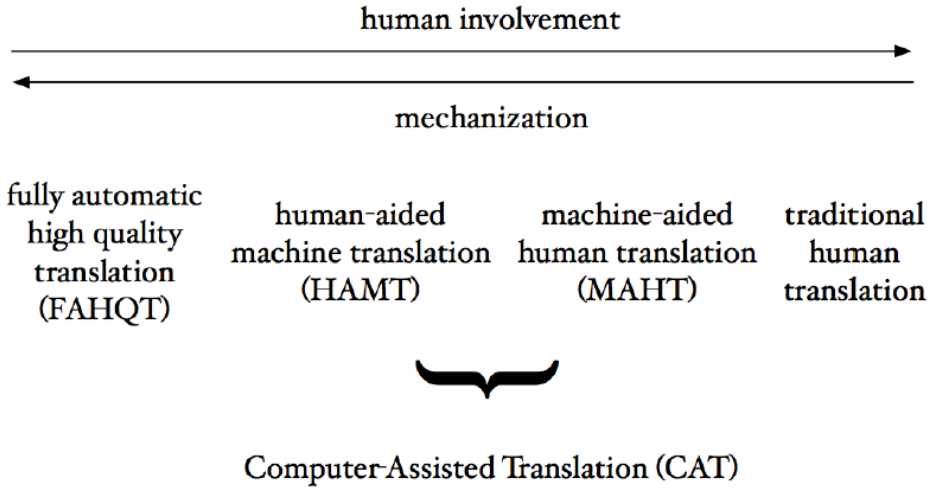


Abbildung 1: Grad der Arbeitsteilung zwischen Mensch und Maschine (Hutchins & Somers 1992: 148)

Abbildung 1 zeigt das vierstufige Spektrum der Kooperationsgrade zwischen Mensch und Maschine. Bei der vollautomatischen MÜ (engl. *Fully Automatic High Quality Translation / FAHQT*) findet der sprachliche Transferprozess von einer Ausgangssprache (AS) in eine Zielsprache (ZS) ausschließlich durch die Maschine, also vollkommen ohne menschlichen Eingriff in diesen Prozess, statt.

In der Wissenschaft besteht Einigkeit darüber, dass FAHQT nicht möglich ist, und jede qualitative rein maschinell erstellte Übersetzung eines gewissen Grads an menschlichen Eingreifens bedarf (Chan 2018: 1; Hutchins & Somers 1992: 3; Tenney 1983: 106). Diese Formen der Mensch-Maschine-Kollaboration befinden sich in Abbildung 1 in der Mitte der Skala, wobei hier auf Basis des Grads des menschlichen Eingreifens in die maschinell erstellte Übersetzung zwischen der anwender*innenunterstützen maschinelle Übersetzung (engl. *Human-Aided Machine Translation / HAMT*) und der maschinenunterstützen Humanübersetzung (engl. *Machine-Aided Human Translation / MAHT*) differenziert wird. Bei der HAMT wird die eigentliche Übersetzung unter Verwendung eines MÜ-Systems komplett maschinell angefertigt und die menschliche Arbeit beschränkt sich rein auf Tätigkeiten vor, während, oder nach dem maschinellen Übersetzungsprozess, in Form von Pre-Editing, also der Überarbeitung des Ausgangstexts, um ihn für das MÜ-System besser übersetzbar zu machen, oder Post-Editing, der

Überarbeitung des Zieltexts nach definierten Qualitätskriterien (Hutchins & Somers 1992: 151f.).

Obwohl MÜ demnach sowohl vollautomatische Übersetzungsprozesse ohne menschliches Eingreifen in den Übersetzungsprozess, als auch automatisierte Übersetzungen inklusive menschlicher Mitwirkung umfasst, ist sie dennoch klar von der MAHT bzw. CAT abzugrenzen, bei der die Übersetzung von menschlichen Übersetzer*innen unter Zuhilfenahme von Software bzw. CAT-Tools, wie beispielsweise Terminologiedatenbanken, Translation-Memory-Systemen oder Onlinewörterbüchern angefertigt wird, und diese Tools lediglich unterstützend zur Humanübersetzung verwendet werden (Forcada 2010: 215; Hutchins 1995: 431).

2.1.2 Anwendungsbereiche der Maschinellen Übersetzung

MÜ spielt aus vielen Gründen eine zentrale Rolle. Aus wirtschaftlicher Sicht sind Übersetzungen bedeutsam, da sie den Absatz von Produkten im Ausland ermöglichen, indem Bedienungsanleitungen in den Erstsprachen der Kund*innen verfügbar gemacht werden. Gesellschaftspolitisch betrachtet ermöglichen sie die Kommunikation in vielsprachigen Gemeinschaften, während die Verwendung einer Lingua franca oft die problematische Bevorzugung dieser Sprache mit sich bringt. Trotz des wichtigen Rechts auf Informationen in der Erstsprache sind menschliche Übersetzer*innen in zu geringer Zahl vorhanden, um den steigenden Bedarf in verschiedenen Sprachen und Sprachenpaaren zu decken, da ihre Produktivität begrenzt ist. Ein*e professioneller*r Übersetzer*in kann kaum mehr als 1.500 – 2.000 Wörter pro Tag übersetzen, was angesichts der zusätzlich benötigten Zeit für Recherche, Terminologiearbeit und Überarbeitung oft deutlich weniger ausfällt (Arnold 1994: 4f.; Krenz & Ramlow 2008: 25).

MÜ ist aber nicht nur im öffentlichen Bereich von großer Bedeutung. Krenz und Ramlow (2008: 26f.) unterscheiden zwischen verschiedenen Anwendungsebenen der MÜ. Neben der institutionalisierten Anwendung durch Unternehmen und Übersetzungsagenturen, die zur Produktivitätssteigerung die Arbeit um MÜ erweitert haben und der öffentlichen Anwendung, wie beispielsweise in Form von mehrsprachigen Informationsterminals für Kund*innen, erfreut sich die MÜ auch bei privaten Anwender*innen großer Beliebtheit. Die leistbaren MÜ-Systeme und die leistungsstärkeren PCs ermöglichten nahezu allen Menschen ab den 1990er Jahren die private Anwendung von MÜ. Neben der Verwendung als Hilfsmittel durch professionelle Übersetzer*innen wurden MÜ-Systeme ebenso von Lai*innen verwendet, wenngleich sich auch die Art und der Grad der Verwendung zwischen den beiden Gruppen unterscheidet und Lai*innen MÜ oft nur oberflächlich für die Extraktion wesentlicher Informationen aus einem

fremdsprachigen Text verwenden (2008: 26f.). Wissenschaftlich werden diese unterschiedlichen Anwendungszwecke von MÜ als Dissemination und Assimilation bezeichnet (Bennett & Gerber 2003: 180f.; Forcada 2010: 217).

Bei der Assimilation handelt es sich um die maschinelle Anfertigung einer Übersetzung für den Informationsgewinn. Für gewöhnlich sind die Anwender*innen der Ausgangssprache nicht mächtig und ihr Ziel besteht darin, den Kern eines Textes zu erfassen und eine ungefähre Vorstellung vom Inhalt zu erhalten, wie beispielsweise beim Browsen von fremdsprachigen Internetseiten, wo mithilfe eines MÜ-Systems der originalsprachige Inhalt in die gewählte Zielsprache übersetzt wird. Da die Übersetzungen für den privaten Gebrauch angefertigt werden spielen Übersetzungsfehler wie beispielsweise grammatikalische und stilistische Fehler bei der Assimilation keine allzu große Rolle, solange die Übersetzung den groben Inhalt und allgemeinen Sinn des Textes korrekt wiedergibt (Forcada 2010: 217).

Bei der Dissemination ist die maschinelle Übersetzung nur ein Zwischenschritt bei der Erstellung eines druckreifen Translats in der Zielsprache, welches nicht primär dem Eigengebrauch dient, sondern veröffentlicht bzw. kommerziell verwendet werden soll. Hierfür wird die maschinelle Rohübersetzung von Translator*innen oder professionellen Editor*innen überprüft und mittels Pre-Editing und/oder Post-Editing überarbeitet (2010: 217). Damit MÜ kommerziell erfolgreich genutzt werden kann, definieren Bennett und Gerber (2003: 186f.) drei Hauptfaktoren: Das System muss die geforderte(n) Sprachrichtung(en) unterstützen, der Ausgangstext muss von ausreichender Qualität sein, um nicht aufgrund der schlechten Qualität einen unverhältnismäßigen und teuren Post-Editing Aufwand zu generieren und es müssen ausreichend spezifische Terminologieressourcen und Fachwörterbücher zur Verfügung stehen.

2.1.3 Geschichtliche Entwicklung der Maschinellen Übersetzung

Obwohl die ersten konkreten Versuche zur Entwicklung und Umsetzung eines funktionsfähigen MÜ-Systems erst in der Nachkriegszeit der 1940er Jahre unternommen wurden (Kenny 2020: 305; Stein 2009: 5) reicht das der MÜ zugrundeliegende Konzept der Nutzung mechanischer Wörterbücher zur Überbrückung von Sprachbarrieren bis in das 17. Jahrhundert zurück. Die Basis dieser Wörterbücher bildeten universelle numerische Codes, denen die Vorstellung einer Universalsprache zugrunde liegen. Die Grundidee dieser Universalsprache war die Entwicklung einer auf eindeutigen logischen Strukturen und Symbolen basierenden Sprache, mit der die gesamte Menschheit miteinander kommunizieren kann. Über die Jahrhunderte hinweg gab es verschiedene andere Konzepte, die die Idee einer Universalsprache aufgriffen, wie

beispielsweise Esperanto, jedoch gab es bis in die Mitte des 20. Jahrhunderts nur mehr wenige Versuche Übersetzungen zu mechanisieren (Hutchins & Somers 1992: 5).

Die Einreichung zweier Patente im Jahr 1933 markierte einen entscheidenden Meilenstein in der frühen Geschichte der MÜ. In diesem Jahr entwickelten Georges Artsrouni (Frankreich) und Petr Trojanskij (Russland) unabhängig voneinander elektromechanische Geräte, die als mehrsprachige Übersetzungswörterbücher verwendet werden konnten (Kenny 2019: 429). Trojanskijs Erfindung war die fortschrittlichere der beiden und beruhte auf der Idee, dass ein*e monolinguale*r menschliche*r Anwender*in den Ausgangstext mithilfe eines universalen Schemas analysiert und anschließend mit der als Übersetzungswörterbuch fungierenden Maschine den relevanten grammatikalischen Code für jedes Wort findet und hinzufügt. Ein*e monolinguale*r humane*r Anwender*in mit Kenntnissen der Zielsprache erstellt daraufhin basierend auf den von der Maschine ausgegebenen äquivalenten zielsprachigen Wörtern und dem grammatikalischen Code einen morphologisch und linguistisch korrekten Zieltext (2019: 429f.).

Die Idee, Computer für Übersetzungen zu verwenden, entstand erstmals zwischen Warren Weaver, einem Mathematiker und Mitglied der Rockefeller Foundation und Andrew D. Booth, einem britischen Kristallografen (Hutchins & Somers 1992: 5). Im Jahr 1949 veröffentlichte Weaver ein Memorandum, in dem er feststellte, dass die große Anzahl an verschiedenen Sprachen den kulturellen Austausch über Landesgrenzen hinweg behindert und schlug Methoden vor dieses Problem mittels computergestützter MÜ zu lösen (Weaver 1949). Zu den vorgeschlagenen Methoden zählten unter anderem statistische Analysen, die Verwendung von kryptografischen Techniken aus der Kriegszeit und die Erforschung der einer Sprache zugrundeliegende Logik und ihre universellen Merkmale (Hutchins & Somers 1992: 5f.). Dieses Memorandum war die Initialzündung, aufgrund derer die Möglichkeit der MÜ mithilfe von Computern in die allgemeine Wahrnehmung rückte und in weiterer Folge dazu führte, dass in den 1950er Jahren intensive MÜ-Forschung betrieben wurde und es mit Yehoshua Bar-Hillel ab dem Jahr 1951 den ersten vollzeitbeschäftigte MÜ-Forscher gab (Hutchins 1995: 433).

Zu dieser Zeit wurden verschiedene Ansätze thematisiert, wie einerseits die Verwendung von kontrollierten Sprachen oder Teilsprachensystemen und andererseits die Notwendigkeit, menschlicher Mitwirkung beim Übersetzungsprozess in Form von Pre-Editing und Post-Editing. Für einige Wissenschaftler*innen war zu dieser Zeit zunächst der Nachweis der grundsätzlichen technischen Realisierbarkeit maschineller Übersetzung ein vorrangiges Ziel (Hutchins & Somers 1992: 6). Ein bedeutender Durchbruch gelang in diesem Zusammenhang im Jahr

1954 mit einem Projekt von Leon Dostert an der Georgetown University, der in Zusammenarbeit mit IBM die erste öffentliche Demonstration eines funktionstüchtigen maschinellen Übersetzungssystems realisierte (1992: 6). Im Zuge dieses Projekts wurden 49 russische Sätze ins Englische übersetzt, wobei aber nur ein sehr eingeschränkter Wortschatz von 250 Worten und sechs Grammatikregeln verwendet wurden. Wenngleich diese Demonstration nur geringen wissenschaftlichen Wert hatte, führte sie zur Generierung erheblicher finanzieller Mittel für die MÜ-Forschung in den Vereinigten Staaten und zu vielen weiteren MÜ-Projekten weltweit und insbesondere in der Sowjetunion (1992: 6).

Die Fortschritte im Bereich der Computertechnologie und der formalen Linguistik versprachen imposante Qualitätsverbesserungen und führten zur Prognose von raschen Durchbrüchen bei der vollautomatischen MÜ. Auf den Optimismus ein funktionierendes MÜ-System zu entwickeln, dessen Übersetzungen qualitativ nicht von denen menschlicher Übersetzer*innen zu unterscheiden sind, folgte eine Phase steigender Enttäuschung aufgrund der durch die Komplexität von Sprache verursachten Probleme. Bar-Hillel, der zu dieser Zeit einen nicht unmaßgeblichen Einfluss hatte, stellte fest, dass den Maschinen das menschliche Weltwissen fehlt, um korrekte Übersetzungen anfertigen zu können und kritisierte die zu ehrgeizigen Ziele der MÜ-Forschung. Er führte das unrealistische Vorhaben aber nicht auf den damaligen Stand der Forschung im Bereich Computerlinguistik zurück, sondern hielt das Ziel aufgrund des fehlenden menschlichen Weltwissens grundsätzlich für unmöglich (Hutchins 1995: 435).

Daraufhin wurde im Jahr 1964 das Automatic Language Processing Advisory Committee (ALPAC) gegründet, welches zum Ziel hatte, den damaligen Stand der MÜ und die Zukunftsaussichten zu evaluieren. Das ALPAC kam zu dem ernüchternden Ergebnis, dass MÜ ungenauer, langsamer und doppelt so teuer wie Humanübersetzung ist und es daher keine unmittelbare und vorhersehbare Aussicht auf eine funktionierende nützliche MÜ gibt (Hutchins 1995: 435).

> We have already noted that, while we have machine-aided translation of general scientific text, we do not have useful machine translation. Further, there is no immediate or predictable prospect of useful machine translation. (ALPAC 1966: 32)

Obwohl der Bericht aufgrund diverser Mängel, wie beispielsweise dem Ignorieren der Tatsache, dass auch qualitative Humanübersetzungen einer Revision bedürfen, als kurzsichtig und voreingenommen kritisiert wurde, hatte er einen nicht unerheblichen Einfluss auf die damaligen Forschungsmöglichkeiten im Bereich der MÜ, die danach viele Jahre lang als vollständiger Misserfolg betrachtet wurde. Diese neue Sichtweise führte zu drastischen Kürzungen der

staatlichen Förderungen und im weiteren Verlauf zum kompletten Erliegen der MÜ-Forschung in den USA. In Kanada und Europa stieg zur selben Zeit jedoch das Interesse an MÜ (Hutchins 1995: 436).

Mitte der 1970er Jahre erlebte die MÜ eine Art Renaissance (Hutchins 2000: 12) als 1968 das MÜ-System SYSTRAN zum ersten Mal vom Georgetown University MÜ-Forscher Peter Toma vorgestellt wurde (Toma 1986: 5). SYSTRAN verfügte über viele Vorteile, wie beispielweise der Fähigkeit des Systems Konkordanz-Korpora zu erstellen, die den Sprachwissenschaftler*innen ermöglichten spezifische semantische oder syntaktische Phänomene zu untersuchen und daraus Regeln abzuleiten. Darüber hinaus bot das System einheitliche syntaktische Merkmale für alle Sprachen, wobei identische Codes an denselben Stellen verwendet wurden. Dies minimierte den Bedarf an umfangreichen Systemanpassungen bei der Übersetzung und der anschließenden Synthese jeder verarbeiteten Sprache (Toma 1977: 573ff.). SYSTRAN erfreute sich immer größerer Beliebtheit, so begann beispielsweise die US Air Force in den 1970er Jahren damit, SYSTRAN für Übersetzungen zwischen Russisch und Englisch zu nutzen. Im Jahr 1976 erwarb die Europäische Kommission dann eine französisch-englische Version und dem System wurden nach und nach immer weitere europäische Sprachenpaare hinzugefügt (Koehn 2020: 35).

Mit der Kommerzialisierung von MÜ-Systemen brachten die 1980er Jahre weitere rasante Fortschritte, die zu einer Diversifizierung der MÜ-Forschung in viele Richtungen führten (Hutchins 1995: 436; Kenny 2019: 432). Dazu zählte auch die durch die Entwicklung neuer Technologien und die Einführung von Heimcomputern ermögliche Erforschung des Konzepts der Künstlichen Intelligenz und der natürlichen Sprachverarbeitung (engl. *Natural Language Processing / NLP*) zur Verbesserung von MÜ (Hutchins 1995: 439). Überdies brachte die Verbreitung des Internets für die private Nutzung in den 1990er Jahren einerseits eine rasant wachsende sprachliche Diversität und andererseits die Möglichkeit Maschinelle Übersetzung weltweit Millionen neuer Nutzer*innen zugänglich zu machen (Kenny 2019: 433).

2.1.4 Regelbasierte Maschinelle Übersetzung

Die Regelbasierte Maschinelle Übersetzung (RBMÜ) entstand in der Zeit nach dem zweiten Weltkrieg und ist damit die früheste Form der MÜ (Poibeau 2017: 49f.). Bis in die 2000er Jahre waren RBMÜ-Systeme der aktuelle Stand der Technik im Bereich der MÜ. Sie werden auch heute nach wie vor verwendet und weiterentwickelt, besonders für ressourcenarme Sprachen,

also Sprachen, für die es nur wenig, bis keine verfügbaren sprachlichen Ressourcen gibt (Kenny 2019: 434).

RBMÜ-Systeme basieren auf der Simulation von linguistischem Wissen in Form von durch Linguist*innen manuell erstellten grammatikalischen und lexikalischen Regeln (2019: 434) und erstellen maschinelle Übersetzungen in drei Schritten: Analyse, Transfer und Synthese bzw. Generierung (Stein 2009: 8). Dabei erfolgt eine Sprachanalyse auf verschiedenen linguistischen Ebenen. Auf syntaktischer Ebene werden beispielsweise Informationen zur Anordnung und Position der verschiedenen Wörter innerhalb des Satzes analysiert, um Informationen zu Rolle und Beziehungen der Wörter zueinander zu erhalten. Auf morphologischer Ebene erfolgt eine Untersuchung kasusbedingter substantivischer und pronominaler Veränderungen, die Informationen über die Rolle eines Worts liefern und damit Rückschlüsse auf dessen grammatikalische Funktion erlauben. Abhängig von der gewählten MÜ-Strategie wird die Ausgangssprache anschließend in eine mehr oder weniger abstrakte Repräsentation umgewandelt, auf der anschließend Übersetzungsregeln angewendet werden (Werthmann & Witt 2014: 87). Die RBMÜ wird in drei solche MÜ-Strategien bzw. Stufen unterteilt: direkte Übersetzung, Transfer und Interlingua (Hutchins & Somers 1992: 107; Kenny 2019: 434).

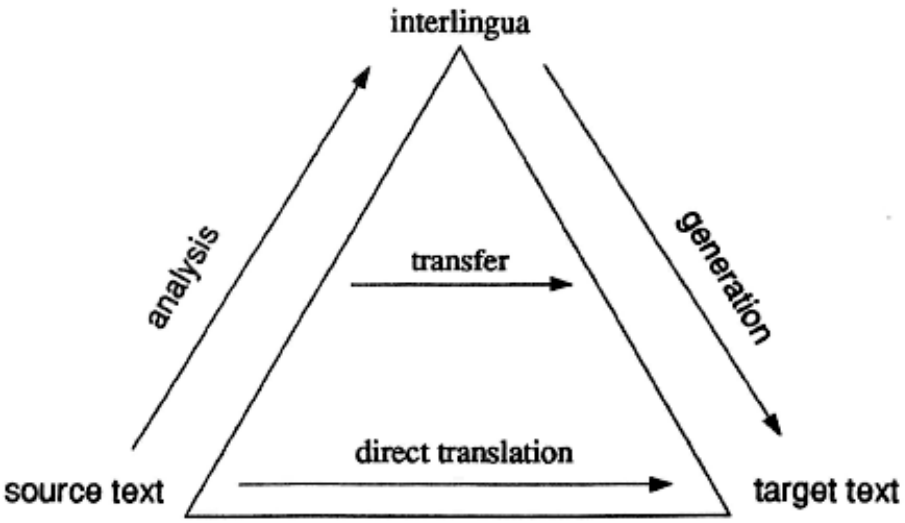


Abbildung 2: Vauquois‘ Dreieck (Hutchins & Somers 1992: 107)

Abbildung 2 zeigt das Architekturschema dieser MÜ-Strategien, die sich hauptsächlich in Bezug auf die Anzahl der bei der Übersetzung berücksichtigten Sprachenpaare, der Analysemethode der Ausgangssprache und dem Grad bzw. der Art und Weise der Informationsabstraktion für die Übersetzung unterscheiden (Werthmann & Witt 2014: 87).

Die **direkte Übersetzung** ist der einfachste RBMÜ-Ansatz, bei dem Wörter oder Phrasen aus der Ausgangssprache mit den entsprechenden Einträgen im bilingualen Lexikon oder Wörterbuch als simple Wort-zu-Wort-Übertragung ohne umfassende strukturelle oder semantische Analyse übersetzt werden (2014: 88). Über eine einfache syntaktische Komponente erfolgt eine oberflächliche Anpassung der ausgangssprachlichen Wörter an die Satzstruktur der Zielsprache. Aufgrund der Problematik von durch Leerzeichen getrennten, aber syntaktisch zusammenhängenden Mehrwortlexem, die nicht wörtlich übersetzt werden können und der generellen Ambiguität von Sprache und den damit verbundenen Qualitäts- und Genauigkeitseinbußen sind Übersetzungen, die mithilfe von direkter Übersetzung produziert wurden, nur begrenzt brauchbar. Darüber hinaus weist sie eine deutlich geringere Komplexität auf als die indirekten Übersetzungsansätze Transfer und Interlingua (Stein 2009: 8).

Beim **Transfer** passiert die sprachliche Übertragung nicht ausschließlich auf der Wortebene, sondern wird auf die Satzebene ausgeweitet und zusätzlich werden morphologische und semantische Informationen extrahiert. Hierfür wird der Ausgangstext in Sätze und Satzteile segmentiert. Anschließend werden die dadurch gewonnen abstrakten Informationen analysiert und in abstrakte zielsprachliche Strukturen transferiert. Im letzten Schritt wird aus den übertragenen abstrakten Strukturen ein natürlichsprachlicher Text in der Zielsprache generiert. Abbildung 3 zeigt diese Schritte der Transfer-Übersetzung. Die Ausweitung auf die semantische Ebene im Zuge der Übersetzung führt zu einer Qualitätsverbesserung gegenüber der direkten Übersetzung (Werthmann & Witt 2014: 89f.).

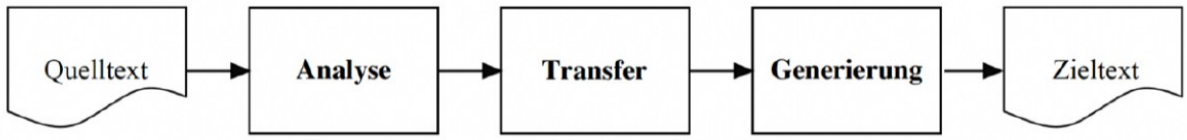


Abbildung 3: Prozess des Transfers (Carstensen et al. 2010: 646)

Die **Interlingua**, die komplexeste Form der RBMÜ, basiert auf der Annahme einer sprachunabhängigen Universalkodierung von linguistischen Informationen und setzt als Basis für alle Übersetzungen die Existenz einer universellen auf alle Sprachen anwendbaren abstrakten Repräsentation, also eine Art Zwischenrepräsentation, die s. g. Interlingua, voraus. Ähnlich wie beim Transfer erfolgt auch bei der Interlingua in der Analysephase eine Segmentierung des Ausgangstexts. Die im Zuge der Analyse gewonnenen Informationen werden daraufhin als Interlingua dargestellt. Somit geht man von der Möglichkeit der Zieltextgenerierung auf

Grundlage einer von der Ausgangssprache unabhängigen sprachlichen Informationsextraktion aus. Die besondere Schwierigkeit liegt darin, dass so eine Universalsprache, die dieser Idee gerecht wird, bis heute noch nicht entdeckt bzw. entwickelt wurde (Stein 2009: 8; Werthmann & Witt 2014: 90f.). Abbildung 4 zeigt die Schritte der Interlingua. Der Schritt Transfer entfällt aufgrund der Annahme einer universellen Interlingua, die den Transfer der in der Analysephase extrahierten, sprachspezifischen abstrakte Informationen zwischen den Sprachen obsolet macht.

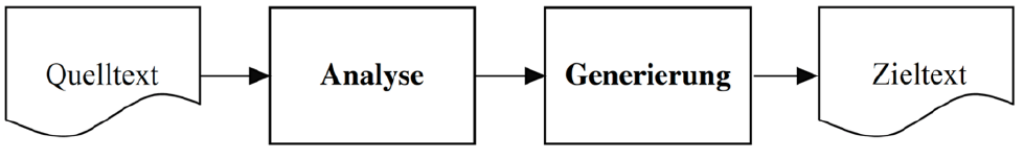


Abbildung 4: Prozess der Interlingua (Carstensen et al. 2010: 646)

2.1.5 Statistische Maschinelle Übersetzung

Da die RBMÜ aufgrund der Notwendigkeit der Erstellung grammatikalischen Regeln und den für die Wort-zu-Wort-Übersetzungen benötigten Wörterbüchern einen hohen Arbeitsaufwand verursacht, trat der korpusbasierte Ansatz im Bereich der MÜ-Systeme ab dem Jahr 1989 immer mehr in den Vordergrund, da er durch seine hohe Genauigkeit andere Methoden übertroffen hat (Sarkhel & Tripathi 2010: 390; Werthmann & Witt 2014: 91). Bei der korpusbasierten MÜ erfolgt die Informationsextraktion aus Parallelkorpora, die den ausgangssprachlichen Text und dessen alignierte Übersetzungen beinhalten. Die Alignierung kann manuell, automatisch oder halbautomatisch erfolgen, wobei die manuelle Alignierung die beste Qualität liefert, aber aufgrund ihrer Ineffizienz für größere Textmengen, die jedoch für die Übersetzung benötigt werden, ungeeignet ist (Werthmann & Witt 2014: 91f.). Bei der automatischen Alignierung werden Ausgangstext und Übersetzung in Übersetzungseinheiten segmentiert und mithilfe von Algorithmen zueinander in Relation gebracht. Die Methode ist effizient aber aufgrund von möglichen Überlappungen in den Segmenten fehleranfälliger als die manuelle Alignierung. Diese Problematik wird durch die halbautomatische Alignierung behoben, bei der ein Mensch die Alignierung überwacht und bei Bedarf nachbessert, um möglichst qualitative alignierte Parallelkorpora zu erhalten (2014: 91f.).

Es existieren mittlerweile zahlreiche Parallelkorpora, wie z. B. das Europarl-Korpus (Koehn 2005), das zwischen 400.000 und zwei Millionen Sätze aus Texten des Europäischen Parlaments in 21 Sprachen umfasst. Viele Korpora, wie beispielsweise das ParaCrawl-Korpus

(Bañón et al. 2020), werden mit Web Scraping, also der systematischen, automatisierten Extraktion von Daten aus allgemein verfügbaren Websites, erstellt. ParaCrawl enthält 223 Millionen alignierte Satzpaare zwischen Englisch und 23 verschiedenen EU-Sprachen (Jurafsky & Martin 2025: 270). Diese Parallelkorpora bilden die Grundlage für die SMÜ bei der Übersetzungsentscheidungen auf Basis von statistischen Wahrscheinlichkeiten getroffen werden. Zur Berechnung der Wahrscheinlichkeiten wird neben den alignierten Parallelkorpora, dem so genannten Übersetzungsmodell, noch ein monolinguales Textkorpus in der Zielsprache, das s. g. Sprachmodell benötigt, das den korrekten Satzbau und die Struktur der Zielsprache abbildet (Stein 2009: 10).

Die Grundannahme der SMÜ ist, dass jeder ausgangssprachliche Satz *e* in eine Vielzahl an möglichen zielsprachlichen Sätzen *f* übersetzt werden kann, aber nicht alle möglichen Übersetzungen gleich wahrscheinlich sind, oder mathematisch formuliert als $P(e|f)$. Ziel der SMÜ ist die Identifizierung der wahrscheinlichsten Übersetzung. Um dieses Ziel zu erreichen, werden Übersetzungs- und Sprachmodell auf unterschiedlichen Ebenen abstrahiert. Um die Dimension der möglichen Sätze zu reduzieren, bewegt man sich beim Sprachmodell auf der Wort- bzw. Wortsequenzebene. Das Übersetzungsmodell wird im Zuge der Abstraktion in ein Lexikonmodell für die Korrektheit der Wortübersetzungen und ein Alignierungsmodell für die Korrektheit der Satzstellungen unterteilt. Da in allen drei Fällen ein möglichst hoher Wert angestrebt wird, ermittelt ein probabilistischer Algorithmus anschließend den Satz mit dem höchsten Ergebnis aus dem Produkt der Werte für Satzstellung (Alignierungsmodell), Wortübersetzung (Lexikonmodell) und Satzgültigkeit (Sprachmodell). Der errechnete Satz ist die wahrscheinlichste Übersetzung basierend auf den Werten der drei Modelle (2009: 9ff.). Abbildung 5 zeigt die schematische Darstellung dieser Maximalwertermittlung mithilfe der drei Modelle.

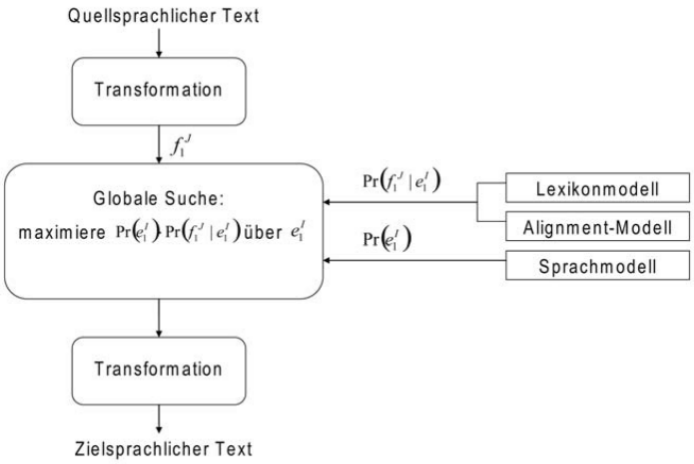


Abbildung 5: Aufbau eines SMÜ-Modells unter Berücksichtigung der drei Modelle (Stein 2009: 11)

2.1.6 Neuronale Maschinelle Übersetzung

Obwohl dank statistischer Methoden maßgebliche Verbesserungen im Bereich der MÜ erzielt werden konnten, gab bzw. gibt es linguistische Probleme, insbesondere in den Bereichen Kongruenz und Anaphorik, die über die Möglichkeiten der SMÜ hinausgehenden Lösungsansätze benötigen (Poibeau 2017: 175f.). Dies führte in den letzten Jahren zu einem bedeutenden Anstieg des allgemeinen Interesses an einem neuen MÜ-Paradigma, welches die SMÜ im Laufe der Zeit verdrängte, die neuronale maschinelle Übersetzung, die auf künstlichen neuronalen Netzen basiert (Forcada 2017: 291f.).

2.1.6.1 Aufbau und Funktionsweise eines KNN

Auch wenn es zwischen biologischen und künstlichen Neuronen gewisse Unterschiede in der Art und Weise der Informationsverarbeitung gibt, leitet sich die Bezeichnung „neuronal“ dennoch von den im menschlichen Gehirn vorhandenen Neuronen ab (Koehn 2020: 31). Abbildung 6 zeigt den Aufbau und die Kernelemente eines biologischen Neurons.

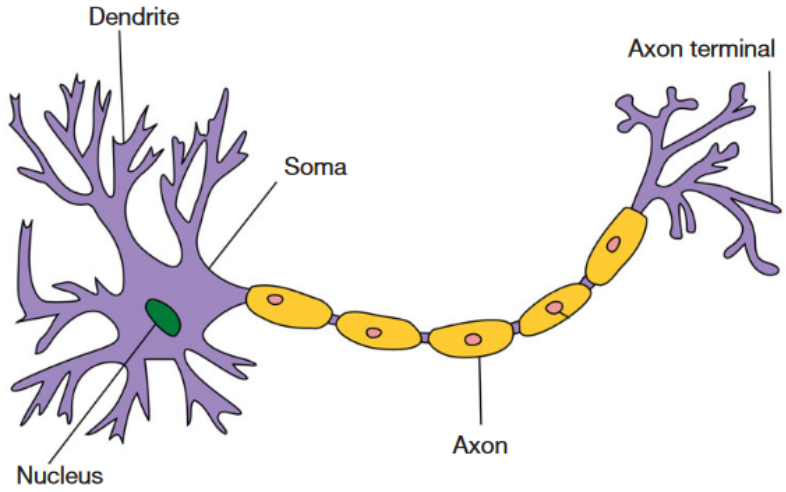


Abbildung 6: Kernelemente eines biologischen Neurons (Koehn 2020: 31)

Vereinfacht ausgedrückt empfängt jedes Neuron über seine Dendriten Eingangssignale von anderen Neuronen und gibt, sofern die empfangenen Signale stark genug sind und das Neuron aktiviert wird, ein Signal über sein Axon an die verschiedenen Axonterminale weiter, die ihrerseits mit den Dendriten anderer Neuronen kommunizieren. Künstliche Neuronen greifen diese grundlegende Idee der biologischen Impulsweiterleitung zwischen natürlichen Neuronen auf, entwickeln sich aber im Gegensatz zu ihren biologischen Namensvettern nicht von selbst, sondern müssen trainiert werden (Koehn 2020: 30f.).

Ein künstliches neuronales Netz ist ein Zusammenschluss vieler miteinander interagierender Neuronen, die in Schichten organisiert sind und so komplexe Sachverhalte modellieren können (Kelleher 2019: 67). Auf Basis ihres Aufbaus werden sie in einschichtige und mehrschichtige KNNs unterteilt. In den einschichtigen Netzen wird der Input direkt über eine Variation einer linearen Funktion auf einen Output gemappt. In mehrschichtigen KNNs befinden sich zwischen Input- und Output-Schicht eine oder mehrere versteckte Schichten (Aggarwal 2018: 4). Da Benutzer*innen keinen Einblick in die genauen Vorgänge und Berechnungen in diesen versteckten Schichten haben, werden sie auch als Blackbox bezeichnet (2018: 17). Abbildung 7 zeigt den Aufbau eines mehrschichtigen Feed-Forward-Netzwerks als Beispiel mit einer Input-Schicht, einer Output-Schicht und drei versteckten Schichten dazwischen. Die Kreise repräsentieren Neuronen, die die Informationen innerhalb des Netzwerkes verarbeiten.

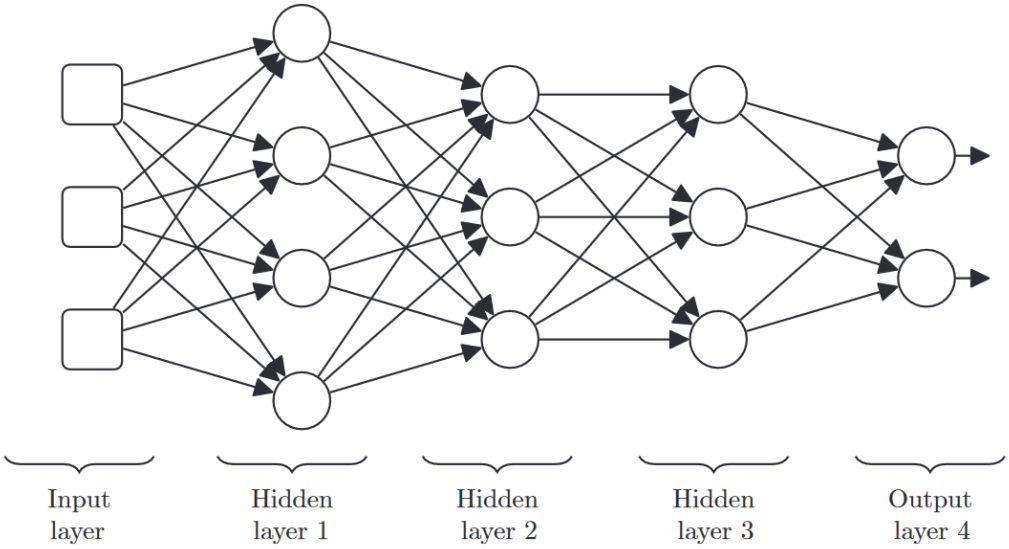


Abbildung 7: Darstellung eines Feed-Forward-Netzwerks mit drei versteckten Schichten (Kelleher 2019: 68)

In einem Feed-Forward-Netzwerk wird Information sequenziell verarbeitet, indem sie von der Input-Schicht über mehrere versteckte Schichten zur Output-Schicht gelangt. Dies geschieht in einem zweistufigen Prozess, bestehend aus der Berechnung der gewichteten Summe der Input-Werte und der Aktivierung der relevanten Neuronen mithilfe einer nichtlinearen Aktivierungsfunktion. Dieser Wert wird anschließend an Neuronen in darauffolgenden versteckten Schichten weitergegeben. Dieser Vorgang wird so lange wiederholt, bis die Output-Schicht erreicht ist. Die Neuronen der Output-Schicht erzeugen schlussendlich den finalen Output des kompletten Netzwerks. Feed-Forward-Netzwerke lassen sich so zur Bewältigung verschiedenster Aufgaben, wie Klassifikations- und Regressionstasks verwenden (Kelleher 2019: 70ff.).

Das volle Potenzial der KNNs liegt aber nicht ausschließlich in der Vorwärtsbewegung, sondern entfaltet sich vielmehr erst durch die Möglichkeit das Netz „rückwärts“ mithilfe des Backpropagation-Algorithmus zu trainieren. Dieser Prozess ermöglicht eine Fehlerminimierung auf Grundlage einer Gewichtsneujustierung durch die Rückwärtsbewegung. Backpropagation nutzt hierfür die Kettenregel der Differentialrechnung, um die Fehlergradienten effizient zu berechnen und besteht aus der oben beschriebenen Vorwärtsphase, an deren Ende der erhaltene Output mit dem idealen Output verglichen wird. Die Differenz der Outputs wird anschließend als Fehler in einer Rückwärtsphase über die diversen Schichten des KNNs zurück propagiert und die Gewichtungen werden in Relation zum Fehlerwert neu berechnet (Aggarwal 2018: 21; Rumelhart et al. 1986: 318ff.).

2.1.6.2 Encoder-Decoder-Modell

MÜ zählt zu den sequenziellen Aufgabenstellungen (Sutskever et al. 2014: 1), d. h. eine Sequenz an Wörtern in der Ausgangssprache soll in eine Sequenz an Wörtern in der Zielsprache übertragen werden. Eine von einem NMÜ-System produzierte Übersetzung ist folglich das Ergebnis einer für KNNs verarbeitbaren Kodierung (engl. *encoding*) des Ausgangstexts und einer kompatiblen Dekodierung (engl. *decoding*) des Zieltexts zurück in natürliche Sprache (Forcada 2017: 297). Eine Standardarchitektur für maschinelle Übersetzung ist die Encoder-Decoder-Architektur. Das sind neuronale Netze, die in der Lage sind Ausgabesequenzen beliebiger Länge auf Basis von Eingabesequenzen zu generieren. Encoder-Decoder-Modelle werden für eine Vielzahl von Machine Learning-Aufgaben, wie beispielsweise Textzusammenfassung oder Question-Answering verwendet, eignen sich aber insbesondere für MÜ, bei der Eingabe- und Ausgabesequenzen oft unterschiedliche Längen haben (Jurafsky & Martin 2025: 174).

Um diesen Kodierungsprozess mithilfe von KNNs zu ermöglichen, ist es notwendig natürliche Sprache und ihre diskreten Elemente (Wörter) in eine geeignete, also eine für das KNN verarbeitbare, Darstellungsform zu bringen (Koehn 2020: 104). Moderne NMÜ-Systeme arbeiten hierfür mit subwortbasierter Tokenisierung. Für Ausgangs- und Zielsprachen wird ein gemeinsames Vokabular verwendet, was den Transfer bestimmter Token, wie Eigennamen, erleichtert. Subwort-Tokenisierung ermöglicht außerdem eine Verarbeitung von Sprachen mit und ohne Leerzeichen, wie Englisch und Chinesisch. Im Kontext von MÜ wird der Ausgangstext mithilfe eines Tokenisierungsalgorithmus, wie beispielsweise Byte Pair Encoding (BPE), wordpiece, oder SentencePiece (Kudo & Richardson 2018) in Token, also eine Serie von Zeichensequenzen, die aus ganzen Wörtern, Wortteilen oder einzelnen Zeichen bestehen können,

umgewandelt (Jurafsky & Martin 2025: 269). Abbildung 8 veranschaulicht die resultierenden Token nach Tokenisierung mit wordpiece.

- **Word**: Jet makers feud over seat width with big orders at stake
- **wordpieces**: _J et _makers _fe ud _over _seat _width _with _big _orders _at _stake

Abbildung 8: Resultat der wordpiece-Tokenisierung mit GoogleMT (Wu et al. 2016: 7)

Jedem Token wird anschließend eine numerische ID aus dem Vokabular zugewiesen. Danach wird jeder Token der Eingabesequenz in eine numerische Repräsentation (Vektor), transformiert, die als Input für den Encoder dient (Jurafsky & Martin 2025: 185). Ziel des Encoders ist anschließend die Erzeugung kontextualisierter Repräsentationen der Eingabesequenz, die die dynamischen, kontextabhängigen semantischen und syntaktischen Relationen abbilden. Für Encoding und Decoding können verschiedene neuronale Architekturen verwendet werden, die sich im Laufe der Zeit und mit den besseren technischen Möglichkeiten weiterentwickelten. Zwei dieser Architekturen, die derzeit dafür verwendet werden, sind Rekurrente Neuronale Netze (engl. *Recurrent Neural Networks* / *RNN*) und Transformer-Architekturen (Vaswani et al. 2017), der derzeitige Stand der Technik im Bereich der neuronalen maschinellen Übersetzung (Jurafsky & Martin 2025: 175).

In einem RNN wird die Eingabesequenz sequenziell verarbeitet, wobei bei jedem Verarbeitungsschritt eine Aktualisierung auf Basis der letzten Information sowie der zuvor akkumulierten Informationen erfolgt. Aufgrund der rekursiven Architektur ist das Modell theoretisch in der Lage, Kontextinformationen sowohl aus lokal benachbarten als auch aus weiter zurückliegenden Sequenzanteilen zu berücksichtigen. Nach Abschluss der Verarbeitung resultiert eine Repräsentation, die die aggregierte Kontextinformation der gesamten Eingabesequenz enthält und typischerweise als Grundlage für den Decoder dient (Forcada 2017: 297f.).

Obwohl die RNN-Architektur prinzipiell in der Lage ist, Informationen aus weiter zurückliegenden Sequenzanteilen zu berücksichtigen, erwies sie sich in der Praxis als wenig geeignet, langfristige Abhängigkeiten effektiv zu modellieren, was zur Entwicklung und Nutzung weiterer, verbesserter Methoden, wie u. a. Attention-Mechanismen (Bahdanau et al. 2015) geführt hat. Diese Attention-Mechanismen haben in vielen Anwendungsbereichen der natürlichen Sprachverarbeitung, darunter auch NMÜ, zu einer starken Leistungsverbesserung beigetragen (Kardakis et al. 2021), indem sie dem Decoder ermöglichen, nicht nur auf die letzte vom Encoder erzeugte Repräsentation zuzugreifen, sondern für jeden Ausgabeschritt einen gewichteten Kontextvektor aus allen Encoder-Zuständen zu berechnen (Forcada 2017: 299).

Der Einfluss von Attention auf MÜ reicht über die reine Verbesserung der maschinellen Übersetzung von langen Sätzen hinaus. Seit ihrer Einführung hat sie sich zu einem wesentlichen Bestandteil verschiedenster NMÜ-Architekturen etabliert. Zu diesen Architekturen zählt auch die Transformer-Architektur, die auf dem Prinzip von Attention beruht. Im Gegensatz zur sequenziellen Verarbeitung in RNN-Architekturen werden bei Transformer-Architekturen alle Token der Eingabesequenz parallel verarbeitet. Aufgrund der parallelen Verarbeitung werden zusätzliche Positionskodierungen benötigt, um die Reihenfolge der Token in der Sequenz zu behalten. Anstelle der Rekursion verwendet die Transformer-Architektur die so genannte Self-Attention, die es ermöglicht, Abhängigkeiten unabhängig von der Position und Distanz der Token, effektiv zu erfassen (Jurafsky & Martin 2025: 197f.). Der Self-Attention-Mechanismus berechnet für jeden Token eine Gewichtung, die die Relevanz der anderen Token in der Eingabesequenz für die Erzeugung der aktuellen Token-Repräsentation bestimmt, um relevante Kontextinformationen zu extrahieren. Auf diese Weise enthält die Repräsentation eines Tokens nicht nur die Information des Tokens selbst, sondern auch kontextuelle Informationen, die aus der gesamten Eingabesequenz stammen. Die Multi-Head Attention erfasst parallel verschiedene Aspekte der Beziehungen zwischen den Token, wodurch eine umfassendere Betrachtung der Eingabesequenz ermöglicht wird. Im Anschluss werden die kontextualisierten Repräsentationen durch ein Feed-Forward-Netzwerk weiterverarbeitet und zur Generierung der Ausgabe verwendet (Jurafsky & Martin 2025: 184ff.).

Sowohl RNN- als auch Transformer-Decoder erzeugen die Zielsequenz schrittweise und autoregressiv, wobei jeder neue Token auf Grundlage der bisherigen Ausgabe und der vom Encoder bereitgestellten Repräsentation berechnet wird. Für jeden Schritt berechnet das Modell einen Wahrscheinlichkeitsvektor über das Vokabular aus dem in der Regel der wahrscheinlichste Token ausgewählt wird (2025: 177f.). Während RNN-Decoder dabei auf rekursiv aktualisierte Zustände zurückgreifen, nutzen Transformer-Decoder Self-Attention, um die Beziehungen zwischen den bisher generierten Tokens parallel und ohne sequenzielle Verarbeitung zu berücksichtigen, was eine effizientere Modellierung von Abhängigkeiten ermöglicht (2025: 272ff.).

2.1.6.3 Verfügbare NMÜ-Systeme

Im Bereich der maschinellen Übersetzung gibt es viele unterschiedliche Systeme, die je nach Einsatzbereich verschiedene Anforderungen erfüllen. Während spezialisierte Lösungen auf bestimmte Fachgebiete ausgerichtet sind, bieten große Anbieter wie Google, Microsoft, DeepL

oder Amazon breit einsetzbare Übersetzungsdienste an. Diese richten sich an eine breite Nutzer*innengruppe, wie Reisende, Social-Media-Nutzer*innen oder Privatpersonen. Ihr gesellschaftlicher Nutzen ist erheblich, da sie sprachliche Barrieren abbauen und besonders auch Sprecher*innen ressourcenarmer Sprachen den Zugang zu globalen Informationen erleichtern (Schmalz 2019: 197).

Drei NMÜ-Systeme, die mit Attention-Mechanismen und der Transformer-Architektur arbeiten sind Google Translate, DeepL und SYSTRAN. Im Jahr 2016 (Wu et al. 2016) führte Google das **Google Neural Machine Translation-System** (GNMT) ein, das anfänglich auf RNNs basierte und einen bedeutenden Fortschritt in der maschinellen Übersetzung darstellte. Im Gegensatz zu früheren regelbasierten Ansätzen analysierte GNMT ganze Sätze als semantisch zusammenhängende Einheiten, wodurch Übersetzungen deutlich natürlicher und kohärenter wurden. Diese Entwicklung markierte für Google einen grundlegenden Paradigmenwechsel von regelbasierten Systemen hin zu neuronalen Modellen, die mit deutlich geringerem manuellem Aufwand auskommen (Quoc & Schuster 2016). In den darauffolgenden Jahren entwickelte sich das System weiter. Neuere Modelle kombinierten die Stärken verschiedener Architekturen. Google ersetzte das ursprüngliche GNMT durch ein hybrides System mit einem Transformer-Encoder und einem RNN-Decoder. Der Transformer-Teil verbessert die Qualität durch tieferes Sprachverständnis, während der RNN-Decoder eine schnellere Übersetzung ermöglicht. Gleichzeitig verbesserte Google die Qualität der Trainingsdaten, indem es einen neuen, präziseren Web-Crawler einführte und mithilfe von Techniken zur Rauschunterdrückung irrelevante oder fehlerhafte Daten ausfilterte (Caswell & Bowen 2020).

Im August 2017 (DeepL SE 2025) präsentierte das deutsche Unternehmen **DeepL** einen neuartigen maschinellen Übersetzungsdienst, der erstmals auf Convolutional Neural Networks (CNNs) anstelle der bis dahin RNNs basiert (Schmalz 2019: 200). Diese Architektur, bisher primär in der Bildverarbeitung eingesetzt, ermöglicht eine parallele Verarbeitung von Wörtern und nutzt bereits optimierte Rechenbibliotheken. Ein wesentlicher Qualitätsfaktor ist der Zugriff auf umfangreiche, qualitativ hochwertige Trainingsdaten aus der früheren Übersetzungssuchmaschine Linguee (2019: 199ff.). Mittlerweile verwendet DeepL Elemente der gängigen Transformer-Architektur, wie etwa Attention-Mechanismen, hat jedoch signifikante strukturelle Modifikationen an der Topologie der neuronalen Netze vorgenommen. Darüber hinaus verfolgt DeepL einen selektiven Ansatz für die Gewinnung von Trainingsdaten. Mithilfe spezialisierter Crawler werden qualitativ hochwertige Paralleltexte gezielt identifiziert und

gesammelt. Dies ermöglicht ein Training auf domänenspezifisch relevanten und sprachlich hochwertigen Daten (DeepL 2021).

SYSTRAN ist eine Softwarefirma mit über 50 Jahren Erfahrung in der Entwicklung von (N)MÜ-Technologien und arbeitet mit der OpenNMT Open Source Übersetzungstechnologie, die weltweit von Expert*innen aus Wissenschaft und Industrie verwendet und trainiert wird (SYSTRAN 2025a). SYSTRAN translate basiert auf der Transformer-Architektur und das SYSTRAN translate Model Studio erlaubt es mit eigenen vorhandenen Ressourcen, wie Translation Memories und Terminologiedatenbanken, ein benutzerdefiniertes NMÜ-Modell selbst zu trainieren und dadurch auf die eigene(n) Domäne(n) und Daten anzupassen (SYSTRAN 2025b).

2.1.6.4 Grenzen Neuronaler Maschineller Übersetzung

Obwohl die Transformer-Architektur immense Verbesserungen für die NMÜ brachte, gilt es immer noch zahlreiche Herausforderungen zu bewältigen. Trotz einiger Versuche, bestehende Modelle weiterzuentwickeln, blieben weitere wesentliche Fortschritte und Weiterentwicklungen im Bereich der NMÜ-Architekturen bisher aus (Zhang & Zong 2020: 2030f.).

Viele Herausforderungen bei NMÜ-Systemen resultieren aus ihrer schlechten Generalisierungsfähigkeit, also der Fähigkeit gute Ergebnisse auch bei Daten zu erzielen, die sich signifikant von den Trainingsdaten unterscheiden. Aufgrund dieser mangelnden Robustheit können die Systeme neue Domänen nicht, oder nur schlecht bewältigen. Einer der Hauptgründe dafür ist, dass die für effizientes Training benötigten großen Mengen an Trainingsdaten oft nicht bzw. nur in fremden oder unpassenden Domänen verfügbar sind. Eine Domäne bezeichnet eine Sammlung von Texten mit ähnlichen Eigenschaften, wie Thema, Grad an Formalität oder Stil (Koehn 2020: 239). Da Wörter aber je nach Domäne unterschiedlich übersetzt werden können, führt dies oft zu unpassenden bzw. falschen Übersetzungen. Im Englischen wird dieses Phänomen als *domain mismatch* bezeichnet. Solche *domain mismatches* können Fehlübersetzungen produzieren, die zwar durchaus flüssig klingen können, aber aufgrund der unterschiedlichen Domänen oft wenig bis gar nicht mit dem ursprünglichen Ausgangstext übereinstimmen (2020: 293ff.).

Die **Domänenadaption** versucht dieses Problem zu minimieren, indem ein MÜ-System, zusätzlich zum Training mit großen Mengen allgemeinen Trainingsdaten, durch gezieltes Training mit domänenspezifischen Daten, an die sprachlichen Besonderheiten einer bestimmten Domäne angepasst wird. Dadurch wird eine Erhöhung der Relevanz domänenspezifischer

Information erreicht und das MÜ-System für ein bestimmtes Fachgebiet optimiert (2020: 239ff.). Darüber hinaus ist ebenso die Beschaffung von ausreichend **bilingualen Trainingsdaten**, besonders für ressourcenarme Sprachenpaare, eine weitere Herausforderung im Bereich der NMÜ. Aktuell werden viele Versuche unternommen die NMÜ für ressourcenarme Sprachen zu verbessern. Darunter fallen beispielsweise Strategien wie Wissenstransfer und halb-überwachtes sowie unüberwachtes Lernen. In diesem Kontext ist ein Ansatz mithilfe von Wissenstransfer, Sprachmuster und -strukturen, die aus einer ressourcenreichen Sprachen erlernt wurden, auf ressourcenarme Sprachen zu transferieren (Limkonchotiwat et al. 2022: 2143). Halb-überwachte Ansätze versuchen Verbesserungen mittels Kombination aus limitierten annotierten Trainingsdaten und einer großen Menge an monolingualen Daten, also große nicht annotierte Trainingskorpora, zu erzielen und unüberwachte Ansätze verwenden ausschließlich monolinguale Daten und streben Verbesserungen im Bereich der NMÜ komplett ohne annotierte bilinguale Trainingsdaten an (Zhang & Zong 2020: 2035f.)

Auf Dokumentenebene müssen Übersetzungen konsistent sein und Wörter mit gleicher Bedeutung müssen im gesamten Ausgangstext gleich übersetzt werden. Sprache ist jedoch ambig und deshalb können Wörter innerhalb eines Satzes mehrdeutig sein. Oftmals erschließt sich der Sinn dieser ambigen Wörter nur aus dem Kontext der umgebenden Sätze oder Absätze. Darüber hinaus können Diskursphänomene, wie beispielsweise Koreferenzen in kontextunabhängigen Satz-für-Satz-Übersetzungen nicht immer korrekt aufgelöst werden (2020: 2031). Die **Coreference Resolution** ist sowohl für Menschen als auch für automatische Übersetzungsmaschinen überdies eng mit der Frage nach dem Geschlecht verbunden.

Der geschlechtsspezifische Bias, speziell bei Berufen (Rudinger et al. 2018: 8) oder allgemeinen Tätigkeiten wie Kochen (Zhao et al. 2017: 2979f.), ist ein wachsendes Problem, welches das NMÜ-System nicht zuletzt von verzerrten Daten lernt und somit den menschengemachten Bias übernimmt, oder im schlimmsten Fall reproduziert und damit sogar verstärkt (Rudinger et al. 2018: 8). Trotz einiger Forschungsbestrebungen sind diese genannten Phänomene nach wie vor bestehende Herausforderungen in der NMÜ.

2.1.6.5 Neuronale Sprachmodelle für NMÜ

Aktuelle Weiterentwicklungsbestrebungen im Bereich der NMÜ konzentrieren sich vermehrt auf die Erforschung neuronaler, auf der Transformer-Architektur basierender Sprachmodelle, wie beispielsweise *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al. 2019) oder *Generative Pre-Trained Transformer* (GPT). Diese Modelle können mithilfe

von selbstüberwachtem Lernen aus umfangreichen, nicht annotierten Textkorpora „selbstständig“ kontextualisierte Sprachrepräsentationen lernen. Selbstüberwachte Trainingsansätze benötigen daher, im Gegensatz zu überwachtem Lernen, keine großen Mengen annotierter Trainingsdaten, da das Modell die Merkmale selbst aus den nicht annotierten Daten extrahieren kann (Jaiswal et al. 2021: 1). Anschließend können diese Modelle inklusive der zuvor erlernten generellen Sprachrepräsentationen für spezifische Aufgabenstellungen durch Fine-Tuning, wie beispielsweise bei der in Kapitel 2.1.6.4 erwähnten Domänenadaption durch gezieltes Training mit domänenspezifischen Daten, weiterverwendet werden (Guo et al. 2021: 1740).

2.2 Qualität von Übersetzungen

Übersetzen ist eine komplexe kognitive, soziale, kulturelle, sprachliche und technologische Tätigkeit. Die Qualitätsmessung von Übersetzungen (engl. *Translation Quality Assessment / TQA*) gestaltet sich aufgrund dieser Komplexität oft sehr schwierig, denn für eine sinnvolle Definition von Übersetzungsqualität und einer formalen Methode zur Messung und Evaluierung ebenjener, müssen all diese verschiedenen Dimensionen und ihre Wechselwirkungen erfasst werden (Castilho et al. 2018: 10).

Im Laufe der Zeit wurden in der Translationswissenschaft viele Bestrebungen unternommen Übersetzungsqualität zu definieren, beispielsweise beschreiben Koby et al. (2014: 416f.) eine qualitativ hochwertige Übersetzung allgemein als eine Übersetzung, die sowohl die Zwecke und Bedürfnisse der Endnutzer*innen mit der erforderlichen Genauigkeit und Flüssigkeit erfüllt, als auch sämtlichen vorab getroffenen Vereinbarungen zwischen Anbieter*innen und Kund*innen entspricht und betonen dabei die Messbarkeit der beiden Parameter Genauigkeit und Flüssigkeit. Die Definition wird weiter eingeschränkt, sodass eine Übersetzung im engeren Sinn als qualitativ hochwertig gilt, wenn die Kernaussagen des Ausgangstexts präzise, mit korrekter Konnotation, Denotation, Granularität und Stil und unter Berücksichtigung der Zielkultur in die Zielsprache übertragen werden, sodass beim Lesen das Gefühl entsteht, die Übersetzung sei von Erstsprecher*innen der Zielsprache für Leser*innen der Zielkultur verfasst worden.

Die Skopostheorie (Reiß & Vermeer 1984) betont, dass Übersetzungen einen bestimmten Zweck (Skopos) erfüllen sollen und dass die Qualität einer Übersetzung durch ihre Funktionalität, ihre Anpassung an kulturelle Unterschiede und ihre Relevanz für den Zweck bestimmt wird. Der beabsichtigte Zweck soll effektiv erreicht werden, indem der Fokus auf das kommunikative Handeln zwischen Sender*in und Empfänger*in gelegt wird. Darüber hinaus wird die

funktionale Äquivalenz, also die Bestrebung, dass die Übersetzung die gleiche Funktion erfüllen soll wie der Ausgangstext, auch wenn unterschiedliche sprachliche Mittel zur Erreichung dieses Ziels verwendet wurden, hervorgehoben, die den Übersetzer*innen ermöglicht, die Bedürfnisse des Zielpublikums bei der Übersetzung zu berücksichtigen.

Basierend auf dem Konzept der Skopostheorie definiert Nord (2006: 15f.) Übersetzungsqualität aus funktionaler Perspektive. Qualität wird dabei anhand des Zwecks der Translationshandlung festgelegt, der sich aus der kommunikativen Situation ergibt, für die das Translat bestimmt ist. In dieser Perspektive spielen die Funktion des Textes und die Rezipient*innen eine entscheidende Rolle, da diesen bei der Rezeption eine spezifische Funktion zugeschrieben wird. Da Texte oft mehrere Funktionen erfüllen, müssen je nach kultureller Distanz zwischen Ausgangs- und Zielsituation adaptive Prozeduren unterschiedlicher Qualität und Quantität angewendet werden. Der Text selbst hat dabei keine feste Funktion, sondern erhält diese erst in der Rezeptionssituation. Die Qualität des Translats wird demnach anhand seiner Funktionsgerechtigkeit und der Loyalität gegenüber den Erwartungen der Handlungspartner*innen gemessen.

Übersetzungsqualität ist jedoch nicht nur in der Translationswissenschaft, sondern auch und besonders in der Sprachindustrie von zentraler Bedeutung, wenngleich sich die Vorstellungen darüber zwischen Wissenschaft und Industrie unterscheiden (Castilho et al. 2018: 16). Im Gegensatz zum theoretisch-wissenschaftlichen Qualitätsansatz der Translationswissenschaft, definiert die DIN EN ISO 17100 (Österreichisches Normungsinstitut, Institut, & Austrian Standards Institute 2016) für die Sprachindustrie die Anforderungen an qualitative Übersetzungs**dienstleistungen**. Die Qualität von Dienstleistungen wird an der Fähigkeit der Anbieter*innen die Leistung auf einem bestimmten Niveau und im Einklang mit den oft sehr subjektiven Erwartungen der Kund*innen zu erbringen gemessen und ist im Vergleich zu translationswissenschaftlichen Definitionsversuchen flexibler und stark auf die Perspektiven und Erwartungshaltungen von Kund*innen bezogen (Bruhn 2020: 37).

Trotz fehlenden Konsenses über die genaue Definition von Übersetzungsqualität gibt es dennoch einige für eine qualitative Übersetzung relevante Punkte, in denen in Wissenschaft und Praxis grundsätzliche Einigkeit besteht. Koby et al. (2014: 415f.) definieren fünf solcher Bereiche, nämlich die definitorische Notwendigkeit die breite Gruppe der Kund*innen klar in Auftraggeber*innen und Endnutzer*innen zu unterscheiden, die Notwendigkeit sicher zu stellen, dass Übersetzungen korrekt und flüssig sind, die Akzeptanz eines Spannungsverhältnisses und damit einhergehend die Akzeptanz über die Unmöglichkeit von Perfektion zwischen perfekter

Genauigkeit und fließendem Sprachklang, die Notwendigkeit Zweck und Zielgruppe der Übersetzung zu verstehen und die Tatsache, dass eine eindeutige Definition von Übersetzungsqualität nur mithilfe einer objektiven Messmethode sowohl für die Translationswissenschaft, als auch für die Sprachindustrie erzielt werden kann.

2.2.1 Qualität von MÜ

Ziel von MÜ ist es, Übersetzungen zu produzieren, die qualitativ mit Humanübersetzungen vergleichbar sind. Die Kriterien für die Bewertung der Qualität von maschinell erstellten Übersetzungen entsprechen daher im Wesentlichen den Qualitätsansprüchen und -kriterien von Humanübersetzungen (Chatzikoumi 2020: 138), mit dem Unterschied, dass MÜ einen breiteren Anwendungsbereich abdeckt (vgl. dazu Kapitel 2.1.2) und ein weitaus diverseres Zielpublikum anspricht, welches von gelegentlichen Internetnutzer*innen bis zu professionellen Übersetzer*innen reicht (Krenz & Ramlow 2008: 26ff.). Darüber hinaus ist die Qualität einer maschinell erstellten Übersetzung auch stark von den vorhandenen Trainingsdaten, insbesondere von deren Umfang und Qualität, dem Sprachenpaar, der Textsorte, dem Fachbereich und des gewählten MÜ-Systems bzw. Algorithmus abhängig (Schmalz 2019: 202).

2.2.2 Translation Quality Assessment

Das TQA ist ein wichtiges Instrument zur Qualitätssicherung von Übersetzungen und zielt in der Sprachindustrie hauptsächlich darauf ab, ein bestimmtes Qualitätsniveau zu gewährleisten und Kund*innen, Benutzer*innen, Leser*innen und anderen Zielgruppen qualitativ hochwertige Übersetzungen zu liefern (Doherty 2017: 131). Die Festlegung auf eine universell anwendbare Messmethode gestaltet sich dabei ähnlich komplex und schwierig, wie die in Kapitel 2.2. erwähnte Festlegung auf verbindliche Kriterien und eine einheitliche Definition von Übersetzungsqualität, da es auch im Bereich der Qualitätsmessung keine universellen Standards und keine einheitliche objektive Messmethode gibt. Obwohl die Translationswissenschaft die Notwendigkeit der Evaluierung von Übersetzungsqualität für die Sprachindustrie anerkennt, wird dennoch oft dieses Fehlen eines klar definierten formalen, nachvollziehbaren und einheitlichen Rahmens implizit kritisiert (Drugan 2013: 35).

Diese Unmöglichkeit eine objektive Messmethode zu definieren ist mitunter darin begründet, dass die Beurteilung der Übersetzungsqualität verschiedene Zwecke erfüllen kann und dabei unterschiedliche Modelle verwendet werden (müssen), die auf den jeweiligen Zweck der Qualitätsmessung abgestimmt sind. Demnach unterscheidet sich die Herangehensweise zwischen der Durchführung eines TQA im Rahmen von Produktions- und Industrieprozessen

deutlich von dem in bzw. für die wissenschaftliche Forschung durchgeführten TQA. Während in der Industrie das Hauptziel darin liegt, sicherzustellen, dass das gelieferte Produkt einem vordefinierten Standard entspricht und die Qualitätsanforderungen der Endkund*innen erfüllt, liegt der Zweck des TQA im akademischen Umfeld in der Regel darin, nachweisbare Fortschritte zu früheren Übersetzungsansätzen und Forschungsmethoden zu erreichen (Castilho et al. 2018: 11). Darüber hinaus spielt TQA in der Translationswissenschaft für die formale, akademische Ausbildung von Übersetzer*innen eine Rolle, wohingegen sich der Zweck des TQA in der Sprachindustrie auf die Klassifizierung von diversen beruflichen Zertifizierungen für ausgebildete Translator*innen bezieht (Doherty 2017: 131).

Auch wenn zwischen Wissenschaft und Industrie Einigkeit über die fehlende Möglichkeit eines objektiven TQA besteht, müssen Übersetzer*innen, Lektor*innen, Revisor*innen, Kund*innen und andere Personen bei ihrer täglichen Arbeit solche Qualitätsevaluierungen vornehmen, oft mit dem Ergebnis, dass keine objektive Messmethode über die Qualität entscheiden, sondern sie letztlich vom Markt bzw. den Rezipient*innen bestimmt wird (Drugan 2013: 35).

2.2.3 Fehlertypologien und -metriken

Aufgrund der kontextabhängigen und zweckorientierten Natur von Übersetzungen und der damit verbundenen Schwierigkeiten einen einheitlichen extrinsischen Qualitätsstandard festzulegen (Sager 1989: 91), erfolgt die Messung der Übersetzungsqualität oft auf Basis von intrinsischen, quantifizier- und klassifizierbaren textuellen Eigenschaften in der Übersetzung (Secară 2005: 39). Diese quantitativen Qualitätsmesssysteme, s. g. Metriken, basieren vereinfacht ausgedrückt auf einer Fehlertypologie mit Punktabzug nach Art und Schwere der Fehler. Erst werden die gesammelten Fehlerpunkte aggregiert und anschließend von etwaigen Bonuspunkten abgezogen. Dieser Prozess resultiert in einer Punktzahl, die eine Klassifizierung der Übersetzung auf einer Qualitätsskala ermöglicht und anhand dieser Rückschlüsse darauf erlaubt, wie gut die vordefinierten Qualitätsstandards erfüllt werden. Zu diesen Metriken zählen beispielsweise das LISA QA Modell, der SAE J2450 Qualitätsstandard und die Multidimensional Quality Metrics (MQM) (Lommel et al. 2014).

2.2.3.1 LISA QA-Modell

Das 1995 von der Localization Industry Standards Association (LISA) eingeführte LISA-QA-Modell war ein System zur Qualitätsbewertung und -sicherung von Lokalisierungsprojekten. Obwohl das Modell nie formal standardisiert wurde, diente es, seit seiner Einführung bis zu

seiner Auflösung im Jahr 2011, als de-facto-Standard für die Qualitätsbeurteilung und -messung von Produktdokumentation, Benutzeroberflächen und anderer Lokalisierungssoftware (Castilho et al. 2018: 15; Lommel 2018: 112f.). Das Modell basierte auf mehreren Elementen, wie einer vordefinierten Liste von Fehlerschweregraden, einer Auflistung von verschiedenen Fehlerkategorien, einem Aufgabenkatalog für Revisor*innen und einer Vorlage, die es anhand eines Akzeptanzschwellwerts ermöglichte, die Übersetzung freizugeben oder abzulehnen (Mateo 2014: 77).

Da das Modell für die Lokalisierungsindustrie entworfen wurde, spielten neben sprachlichen Qualitätsaspekten besonders auch funktionale Aspekte, die mit Testroutinen einfacher objektiv gemessen werden konnten (Dunne 2006: 95) eine zentrale Rolle und nur einer der insgesamt acht Evaluierungsbereiche befasste sich explizit mit Sprache. Die sprachliche Fehlertypen waren Fehlübersetzung, Genauigkeit, Terminologie, Sprache, Stil, Land und Konsistenz (Jiménez-Crespo 2015: 74). Die Klassifizierung der Fehlertypen erfolgte in drei vordefinierten Schweregraden: geringfügig (engl. *minor*), schwerwiegend (engl. *major*) und kritisch (engl. *critical*). Der Schweregrad wird durch Faktoren, wie die Bedeutung des Fehlers, seine Sichtbarkeit im Dokument und sein Potenzial, Probleme zu verursachen, bestimmt (Mateo 2014: 78).

Das Modell umfasste 18 bzw. 21 Fehlerkategorien zur Beschreibung der Anforderungen an ein Übersetzungsprojekt. Die Diskrepanz bei den Fehlerkategorien ist auf Unstimmigkeiten zwischen dem Benutzerhandbuch und dem eigentlichen Interface zurückzuführen (Lommel 2018: 112). Die Bewertungspunkte wurden nach Anzahl und Art der gefundenen Fehler vergeben und die gesamte Übersetzung wurde anhand eines durch Bewerter*innen festgelegten Schwellenwerts als akzeptiert (engl. *Pass*) oder nicht akzeptiert (engl. *Fail*) klassifiziert (Castilho et al. 2018: 15) und durfte, um akzeptiert zu werden, jedenfalls keine Fehler der Kategorie kritisch enthalten (Mateo 2014: 78). Die Kriterien für die Einstufung von Fehlern als geringfügig, schwerwiegend oder kritisch waren jedoch nur unzureichend definiert, was aufgrund dieser mangelnden objektiven Bewertungsmöglichkeit oft zu erheblichen Unterschieden zwischen den Bewerter*innen führte (Jiménez-Crespo 2015: 73).

2.2.3.2 SAE J2450

Im Jahr 2001 wurde von einer Arbeitsgruppe, bestehend aus Vertreter*innen der Society of Automotive Engineers (SAE) und General Motors (GM), eine Metrik zur Qualitätssicherung von Übersetzungen entwickelt. Der Fokus, der ursprünglich als empfohlene Praxis für

Oberflächenfahrzeuge eingeführten Metrik (Mateo 2014: 78), lag anfänglich auf der Etablierung eines objektiv messbaren und standardisierten Übersetzungsstandards von KFZ-Serviceinformationen (SAE International J2450 2001: 1), der sowohl sprachunabhängig, im Sinne der Unabhängigkeit von Ausgangs- und Zielsprache, als auch methodenunabhängig, d. h. unabhängig von der eingesetzten Übersetzungsmethode, sei sie menschlich oder maschinell, sein sollte. Im Jahr 2005 wurde die Metrik dann in eine Norm umgewandelt (Mateo 2014: 78).

Die SAE J2450 wurde für ein Zielpublikum, welches hauptsächlich aus KFZ-Servicetechniker*innen bestand entwickelt und bewertet daher Übersetzungen ausschließlich auf linguistischer, insbesondere syntaktischer Ebene, ohne Berücksichtigung etwaiger Stil- oder Registerfehler. Folglich kann eine Übersetzung, die keinerlei SAE J2450-Fehler enthält, aus anderen Gründen für andere Bereiche inakzeptabel sein (Woyde 2001: 39), beispielsweise in Fachbereichen, für die stilistische Korrektheit von zentraler Bedeutung ist. Ungeachtet dessen fand sie dennoch in verschiedenen anderen Bereichen, wie beispielsweise in der Pharmazie, Anwendung. Trotz ihrer Anpassungsfähigkeit weist die Norm keine festen Akzeptanzschwellen oder Anleitungen für die Bewertung von Ergebnissen auf, da ihr Fokus ausschließlich auf der Erkennung von sprachlichen Fehlern, auch jener Fehler, die von einem fehlerhaften Ausgangstext stammen, liegt. Diese sprachlichen Fehler werden von menschlichen Annotator*innen nach Art und Schweregrad bewertet und der daraus ermittelte gewichtete numerische Wert ergibt die insgesamte Qualität der Übersetzung (Mateo 2014: 78f.).

Die SAE J2450 besteht aus vier Elementen, nämlich sieben Fehlerkategorien, zwei Subkategorien für die Schwergrade, zwei Regeln als Entscheidungshilfe bei Zweifelsfällen in Bezug auf die Fehlerzuordnung in Haupt- und Subkategorien, und dem gewichteten numerischen Wert (SAE International J2450 2001: 1). Tabelle 1 zeigt den Aufbau und die Elemente der SAE J2450 Qualitätsmetrik und gibt einen detaillierten Überblick über die einzelnen Kategorien.

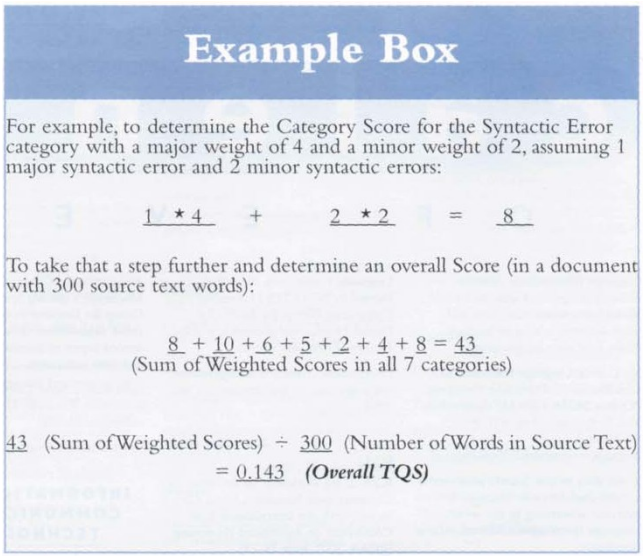
Tabelle 1: Fehlerkategorien und Subkategorien der SAE J2450 (Mateo 2014: 79)

Main Category	**(abb.)**	**Sub-Category** (abbreviation)	**Weight** serious-minor
Wrong Term	(WT)	serious (s)	5/2
Syntactic Error	(SF)	minor (m)	4/2
Omission	(OM)		4/2
Word Structure or Agreement Error	(SA)		4/2
Misspelling	(SP)		3/1
Punctuation Error	(PE)		2/1
Miscellaneous Error	(ME)		3/1

Zu den sieben in der SAE J2450 verwendeten und absteigend nach Schweregraden geordneten Fehlerkategorien zählen: falsche Bezeichnung (engl. *Wrong Term*), Syntaxfehler (engl. *Syntactic Error*), Auslassungen (engl. *Omission*), morphologische Fehler (engl. *Word Structure or Agreement Error*), Orthografiefehler (engl. *Misspelling*), Interpunktionsfehler (engl. *Punctuation Error*) und sonstige Fehler (engl. *Miscellaneous Error*) (Mateo 2014: 79; SAE International J2450 2001: 5ff.).

Die Zuordnung der Fehler in die jeweiligen Kategorien erfolgt durch menschliche Annotator*innen auf Basis zweier als Entscheidungshilfe entwickelter Metaregeln. Dadurch werden die Fehler nicht nur den sieben Hauptfehlerkategorien zugeordnet, sondern auch einer von zwei Unterkategorien, nämlich geringfügig (engl. *minor*) und schwerwiegend (engl. *serious*). Die beiden Regeln legen im Zweifel den strengeren Maßstab an und besagen im Falle von Zweifeln bei der Auswahl der Fehlerkategorie immer die höher gelistete, und demnach, aufgrund der absteigenden Reihung nach Schweregraden, schwerwiegendere, Kategorie, und im Falle von Zweifeln bei der Auswahl der Fehlerschwere in der Subkategorie, immer eher schwerwiegend als geringfügig zu wählen (Mateo 2014: 79; Woyde 2001: 39).

Nach Zuordnung aller Fehler in die sieben Fehlerkategorien und die beiden Subkategorien für Schweregrade, werden sie mit der vorgegebenen Gewichtung (Tabelle 1, Spalte 4 „*Weight serious-minor*“) multipliziert und die daraus resultierenden Werte addiert. Für die Ermittlung der Gesamtpunktezahl des gesamten Dokuments (engl. *Overall Translation Quality Score / Overall TQS*) wird dieser Wert anschließend durch die Gesamtanzahl der Wörter des zu übersetzenden Ausgangstext dividiert (Woyde 2001: 39). Abbildung 9 illustriert beispielhaft eine solche Berechnung mit der SAE J2450 Qualitätsmetrik.


Example Box

For example, to determine the Category Score for the Syntactic Error category with a major weight of 4 and a minor weight of 2, assuming 1 major syntactic error and 2 minor syntactic errors:

1 * 4 + 2 * 2 = 8

To take that a step further and determine an overall Score (in a document with 300 source text words):

8 + 10 + 6 + 5 + 2 + 4 + 8 = 43
(Sum of Weighted Scores in all 7 categories)

43 (Sum of Weighted Scores) ÷ 300 (Number of Words in Source Text)
= 0.143 ***(Overall TQS)***

Abbildung 9: Beispiel einer Rechnung zur Ermittlung des *Overall TQS* einer Übersetzung (Woyde 2001: 39)

Während diese Norm als solide Grundlage für die Reduzierung von linguistischen bzw. syntaktischen Fehlern dient, gibt es einige Aspekte im TQA-Bereich, die nicht mithilfe von Metriken messbar sind (Woyde 2001: 39). Die konsistente Erstellung hochqualitativer Übersetzungen bedarf einer Kombination aus unterschiedlichen Maßnahmen, und die Metrik als solche, stellt nur ein Glied in dieser Kette von Maßnahmen zur Gewährleistung, Überprüfung und Verbesserung der Übersetzungsqualität dar (Mateo 2014: 78; Woyde 2001: 39).

2.2.3.3 Multidimensional Quality Metrics (MQM)

Die MQM leitet sich zu einem großen Teil aus dem LISA QA-Modell ab und basiert folglich auf einigen Prinzipien dieses Modells (Castilho et al. 2018: 17). Im Unterschied zu früheren Qualitätsmetriken wurde die MQM jedoch von Beginn an flexibel konzipiert, einerseits, um mit anderen Standards kompatibel zu sein und andererseits, um sowohl die Nachteile von starren Einheitsmodellen als auch die Nachteile von zu hoher Fragmentierung aufgrund nahezu unbegrenzter Anpassungsmöglichkeiten zu vermeiden (Castilho et al. 2018: 17; Lommel et al. 2014: 458). Darüber hinaus ist die MQM der Versuch ein Rahmenwerk für die Evaluierung von Humanübersetzungen einerseits und MÜ andererseits zu schaffen. Mit der MQM wird also vornehmlich das Übersetzungsprodukt bewertet, unabhängig davon, ob es von Menschen oder Maschinen erstellt wurde, und sie lässt sich somit nicht nur auf professionelle Übersetzungen, sondern auch auf alle Arten der maschinellen Übersetzungen anwenden (Castilho et al. 2018: 17).

Die einzelnen Fehlerkategorien der MQM entstanden aus einem detaillierten Vergleich, inklusive fehlertypologischer Harmonisierung, von insgesamt neun unterschiedlichen Fehlertypologien bzw. Metriken (siehe Lommel 2018: 114), darunter u. a. das LISA QA-Modell und die SAE J2450 Qualitätsmetrik. Es wurden 145 Fehlerpunkte aus den verschiedenen Typologien abgeleitet, wobei nur ein Fehlertyp in allen Metriken und 23 in mehr als einer gefunden wurden. Zu den häufigsten Fehlerarten zählten Terminologie/Glossar, Auslassungen, Interpunktion, Konsistenz, Grammatik/Syntax, Orthografie, Übersetzungsfehler und Stil (Lommel 2018: 114). Die MQM zeichnet sich durch eine hierarchische Baumstruktur aus und besteht aus vier Ebenen mit steigendem Grad an Spezifität. Abbildung 10 visualisiert die verschiedenen Dimensionen der obersten Ebenen der MQM, die aus acht Hauptdimensionen, nämlich Genauigkeit (engl. *Accuracy*), Design (engl. *Design*), Flüssigkeit (engl. *Fluency*), Internationalisierung (engl. *Internationalisation*), Lokalkonventionen (engl. *Locale convention*), Stil (engl. *Style*), Terminologie (engl. *Terminology*) und Faktentreue (engl. *Verity*) besteht (Lommel 2018: 116ff.).

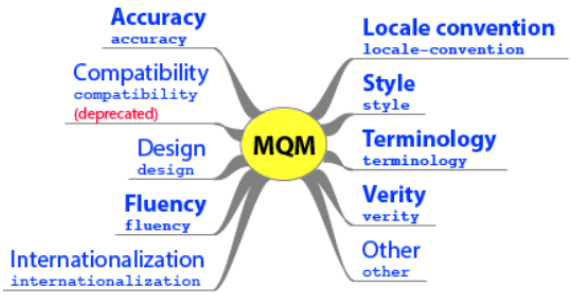


Abbildung 10: Dimensionen der obersten Ebenen der MQM (Lommel et al. 2015: o. S.)

Jede dieser Dimensionen enthält wiederum weitere Aspekte auf ihren spezifischeren Unterebenen (siehe Lommel et al. 2015) . In der Regel konzentriert sich der Großteil der Hierarchie auf die zweite oder dritte Ebene. Aspekte der zweiten Ebene umfassen in der Dimension Flüssigkeit beispielsweise Grammatik (engl. *Grammar*) und Inkonsistenz (engl. *Inconsistency*) und auf dritter Ebene (spezifischere Unterebene von Grammatik) zum Beispiel Wortform (engl. *Word form*) und Wortstellung (engl. *Word order*). Jedes Element verfügt dabei über zwei Bezeichnungen, nämlich einen menschenlesbaren Namen für die Oberfläche, der übersetzt werden kann, sowie eine konstante ID für XML und die meisten Programmiersprachen (2018: 116ff.).

Abbildung 11 zeigt die weiteren drei Ebenen der hierarchischen Baumstruktur der MQM am Beispiel der Dimension Flüssigkeit. Die zweite Ebene ist in blauer Schrift, die dritte in grüner Schrift und die vierte in roter Schrift dargestellt.

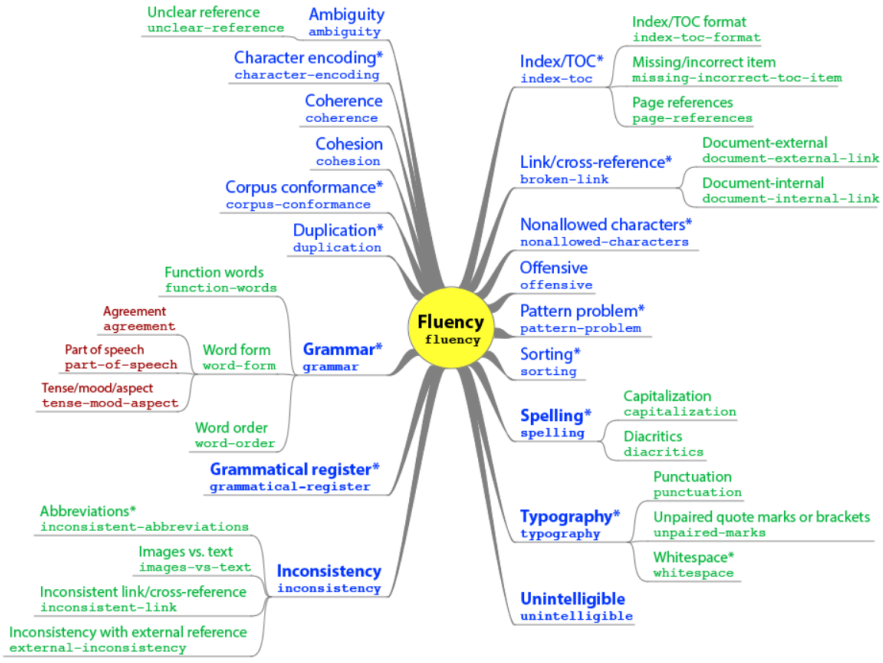


Abbildung 11: Baumstruktur der MQM am Beispiel der Dimension Flüssigkeit (Lommel et al. 2015: o. S.)

Für die Evaluierung der Übersetzungsqualität mithilfe der MQM spielen neben der Anzahl der auftretenden Fehler auch deren Schweregrade (engl. *severity*) und ihre Relevanz (engl. *importance*) für das jeweilige Übersetzungsprojekt eine entscheidende Rolle. Die auftretenden Fehler werden von Annotator*innen nach Schweregrad in eine von vier Kategorien, nämlich *critical*, *major*, *minor* und *null* eingeteilt, wobei jeder Fehler basierend auf der Kategorie mit bestimmten Strafpunkten versehen ist, die zur Gesamtbewertung beitragen. Standardmäßig werden 100 Strafpunkte für Fehler der Kategorie *critical*, 10 für Fehler der Kategorie *major*, 1 für Fehler der Kategorie *minor* und 0 für Fehler der Kategorie *null* (also keine Strafpunkte) vergeben. Diese Standardeinstellung dient der Schaffung einer klaren Unterscheidung zwischen den Auswirkungen kritischer Fehler im Vergleich zu schwerwiegenden oder geringfügigen Fehlern (Lommel 2018: 120f.).

Die Relevanz bezieht sich im Gegensatz zum Schweregrad nicht auf individuelle Fehler, sondern auf die relative Gewichtung der verschiedenen Dimensionen, wobei hier standardmäßig alle Dimensionen mit 1,0 gewichtet werden. Je nach Übersetzungsprojekt kann dieser Standardwert für jede Dimension individuell von Annotator*innen festgelegt bzw. verändert werden, um Dimensionen mit höheren Werten mehr Bedeutung und Dimensionen mit niedrigeren Werten weniger Bedeutung zu geben. Ist beispielsweise die korrekte Einhaltung der Terminologie für ein Projekt von besonderer Wichtigkeit, können Annotator*innen der Dimension Terminologie eine Gewichtung von 1,5 zuweisen. Dies hat zur Folge, dass Terminologiefehler um 50 % mehr zur Gesamtbewertung beitragen als Fehler in Dimensionen mit dem Standardwert 1,0. Dieses System lässt sich umgekehrt auch auf Dimensionen mit weniger relativer Bedeutung für das konkrete Projekt anwenden. Ist beispielweise die Dimension Stil für das konkrete Übersetzungsprojekt von nachrangiger Bedeutung, kann die Relevanz auf 0,5 reduziert werden.

Um über die reine Fehleridentifizierung hinaus die Gesamtbewertung (engl. *Score*) für eine Übersetzung zu ermitteln wird im ersten Schritt jeder Fehler mit seinem Schweregrad und seiner Relevanz multipliziert. Anschließend werden die daraus resultierenden Strafpunkte (engl. *penalties* bzw. *penalty points*) addiert, um den Gesamtwert der Strafpunkte zu ermitteln (2018: 121f.).

Tabelle 2 zeigt eine vereinfachte bespielhafte Berechnung der Strafpunkte anhand von drei Dimensionen mit dem standardmäßigen Schweregrad für die Kategorien *null* (0), *minor* (1), *major* (10) und *critical* (100) und einer individuellen Relevanz von 1,5 in der Dimension Terminologie, 1,0 in der Dimension Genauigkeit und 0,5 in der Dimension Stil.

Tabelle 2: Beispielhafte Darstellung der Berechnung des Gesamtwerts der Strafpunkte mittels MQM

		Gewichtung	*null* (0)	*minor* (1)	*major* (10)	*critical* (100)	Ergebnis
Dimensionen	Terminologie	1,5		2	1		18 (3 + 15)
	Genauigkeit	1,0				1	100
	Stil	0,5		3			1,5
						Gesamt	119,5

Um die finale Gesamtbewertung zu erhalten, wird im nächsten Schritt der Gesamtwert der Strafpunkte durch die Anzahl der Wörter des gesamten Texts dividiert und die daraus resultierende Dezimalzahl, eine Zahl zwischen 0 und 1, von 1 subtrahiert. Formel 1 zeigt die Formel zur Berechnung der Gesamtbewertung einer Übersetzung auf Basis des zuvor ermittelten Werts der Strafpunkte mithilfe der MQM (2018: 122).

Formel 1: Formel zur Berechnung der finalen MQM-Gesamtbewertung

$$Score = 1 - \frac{Penalties}{WordCount}$$

Ein beispielhafter Text mit 2.000 Wörtern und dem zuvor ermittelten Gesamtwert der Strafpunkte von 119,5 ergibt, nach Anwendung der Formel einen Gesamtwert (engl. *Score*) von 94 % (0,94025) für die gesamte Übersetzung. Abschließend wird der errechnete Gesamtwert mit den individuell festgelegten Schwellenwerten verglichen, um zu entscheiden, ob die Gesamtübersetzung akzeptiert werden kann. Diese Schwellenwerte basieren beispielsweise auf Referenzübersetzungen, die von den Auftraggeber*innen zur Verfügung gestellt werden. Sie umfassen sowohl zufriedenstellende als auch unzufriedenstellende Beispielübersetzungen und dienen als Maßstäbe für künftige Übersetzungen, indem ihre Punktzahlen als Benchmarks für Bewertungen fungieren (2018: 122).

Dieses System veranschaulicht die flexible Natur der MQM. Sie lässt sich unkompliziert an unterschiedliche Zwecke und Übersetzungsprojekte anpassen, wodurch Nutzer*innen die Möglichkeit erhalten, rasch und effizient qualitätsmindernde Faktoren zu identifizieren und gezielte Verbesserungen in diesen Bereichen vorzunehmen. Wenn beispielswiese Rechtschreibfehler ein besonderes Problem darstellen, können zusätzliche Rechtschreibprüfungen vor der Übermittlung der finalen Übersetzung integriert werden (2018: 122).

2.3 Qualitätsevaluierung von MÜ

Der Prozess der Qualitätsevaluierung von maschinell erstellten Übersetzungen umfasst die Analyse der Übersetzungen und die Beurteilung ihrer Korrektheit mittels sinnvoller und messbarer Methoden. Diese Evaluierung kann mithilfe von Qualitätsmetriken **automatisiert** oder von Menschen **manuell** durchgeführt werden. Obwohl die manuelle Evaluierung, bei der Annotator*innen maschinell erstellte Übersetzungen qualitativ bewerten, äußerst zuverlässig ist, stellt sie aufgrund des hohen Arbeitsaufwands eine kosten- und zeitintensive Methode dar. Im Gegensatz dazu bedient sich die automatisierte Evaluierung spezieller Metriken, die qualitative Merkmale mit geringem Aufwand als numerische Werte darstellen. Im Vergleich zur manuellen Evaluierung ist diese Methode zwar weniger präzise, wird aber aufgrund der Kostenersparnis durch nahezu unbegrenzt anwendbare Iterationen besonders im Entwicklungsprozess vermehrt eingesetzt (Avramidis 2019: 2).

Die Metriken für die automatisierte Evaluierung berechnen Korrektheit und Qualität von maschinell erstellten Übersetzungen und nutzen hierfür von Menschen angefertigte Übersetzungen als Goldstandard bzw. Referenzdaten (Avramidis 2019: 2; Chatzikoumi 2020: 139). Ein weiterer Ansatz zur Qualitätsbewertung von maschinellen Übersetzungen ist die Quality Estimation (QE), bei der die Übersetzungsqualität ohne den Vergleich mit Goldstandard- oder Referenzübersetzungen geschätzt wird. Ziel der QE ist es, die Übersetzungsqualität direkt vorherzusagen und sich von der oft variierenden Qualität menschlicher Referenzen unabhängig zu machen (Specia et al. 2010: 10). Der gesamte Evaluierungsprozess dient nicht nur der Qualitätsverbesserung der Übersetzungen, sondern darüber hinaus auch der Erforschung und Weiterentwicklung von MÜ-Systemen, da er messbare Vergleiche zwischen verschiedenen MÜ-Systemen, angewandten Methoden und verwendeten Daten ermöglicht (Avramidis 2019: 2).

2.3.1 Automatisierte Evaluierung von MÜ

Kontinuierliche Tests zur Qualitätsüberprüfung sind ein elementarer Bestandteil der Entwicklung von MÜ-Systemen. Die automatisierte Evaluierung von MÜ-Resultaten wird daher nicht zuletzt aus Gründen der Kostenersparnis und der nahezu unbegrenzten Wiederverwendbarkeit automatisierter Testmetriken insbesondere im Entwicklungsprozess eingesetzt (Avramidis 2019: 2). Sie basiert auf Metriken, also „scripts or software programs that perform TQA using an explicit and formalized set of linguistic criteria to evaluate the MT output, typically against a human reference translation or ‘gold standard’“ (Doherty 2017: 134). In Anbetracht der Tatsache, dass für jeden Ausgangssatz mehrere korrekte zielsprachliche Übersetzungen existieren

können (Farrús et al. 2010), stellt diese Goldstandard-Humanübersetzung lediglich eine Stichprobe an möglichen korrekten Übersetzungen für einen bestimmten Ausgangstext dar (Dorr et al. 2011: 759).

Eine der frühesten Evaluierungsmetriken, die auf der Messung der Ähnlichkeit zwischen zwei Zeichenketten basiert, ist die Editierdistanz nach Levenshtein (1966). Sie misst die Mindestanzahl an notwendigen Änderungen, um eine Zeichenfolge *a* in eine Zeichenfolge *b* umzuwandeln. Diese Änderungen können in Form von Einfügungen, Löschungen oder Substitutionen einzelner oder mehrerer Zeichen einer Zeichenfolge erfolgen. Umgelegt auf die Evaluierung einer Übersetzung umfasst die Levenshtein-Distanz folglich die Anzahl an notwendigen Änderungen, um das MÜ-Resultat der Referenzübersetzung anzugleichen und bezieht sich nicht auf einzelne Zeichen, sondern auf die Wörter eines Satzes (Chatzikoumi 2020: 140).

Neben der Levenshtein-Distanzmessung existiert eine Vielzahl weiterer gängiger automatisierter Evaluationsmetriken, wie beispielsweise die auf dem Konzept der Levenshtein-Distanz basierende Word Error Rate (WER), bzw. eine Erweiterung davon die Multi-Reference WER (MWER) (Nießen et al. 2000), die ebenfalls von der WER abgeleitete Translation Edit Rate (TER) (Snover et al. 2006), die Bilingual Evaluation Understudy (BLEU) (Papineni et al. 2002) und die Metric for Evaluation of Translation with Explicit ORdering (METEOR) (Banerjee & Lavie 2005). Damit diese Metriken aber effizient genutzt werden können müssen sie bestimmte Kriterien erfüllen, wie beispielsweise eine hohe Korrelation mit menschlicher Auffassung von Qualität (Chatzikoumi 2020: 140), Konsistenz, Zuverlässigkeit, Benutzerfreundlichkeit, Adaptationsfähigkeit auf verschiedene Bereiche (Banerjee & Lavie 2005: 66) und eine kostensparende Anwendung (Koehn 2010: 220).

2.3.1.1 (M)WER

Die WER ist eine automatisierte Metriken, die häufig in der automatischen Spracherkennung (engl. *Automatic Speech Recognition* / *ASR*) verwendet wird (Arif et al. 2025; Sheshadri et al. 2021). Im Kontext von MÜ ist die WER die auf Wortebene berechnete Levenshtein-Distanz zwischen den Wörtern einer maschinellen Übersetzung und den Wörtern der dazugehörigen Referenzübersetzung und wird berechnet, indem die ermittelte Levenshtein-Distanz durch die Gesamtzahl aller Wörter der Referenzübersetzung dividiert wird (Dorr et al. 2011: 760).

Hierbei ermöglicht die Verwendung von dynamischer Programmierung die Ermittlung der optimalen Zuordnung zwischen den Wörtern der MÜ und denen der Referenzübersetzung. Jedes Wort des MÜ-Resultats wird entweder einem oder keinem Wort der Referenzübersetzung

zugeordnet. Existiert für ein Wort der Referenzübersetzung keine Entsprechung im MÜ-Resultat spricht man von einer Löschung. Im Gegensatz dazu ist eine Einfügung das Fehlen eines Äquivalents in der Referenzübersetzung für ein Wort im MÜ-Resultat. Kann ein Wort der Referenzübersetzung direkt einem Wort des MÜ-Resultats zugeordnet werden, spricht man von einer Übereinstimmung, andernfalls von einer Substitution (2011: 760).

Die Berechnung der WER erfolgt auf Basis der Anzahl aller für die Angleichung des MÜ-Resultats an die Goldstandard-Referenzübersetzung notwendigen Änderungen, indem die Summe aller Substitutionen (*S*), Einfügungen (*I*) und Löschungen (*D*) durch die Gesamtzahl aller Worte der Referenzübersetzung (*N*) geteilt wird (2011: 760). Formel 2 veranschaulicht die Berechnung der WER.

Formel 2: Formel zur Berechnung der WER

$$WER = \frac{S + I + D}{N}$$

Nießen et al. (2000) schlagen die MWER, eine erweiterte Version der WER, vor. Die MWER bezieht sich nicht ausschließlich nur auf eine einzelne Referenzübersetzung, sondern verwendet zur Berechnung alle Referenzübersetzungen, die mit der höchsten Punktezahl (10) als perfekt bewertet wurden (daher *multi-reference*). Dabei wird der WER-Wert zwischen einer maschinell erstellten Übersetzung und jeder in Frage kommenden Referenzübersetzung ermittelt. Die Berechnung der MWER basiert auf dem niedrigsten WER-Wert, also dem Wert mit der größten Ähnlichkeit zwischen MÜ-Resultat und Referenzübersetzung (Dorr et al. 2011: 760).

2.3.1.2 Genauigkeit und Trefferquote

Genauigkeit und Trefferquote sind zwei gängige Metriken zur Evaluierung von NLP-Systemen (Turian et al. 2003). In der Qualitätsevaluierung von maschinell erstellten Übersetzungen misst, wie in Formel 3 dargestellt, Genauigkeit (*P*) den Anteil korrekter Wörter im MÜ-Ergebnis (*h*) im Verhältnis zur Vergleichsreferenz (*r*) bzw. die Genauigkeit der Übersetzung, während die Trefferquote (*R*) den Anteil der Wörter in der Vergleichsreferenz (*r*) ermittelt, die mit den Wörtern im MÜ-Ergebnis (*h*) übereinstimmen bzw. die Vollständigkeit der Übersetzung abbildet (Munkova et al. 2020: 1329).

Formel 3: Formel zur Berechnung von Genauigkeit (links) und Trefferquote (rechts)

$$P(h|r) = \frac{|h \cap r|}{|h|} \quad R(h|r) = \frac{|h \cap r|}{|r|}$$

Genauigkeit wird berechnet, indem die Anzahl aller Wörter des MÜ-Resultats, die in den Referenzübersetzungen vorkommen durch die Gesamtwortzahl des MÜ-Resultats dividiert wird (Papineni et al. 2002: 312). Die Berechnung der Trefferquote erfolgt durch Division der Anzahl aller Wörter des MÜ-Resultats, die in den Referenzübersetzungen vorkommen, durch die Gesamtwortzahl der Vergleichsreferenz (Lavie et al. 2004: 137). Viele der heutzutage verwendeten MÜ-Evaluationsmetriken, darunter auch BLEU und METEOR, basieren auf dem Interplay zwischen Genauigkeit und Trefferquote (Popović 2018: 130).

2.3.1.3 BLEU

BLEU ist eine genauigkeitsorientierte (Papineni et al. 2002: 312) Evaluationsmetrik für automatisierte Qualitätsbewertung von MÜ-Resultaten und bedient sich, ähnlich wie die MWER, mehrerer unterschiedlicher Vergleichsreferenzen. Der BLEU-Wert wird auf Basis der Genauigkeit bzw. Übereinstimmung von *n*-Grammen oder Wortsequenzen im Ergebnis einer maschinell erstellten Übersetzung im Vergleich mit Referenzübersetzungen ermittelt. Der Fokus der Bewertung mittels BLEU liegt demnach nicht auf der Überprüfung der vollständigen Wiedergabe der Referenzübersetzung in der MÜ (Trefferquote), sondern misst, in welchem Ausmaß das MÜ-Resultat korrekt ist (Genauigkeit).

Damit aufgrund der genauigkeitsorientierten Struktur der BLEU-Metrik das System des *n*-Gramm-Abgleichs nicht mittels Generierung sehr kurzer und daher sehr präziser *n*-Gramme zur Verzerrung des BLEU-Werts ausgenutzt werden kann, inkludiert die Formel zur Berechnung des BLEU-Werts ein Kontrollinstrument, die s. g. *brevity penalty (BP)*, das Strafpunkte vorsieht, wenn das MÜ-Resultat kürzer ist als die Vergleichsreferenzen (Callison-Burch et al. 2006: 250; Dorr et al. 2011: 760f.). Formel 4 veranschaulicht die Berechnung der *brevity penalty*, wobei *c* die Länge des MÜ-Resultats und *r* die Länge des Referenzkorpus darstellt (Papineni et al. 2002: 315).

Formel 4: Formel zur Berechnung der *brevity penalty*

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

Der BLEU-Wert wird dann berechnet, indem zunächst die modifizierte *n*-Gramm-Genauigkeit (p_n) der übereinstimmenden *n*-Gramme zwischen dem MÜ-Resultat und den Vergleichsreferenzen in einem Testkorpus ermittelt wird. Die *n*-Gramm-Übereinstimmungen werden zunächst

für jeden Ausgangssatz und die entsprechenden Zielsätze ermittelt. Um zu verhindern, dass Übersetzungen mit einer übermäßigen Anzahl an identischen *n*-Grammen bevorzugt werden, wird eine Clipping-Technik verwendet, die sicherstellt, dass die Anzahl der *n*-Gramme im MÜ-Resultat die Maximalanzahl derselben *n*-Gramme in den Vergleichsreferenzen nicht übersteigt. Die daraus resultierenden *n*-Gramm-Zahlen für alle Sätze des MÜ-Resultats werden anschließend summiert und der modifizierte Genauigkeitswert wird mittels Division dieser Summe durch die Gesamtanzahl der *n*-Gramme des MÜ-Resultats im gesamten Testkorpus berechnet. Formel 5 zeigt die finale Berechnung des gesamten BLEU-Wertes, der sich durch Multiplikation des geometrischen Mittels der modifizierten Genauigkeitswerte unter Berücksichtigung einer Gewichtung w_n und der *brevity penalty* ergibt (2002: 313ff.).

Formel 5: Formel zur Berechnung des BLEU-Werts

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

In der ursprünglichen Form resultierte die Berechnung des finalen BLEU-Werts in einer Zahl zwischen 0 und 1, wobei 1 die komplette Übereinstimmung des MÜ-Resultats mit der Referenzübersetzung bedeutet (2002: 315). Je näher der berechnete Wert also an der Zahl 1 liegt, desto höher ist die Qualität der Übersetzung. Heutzutage ist es in der Praxis üblich den BLEU-Wert mit 100 zu multiplizieren, um BLEU-Punkte zu erhalten. Dadurch wird nicht mehr nur ein Richtwert abgebildet, sondern es werden darüber hinaus interpretierbarere Ergebnisse in Form einer Punktewertung ermöglicht, die auch die Fortschrittsmessung vereinfachen (Way 2018: 163).

Obwohl Papineni et al. (2002: 318) bei der Evaluierung von Übersetzungen mit BLEU eine hohe Übereinstimmung mit den Ergebnissen menschlicher Bewertung, insbesondere durch monolinguale Annotator*innen, feststellten, gibt es unter MÜ-Forscher*innen einige kritische Stimmen. Callison-Burch et al. (2006: 255) stellten sowohl theoretisch als auch praktisch fest, dass ein verbessertes BLEU-Ergebnis nicht zwangsläufig eine echte Verbesserung der Übersetzungsqualität bedeutet. Die BLEU-Metrik sieht keine Einschränkungen hinsichtlich der Reihenfolge übereinstimmender *n*-Gramme vor und ermöglicht somit eine große Flexibilität in den MÜ-Resultaten, was bei Vorhandensein mehrerer Vergleichsreferenzen dazu führen kann, dass viele Übersetzungen denselben BLEU-Wert ergeben können (Way 2018: 168).

Die ungenaue BLEU-Metrik kann aufgrund der vielen zulässigen Übersetzungsvariationen zu einer Diskrepanz zwischen höheren BLEU-Werten und der tatsächlichen, vom

Menschen wahrgenommenen Übersetzungsqualität führen, da nicht alle Übersetzungsvarianten mit hohen BLEU-Werten gleichermaßen grammatikalisch und semantisch korrekt sind (Callison-Burch et al. 2006: 255f.) und es nicht zuletzt deshalb unwahrscheinlich scheint, dass menschliche Annotator*innen ähnlich viele Übersetzungen als qualitativ identisch betrachten würden (Way 2018: 168). Um den Nachteilen des genauigkeitsorientierten BLEU-Ansatzes entgegenzuwirken wurde 2005 die METEOR-Evaluationsmetrik entwickelt (Banerjee & Lavie 2005), die auf Wortübereinstimmungen, Trefferquote und der Miteinbeziehung weiterer linguistischer Elemente beruht (Popović 2018: 130).

2.3.1.4 METEOR

Im Gegensatz zur genauigkeitsorientierten BLEU-Metrik, berechnet die METEOR-Metrik sowohl Genauigkeit als auch Trefferquote und kombiniert beide Werte in einer Formel mit Fokus auf Trefferquote zur Berechnung des harmonischen Mittelwerts (Dorr et al. 2011: 762). Die Bewertung der Übersetzungen erfolgt dabei auf Basis expliziter Wort-zu-Wort-Übereinstimmungen zwischen MÜ-Ergebnis und Vergleichsreferenz. Die Wortübereinstimmungen werden mithilfe automatisierter Module zur Wortzuordnung schrittweise in einem dreistufigen, aufeinander aufbauenden Alignierungsprozess, also der Zuordnung von Wörtern, wo jedes Wort des MÜ-Ergebnisses höchstens einem Wort der Vergleichsreferenz entspricht, ermittelt. Im ersten Schritt erfolgt ein exakter Abgleich von Zeichenketten mittels *exact*-Modul, gefolgt von einem morphologischen Abgleich der Wortstämme mithilfe des *stemmer*-Moduls und einem abschließenden semantischen Synonymabgleich mittels *synonymy*-Modul (Agarwal & Lavie 2008: 115f.).

Auf Basis der daraus resultierenden Zuordnung wird zunächst die Anzahl der zugeordneten Wörter bzw. Unigramme (m) bestimmt, wobei die Gesamtzahl der Wörter in der Übersetzung (t) und in der Referenz (r) berücksichtigt wird. Genauigkeit (P) ist das Verhältnis der korrekt zugeordneten Wörter zur Gesamtzahl in der Übersetzung ($P = m/t$), während Trefferquote (R) das Verhältnis der korrekt zugeordneten Wörter zur Gesamtzahl in der Referenz beschreibt ($R = m/r$). Formel 6 veranschaulicht die Berechnung des parametrisierten harmonischen Mittelwerts (F_{mean}) aus Genauigkeit, Trefferquote und dem Parameter α, der als Gewichtungselement zwischen P und R fungiert (Agarwal & Lavie 2008: 116; Dorr et al. 2011: 814).

Formel 6: Formel zur Berechnung des parametrisierten harmonischen Mittelwerts

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1 - \alpha) \cdot R}$$

Da sich Genauigkeit, Trefferquote und F_{mean} nur auf die Übereinstimmung einzelner Wörter zwischen dem MÜ-Ergebnis und Vergleichsreferenz beziehen, sieht die METEOR-Metrik überdies Strafpunkte (engl. *Penalty / Pen*) vor, die es ermöglichen auch die Wortreihenfolge in die Evaluierung miteinzubeziehen. Hierfür wird zunächst auf Basis der übereinstimmenden Wortsequenzen die kleinstmögliche Anzahl an Chunks (*ch*) ermittelt. Ein Chunk ist dabei ein Textsegment, in dem die übereinstimmenden Wörter in identischer Reihenfolge nebeneinander angeordnet sind. Anschließend wird der Fragmentierungsanteil (*frag*) mittels Division der Chunks durch die Anzahl der Wortübereinstimmungen zwischen MÜ-Ergebnis und Vergleichsreferenz (*m*) berechnet (*frag* = *ch*/*m*). Die Strafpunkte (*Pen*) werden anschließend mit dem Fragmentierungsanteil und den beiden Parametern γ and β berechnet (Agarwal & Lavie 2008: 116), wobei β steuert, wie sich die Strafpunkte in Abhängigkeit vom Grad der Fragmentierung verändern und γ bestimmt, wie viel Bedeutung bzw. Gewicht den Strafpunkten beigemessen wird (Dorr et al. 2011: 814f.). Formel 7 zeigt die Berechnung der Strafpunkte unter Berücksichtigung des Fragmentierungsanteils und der Parameter γ und β.

Formel 7: Formel zur Berechnung der Strafpunkte

$$Pen = \gamma \cdot frag^{\beta}$$

Der finale Wert (engl. *score*) ergibt sich schließlich auf Basis der zuvor ermittelten Werte *Pen* und F_{mean}, wobei die Parameter α, β und γ so gewählt werden, dass eine maximale Korrelation mit menschlichen Bewertungen erzielt wird (Agarwal & Lavie 2008: 116). Formel 8 zeigt die Berechnung des finalen METEOR-Werts.

Formel 8: Formel für die Berechnung des finalen METEOR-Werts

$$score = (1 - Pen) \cdot F_{mean}$$

Im Gegensatz zu BLEU ist es mit METEOR nicht möglich direkt Informationen aus mehreren Vergleichsreferenzen für die Ermittlung des Werts zu nutzen, jedoch wird, wenn mehrere Referenzen zur Verfügung stehen, für jedes Segment die Referenzübersetzung ausgewählt, die den höchsten METEOR-Wert erzielt. Der Vorteil gegenüber BLEU besteht in der Möglichkeit Wortstämme und Synonyme abzugleichen, was zu einer Erhöhung von Zuverlässigkeit und Robustheit führen kann (Dorr et al. 2011: 763).

Die tatsächliche Effektivität der METEOR-Evaluierungsmetrik hängt hierbei jedoch stark von der Verfügbarkeit geeigneter semantischer Ressourcen ab, die besonders für ressourcenarme Sprachen nicht immer zur Verfügung stehen. Obwohl die METEOR-Evaluationsmetrik im Vergleich zur BLEU-Metrik eine höhere Korrelation mit menschlichen Evaluierungsergebnissen aufweist, ist sie aufgrund der vielen Möglichkeiten der Einbeziehung linguistischer Ressourcen komplexer in der Anwendung und Ergebnisinterpretation und wird aufgrund dieser Nachteile weniger häufig für die automatisierte Evaluierung von MÜ-Ergebnissen eingesetzt als die BLEU-Metrik (Poibeau 2017: 207). Da Übersetzungsqualität sehr subjektiv ist, stellt trotz aller automatisierter Möglichkeiten die menschliche Evaluierung durch bilinguale Annotator*innen, die das zielsprachliche MÜ-Ergebnis mit dem Ausgangstext vergleichen, die Goldstandard-Methode für die Evaluierung von MÜ-Ergebnissen dar (Dorr et al. 2011: 750).

2.3.2 Menschliche Evaluierung von MÜ

Die mit den weitreichenden Möglichkeiten des Internets und der Verbreitung von MÜ-Systemen für den privaten Gebrauch (Krenz & Ramlow 2008: 26) einhergehende „power of the crowd“ (Jiménez-Crespo 2018: 70), also die Möglichkeit der Auslagerung von Aufgaben an eine große Masse von Nichtfachleuten, die kollaborativ an einem Projekt arbeiten, hat auch in die Translationswissenschaft in Form von Translation Crowdsourcing Einzug gefunden. Dieser Paradigmenwechsel hat zugleich eine Bandbreite an neuen Übersetzungsleistungen mit sich gebracht. Neben professionellen Übersetzungen, die von ausgebildeten Übersetzer*innen angefertigt werden, existiert mittlerweile auch eine Vielzahl an nicht-professionellen, maschinell erstellten, posteditierten oder mittels Translation Crowdsourcing produzierten Übersetzungen (2018: 70).

Mit den neuen Übersetzungsdienstleistungen und der Tatsache, dass nahezu jeder Mensch mithilfe von MÜ-Systemen Übersetzungen für den privaten Gebrauch anfertigen kann, variieren auch die Qualitätsauffassungen zwischen Lai*innen und Fachleuten und reichen von vollkommener grammatikalischer und stilistischer Korrektheit für professionelle Übersetzungen bis zur ausreichend sinnvollen Wiedergabe des Ausgangstexts ohne Berücksichtigung von Grammatik und Stil für private Übersetzungen (Forcada 2017: 217). Da sich die Qualitätsdefinition von Dienstleitungen stark auf die subjektiven Perspektiven und Erwartungshaltungen der Endkund*innen bezieht (Bruhn 2020: 140), ist demnach, je nach Aufgabenstellung bzw.

Übersetzungsauftrag, die Miteinbeziehung dieser Endkund*innen bzw. Lai*innen in die Bewertung von Übersetzungen durchaus sinnvoll.

Im Gegensatz zu herkömmlichen TQA-Ansätzen, die sich hauptsächlich auf das Endprodukt konzentrieren und bei denen ausschließlich eine Gesamtevaluierung der finalen Übersetzung durchgeführt wird, fokussiert beispielsweise der nutzer*innenzentrierte Übersetzungsansatz (engl. *User-Centered Translation / UCT*) auf die aktive Beteiligung der Endnutzer*innen am gesamten Übersetzungsprozess und versucht dadurch die Nachteile traditioneller TQA-Methoden zu kompensieren (Suojanen et al. 2015). Fehler, insbesondere Übersetzungsfehler, werden in der UCT basierend auf ihrer Relevanz für Funktionalität und Benutzer*innenfreundlichkeit direkt von den Endnutzer*innen während des gesamten Prozesses bewertet. Am Evaluierungsprozess von Übersetzungen können also nicht nur professionell ausgebildete Übersetzer*innen, sondern auch Lai*innen beteiligt sein (Castilho et al. 2018: 23).

Die Sicherstellung, dass die Fähigkeiten und Eigenschaften der Annotator*innen mit der konkreten Evaluationsaufgabe übereinstimmen, ist von großer Bedeutung und sollte im Zuge der Analyse der Aufgabenstellung erfolgen (Doherty 2017: 140). Obwohl Übersetzer*innen und Linguist*innen in der Sprachindustrie als Goldstandard unter den Annotator*innen angesehen werden und daher vornehmlich translationswissenschaftliche Fachleute für das TQA herangezogen werden, können Lai*innen bei einigen TQA-Aufgaben ähnlich hilfreich sein wie professionelle Übersetzer*innen. Besonders im Bereich der Qualitätsevaluierung von Übersetzungen, die mittels Translation Crowdsourcing erstellt worden sind, werden vermehrt Lai*innen für die Evaluierung herangezogen und auch in der MÜ-Forschung sind professionelle Annotator*innen eher die Ausnahme als die Regel (Castilho et al. 2018: 23). Insbesondere für ressourcenarme Sprachen ist es überdies nicht immer möglich das TQA von professionell ausgebildeten Übersetzer*innen durchführen zu lassen. Es wird daher in vielen Fällen nicht von Übersetzer*innen, sondern von Personen, denen man sprachliche Kompetenz zuschreibt, wie beispielsweise Studierende der Sprach- oder Translationswissenschaft, durchgeführt (Doherty 2017: 140).

Die Evaluierung von MÜ durch menschliche Annotator*innen gestaltet sich aber nicht zuletzt aufgrund der Subjektivität der Annotator*innen als schwierig (2017: 134). Schmalz (2019: 202) argumentiert die Problematik der subjektiven menschlichen Qualitätsevaluierung auf Basis der Feststellung, dass selbst Humanübersetzung, wenn sie von Menschen evaluiert werden, nicht unbedingt die volle Punktzahl erreichen. Diese Evaluierungsdiskrepanz basiert auf der Tatsache, dass dieselbe Übersetzung von verschiedenen Annotator*innen mit

unterschiedlichen Punktezahlen bewertet werden kann. Diesem Problem versucht man mittels Untersuchung der so genannten Inter-Rater-Reliabilität (engl. *inter-rater reliability*) entgegenzuwirken (Doherty 2017: 139).

Unter Reliabilität versteht man allgemein den Grad an Konsistenz von Messergebnissen zwischen Messinstrumenten, wenn sie unter ähnlichen Bedingungen eingesetzt werden (Clifford 2001: 374). Reliabilität wird anhand der Überprüfung der Testergebnisse auf Konsistenz ermittelt. Wird beispielsweise dieselbe Übersetzung zu verschiedenen Zeitpunkten unter denselben TQA-Kriterien evaluiert, sollten die Ergebnisse der Bewertung ähnlich, wenn nicht sogar ident sein. Die Inter-Rater-Reliabilität ist eine Erweiterung dieses Reliabilitätsbegriffs und beschreibt eine Situation in welcher dieselbe Übersetzung von mehreren Annotator*innen evaluiert wird. Folglich sollten ebenso die Bewertungsergebnisse derselben Übersetzung von verschiedenen Annotator*innen ähnlich sein, wenn die Übersetzung anhand identischer TQA-Kriterien evaluiert wurde (Doherty 2017: 139).

2.3.2.1 Fehlerklassifizierung

Die manuelle Fehlerklassifizierung, also die von Menschen durchgeführte Fehlerannotation, zählt zu den gängigsten, aber zugleich auch zeit- und ressourcenaufwändigsten Evaluierungsmethoden, die überdies die Gefahr einer geringen Inter-Rater-Reliabilität birgt, nämlich insbesondere dann, wenn den Annotator*innen eine Vielzahl an unterschiedlichen Fehlerkategorien für die Bewertung zur Verfügung steht. Ziel der Fehlerklassifizierung ist die Erstellung eines Fehlerprofils für jede Übersetzung, welches sämtliche Fehler in einer Fehlertypologie aufschlüsselt und dadurch die gesamte Qualität der Übersetzung abbildet. Obwohl es keine allgemeinen Regeln für die Definition der Fehlerkategorien gibt, sollen damit aber sowohl translationsrelevante als auch linguistische Aspekte abgedeckt werden. Neben der Verwendung für die Qualitätsevaluierung einer Übersetzung kann dieses Fehlerprofil darüber hinaus auch dazu verwendet werden, verschiedene Übersetzungen miteinander zu vergleichen, oder den Einfluss unterschiedlicher Fehlerkategorien auf den Post-Editing-Aufwand zu untersuchen (Popović 2018: 129ff.).

Farrús et al. (2010) beschreiben ein sehr einfaches Fehlerklassifizierungsschema für SMÜ, das aus fünf großen Fehlerkategorien besteht. Diese fünf Fehlerkategorien sind: morphologische Fehler, lexikalische Fehler, orthografische Fehler, syntaktische Fehler und semantische Fehler. Die Fehleranalyse wurde von bilingualen Annotator*innen in der Sprachkombination Spanisch-Katalanisch durchgeführt und zielte darauf ab die linguistischen Hauptprobleme

der SMÜ-Systeme zu eruieren. Die Annotator*innen befolgten dabei spezifische Richtlinien, um eine umfassende linguistische Analyse zu gewährleisten. Neben der linguistischen Analyse wurde auch eine Wahrnehmungsevaluierung durchgeführt, bei der festgestellt wurde, dass lexikalische und semantische Fehler mehr Einfluss auf die Qualitätswahrnehmung der menschlichen Annotator*innen haben, als Fehler in anderen Kategorien.

Stymne und Ahrenberg (2012) verwenden ein detaillierteres hierarchisches Fehlerklassifizierungsschema bestehend aus zwei Ebenen. Der Schwerpunkt der Studie liegt auf der Erforschung der Inter-Rater-Reliabilität, die anhand der Evaluierung von maschinell mit SMÜ-Systemen erstellten Übersetzungen untersucht wird. Die Fehlerkategorie Ling (erste Ebene) besteht beispielsweise aus den linguistischen Unterkategorien Orthografie, Semantik und Syntax (zweite Ebene). Darüber hinaus ermöglicht dieses Fehlerklassifizierungssystem den Annotator*innen auch die Beurteilung der gefundenen Fehler auf einer vierstufigen Skala nach Schweregrad von unbedeutend (engl. *insignificant*) zu schwerwiegend (engl. *serious*). Die Studie zeigt, dass eine höhere Inter-Rater-Reliabilität entweder durch die Verwendung einer einfacheren Fehlerklassifizierung oder durch die Verwendung einer ausführlicheren Taxonomie in Verbindung mit klaren Richtlinien für die Annotator*innen erzielt werden kann (2012: 1789).

Neben der Verwendung für die Evaluierung von SMÜ-Resultaten eignet sich die Fehlerklassifizierung darüber hinaus auch zur Evaluierung von Übersetzungen, die mithilfe von NMÜ angefertigt wurden. Castilho et al. (2017) untersuchten die qualitativen Unterschiede zwischen SMÜ und NMÜ und ließen hierfür die Übersetzungsresultate von SMÜ- und NMÜ-Systemen von professionell ausgebildeten Translator*innen u. a. mittels Fehlerklassifizierung bewerten. Die Bewertung umfasste die vier Sprachenpaare Englisch-Deutsch, Englisch-Portugiesisch, Englisch-Russisch und Englisch-Griechisch und die Fehlerkategorien Flexionsmorphologie, Wortstellung, Auslassungen, Ergänzungen und Fehlübersetzung. Die Ergebnisse zeigen, dass NMÜ in allen vier Sprachkombinationen bessere Resultate liefert als SMÜ, insbesondere aufgrund einer geringeren Anzahl an morphologischen Fehlern und Wortstellungsfehlern.

Yamada (2019: 96) stellte ähnlich dazu fest, dass Übersetzungen, die mit SMÜ-Systemen erstellt wurden hauptsächlich lexikalische und morphologische, also für Menschen leicht erkenn- und korrigierbare Fehler, aufweisen, während Übersetzungen, die mit NMÜ-Systemen erstellt wurden, ein Fehlermuster aufweisen, das dem von Humanübersetzungen ähnlich ist. Flanagan (1994) entwickelte ein weithin anerkanntes und flexibles Fehlerklassifizierungsschema, das 19 Kategorien umfasst, darunter Kategorien wie beispielsweise Rechtschreibung, nicht gefundene Wörter, Flexion, Pronomen, Artikel, Präposition, Wortwahl und Ausdruck. Die

Flexibilität des Modells zeigt sich in der Möglichkeit verschiedene Fehlerkategorien hinzuzufügen, zu entfernen und nach Priorität zu reihen (1994: 71). Dieses Fehlerklassifikationsschema ist auch im 21. Jahrhundert noch für die Evaluierung und Fehlerklassifizierung von maschinell erstellten Übersetzungen relevant und nützlich (Stankevičiūtė et al. 2017: 77).

2.3.2.2 Ranking

Eine weitere Methode zur manuellen Evaluierung von maschinell erstellten Übersetzungen, die insbesondere im Forschungskontext zur vergleichenden Bewertung von Übersetzungen verwendet wird, ist das Ranking. Bei diesem Ansatz werden den Annotator*innen Übersetzungen von verschiedenen MÜ-Systemen zur Evaluierung vorgelegt, die alle aus demselben Ausgangstext generiert und anonym in einer zufällig verschlüsselten Reihenfolge aufgelistet wurden. Die Übersetzungen werden von den Annotator*innen auf Basis diverser Kriterien, wie beispielsweise Flüssigkeit bewertet. Es gibt verschiedene Ranking-Durchführungsansätze, so können Annotator*innen z. B. entweder aus allen vorhandenen Übersetzungen das beste Resultat auswählen, oder die Ergebnisse in einer bestimmten Reihenfolge, üblicherweise von der besten zur schlechtesten Übersetzung, reihen. Dieser Ansatz ist effizient und bietet leicht zu interpretierende Ergebnisse (Castilho et al. 2018: 21). Gleichzeitig stellt dieser vergleichende und reihende Ansatz auch die Basis für die BWS-Methode dar, bei der Annotator*innen aus einem Set von möglichen Optionen die jeweils beste und schlechteste Option auswählen (Louviere et al. 2015: 3).

2.4 Best-Worst-Scaling

Die BWS-Methode, auch *maximum difference scaling* bzw. *maxdiff* genannt (Louviere et al. 2015: 11), wurde 1987 von Jordan J. Louviere entwickelt. Er verfolgte dabei das Ziel, mögliche Vorteile gegenüber herkömmlicher diskreter Auswahlexperimente (engl. *Discrete Choice Experiments / DCE*), die sich ausschließlich auf die besten Optionen konzentrieren, zu erforschen, da die BWS-Methode die Möglichkeit eines zusätzlichen Informationsgewinns durch Miteinbeziehung der schlechtesten Auswahlmöglichkeit bietet (2015: 3). Gemeinsam mit anderen Forscher*innen widmete er sich daraufhin in den 1990er und frühen 2000er Jahren der Weiterentwicklung der BWS-Methode, die in vielen Ländern Anklang fand und besonders für Marktforschungszwecke, z. B. in der Gesundheitsökonomie oder der Mund- und Zahnpflege übernommen, und für öffentliche Umfragen zur Lebensmittelsicherheit verwendet wurde (2015: 4f.).

Obwohl DCE in vielen verschiedenen Disziplinen verwendet werden, weisen sie einige Nachteile bzw. Einschränkungen auf, wie beispielsweise Granularitätsprobleme, Bias bzw. Tendenzen der Annotator*innen zu einem bestimmten Bereich der Skala sowie mangelnde Inter-Rater-Reliabilität. In Bezug auf Granularität können zu grobgranulare Auswahlmöglichkeiten zu Entscheidungsrestriktionen, und zu feingranulare Optionen zu Überforderung bei den Annotator*innen führen. Das Problem der mangelnden Inter-Rater-Reliabilität zeigt sich außerdem nicht nur zwischen den verschiedenen Annotator*innen, sondern, mit fortschreitender Dauer eines Experiments, auch bei derselben Annotatorin bzw. demselben Annotator (Kiritchenko & Mohammad 2017: 465).

Der paarweise Vergleich von Auswahlmöglichkeiten ist eine vergleichende Methode, bei der Annotator*innen die Optionen paarweise präsentiert werden und anschließend die Option mit dem höheren Interessenswert ausgewählt werden soll. Dadurch lassen sich einige Limitationen der klassischen DCE-Methoden vermeiden. Die Ergebnisse dieses paarweisen Vergleichs können anschließend gereiht, und das proportionale Verhältnis zwischen Optionen und Interessenswerten quantifiziert und als reale Werte abgeleitet werden. Die BWS-Methode stellt eine Variante dieses vergleichenden Ansatzes dar, bei der Annotator*innen aus n-Optionen, also n-Tupel mit $n > 1$ und standardmäßig $n = 4$, die jeweils beste und schlechteste Option auswählen. Alle zu bewertenden Elemente werden in ein Set von m 4-Tupel ($m \geq N$, wobei N die Anzahl der Optionen ist) unterteilt, sodass jedes Element mehrmals in verschiedenen 4-Tupeln evaluiert wird. Anschließend können reale Punktewerte mithilfe eines einfachen Zählverfahrens ermittelt werden (2017: 465f.).

Die BWS-Methode kann in drei verschiedene Anwendungsfälle mit steigender Komplexität unterschieden werden, nämlich den Objektfall, den Profilfall und den Multiprofilfall (siehe Louviere et al. 2015). Die beste und die schlechteste Option sind bei der BWS-Methode zwei Extreme eines subjektiven Kontinuums, die aber nicht mit akzeptabel und inakzeptabel gleichzusetzen sind. Die Extreme können verschieden ausgestaltet sein, so wären beispielsweise bei einer Gruppe von Menschen die Attribute am größten und am kleinsten ebenso zulässige Extrempunkte des subjektiven Kontinuums (Louviere et al. 2015: 5f.). Abbildung 12 zeigt beispielhaft den Aufbau der BWS-Methode. Die vier Fluglinien repräsentieren das Set an Optionen. Die Proband*innen wählen davon die jeweils beste und schlechteste Option aus.

I think that this is the best Airline (☑ one)	Airline	I think that this is the worst Airline (☑ one)
☐	American	☐
☐	United	☑
☑	Qantas	☐
☐	Delta	☐

Abbildung 12: Beispielhafte Darstellung des Objektfalls der BWS-Methode (Louviere et al. 2015: 7)

Sowohl DCE als auch BWS werden üblicherweise mit einem multinomialen Logit-Modell (MNL), das auf der Zufallsnutzen-Theorie (engl. *Random Utility Theory / RUT*) von Thurstone (1927) basiert, analysiert. Dabei wird unter anderem davon ausgegangen, dass die Wahlmöglichkeiten für alle Proband*innen konsistent sind, und Entscheidungen unvoreingenommen, also unbeeinflusst durch die Anordnung der auszuwählenden Optionen, getroffen werden (Krucien et al. 2017), jedoch unvorhersehbare Faktoren, wie individuelle Präferenzen und Wahrnehmungen die Entscheidungsprozesse beeinflussen können (Hess et al. 2018: 182). Zudem geht die RUT davon aus, dass die Häufigkeit von Entscheidungen nach wiederholten Entscheidungsinstanzen direkt auf den Wert hinweist, den sie den verfügbaren Optionen zuschreibt, also dass wiederholte gleiche Entscheidungen die Präferenzen der Proband*innen unter den gegebenen Alternativen reflektieren (Louviere et al. 2015: 7).

2.5 MTranslatability

Der Begriff MTranslatability wurde von Bernth und Gdaniec (2001) geprägt und bezeichnet die Übersetzbarkeit von Texten durch MÜ-Systeme. Die Qualität von rohen maschinellen Übersetzungen variiert abhängig von verschiedenen Faktoren, wie beispielsweise dem verwendeten MÜ-System, dem Sprachenpaar und dem Fachgebiet, kann aber selten ohne weitere menschliche Interventionen, in der Regel in Form von Post-Editing bzw. dem manuellen menschlichen Nachbearbeiten, als perfekt betrachtet werden. Dabei sind viele der Faktoren, die zur Verschlechterung der MTranslatability beitragen, auf die Autor*innen der Ausgangstexte zurückzuführen und ließen sich größtenteils beheben, wenn die Texte von den Ersteller*innen mit dem Bewusstsein über die Limitationen der MÜ-Systeme und mit „‘MTranslatability‘ in mind“ (Bernth & Gdaniec 2001: 175) verfasst worden wären.

Zu diesen Verschlechterungsfaktoren zählen unter anderem banale Dinge wie ambige Satzkonstruktionen oder Orthografie- und Grammatikfehler, wie beispielsweise die inkorrekte Verwendung von *they’re* und *there*. Die meisten dieser Faktoren können durch

Berücksichtigung einiger einfacher Regeln beim Verfassen des Ausgangstexts eliminiert werden. Obwohl diese Regeln, die sich auf Faktoren wie Ambiguität, Stil, Orthografie, oder Grammatik beziehen, für englischsprachige Ausgangstexte entwickelt wurden, lassen sie sich auch auf andere Sprachen anwenden (2001: 175ff.). Bernth und Gdaniec (2001) schlagen insgesamt 26 verschiedene Regeln zur Verbesserung der MTranslatability vor, die von Rechtschreibprüfung und Komma- bzw. Zeichensetzung über Empfehlungen zu Satzlänge und Verwendung von Sonderzeichen, bis zum empfohlenen Umgang mit Personal- und Relativpronomen reichen. Tabelle 3 fasst die für diese Arbeit relevanten sechs Regeln übersichtlich zusammen (für alle 26 Regeln siehe Bernth und Gdaniec 2001)

Tabelle 3: Übersicht über die für diese Arbeit relevanten sechs Regeln nach Bernth und Gdaniec (2001)

REGEL	BESCHREIBUNG
3	Use articles with ing-words when they are used as nouns, or use infinitives instead of ing-words, depending on what you mean
4	Rewrite ing-words that follow an object as a relative clause or add a suitable preposition, depending on what you mean
12	Use one-word verbs instead of verb + particle whenever possible
14	Avoid metaphors, idioms, slang and, dialect
15	Avoid ellipsis
16	Avoid passive constructions if possible

Eine weitere Möglichkeit die MTranslatability zu verbessen, sehen Bernth und Gdaniec (2001: 176) in der Verwendung einer so genannter kontrollierten Sprache (engl. *Controlled Language* / *CL*). CLs leiten sich aus der Allgemeinsprache einer natürlichen Sprache ab und unterliegen, im Gegensatz zur Allgemeinsprache, strengeren grammatikalischen und lexikalische Regeln. Die Texterstellung mit einer CL erfolgt auf Basis eigens dafür konzipierter Redaktionsleitfäden und Wortlisten. Diese linguistische Vereinfachung bzw. Standardisierung soll ein besseres Textverständnis und damit einhergehend eine Erhöhung der MTranslatability ermöglichen (Mügge 2002: 110f.). Roturier (2004) untersuchte die Effektivität von verschiedenen CL-Regeln auf den Post-Editing-Aufwand und erarbeitete eine Liste an Regeln, die einen besonders positiven Effekt auf die MÜ-Qualität und die MTranslatability haben und zu einer signifikanten Reduktion des Post-Editing-Aufwands geführt haben. Abbildung 13 zeigt eine Übersicht der erarbeiteten Regeln.

Always write a verb next to its particle. (46)

Do not use slashes to list lexical items (except for product names). (44)

Repeat the head noun with conjoined articles and prepositions. (37)

Avoid footnotes in the middle of a segment. Turn footnotes into independent segments. (37)

Do not omit words within lexical items, even when the term has already been used in the sentence. (29)

Make sure that every segment can stand alone syntactically. (27)

Only use the modal 'could' when the sentence contains 'if', otherwise use 'can'. (26)

Be very careful with the -ing words: If it is a gerund, use an article in front of it. (14). If it is introducing a new clause, use 'by' in front of it (23). If it is modifying a noun in a non-finite clause, replace it with a relative clause. (16)

Abbildung 13: Identifizierte CL-Regeln (Roturier 2004: 11)

Um wiederkehrende Fehler effizienter zu beheben und dadurch einerseits Zeit zu sparen und andererseits den Post-Editor*innen die Frustration über dieselben wiederkehrenden Fehler zu ersparen, schlägt Roturier (2004) als Schlussfolgerung seines Experiments einen so genannten „*batch pre-post-editing process*“ (2004: 12) vor, also die systematische Nachbearbeitung der Texte in Batches vor und nach der MÜ.

2.6 Pre-Editing

Obwohl Pre-Editing und Post-Editing (PE) von rohen maschinell erstellten Übersetzungen bereits in den frühen 1980er Jahren in wichtigen Organisationen, wie beispielweise der Europäischen Kommission, angewandt wurden, haben sich diese Konzepte, insbesondere das Post-Editing von MÜ (engl. *Machine Translation Post Editing / MTPE*), erst in den letzten 15 Jahren als massentaugliche Übersetzungsworkflowprozesse in der weltweiten Lokalisierungsindustrie etabliert. Während MTPE jegliche Prozesse beschreibt, die sich auf Korrekturen des maschinell erstellten Zieltextes konzentrieren, also jedwede Korrekturen, die **nach** dem maschinellen Übersetzungsprozess stattfinden, versteht man unter Pre-Editing die Anwendung von stilistischen und terminologischen Regeln zur Vorbereitung des Ausgangstextes **vor** dem maschinellen Übersetzungsprozess, mit dem Ziel, dadurch ein qualitativ höherwertiges MÜ-Resultat zu erhalten (Guerberof Arenas 2019: 333f.).

Pre-Editing kann dabei auf verschiedene Arten stattfinden. Eine Möglichkeit ist die Verwendung eines formalen Regelkatalogs, der auf dem Konzept von CLs basiert und grammatikalische und semantische Strukturen und erlaubte und verbotene Elemente regelt. Eine weitere Möglichkeit besteht in der Anwendung einiger einfacher Textkorrekturen, wie beispielsweise der Korrektur von Rechtschreib- und Orthografiefehlern. Das Ziel ist unabhängig von den konkret verwendeten Techniken aber immer das Streben nach einem qualitativeren MÜ-Resultat

(Sánchez-Gijón & Kenny 2022: 81f.). Darüber hinaus kann Pre-Editing entweder monolingual oder bilingual durchgeführt werden. Beim bilingualen Pre-Editing erfolgt die Bearbeitung des Ausgangstexts unter Berücksichtigung und Miteinbeziehung des MÜ-Resultats und erfordert daher Kenntnisse von Ausgangs- und Zielsprache. Monolinguales Pre-Editing bezieht sich ausschließlich auf den Ausgangstext und macht daher zielsprachliche Kenntnisse obsolet (Hiraoka & Yamada 2019: 66).

2.6.1 Pre-Editing und NMÜ

Die Entwicklung von NMÜ regte ein Um- bzw. Neudenken der traditionellen Pre-Editing-Praxis an. Insbesondere bei RBMÜ und SMÜ war Pre-Editing ein entscheidender Qualitätsfaktor. Ein tiefes Verständnis von MÜ und MÜ-Systemen ermöglichte es den Anwender*innen, systematische Flüssigkeits- und Adäquatheitsfehler im Ausgangstext zu identifizieren, die in weiterer Folge zu Schwierigkeiten im maschinellen Übersetzungsprozess führen, und in Fehlübersetzungen resultieren könnten. Im Gegensatz dazu ist das Fehlen solcher systematischen Fehler charakteristisch für NMÜ und führt folglich zu Schwierigkeiten, die Art der auftretenden Fehler mit einer gewissen Sicherheit zu bestimmen (Sánchez-Gijón & Kenny 2022: 82f.).

Diese Änderung des Fehlermusters zwischen SMÜ und NMÜ führt folglich dazu, dass Übersetzungen, die mittels neuronaler Systeme erstellt wurden, zwar oftmals flüssig klingen, aber dennoch schwerwiegende Fehler, wie Fehlübersetzungen, Auslassungen oder Einfügungen, enthalten können, die aber, aufgrund der feineren, nuancierten Fehlerart weniger vorhersehbar sind als traditionelle, systematische SMÜ-Fehler. Demzufolge bedarf es auf NMÜ angepasste Pre-Editing-Methoden und insbesondere neues, zusätzliches Wissen über Aspekte wie beispielsweise NMÜ-Fehlertypologien, Besonderheiten der NMÜ-Terminologie, oder dem effizienten Umgang mit langen Sätzen (Guerberof Arenas 2019: 349).

Darüber hinaus sind aufgrund der rasanten technologischen Weiterentwicklung die MÜ-Limitationen nicht mehr länger technischer Natur, sondern liegen vielmehr in der Qualität des rohen Ausgangstextes, und damit verbunden der Sicherstellung fehlerfreier Ausgangstexte, und der Angemessenheit der Übersetzung in Bezug auf die Erfüllung der kommunikativen Funktion des Zieltexts, wie z. B. Register oder Genrekonventionen. Für Textarten, wie technische Dokumentationen, die in der Regel strengen Konventionen folgen und über wenig bis keine referenziellen oder stilistischen Merkmale verfügen, reicht oftmals ein einfacher Pre-Editing-Prozess, der sich auf eine Rechtschreibprüfung beschränkt aus, um qualitativ hochwertige Ergebnisse mit NMÜ zu erzielen. Textarten wie „Unboxing"-Videos, also Videos, in denen jemand ein

neues Produkt auspackt und seinen ersten Eindruck davon teilt, die lehrreiche (informativ) und unterhaltsame (expressiv, appellativ) Elemente vereinen und daher mehrere kommunikative Funktionen erfüllen, oder Textsorten, die kulturelle oder soziale Aspekte der Ausgangskultur enthalten, mit denen sich zwar Leser*innen der Ausgangskultur identifizieren können, jedoch Leser*innen der Zielkultur nicht, stellen nach wie vor eine Herausforderung für die NMÜ dar (Sánchez-Gijón & Kenny 2022: 88f.).

Wenngleich also neuronal erstellte Übersetzungen oftmals flüssig und grammatikalisch korrekt klingen, ist es möglich, dass selbst grammatikalisch korrekte Übersetzungen den stilistischen, rhetorischen oder kommunikativen Zwecken nicht entsprechen. Pre-Editing bietet hier eine Möglichkeit sicherzustellen, dass eine Übersetzung den Erwartungen eines globalen Publikums entspricht. Bereits in der Vergangenheit haben große Unternehmen Pre-Editing mit SMÜ oder RBMÜ genutzt, um wiederkehrende systematische Fehler zu eliminieren. Mit der Einführung von NMÜ und den damit einhergehenden neuen Anforderungen, könnte sich der Einsatz von Pre-Editing zukünftig ausweiten (2022: 89).

2.6.2 Pre-Editing-Richtlinien für NMÜ

Der grundsätzliche Erfolg des Pre-Editings von NMÜ wird von zwei Hauptfaktoren bestimmt: der Textfunktion und den Fehlerarten, die ein bestimmtes MÜ-System in Rohübersetzungen produziert. Im Bereich der informativen Texte, also Texte, deren Kommunikationsziel in erster Linie darin besteht die Leser*innen zu informieren oder zu instruieren, lassen sich die gängigsten Pre-Editing-Regeln für NMÜ in drei große Gruppen zusammenfassen, nämlich in Wortwahl, Struktur und Stil und Textreferenzen (Sánchez-Gijón & Kenny 2022: 90).

Bei der NMÜ wird die Verarbeitung jedes Worts bzw. jeder Bedeutungseinheit von seinem jeweiligen Kontext beeinflusst. Kontext und Wörter stehen also in einer wechselseitigen Beziehung zueinander. Die lexikalischen Entscheidungen in einem Text betreffen ein breites Spektrum an Kontexten, in denen dieselben Wortwahlentscheidungen getroffen werden. Wenn also beispielsweise ein Ausgangstext schnell in mehrere Zielsprachen übersetzt werden soll, trägt eine sorgfältige Wortwahl nicht nur zur Vermeidung von Übersetzungsfehlern bei, sondern fördert auch die Einhaltung sprachlicher Normen, die auf der Funktion und dem Zweck des Textes basieren. Zu den gängigsten lexikalischen Pre-Editing-Richtlinien zählen unter anderem die Verwendung einer konsistenten Terminologie und die Vermeidung ungebräuchlicher Abkürzungen (2022: 90f.).

Die Art und Weise, wie Texte auf Text- und Satzebene formuliert sind und wie Ideen satzintern, satzübergreifend und sogar intertextuell angeordnet sind, ist für die Textverständlichkeit von zentraler Bedeutung und beeinflusst das Verständnis der Leser*innen erheblich. Bei der NMÜ können die im Ausgangstext getroffenen diesbezüglichen Entscheidungen die Übersetzungsoptionen positiv oder negativ beeinflussen. Komplexe und ambige Textstrukturen erhöhen die Wahrscheinlichkeit, dass NMÜ-Systeme inkohärente oder unverständliche Übersetzungen vorschlagen. Die Einhaltung von Pre-Editing-Richtlinien auf stilistischer und struktureller Ebene hilft nicht nur die Leistung des NMÜ-Systems zu verbessern, sondern gibt dem Text darüber hinaus auch mehr Klarheit und trägt dadurch dazu bei, die Bedeutung des Ausgangstexts im Translat akkurat wiederzugeben. Zu den gängigsten Pre-Editing-Richtlinien auf strukturell-stilistischer Ebene zählen beispielsweise die Verwendung von einfachen, kurzen, vollständigen Sätzen und die Verwendung derselben syntaktischen Strukturen (2022: 91f.). Tabelle 4 zeigt die gängigsten lexikalischen und strukturell-stilistischen Pre-Editing-Richtlinien nach Sánchez-Gijón und Kenny (2022).

Tabelle 4: Lexikalische und strukturell-stilistischen Richtlinien nach Sánchez-Gijón und Kenny (2022)

LEXIKALISCH	Avoid lexical shifts in register
	Avoid uncommon abbreviations
	Avoid unnecessary words
	Be consistent
STRUKTURELL-STILISTISCH	Short and simple sentences
	Complete sentences
	Use parallel structures in related sentences
	Active voice
	Homogenous style

Textreferenzen bzw. Referenzelemente sind linguistische Komponenten, die ein anderes Element ersetzen oder auf ein anderes Element verweisen. Pronomen wie *ich*, *er*, *sie* und Possessivpronomen wie *mein* oder *sein* sind Beispiele solcher Referenzelemente. Pronomen, die sich auf Elemente innerhalb desselben Satzes beziehen, oder konsistent auf dasselbe Element verweisen, verursachen bei der NMÜ typischerweise weniger Probleme als Referenzen auf unterschiedliche Elemente oder wechselnde Referenten, da NMÜ-Systeme bei wechselnden Referenzelementen Konsistenzschwierigkeiten haben und im Übersetzungsverlauf dazu neigen auf die gängigsten Pronomen zurückzugreifen. Dies geschieht insbesondere dann, wenn der Korpus, der für das Training des jeweiligen NMÜ-Systems verwendet wurde, nicht über ausreichend diverse und nuancierte referenzielle Unterscheidungen verfügt. Diese Problematik kann

im Zuge des Pre-Editings mittels Verwendung einfacher Sätze mit klaren pronominalen Bezügen minimiert werden (Sánchez-Gijón & Kenny 2022: 94f.).

2.7 Fachbereich Astronautik

Angelo (2006: 61) definiert Astronautik im weitesten Sinne als „[t]he branch of engineering science dealing with spaceflight and the design and operation of space vehicles." Die Raumfahrt ist jedoch ein stark interdisziplinäres Forschungsgebiet, das sich aufgrund der komplexen wissenschaftlichen und technischen Überlappung verschiedener Disziplinen nur schwer genau eingrenzen lässt (Almár & Both 2002: 83). Neben der Aeronautik bzw. Luftfahrttechnik sowie einigen naturwissenschaftlichen Disziplinen definiert Campbell (1962: 9) drei weitere zentrale Kernbereiche als Wurzeln der Astronautik, nämlich die Astronomie, die das Grundlagenwissen liefert, die Kommunikation und Philosophie, die den Austausch, die Verbreitung und die Förderung von Ideen ermöglicht und die Raketentechnik, die überhaupt erst die technische Grundlage für den Weltraumflug schafft.

Almár und Both (2002: 84f.) schlagen eine ähnliche Klassifizierung der Astronautik in mehrere zentrale Themenbereiche vor, darunter Naturwissenschaften, Geisteswissenschaften, Technik und anwendungsorientierte Tätigkeitsfelder. Abbildung 14 veranschaulicht die vorgeschlagene Einteilung. In der Mitte befindet sich die Astronautik, kreisförmig umringt von allen mit ihr verwandten wissenschaftlichen, technischen und anwendungsorientierten Disziplinen, darunter klassische Naturwissenschaften wie beispielweise Physik, Chemie, Astronomie und Materialwissenschaften, aber auch technische Bereiche wie Aeronautik und Raketentechnik, oder Tätigkeiten wie beispielsweise Erdbeobachtung, Meteorologie oder Navigation.

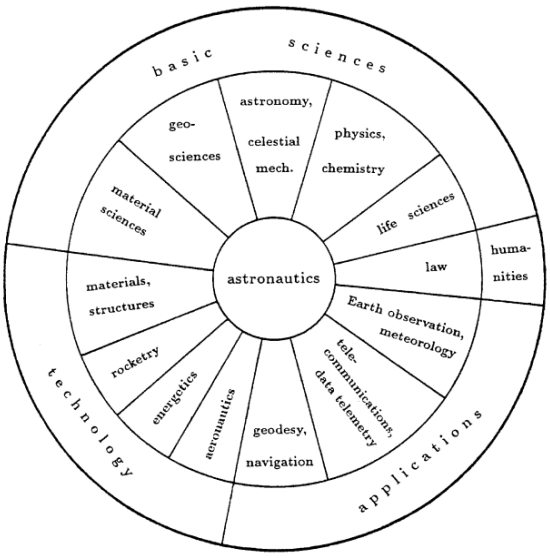


Abbildung 14: Schema zur Visualisierung der Interdisziplinarität von Astronautik (Almár & Both 2002: 85)

Alle diese Disziplinen entwickeln kontinuierlich ihre eigene Terminologie, die oft unverändert von der Astronautik übernommen wird. In einigen Fällen wird jedoch die Bedeutung der übernommenen Begriffe angepasst oder es entstehen ganz neue Begriffe (Almár & Both 2002: 85). Als eine der ältesten naturwissenschaftlichen Disziplinen (Pannekoek 1989: 13) und als jene, die das Grundlagenwissen liefert (Campbell 1962: 9) spielt die Astronomie in der Astronautik eine zentrale Rolle. Der Begriff *orbit* stammt beispielsweise aus der Astronomie, mit *to orbit* wurde darauf basierend für die Raumfahrt ein neuer spezifischer Fachbegriff der Astronautik abgeleitet (Almár & Both 2002: 85).

2.7.1 Fachsprache der Astronomie

Astronomie (αστρον griech. Stern) ist die Wissenschaft von den Galaxien, Planeten und Sternen und befasst sich unter anderem mit der Frage nach der Entstehung und der zukünftigen Weiterentwicklung des Universums (Hanslmeier 2020: 1). Mit ihrem Ursprung in prähistorischen Zeiten, und damit viel früher als andere Naturwissenschaften, gilt die Astronomie als eine der ältesten wissenschaftlichen Disziplinen und hat daher einen besonders hohen Stellenwert in der Geschichte der menschlichen Kultur (Pannekoek 1989: 13).

2.7.1.1 Entstehung der Fachsprache der Astronomie

Bereits früheste geschichtliche Aufzeichnungen belegen, dass die Menschheit sehr bald ein Interesse an astronomischen Phänomenen, wie dem Tag- und Nachtzyklus und den Jahreszeiten entwickelte. Diese Phänomene waren ein wichtiger Teil der Lebensumwelt der Menschen und sie zu erforschen und diese Kräfte der Natur für sich zu nutzen war ein wichtiger Aspekt des menschlichen Überlebens. Lange vor der Erfindung der Uhr konnten die Menschen an sonnigen Tagen anhand des Sonnenstandes die Tageszeit mit hoher Genauigkeit bestimmen und Jäger*innen, Fischer*innen, Hirt*innen und Landwirt*innen haben ihr Arbeitsverhalten (Jagd, Aussaat, Ernte, Treiben der Herden, etc.) den jahreszeitenbedingten Änderungen der Natur angepasst.

Aufgrund von Schifffahrt, und der damit einhergehenden reisenden Kaufleute, gewannen auch Navigation und Zeitmessung bzw. Zeitberechnung an immer größerer Bedeutung und zählen zu den frühesten spezialisierten, astronomischen Anwendungen, aus denen die wissenschaftliche Disziplin der Astronomie entstand (Pannekoek 1989: 19f.). Da die zunehmende Spezialisierung der menschlichen Tätigkeiten, insbesondere auch jener überlebensnotwendiger Aktivitäten wie Jagd oder Ernte, und die damit einhergehende Arbeitsteilung eine wichtige

Grundlage für die Entwicklung einer Fachsprache darstellen und das Vorhandensein von Fachsprachen bereits in früheren und weniger komplexen Gesellschaften angenommen wird (Mayer 2013: 15) entstand folglich zusätzlich zur wissenschaftlichen Disziplin der Astronomie auch eine erste einfache astronomische Fachsprache.

2.7.1.2 Moderne Fachterminologie der Astronomie

Aus etymologischer Sicht entstanden die heutigen astronomischen Fachtermini über viele Jahrhunderte aus der Adaptation von allgemeinsprachlichen Wörtern auf technische Sachverhalte, die dem wörtlichen Sinn mehr oder weniger entsprachen. Die heute gebräuchliche astronomische Fachterminologie spiegelt somit die sprachlichen Einflüsse aller Zivilisationen wider, die über die Jahrhunderte unweigerlich zur Entwicklung der modernen Astronomie beigetragen haben (Woolard 1949: 482f.).

Im neunten und zehnten Jahrhundert erlebte die islamische Welt in Wissenschaft und Kultur eine Blütezeit. Während dieses Zeitraums konzentrierten sich die astronomischen Aktivitäten hauptsächlich auf den Nahen Osten, Nordafrika und das maurische Spanien, während Europa wissenschaftlich brach lag (Pannekoek 1989: 173). Die Zeit zwischen dem achten und dem vierzehnten Jahrhundert wird in der Astronomie von Historikern daher oft die Islamische Periode genannt (Gingerich 1986: 74). In Bagdad wurden anfangs persische und indische Quellen und später griechische Wissenschaftstexte studiert und in die arabische Sprache übersetzt (van Dalen 2013: 28f.), darunter auch Ptolemäus‘ Werk Syntaxis mathematike, ein Hauptwerk der griechischen Astronomie, welches bis heute jedoch hauptsächlich unter seinem arabischen Namen Almagest bekannt ist (Gingerich 1986: 74).

Im Gegensatz zu einigen lateinischen Begriffen, die über die Jahrhunderte durch präzisere Benennungen ersetzt wurden (z. B. *Deklination* der Sonne statt *obliquatio*), werden viele der ursprünglichen allgemeinsprachlichen arabischen Begriffe heute nach wie vor unverändert in der modernen astronomischen Fachterminologie verwendet, darunter beispielsweise die modernen Fachbegriffe Azimut, Zenit oder Nadir, oder die heute gebräuchlichen Namen von Himmelskörpern, wie Algol vom arabischen raʾs al-ġūl (Pannekoek 1989: 174). Neben griechischen, lateinischen und arabischen Einflüssen finden sich auch Spuren anderer Sprachen und Sprachvarietäten in den astronomischen Fachbegriffen der verschiedenen modernen Sprachen. In vielen Fällen hat sich der Gebrauch der Originalbegriffe über die Jahrhunderte stark verändert und aufgrund der wenigen Aufzeichnungen, die oft nicht mehr Informationen als die reine Benennung, ohne Begründung oder Erklärung, warum gewisse Benennungen für Begriffe

verwendet wurden, beinhalten, lässt sich Vieles heute nicht mehr nachvollziehen. Andere Begriffe wiederum überdauerten die Jahrhunderte und werden auch heute noch in modernen Sprachen verwendet. Das Wort *horizon*, die lateinische Transkription des griechischen Wortes für den Himmelskreis wird, teilweise mit kleinen sprachspezifischen orthografischen und phonetischen Änderungen, beispielsweise im Französischen (*horizon*), im Deutschen (*Horizont*), im Spanischen (*horizonte*) und dem Italienischen (*orizzonte*) verwendet (Woolard 1949: 482f).

2.7.2 Sprachliche Merkmale englischer wissenschaftlich-technischer Sprache

Die englische wissenschaftlich-technische Sprache (engl. *English for Science and Technology / EST*) wird vornehmlich in wissenschaftlichen Publikationen, Fachartikeln, Lehrbüchern und technischen Berichten verwendet und ist durch spezifische linguistische Merkmale charakterisiert. Sie dient der präzisen systematischen Darstellung physikalischer und naturwissenschaftlicher Phänomene, ihrer Prozesse, Eigenschaften, Merkmale, Gesetzmäßigkeiten sowie deren praktischer Anwendung (Li & Li 2015: 161).

Zu ihren lexikalischen Besonderheiten zählt der häufige Gebrauch spezifischer Fachwörter, präziser Verben und deskriptiver Adjektive (Li & Li 2015: 161f.) sowie Abkürzungen und Akronyme (Jordan 1999). Die spezifischen Fachwörter werden je nach Verwendung in vier Kategorien eingeteilt, nämlich in die (i) streng fachgebundenen monosemischen, meist aus lateinischen oder griechischen Morphemen gebildeten Begriffe, die (ii) monosemischen, aber in verschiedenen Fachbereichen verwendeten Begriffe, die (iii) aus der Allgemeinsprache stammenden und je nach Fachbereich unterschiedliche Bedeutungen annehmenden Begriffe und die (iv) Akronyme bzw. die durch Wortbildungsprozesse, wie Affixierung oder Komposition entstandenen Begriffe (Li & Li 2015: 161f.). Tabelle 5 zeigt Fachbegriffe aus dem Themengebiet der Astronautik für alle vier Kategorien.

Tabelle 5: Beispiele für spezifische Fachwörter nach Kategorie im Fachbereich Astronautik

Kategorie	Beispiele aus der Astronautik
(i)	perigee, aphelion
(ii)	navigation, trajectory
(iii)	mission, operation
(iv)	planetesimal, MetOp

Auf syntaktischer Ebene zeichnet sich die englische wissenschaftlich-technische Sprache durch einige Besonderheiten aus, die der Präzision und Objektivität wissenschaftlich-technischer

Texte dienen sollen. Dazu gehört der häufige Gebrauch von Präpositionalphrasen, Adjektiv- und Adverbialphrasen, Partizipien und Attributsätze, oft in komplexen Kombinationen, um präzise Definitionen zu ermöglichen. Ein weiteres zentrales Merkmal sind lange Satzstrukturen, die durch die logische Verknüpfung von Nebensätzen und Phrasen auch anspruchsvolle wissenschaftliche Sachverhalte präzise und nachvollziehbar darstellen können (Li & Li 2015: 162f.). Solche Satzstrukturen sind auch im Fachbereit Astronautik zu finden:

> With the advent of the space shuttle, it will be possible to put an orbiting solar power plant in stationary orbit 24, 000 miles from the earth that would collect solar energy almost continuously and convert this energy either directly to electricity via photovoltaic cells or indirectly with flat plate or focused collectors that would boil a carrying medium to produce steam that would drive a turbine that then in turn would generate electricity. (Li & Li 2015: 163)

Dieser Satz enthält neben vielen technischen Begriffen, die Einzelheiten über Technologien und Prozesse beinhalten, eine Präpositionalphrase, eine Infinitivphrase und mehrere Attributsätze, die den Satz insgesamt sehr komplex machen (Li & Li 2015: 163).

3 Aktueller Stand der Forschung

Chan (2018: 1) stellte fest, dass der Mensch trotz der rasanten technologischen Fortschritte im Bereich der MÜ ein wichtiger Bestandteil ebenjener bleiben wird. Neben der Wahl des MÜ-Systems ist die Korrektur durch Menschen in Form von Pre- und Post-Editing fast schon zur wichtigeren Komponente im maschinellen Übersetzungsprozess geworden (Porsiel 2020: 4). Viele Forscher*innen haben sich bisher mit Pre- und Post-Editing auseinandergesetzt und den Einfluss dieser Bearbeitungsmethoden auf die Qualität der MÜ erforscht.

Zur Verbesserung der Übersetzbarkeit von nutzer*innengenerierten Inhalten wurden in der Forschungsarbeit von Seretan et al. (2014) Pre-Editing-Regeln für die Sprachkombination Englisch-Französisch und den Fachbereichen Gesundheit und Technik entwickelt, anhand derer die Quelltexte überarbeitet werden sollten. Als Quelltexte dienten allgemein zugängliche Foren- bzw. Blogtexte, deren umgangssprachliches, informelles Sprachregister von SMÜ-Systemen nur schwer korrekt verarbeitet werden konnte. Die automatisch oder manuell anwendbaren Pre-Editing-Regeln wurden in Anlehnung an bestehende Acrolinx-Regeln gewählt. Evaluiert wurden die mittels Pre-Editing bearbeiteten und die unbearbeiteten Texte durch Studierende des Masterstudiengangs Translation mit Französisch als Erstsprache anhand einer 5-Punkte-Likert-Skala. Es zeigte sich, dass Pre-Editing zu einer Qualitätsverbesserung des Resultats der SMÜ führt.

Mercader-Alarcón und Sánchez-Matínez (2016) untersuchten die Auswirkungen von Pre-Editing-Regeln auf maschinelle Übersetzungen des MÜ-Systems Lucy LT. Dazu wurden Fehlerannotationen anhand eines Textkorpus durchgeführt und diese dann anschließend qualitativ und quantitativ analysiert. Besonderes Augenmerk wurde dabei auf die linguistischen Aspekte wie Grammatik und Rechtschreibung gelegt, damit am Ende ein lexikalisch und grammatikalisch richtiges Ergebnis produziert wird. Die finalen Pre-Editing-Regeln wurden basierend auf der MQM gewählt. Die Untersuchung zeigte, dass die Anwendung von Pre-Editing-Regeln einen positiven Einfluss auf die Übersetzungsqualität hat und zu einer Reduktion von Wortfehlern und einer insgesamten Qualitätsverbesserung von 11 % führte.

Hiraoka und Yamada (2019) untersuchten den Einfluss von Pre-Editing-Regeln auf die Qualität von MÜ-Resultaten von TED-Untertitelungen in der Sprachkombination Japanisch (AS) und Englisch (ZS). Solche Untertitel unterliegen bestimmten Vorgaben, so muss die Übertragung der Bedeutung des AT in den ZT möglichst akkurat erfolgen und die Lesegeschwindigkeit von 21 Zeichen pro Sekunde darf nicht überschritten werden. Die bewusst einfach gehaltenen drei Pre-Editing-Regeln sollten es auch monolingualen Menschen ohne Kenntnisse

der ZS ermöglichen Pre-Editing effektiv verwenden zu können und betrafen die in Tabelle 6 dargestellten Regeln, nämlich die korrekte Verwendung von Eigennamen und Satzzeichen sowie die Ergänzung von fehlenden Subjekten und Objekten (2019: 65f.).

Tabelle 6: Überblick über die verwendeten Pre-Editing-Regeln (Hiraoka & Yamada 2019: 66)

Rule	Type	Method
1	Punctuation	Compensate missing punctuation (tôten)
2	Subject / Object	Compensate missing subject and/or object
3	Proper Noun	Write proper nouns in target language (English)

Die bearbeiteten und unbearbeiteten Segmente wurden mit einem NMÜ-System maschinell übersetzt und sowohl menschlich, als auch automatisch mittels BLEU-Metrik bewertet. Die menschliche Evaluierung wurde von Personen mit Erstsprache Japanisch mithilfe einer 3-Punkte-Skala, wo 3 der Bewertung *Good*, 2 der Bewertung *Acceptable* und 1 der Bewertung *Nonsense* entspricht, durchgeführt. Die Studie zeigte, dass die Anwendung von nur drei Pre-Editing-Regeln zu Verbesserungen bei neuronal übersetzen Untertiteln von TED Talks in der Sprachkombination Japanisch-Englisch führte. Sowohl die menschliche als auch die automatische Bewertung zeigte eine signifikante Qualitätssteigerung durch verschiedene Kombinationen dieser drei Pre-Editing-Regeln. 41 % der präeditierten Texte wurden als qualitativ besser bewertet als die unbearbeiteten Versionen, und der Anteil der mit 2 bewerteten Textsegmente stieg von 12 % auf 41 % (2019: 67f.).

Die Verwendung von Best-Worst-Scaling als Methode zur Bewertung der Qualität von maschineller Übersetzung stellt eine Innovation innerhalb der Translationswissenschaft dar. Hiebl und Gromann (2024) kombinierten BWS mit einer Likert-Skala, um die Qualität maschineller Übersetzungen im Vergleich zu Humanübersetzungen anhand von Textpassagen aus verschiedenen Sachbüchern aus den Gebieten Geschichte, Politik, Finanzwesen, Biologie und Physik in der Sprachkombination Englisch (AS) und Deutsch (ZS) zu evaluieren. Im Rahmen des BWS wählten die teilnehmenden Expert*innen, allesamt fortgeschrittene Studierende der Translationswissenschaft mit ausgezeichneten Englisch- und Deutschkenntnissen, aus vier deutschen Übersetzungen eines englischsprachigen Ausgangstexts jeweils die beste und schlechteste Übersetzung aus. Eine der vier Übersetzungen stellt dabei die offizielle

menschliche Übersetzung dar, die anderen drei Übersetzungen wurden maschinell mit Google Translate, DeepL und Microsoft Bing Translator erstellt.

Anschließend wurden die ausgewählten Übersetzungen mittels Likert-Skala hinsichtlich ihrer Qualität bewertet, wobei die jeweils beste Option von 4 (höchste Qualität) bis 0 (niedrigste Qualität) und die schlechteste Option von 0 (höchste Qualität) bis -4 (niedrigste Qualität) bewertet wurde. Diese kombinierte Vorgehensweise ermöglichte sowohl eine relative Rangordnung der Übersetzungen als auch eine differenzierte Einschätzung ihrer wahrgenommenen Qualität. Die Auswertung erfolgte durch Aggregation der BWS-Wertungen, der Auswahlhäufigkeiten sowie der Likert-Skalenbewertungen und wurde durch qualitative Analysen der freien Kommentare ergänzt.

Die Humanübersetzungen erzielten mit 33 Punkten die höchste Gesamtsumme auf der Likert-Skala, gefolgt von DeepL mit 25 Punkten. Google Translate und Microsoft Bing Translator wurden überwiegend negativ bewertet und erreichten -25 bzw. -20 Punkte. Obwohl DeepL eine etwas höhere durchschnittliche Bewertung als die Humanübersetzungen erzielte, wurden letztere insgesamt häufiger als beste Übersetzungsoption ausgewählt (48 % der Fälle gegenüber 30 % bei DeepL). Die Kombination aus BWS und Likert-Skala ermöglicht eine umfassende, differenzierte und nuancierte Bewertung der Qualität maschineller Übersetzungen und erwies sich als effektiv, da damit sowohl die relative Präferenz der Teilnehmer*innen als auch deren graduelle Qualitätswahrnehmung erfasst werden konnte.

Finetti (2020) und Göschl (2022) kombinierten Pre-Editing zur Optimierung des maschinellen Übersetzungsresultats mit BWS zur Evaluierung der Übersetzungsergebnisse und untersuchten die Auswirkung bestimmter Pre-Editing-Methoden auf die Qualität neuronaler maschineller Übersetzungen. Dazu wurden ausgewählte Textpassagen mit Pre-Editing bearbeitet und danach von einem neuronalen maschinellen Übersetzungssystem übersetzt. Insgesamt wurden sechs unterschiedliche Pre-Editing-Regeln verwendet, wovon jeweils zwei Kombinationsregeln, also eine Kombination aus zwei Pre-Editing-Operationen, sind. Die Übersetzungen wurden anschließend Proband*innen vorgelegt, die mittels BWS jeweils die beste und die schlechteste Version aus einem prädeterminierten Set an Übersetzungen wählten und anschließend die beste Option von 0 bis 4 (sehr gut) und die schlechteste von 0 bis -4 (sehr schlecht) bewerteten.

Im Fachbereich Biologie (Biolumineszenz) wurden englischsprachige Ausgangstexte nach Anwendung der Pre-Editing-Methoden mit SYSTRAN ins Italienische übersetzt. Es zeigte sich, dass die Pre-Editing-Methode, die die Verwendung von Ein-Wort-Verben anstelle

von Partikelverben vorsieht, die besten Ergebnisse erzielte und mit +15 Punkten die meisten positiven Bewertungen erhielt. Dagegen erwies sich die Kombinationsregel, die aus der Vermeidung von (Possessiv-)Pronomen und der Bedingung, dass jedes Segment grammatikalischen korrekt sein muss bestand, als am wenigsten effektiv und erhielt die negativste Bewertung (-26 Punkte). Damit wurden mit dieser Methode bearbeitete Textpassagen sogar schlechter bewertet als die Textpassagen ohne Pre-Editing (-19 Punkte). An der Studie nahmen fünf Proband*innen im Alter zwischen 18 und 26 Jahren teil, allesamt Studierende der Universität Wien mit hervorragenden Englisch- und Italienischkenntnissen und einem translationswissenschaftlichen Hintergrund (Finetti 2020: 62ff.).

Im Fachbereich Elektromobilität nahmen insgesamt 23 Personen an der BWS-Umfrage zur Qualitätsbeurteilung der maschinellen Google Translate-Übersetzungen vom Deutschen ins Englische teil. Unter den Teilnehmer*innen waren zehn Sprachexpert*innen und 13 Lai*innen. Bei der Auswertung der Pre-Editing-Methoden zeigte sich, dass insbesondere die Pre-Editing-Methode, die präzise Wortwahl und konsistenter Terminologie vorsieht, sowohl bei den Expertinnen, mit +57 Punkten, als auch bei den Lai*innen, mit +54 Punkten, die besten Ergebnisse erzielte. Auch die Pre-Editing-Methode Explizieren wurde von beiden Gruppen gemeinsam mit insgesamt 17 Punkten positiv bewertet. Die Kombinationsmethode, die zu einer Verbesserung der Satzstruktur führen sollte, wurde zwar von Sprachexpert*innen noch überwiegend positiv (+3), von Lai*innen jedoch kritisch (-6) bewertet. Die zweite Kombinationsregel (Redewendungen umformulieren und keine Wortteile auslassen) wurde mit -25 Punkten bei den Expert*innen und -45 Punkten bei den Lai*innen von beiden Gruppen als qualitätsvermindernd bewertet. Gesamt betrachtet schnitten zwei Pre-Editing-Methoden, nämlich die Vermeidung von Partizipalkonstruktionen (Regel 1) und die Kombinationsregel Metaphern umformulieren und bei Aufzählungen keine Wortteile weglassen (Regel 6K) mit einer Gesamtwertung von -71 (Regel 1) und -70 (Regel 6K) schlechter ab als kein Pre-Editing (-59) (Göschl 2022: 72ff.).

4 Methodik

Im Rahmen dieser Masterarbeit wird eine empirische qualitative Studie durchgeführt, deren Ergebnisse zusätzlichen quantitativ analysiert werden. Das Ziel der vorliegenden Datenerhebung besteht darin, fundierte Erkenntnisse über die Wirksamkeit von Pre-Editing-Regeln zu gewinnen, die auf einer detaillierten Fehleranalyse in einer konkreten Sprachkombination sowie einem spezifischen Fachgebiet basieren. Dazu soll einerseits untersucht werden, welche Pre-Editing-Regeln, und andererseits in welchem Ausmaß diese Regeln zu einer Verbesserung der MTranslatability und in weiterer Folge der maschinellen Übersetzungsqualität führen.

Der Prozess beginnt mit der Auswahl geeigneter Ausgangstextpassagen, gefolgt von der Gestaltung einer auf die Sprachkombination Englisch-Deutsch und das Fachgebiet Astronautik angepassten Fehlertypologie, um aus den ermittelten Übersetzungsfehlern im nächsten Schritt passende Pre-Editing-Regeln abzuleiten und auf die Ausgangstextpassagen anzuwenden. Danach folgt die maschinelle Übersetzung mit SYSTRAN und die menschliche Evaluierung der der resultierenden Übersetzungen mittels BWS. Im letzten Schritt werden die erhobenen Daten ausgewertet und die Ergebnisse analysiert und diskutiert. Abbildung 15 veranschaulicht die verschiedenen Phasen der vorliegenden empirischen Untersuchung.

Abbildung 15: Phasen der vorliegenden empirischen Untersuchung

4.1 Auswahl der Ausgangstexte

Bei der Auswahl der Texte wurde der Themenbereich Astronautik gewählt, da dieser Themenbereich einen konkreten Bezug zur wissenschaftlich-technischen Sprache hat. Als Ausgangstexte dienen kurze Textpassagen aus den öffentlich verfügbaren ESA Highlights (European Space Agency 2025). Das sind visuelle und informative Zusammenfassungen zentraler Aktivitäten und Erfolge der Europäischen Weltraumorganisation (engl. *European Space Agency* / *ESA*), die jährlich in englischer Sprache auf der ESA-Website publiziert werden. Diese Texte sind fachsprachlich geprägt, ohne aber den Grad an Komplexität und Spezialisierung von

typischer Fachliteratur zu erreichen und ohne vertieftes Fachwissen in Astronautik, Astronomie oder komplexe theoretische, mathematische, technische oder naturwissenschaftliche Kenntnisse vorauszusetzen. Die farbenfrohe Gestaltung der wissenschaftlichen Inhalte, ergänzt durch fotografische Darstellungen der ESA-Aktivitäten, trägt dazu bei, die ESA Highlights besonders auch fachfremden Personen zugänglich zu machen. Dadurch wird eine Brücke zwischen wissenschaftlicher Fachkommunikation und populärwissenschaftlicher Darstellung geschaffen, die es erleichtert, astronomische, naturwissenschaftliche und technische Fachbegriffe und Konzepte in einer Form zu präsentieren, die auch für Lai*innen und Translationsexpert*innen ohne spezifische Vorbildung im naturwissenschaftlich-technischen Bereich verständlich und interessant ist.

Die online verfügbaren rohen Textpassagen dargestellt in Anhang 1 wurden ausgewählt, da sie einerseits viele der in Kapitel 2.7.2 erwähnten sprachlichen Merkmale der englischen wissenschaftlich-technischen Sprache aufweisen, darunter auf lexikalischer Ebene eine Vielzahl spezifischer wissenschaftlich-technischer Begriffe und Akronyme aus dem Fachgebiet Astronautik wie z. B. *MetOp*, *ESA*, *Eumetsat*, *Lightning Imager*, *non-zero impact probability* oder *asteroid deflection* und auf syntaktischer Ebene viele Präpositionalphrasen (*With DART and Hera, we will…*), komplexe Strukturen sowie Attributsätze *(…objects for which a non-zero impact probability has been detected*) und andererseits einige sprachliche Komplexitäten und Merkmale aufweisen, die, wie in Kapitel 2.1.6.4 beschrieben, eine besondere Herausforderung für NMÜ-Systeme darstellen können. Hierzu zählen insbesondere idiomatische Wendungen, Metaphern sowie Koreferenzen. Bei den idiomatischen Wendungen *Keeping an eye out for lightning* in Textpassage 2 und *Ready for moon delivery* in Textpassage 11 könnte das MÜ-System Schwierigkeiten haben, die figurative Bedeutung korrekt zu interpretieren und stattdessen eine wörtliche oder semantisch unpassende Übersetzung erzeugen. Textpassage 9 weist eine komplexe, mehrdeutige und langdistanzierte Konstellation von Possessivpronomen auf. Das Possessivpronomen *its* kommt in der Passage zweimal vor und bezieht sich jeweils auf unterschiedliche Elemente. Zudem trennt eine Satzgrenze das zweite *its* von seinem zugehörigen Referenzelement und es liegt mindestens ein weiteres potenzielles Bezugselement dazwischen, was die Zuordnung und Interpretation der Referenz zusätzlich erschwert:

> **ESA's Mars Express** has revisited one of Mars' most mysterious features – the **Medusae Fossae Formation (MFF)** – to clarify **its** composition. **Its** findings suggest layers of water ice stretching several kilometres below ground – the most water ever found in this part of the planet. (European Space Agency 2024: 13); Hervorhebung der Autorin)

Aufgrund der identifizierten Fehlerpotenziale in den Rohtextpassagen, die potenziell zu Schwierigkeiten bei der maschinellen Übersetzung führen, stellen sowohl das Fachgebiet als auch die ausgewählten Textpassagen eine ausgezeichnete Basis für die Anwendung von Pre-Editing-Regeln dar, und bieten eine gute Möglichkeit, die damit verbundenen Veränderungen der MTranslatability sowie die Qualitätsveränderungen der Zieltexte zu untersuchen. Insgesamt wurden für die vorliegende Untersuchung zehn kurze Textpassagen ausgewählt, die im Anhang 1 aufgeführt sind. Da zum Zwecke der Evaluierung der Entscheidungskonsistenz der Proband*innen für diese Untersuchung zwei Passagen wiederholt werden (AT 1 = AT 8 und AT 4 = AT 12), erfolgte die Nummerierung der Textpassagen und der dazugehörigen Ausgangs- und Zieltexte von 1 bis 7 sowie von 9 bis 11.

4.2 Fehlertypologie

Grundlage für die Entwicklung der für die vorliegende Masterarbeit geeigneten Fehlertypologie bildet der Kern der in Kapitel 2.2.3.3 erläuterten MQM. Dafür wurden Übersetzungsfehler, die bei der maschinellen Übersetzung der ausgewählten Rohtextpassagen auftraten, analysiert. Die identifizierten Fehler wurden anschließend systematisch erfasst und in Anlehnung an die MQM-Fehlertypologie kategorisiert. Bestimmte Fehlerkategorie, insbesondere Design, Internationalisierung, Faktentreue und Lokalkonventionen, wurden im Zuge der Analyse ausgeschlossen, da sie für die untersuchten Textpassagen im Fachgebiet Astronautik sowie für die gewählte Sprachkombination Englisch (AS) und Deutsch (ZS) nicht relevant waren. Besonderes Augenmerk wurde hingegen auf die Kategorien Terminologie, Stil, Genauigkeit und Flüssigkeit gelegt, da sich diese als besonders bedeutsam für das vorliegende Untersuchungssetting erwiesen haben. Tabelle 7 zeigt die angepasste Fehlertypologie mit den ermittelten Fehlertypen inklusive zugehöriger MQM-Kategorien.

Tabelle 7: Angepasste Fehlertypologie inkl. Fehlertypen und MQM-Klassifizierungen

KATEGORIE	FEHLERTYPEN
FLÜSSIGKEIT	- Grammatikfehler - Ambiguität - unklare Referenzen
GENAUIGKEIT	- Hinzufügungen - Auslassungen - zu wörtliche Übersetzungen
STIL	- inkonsistenter Stil - holpriger Stil - unidiomatische Ausdrucksweise

TERMINOLOGIE	- unpräzise Fachbegriffe - Verletzung der Domänenkonventionen - inkonsistente Terminologie

4.3 Pre-Editing-Regeln

Die Ableitung der Pre-Editing-Regeln basiert auf der ermittelten Fehlertypologie. Ziel ist es, auf Grundlage einer systematischen Klassifikation entlang der MQM-Fehlerkategorien, Maßnahmen zu definieren und auf den englischen Rohausgangstextpassagen anzuwenden, um die Qualität der Ausgangstexte gezielt zu optimieren und so die MTranslatability und in Folge die maschinelle Übersetzungsqualität zu verbessern.

4.3.1 Auswahl der Pre-Editing-Regeln

Auf Grundlage der erstellten Fehlertypologie und unter Berücksichtigung der Vorschläge von Bernth und Gdaniec (2001), Roturier (2004), Sánchez-Gijón und Kenny (2022) sowie Hiraoka und Yamada (2019) zur Optimierung der MTranslatability für (neuronale) maschinelle Übersetzungssysteme, wurden die entsprechenden finalen Pre-Editing-Regeln für das Fachgebiet Astronautik in der Sprachkombination Englisch-Deutsch ausgewählt. Tabelle 8 veranschaulicht die identifizierten Pre-Editing-Regeln inkl. einer kurzen Beschreibung jeder Regel.

Tabelle 8: Pre-Editing-Regeln inkl. Kurzbeschreibung

PRE-EDITING-REGEL	BESCHREIBUNG
PRE-EDITING-REGEL 1: EINFACHE VERBEN	Partikelverben vermeiden und durch kontextuell passende Ein-Wort-Verben ersetzen
PRE-EDITING-REGEL 2: AKTIVKONSTRUKTIONEN	Passivkonstruktionen in sinnvolle Aktivkonstruktionen umwandeln
PRE-EDITING-REGEL 3: EXPLIZIEREN	Problematische idiomatische Ausdrücke, Metaphern und Ellipsen klar und verständlich explizieren und bei Bedarf notwendigen Kontext erweitern
PRE-EDITING-REGEL 4: PRÄZISE WORTWAHL	Einzelne Wörter durch fachsprachlich präzisere Begriffe ersetzen, überflüssige Wörter streichen und konsistente Terminologie innerhalb der Segmente verwenden
PRE-EDITING-REGEL 5: KLARE PRONOMINALE BEZÜGE	Eindeutige pronominale Bezüge herstellen, fehlende Subjekte und Objekte ergänzen und pronominale Bezüge über Satzgrenzen hinweg vermeiden
PRE-EDITING-REGEL 6: ING-KONSTRUKTIONEN AUFLÖSEN	Ing-Konstruktionen durch Relativsätze ersetzen oder eine passende Präposition (wie z. B. „by“ hinzufügen)

Pre-Editing-Regel 1 basiert auf der Empfehlung von Bernth und Gdaniec (2001: 187), die in ihrer Regel 12 anmerken, dass Verben aus einem Wort anstelle von Verb-Partikel-Konstruktionen verwendet werden sollten, um die Verständlichkeit zu erhöhen. Pre-Editing-Regel 2, die vorsieht Passivkonstruktionen aufzulösen und in Aktivsätze umzuwandeln, stützt sich einerseits auf Regel 16 von Bernth und Gdaniec (2001: 190) sowie andererseits auf die strukturell-stilistischen Richtlinien von Sánchez-Gijón und Kenny (2022: 93), die den Nutzen von Aktivkonstruktion für die Klarheit und Lesbarkeit betonen. Pre-Editing-Regel 3, die das Umschreiben von idiomatischen Wendungen, Metaphern und Ellipsen und gegebenenfalls eine Kontexterweiterung bei problematischen Formulierungen vorschlägt, basiert auf den Empfehlungen von Bernth und Gdaniec (2001: 188f.), die in Regel 14 und 15 davon abraten, Metaphern, idiomatische Ausdrücke und Ellipsen zu verwenden, um eine klare und verständliche Sprache zu fördern. Diese Praxis ist besonders nützlich, um figurative Ausdrücke zu vermeiden und den Ausgangstext verständlicher zu gestalten.

Pre-Editing-Regel 4 zielt darauf ab, im Sinne der domänenspezifischen terminologischen Konsistenz überflüssige Wörter zu entfernen und domänenunspezifische Begriffe durch präzisere Fachtermini zu ersetzen. Diese Regel basiert auf den lexikalischen Richtlinien von Sánchez-Gijón und Kenny (2022: 91), in denen eine konsistente Terminologie und die Vermeidung unnötiger Wörter empfohlen wird, um die Klarheit und Fachsprachlichkeit zu gewährleisten. Diese domänenspezifische Gestaltung des Ausgangstextes sowie die damit verbundene Vermeidung der Vermischung unterschiedlicher Domänen und die Minimierung allgemeinsprachlicher Begriffe in der Zieldomäne verfolgt darüber hinaus das Ziel, die in Kapitel 2.1.6.4 behandelten *domain mismatches* möglichst zu minimieren.

Um pronominale Unklarheiten zu minimieren, empfiehlt Pre-Editing-Regel 5 den Bezug zu präzisieren und fehlende Subjekte und Objekte zu ergänzen. Diese Praxis stützt sich auf die Richtlinie für Referenzelemente von Sánchez-Gijón und Kenny (2022: 94f.), in denen die konsistente Verwendung von Pronomen innerhalb eines Satzes und die Vermeidung von pronominalen Bezügen über Satzgrenzen hinweg empfohlen wird sowie auf die Empfehlungen von Hiraoka und Yamada (2019: 66), die die Wirksamkeit des Ergänzens fehlender Objekte und Subjekte festgestellt haben. Pre-Editing-Regel 6 basiert auf der Handlungsempfehlung von Roturier (2004: 11) zur Auflösung von ing-Konstrukten sowie auf Regel 3 und 4 von Bernth und Gdaniec (2001: 183), die analog dazu entweder die Umformulierung von ing-Konstrukten in Relativsätze oder die Ergänzung einer passenden Präposition empfehlen. Da in ähnlichen Untersuchungen keine signifikanten Qualitätsverbesserungen und zum Teil sogar Qualitäts-

reduktionen mit Kombinationsregeln festgestellt wurden (Finetti 2020: 73; Göschl 2022: 90), wird in dieser Arbeit bewusst auf die Verwendung von Kombinationsregeln verzichtet. Das ermöglicht den Fokus auf die feinere Differenzierung der Auswirkungen von einzelnen Regeln auf die maschinelle Übersetzungsqualität.

4.3.2 Experimentelles Design

Beim experimentellen Design wurde auf eine gleichmäßige Verteilung der Pre-Editing-Regeln geachtet, um sicherzustellen, dass alle Regeln fair getestet werden und eine objektive Analyse der Ergebnisse ermöglicht wird. Jede Rohausgangstextpassage wird dabei mit drei verschiedenen präeditierten Versionen verglichen, d. h. auf jede Textpassage werden drei der sechs verfügbaren Pre-Editing-Regeln angewendet. Die Verteilung der einzelnen Pre-Editing-Regeln erfolgt nicht nur gleichmäßig über alle Textpassagen, sondern auch bei der Gruppierung der einzelnen Regeln wurde auf ein ausgeglichenes Verhältnis geachtet. Jede Gruppierung aus drei Regeln wird so ausgewählt, dass keine einzelne Gruppierung übermäßig dominant ist und jede Regel in unterschiedlichen Sets evaluiert wird. Diese Vorgehensweise dient der Vermeidung von Verzerrungen und stellt sicher, dass alle Regeln unter vergleichbaren Bedingungen bewertet werden können. Ziel ist eine erhöhte Validität und eine ausgewogene Analyse der Effekte einzelner Regeln. Um festzustellen, ob die Entscheidungen der Teilnehmer*innen konsistent sind und nicht von zufälligen oder äußeren Faktoren beeinflusst wurden, erfolgt zusätzlich ein Validierungstest, für den zwei Textpassagen wiederholt werden (AT 8 = AT 1 und AT 12 = AT 4). Tabelle 9 veranschaulicht die experimentelle Anordnung der Pre-Editing-Regeln und dazugehörigen Textpassagen.

Tabelle 9: Experimentelles Design der Textpassagen und Pre-Editing-Regeln

TEXTPASSAGE 1			**TEXTPASSAGE 2**		
AT 1.1	OHNE PRE-EDITING	ZT 1.1	AT 2.1	OHNE PRE-EDITING	ZT 2.1
AT 1.2	PRE-EDITING-REGEL 1	ZT 1.2	AT 2.2	PRE-EDITING-REGEL 5	ZT 2.2
AT 1.3	PRE-EDITING-REGEL 4	ZT 1.3	AT 2.3	PRE-EDITING-REGEL 3	ZT 2.3
AT 1.4	PRE-EDITING-REGEL 2	ZT 1.4	AT 2.4	PRE-EDITING-REGEL 2	ZT 2.4
TEXTPASSAGE 3			**TEXTPASSAGE 4**		
AT 3.1	OHNE PRE-EDITING	ZT 3.1	AT 4.1	OHNE PRE-EDITING	ZT 4.1
AT 3.2	PRE-EDITING-REGEL 5	ZT 3.2	AT 4.2	PRE-EDITING-REGEL 3	ZT 4.2
AT 3.3	PRE-EDITING-REGEL 1	ZT 3.3	AT 4.3	PRE-EDITING-REGEL 6	ZT 4.3
AT 3.4	PRE-EDITING-REGEL 4	ZT 3.4	AT 4.4	PRE-EDITING-REGEL 5	ZT 4.4

TEXTPASSAGE 5			TEXTPASSAGE 6		
AT 5.1	OHNE PRE-EDITING	ZT 5.1	AT 6.1	OHNE PRE-EDITING	ZT 6.1
AT 5.2	PRE-EDITING-REGEL 6	ZT 5.2	AT 6.2	PRE-EDITING-REGEL 2	ZT 6.2
AT 5.3	PRE-EDITING-REGEL 2	ZT 5.3	AT 6.3	PRE-EDITING-REGEL 1	ZT 6.3
AT 5.4	PRE-EDITING-REGEL 4	ZT 5.4	AT 6.4	PRE-EDITING-REGEL 3	ZT 6.4
TEXTPASSAGE 7			**TEXTPASSAGE 8 = TEXTPASSAGE 1**		
AT 7.1	OHNE PRE-EDITING	ZT 7.1	AT 8.1	OHNE PRE-EDITING	ZT 8.1
AT 7.2	PRE-EDITING-REGEL 1	ZT 7.2	AT 8.2	PRE-EDITING-REGEL 1	ZT 8.2
AT 7.3	PRE-EDITING-REGEL 6	ZT 7.3	AT 8.3	PRE-EDITING-REGEL 4	ZT 8.3
AT 7.4	PRE-EDITING-REGEL 2	ZT 7.4	AT 8.4	PRE-EDITING-REGEL 2	ZT 8.4
TEXTPASSAGE 9			**TEXTPASSAGE 10**		
AT 9.1	OHNE PRE-EDITING	ZT 9.1	AT 10.1	OHNE PRE-EDITING	ZT 10.1
AT 9.2	PRE-EDITING-REGEL 4	ZT 9.2	AT 10.2	PRE-EDITING-REGEL 6	ZT 10.2
AT 9.3	PRE-EDITING-REGEL 5	ZT 9.3	AT 10.3	PRE-EDITING-REGEL 1	ZT 10.3
AT 9.4	PRE-EDITING-REGEL 3	ZT 9.4	AT 10.4	PRE-EDITING-REGEL 5	ZT 10.4
TEXTPASSAGE 11			**TEXTPASSAGE 12 = TEXTPASSAGE 4**		
AT 11.1	OHNE PRE-EDITING	ZT 11.1	AT 12.1	OHNE PRE-EDITING	ZT 12.1
AT 11.2	PRE-EDITING-REGEL 6	ZT 11.2	AT 12.2	PRE-EDITING-REGEL 3	ZT 12.2
AT 11.3	PRE-EDITING-REGEL 4	ZT 11.3	AT 12.3	PRE-EDITING-REGEL 6	ZT 12.3
AT 11.4	PRE-EDITING-REGEL 3	ZT 11.4	AT 12.4	PRE-EDITING-REGEL 5	ZT 12.4

4.3.3 Manuelles Pre-Editing

Die Pre-Editing-Regeln wurden anschließend manuell von der Autorin dieser Arbeit auf die Rohtextpassagen angewendet. Infolge dessen entstanden für jede Textpassage drei unterschiedliche präeditierte Varianten, die gemeinsam mit der unbearbeiteten Textpassage im nächsten Schritt einer maschinellen Übersetzung unterzogen wurden. Tabelle 10 veranschaulicht die Anwendung der Pre-Editing-Regeln am Beispiel der Textpassage 5. Diese wurde mit den Pre-Editing-Regeln 6, 2 und 4 bearbeitet. Die Edits im Ausgangstext wurden für jede Regel farblich markiert.

Tabelle 10: Beispielausgangstextpassage ohne und mit Pre-Editing

PRE-EDITING-REGEL	AUSGANGSTEXTE
OHNE PRE-EDITING	Two satellites will be launched together into orbit to maintain formation relative to each other down to a few millimetres, creating an artificial solar eclipse in space. This mission is being put together for ESA by a consortium led by Spain's Sener, with participation by more than 29 companies from 14 countries (European Space Agency 2024: 30).

PRE-EDITING-REGEL 6: ING-KONSTRUKTIONEN AUFLÖSEN	Two satellites will be launched together into orbit to maintain formation relative to each other down to a few millimetres, which will create an artificial solar eclipse in space. This mission is being put together for ESA by a consortium led by Spain's Sener, with participation by more than 29 companies from 14 countries.
PRE-EDITING-REGEL 2: AKTIVKONSTRUKTIONEN	Two satellites will be launched together into orbit to maintain formation relative to each other down to a few millimetres, creating an artificial solar eclipse in space. A consortium led by Spain's Sener, with participation by more than 29 companies from 14 countries, is putting together this mission for ESA.
PRE-EDITING-REGEL 4: PRÄZISE WORTWAHL	Two satellites will be launched into orbit together to maintain formation relative to each other accurate to the millimetre, creating an artificial solar eclipse in space. This mission is being put together for ESA by a consortium led by Spain's Sener, with participation from more than 29 companies from 14 countries.

4.4 Maschineller Übersetzungsprozess mit SYSTRAN

Die Entscheidung für SYSTRAN in dieser Masterarbeit beruht auf der leistungsfähigen, etablierten Technologie mit moderner Transformer-Architektur sowie dem benutzer*innenfreundlichen Online-Interface, das eine unkomplizierte und effiziente Nutzung für alle Menschen mit Internetzugang gleichermaßen ermöglicht. Zudem erfolgte die Wahl bewusst im Hinblick auf das Interesse, ein weniger verbreitetes System im Vergleich zu marktführenden Anbietern wie Google Translate oder DeepL zu untersuchen, um die Auswirkungen von kleinen Textanpassungen auf die Qualität neuronaler maschineller Übersetzungen auch außerhalb der dominierenden Systeme zu analysieren. Abbildung 16 zeigt das SYSTRAN Online-Interface, welches für die vorliegenden Übersetzungen verwendet wurde am Beispiel des Rohtextes der Textpassage 5.

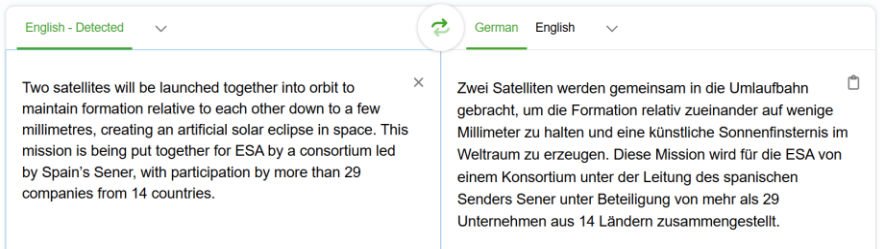


Abbildung 16: Beispielübersetzung der Rohversion von Textpassage 5

Eine detaillierte Übersicht aller Textpassagen, einschließlich der angewendeten Pre-Editing-Regeln sowie der entsprechenden maschinellen Übersetzungen ins Deutsche, findet sich in Anhang 2.

4.5 Menschliche Evaluierung mit Best-Worst-Scaling

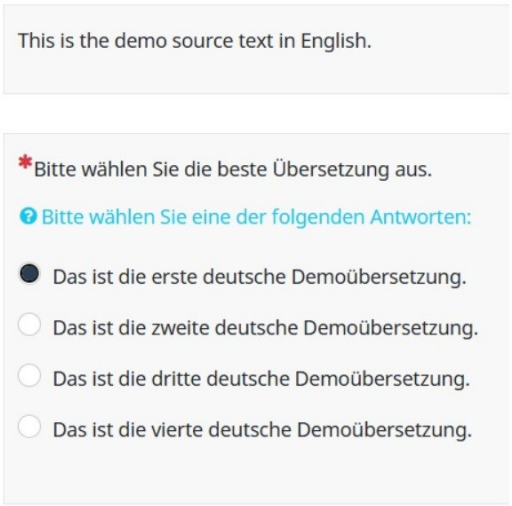
Die menschliche Evaluierung gilt in der Translationswissenschaft nach wie vor als Goldstandard für die Beurteilung der Qualität maschineller Übersetzungen. BWS ist hierbei aber noch ein relativ neuer Ansatz. Aufgrund der in Kapitel 2.1.2 behandelten unterschiedlichen Anwendungszwecke maschineller Übersetzung sowie der breiten Verfügbarkeit und der zunehmenden Nutzung verschiedenster MÜ-Systeme durch Lai*innen, legt diese Arbeit den Fokus auf eine kontrastive Evaluierung der Übersetzungsqualität mittels BWS aus der Perspektive von Translationsexpert*innen und Lai*innen.

Hierfür wurde das in Kapitel 4.3.2 beschriebene experimentelle Design in LimeSurvey (LimeSurvey GmbH 2025) eingebettet und 12 Fragengruppen für die BWS-Methode angelegt. Jede Fragengruppe enthält die BWS-Struktur für eine Textpassage, wobei sich jeweils zwei Textpassagen zum Zwecke der Validierung durch Überprüfung der Entscheidungskonsistenz wiederholen. Pro Fragengruppe wird den Teilnehmer*innen ein englischer Ausgangstext inklusive seiner vier deutschen Übersetzungen präsentiert, wovon diese die beste und schlechteste Übersetzung wählen sollen. Jede gewählte Option wird zusätzlich im Vergleich zu den übrigen Optionen auf einer Skala von 0 (neutral) bis 4 (besonders gut) für die beste Übersetzung bzw. aus 0 (neutral) bis -4 (besonders schlecht) für die schlechteste Übersetzung bewertet. Darüber hinaus gibts es für jede Textpassage ein optionales Feld für Anmerkungen, in dem die Teilnehmer*innen ihre Entscheidungen begründen können.

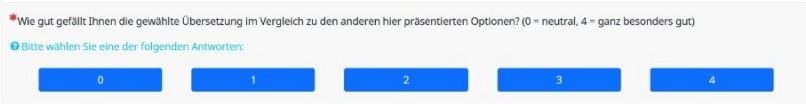
Vor der Bearbeitung der 12 Textpassagen werden die Teilnehmer*innen gebeten ein paar allgemeine und demografische Fragen zu beantworten, darunter auch die Frage, ob ein translationswissenschaftlicher Studienabschluss und/oder translationsrelevante Berufserfahrung, beispielsweise als Übersetzer*in oder Dolmetscher*in, vorliegt. Diese Fragen dienen als Entscheidungsgrundlage für die Unterscheidung zwischen Expert*innen und Lai*innen, wobei der Abschluss eines mindestes dreijährigen translationswissenschaftlichen Studiums auf Bachelorebene oder, äquivalent dazu, eine mindestens dreijährige facheinschlägige Tätigkeit als Translator*in die Voraussetzung für die Kategorie der Translationsexpert*innen darstellt. Darüber hinaus beantworten die Proband*innen Fragen zu Alter, Geschlecht, Sprachniveau in Deutsch und Englisch, basierend auf dem Gemeinsamen Europäischen Referenzrahmen für Sprachen (engl. *Common European Framework of Reference for Languages / CEFR*) (Europarat 2025) und zu ihrer Erfahrung mit MÜ bzw. wie oft sie beruflich oder privat MÜ verwenden. Abbildung 17 zeigt den Leitfaden zur Bearbeitung der Textpassage, der den Teilnehmer*innen vor dem Start der Umfrage gezeigt wurde.

Leitfaden für die Bearbeitung der Textproben

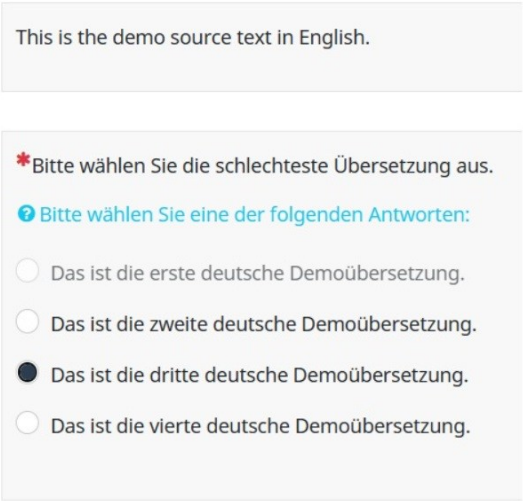
Für jede Textprobe wählen Sie zunächst aus den vier Optionen die aus Ihrer Sicht beste Übersetzung aus.



Im Anschluss bewerten Sie, wie gut Ihnen diese ausgewählte Übersetzung im Vergleich zu den übrigen drei Optionen auf einer Skala von 0 (neutral) bis 4 (ganz besonders gut) gefällt.



Danach wählen Sie bitte für **denselben** englischen Ausgangstext aus den verbleibenden drei Optionen die schlechteste Übersetzung aus.



Bewerten Sie im Anschluss auch diese ausgewählte Übersetzung auf einer Skala von 0 (neutral) bis -4 (ganz besonders schlecht) im Vergleich zu den restlichen Übersetzungsvarianten.

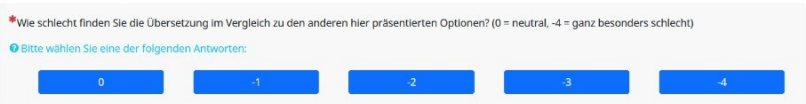


Abbildung 17: Leitfaden für die BWS-Bearbeitung der Textpassagen in der Umfrage

Die Entscheidung für die Durchführung dieser Umfrage LimeSurvey zu verwendet liegt darin begründet, dass kein bekannter Umfragedienstleister gefunden werden konnte, der eine integrierte BWS-Umsetzung anbietet, LimeSurvey es der Autorin dieser Masterarbeit jedoch aufgrund der hohen Anpassbarkeit mittels JavaScript ermöglichte, eine technisch saubere und benutzer*innenfreundliche BWS-Studie zu erstellen, die keine formalen Eingabefehler zulässt. Durch diese Umsetzung wird sichergestellt, dass eine fehlerhafte Beantwortung der Umfrage technisch ausgeschlossen ist.

Vor der Veröffentlichung der Umfrage wurde eine Pilotierung mit zwei fachfremden Personen durchgeführt, um die technische Umsetzung sowie den Aufbau, die Dauer und den Umfang der erstellten Umfrage zu evaluieren. Die Pilotierung ergab, dass die erstellte BWS-Struktur technisch funktioniert, die Bearbeitung der Textpassagen ohne tiefgehendes Wissen im Bereich der Astronautik möglich ist und Dauer und Umfang der Umfrage zwar akzeptabel sind, aber die Bearbeitung stellenweise, vorwiegend aufgrund der starken Ähnlichkeit der einzelnen Übersetzungen, schwierig und mit viel Konzentration verbunden ist. Nach der Pilotierung wurde der Link zur Umfrage am 30. April 2025 über E-Mail, Facebook und diverse IM-Dienste verschickt, um eine möglichst große Gruppe an Translationsexpert*innen und Lai*innen zu erreichen und nach einer Woche, am 7. Mai 2025 wieder offline genommen. Die Ergebnisse dieser Datenerhebung werden zunächst in Kapitel 5 ausgewertet und in Kapitel 6 anschließend diskutiert.

5 Ergebnisse der Datenerhebung

Im Folgenden werden die aus LimeSurvey nach Beendigung der Umfrage extrahierten Daten analysiert, aufbereitet und grafisch dargestellt. Zur Bestimmung der relativen Qualität wird die Summe aller Einzelbewertungen pro Pre-Editing-Regel herangezogen, wobei die höchste Punktzahl auf die qualitativ beste, und die niedrigste Bewertung auf die qualitativ schlechteste Pre-Editing-Regel hinweist. Diese Daten sind wissenschaftlich relevant, da sie Aufschluss darüber geben, welche Regeln die Übersetzungsqualität verbessern und welche sie möglicherweise sogar reduzieren.

Hierfür werden zunächst in Kapitel 5.1. die demografischen Angaben der Teilnehmer*innen ausgewertet und darauf basierend die Profile der Teilnehmer*innen vorgestellt. In Kapitel 5.2. folgt eine detaillierte komparative Auswertung zwischen Translationsexpert*innen und Lai*innen, die Präsentation der Ergebnisse der Validierungstests, eine Gesamtbewertung sämtlicher Pre-Editing-Regeln und deren Auswirkungen auf die wahrgenommene Qualität von MÜ sowie eine qualitative Auswertung der offenen Frage zur Entscheidungsbegründung.

5.1 Profile der Teilnehmer*innen

Die Umfrage wurde von insgesamt 27 Personen vollständig ausgefüllt, darunter 15 Lai*innen und 12 Expert*innen. Zu den Expert*innen zählen all jene Personen, die in der Umfrage angaben, entweder ein mindestens dreijähriges translationswissenschaftliches Studium absolviert zu haben, oder, analog dazu, über eine mindestens dreijährige facheinschlägige Berufstätigkeit zu verfügen. 11 der 12 Expert*innen gaben als höchsten translationswissenschaftlichen Abschluss ein absolviertes Bachelorstudium, und ein Experte bzw. eine Expertin ein Masterstudium an. Ein Laie bzw. eine Laiin gab an, keinen translationswissenschaftlichen Abschluss zu haben, aber über Berufserfahrung im translatorischen Bereich zu verfügen, die aktuell jedoch weniger als drei Jahre beträgt und daher das Kriterium für die Gruppe der Expert*innen nicht erfüllt. Alle teilnehmenden Expert*innen verfügen daher über einen translationswissenschaftlichen Hochschulabschluss mindestens auf Bachelorniveau.

Von den insgesamt 27 Teilnehmer*innen identifizieren sich 14 als weiblich, darunter acht Laiinnen und sechs Expertinnen, sowie 13 als männlich, davon sieben Laien und sechs Experten. Somit ergibt sich sowohl insgesamt als auch in den jeweiligen Gruppen ein nahezu ausgewogenes binäres Geschlechterverhältnis. Etwas mehr als die Hälfte der Umfrageteilnehmer*innen (15 Personen) ist zwischen 26 und 35 Jahre alt, darunter drei Laien, zwei Experten, fünf Laiinnen und fünf Expertinnen. In der Gruppe der Expert*innen gibt es nur eine Person

zwischen 36 und 45 und niemanden über 45. Bei den Lai*innen gibt es ebenso eine Person in der Altersgruppe 36 bis 45, jedoch insgesamt vier weitere Personen, die älter als 45 sind, nämlich drei in der Altersgruppe 46 bis 55 und eine Person, die älter als 55 ist. Von den sechs Personen zwischen 18 und 25 entfallen vier auf die Gruppe der Expert*innen und zwei auf die der Lai*innen. Der Altersschnitt der Lai*innen liegt demnach etwas über dem der Expert*innen. Abbildung 18 veranschaulicht die Geschlechter- und Altersverteilung aller 27 Teilnehmer*innen.

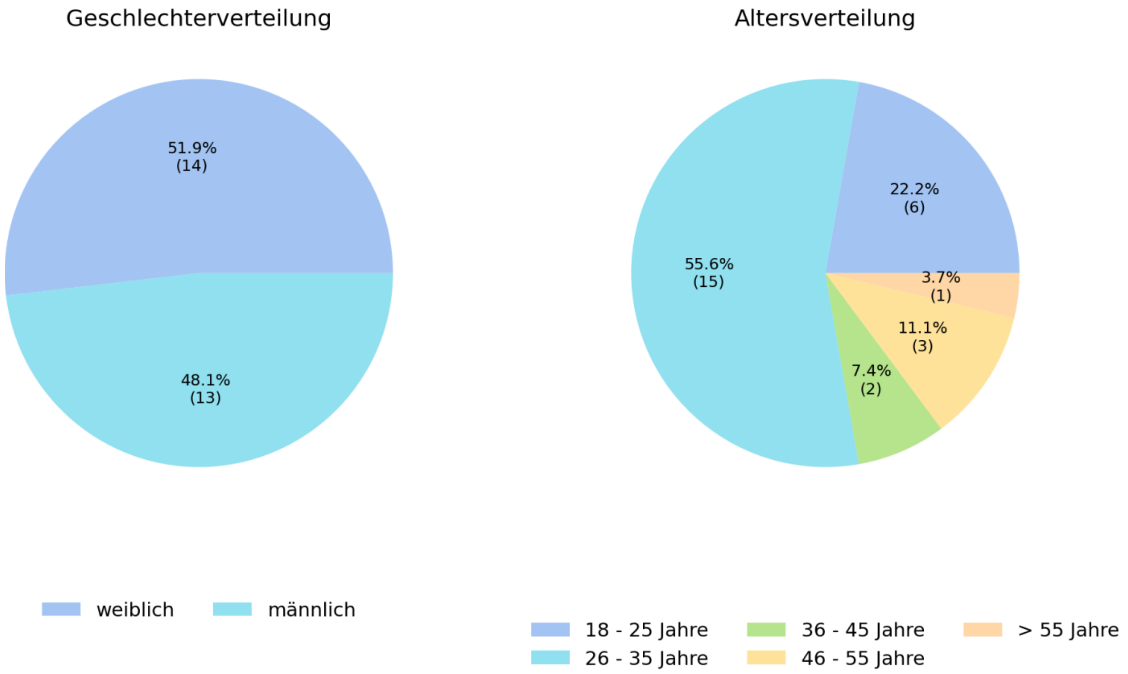


Abbildung 18: Geschlechter- und Altersverteilung aller 27 Teilnehmer*innen

Drei Expert*innen und ein Laie bzw. eine Laiin gaben an, dass Deutsch nicht ihre Erstsprache ist. Eine dieser Personen wählte Englisch als Erstsprache. Alle Teilnehmer*innen bestätigten, Deutsch- und Englischkenntnisse auf Niveau B2 bis C2 zu besitzen. Alle Expert*innen wählten aus, privat oder beruflich MÜ-Systeme zu nutzen. 11 von 12 Expert*innen (91,7 %) gaben sogar an, häufig (ca. ein- bis zweimal pro Woche) bis sehr häufig (fast täglich) MÜ-Systeme zu verwenden. Bei den Lai*innen nutzt zwar immerhin noch die Mehrheit der Teilnehmer*innen regelmäßig MÜ-Systeme, jedoch liegt der Anteil derer, die häufig bis sehr häufig MÜ-Systeme nutzen nur bei 60 %. Zudem gaben zwei Lai*innen (13,3 %) an, nie MÜ-Systeme zu verwenden. Vier Lai*innen (26,7 %) aber nur ein Experte bzw. eine Expertin (8,3 %) wählten aus, dass sie selten privat oder beruflich MÜ-Systeme nutzen. Abbildung 19 veranschaulicht die Häufigkeit der Nutzung von MÜ-Systemen für private und/oder berufliche Zwecke zwischen Translationsexpert*innen (links) und Lai*innen (rechts).

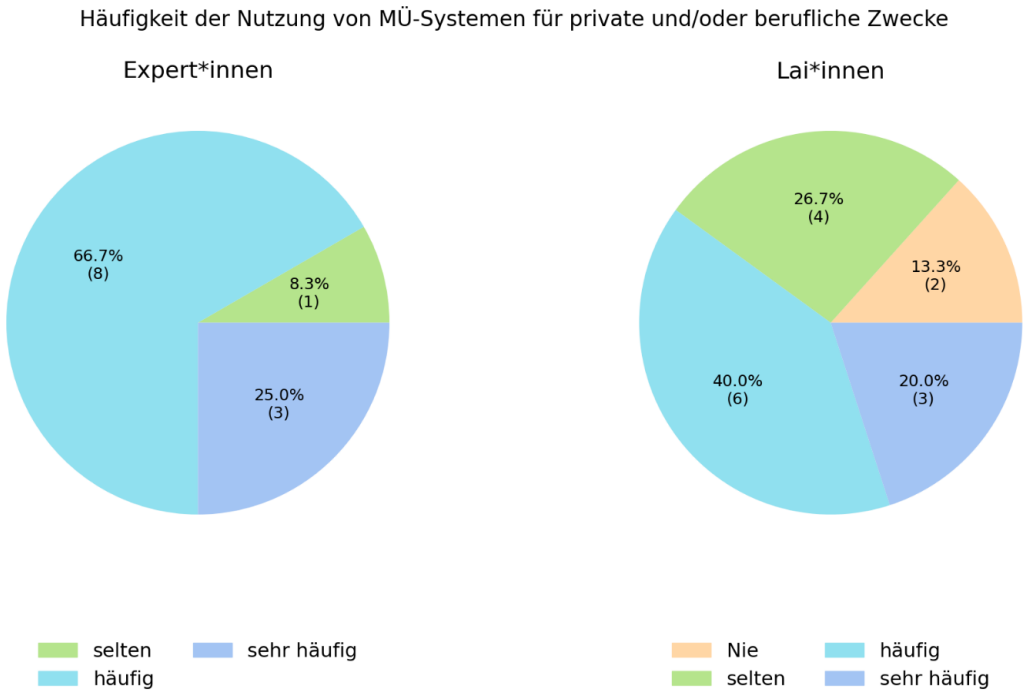


Abbildung 19: Häufigkeit der Nutzung von MÜ-Systemen zwischen Expert*innen und Lai*innen

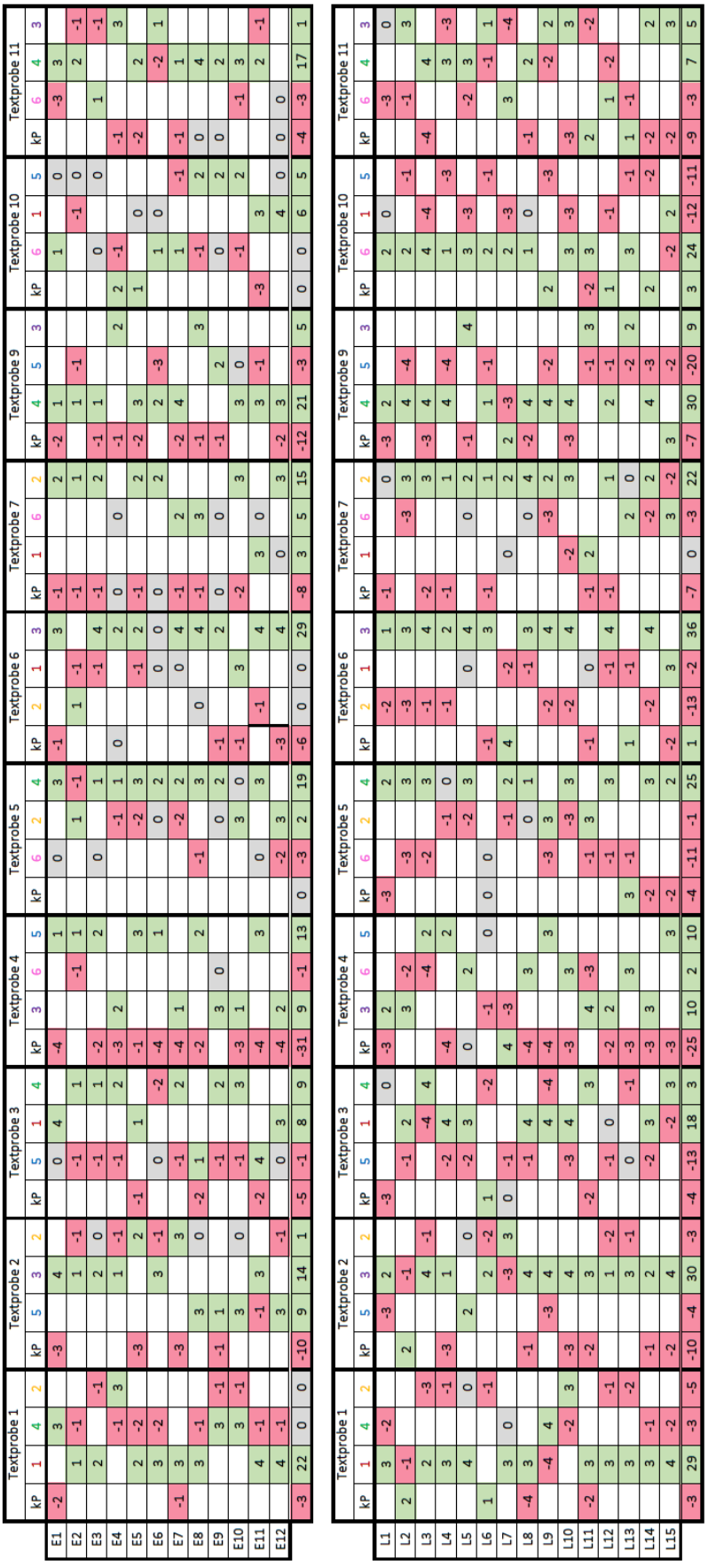
5.2 Bewertungen von Translationsexpert*innen und Lai*innen im Vergleich

Dieses Kapitel präsentiert eine detaillierte vergleichende Auswertung der Ergebnisse von Expert*innen und Lai*innen auf Basis der 10 untersuchten Textpassagen, die sowohl quantitative Einzelbewertungen der verwendeten Pre-Editing-Regeln als auch eine qualitative Analyse der optionalen Frage zur Qualitätswahrnehmung und -begründung umfasst. Zusätzlich werden die Ergebnisse der Validierungstests sowie eine Gesamtbewertung der Pre-Editing-Regeln und deren Einfluss auf die MÜ-Qualität dargestellt.

5.2.1 Einzelbewertungen der Teilnehmer*innen

In Tabelle 11 sind die Einzelbewertungen aller Teilnehmer*innen getrennt nach Expert*innen (E) und Lai*innen (L) für alle 10 Textpassagen sowie die Summe der Bewertungen pro Textpassage detailliert aufgeführt. Die Bewertungen von Textpassage 8 und 12 sind in dieser Übersicht nicht enthalten, da sie lediglich der Validierung dienen und gesondert in Kapitel 5.2.3 mit ihren dazugehörigen identen Gegenstücken, nämlich Textpassage 1 und Textpassage 4, evaluiert werden. Jede Textpassage ist mit ihren jeweiligen Pre-Editing-Regeln dargestellt, wobei kP für kein Pre-Editing, also die unbearbeitete Version steht. Grün eingefärbte Felder stellen positive (1 bis 4), rote Felder negative (-1 bis -4), und grau hinterlegte Felder neutrale Bewertungen (0) dar.

Tabelle 11: Einzelbewertungen von Expert*innen und Lai*innen



5.2.2 Gesamtbewertung der Teilnehmer*innen

Zur Ermittlung der von den Teilnehmer*innen relativ am besten und am schlechtesten bewerteten Pre-Editing-Regeln werden zunächst alle Einzelbewertungen pro Teilnehmer*in aggregiert. Tabelle 12 und Tabelle 13 veranschaulichen das Ergebnis dieser Aggregation differenziert nach Regel sowie nach Expert*innen und Lai*innen.

Tabelle 12: Aggregierte Gesamtbewertungen der Expert*innen nach Pre-Editing Regel

	kP	1	2	3	4	5	6
E1	-13	4	2	7	10	1	-2
E2	-1	-1	2	0	2	-1	-1
E3	-4	1	1	5	3	1	1
E4	-3	0	1	10	2	-1	-1
E5	-9	2	2	2	6	3	0
E6	-4	3	1	4	-2	-2	1
E7	-12	3	1	5	9	-2	3
E8	-6	3	0	7	6	8	1
E9	-3	0	-1	5	9	4	0
E10	-6	3	5	1	12	4	-2
E11	-9	10	-1	6	7	5	0
E12	-9	11	5	6	2	3	-2
Gesamt	-79	39	18	58	66	23	-2

Von den Expert*innen wurden alle präeditierten Textpassagen besser bewertet als die uneditierten Versionen. Von den insgesamt sechs Pre-Editing-Regeln wurden fünf Regeln positiv und eine schwach negativ bewertet. Die positiv bewerteten Pre-Editing-Regeln sind Regel 1 (einfache Verben) mit 39 Punkten, Regel 2 (Aktivkonstruktionen) mit 18 Punkten, Regel 3 (Explizieren) mit 58 Punkten, Regel 4 (präzise Wortwahl) mit 66 Punkten und Regel 5 (klare pronominale Bezüge) mit 23 Punkten. Einzig Pre-Editing-Regel 6 (Vermeidung von ing-Konstruktionen) wurde von den Expert*innen schwach negativ mit -2 Gesamtpunkten bewertet, erzielte aber dennoch ein besseres Ergebnis als kP mit -79 Punkten. Besonders interessant ist, dass kP von allen Expert*innen einheitlich negativ bewertet wurde. Außer bei einem Experten bzw. einer Expertin, der oder die kP mit der gleichen aggregierten Punktezahl bewertete, wie Pre-Editing-Regel 6, 2 und 5 (-1) erzielten alle Pre-Editing-Regeln bei den Expert*innen individuell betrachtet ein besseres Ergebnis als kein Pre-Editing.

Am kritischsten wurde von allen Expert*innen Pre-Editing-Regel 6, also die Vermeidung von ing-Konstruktionen, bewertet. Von allen sechs Regeln wurde Regel 6 sowohl insgesamt am schlechtesten (-2), als auch individuell von den Expert*innen am häufigsten negativ bewertet, nämlich von 5 Expert*innen, und erzielt damit im Vergleich zu den anderen Regeln, die null (Regel 3) bis maximal vier (Regel 5) Negativbewertungen erzielten, die höchste Anzahl

an Negativbewertungen. Besonders bemerkenswert ist darüber hinaus der große Abstand von 77 Punkten zwischen der am schlechtesten bewerteten Pre-Editing-Regel 6 (-2) und kP (-79).

Tabelle 13: Aggregierte Gesamtbewertung der Lai*innen nach Pre-Editing-Regel

	kP	1	2	3	4	5	6
L1	-13	3	-2	5	2	-3	-1
L2	4	1	0	8	2	-6	-7
L3	-9	-6	-2	8	7	2	-2
L4	-8	7	-2	0	15	-7	1
L5	-1	4	0	8	7	0	3
L6	0	0	-2	5	6	-2	2
L7	10	-2	4	-10	-2	-1	5
L8	-12	6	4	7	-1	-1	4
L9	-2	0	3	10	7	-5	-6
L10	-12	-1	1	11	5	-3	6
L11	-8	5	3	8	3	-1	-1
L12	-2	1	-2	7	3	-2	0
L13	2	2	-3	5	-1	-3	6
L14	-6	6	0	11	6	-7	-2
L15	-8	7	-2	7	3	1	1
Gesamt	-65	33	0	90	62	-38	9

Bei den Lai*innen zeigt sich insgesamt betrachtet auf den ersten Blick ein ähnliches Bild wie bei den Expert*innen, denn auch von den Lai*innen wurden insgesamt alle Pre-Editing-Regeln besser bewertet als kP. Bei genauerer Betrachtung zeigen sich aber auch einige Unterschiede zur Bewertung der Expert*innen. Von den sechs Pre-Editing-Regeln wurden 4 positiv, eine neutral und eine negativ bewertet. Die positiv bewerteten Regeln sind, wie bei den Expert*innen, Regel 1 (einfache Verben) mit 33 Gesamtpunkten, Regel 3 (Explizieren) mit 90 Punkten, Regel 4 (präzise Wortwahl) mit 62 Punkten und, konträr zur Bewertung der Expert*innen, Regel 6 (ing-Konstruktionen vermeiden) mit 9 Punkten. Wenngleich der Unterschied bei Regel 6 nicht allzu groß ist (-2 bei den Expert*innen und 9 bei den Lai*innen) zeigt sich ein weitaus deutlicherer Bewertungsunterschied bei der Bewertung von Pre-Editing-Regel 5 (klare pronominale Bezüge), die von den Lai*innen mit -38 Punkten deutlich negativ und von den Expert*innen mit 23 Punkten deutlich positiv bewertet wurde.

Auch bei der Bewertung von kP zeigt sich bei den Lai*innen auf individueller Ebene eine größere Heterogenität als bei den Expert*innen, da es, im Gegensatz zu den Expert*innen, wo kP von niemandem positiv bewertet wurde, bei den Lai*innen für kP drei positive Bewertungen gibt, eine davon sogar mit 10 Punkten deutlich positiv. Im Unterschied zu den Expert*innen gibt es bei den Lai*innen einerseits keine Pre-Editing-Regel, die ausschließlich neutrale oder positive Bewertungen erhielt und andererseits erhielt kP nicht die meisten Negativbewertungen, sondern Pre-Editing-Regel 5 mit 12 Negativbewertungen im Vergleich zu kP

mit 11. Es ist dennoch anzumerken, dass kP bei den Lai*innen trotz einer Negativbewertung weniger als Regel 5, die niedrigste Punktebewertung insgesamt erzielt (-65) und damit immerhin noch 27 Punkte hinter der am schlechtesten bewerteten Pre-Editing-Regel 5 (-38) liegt. Die Differenz zwischen kP und der am schlechtesten bewerteten Pre-Editing-Regel ist bei den Lai*innen somit deutlich geringer als bei den Translationsexpert*innen. Eine Gemeinsamkeit zwischen beiden Gruppen besteht darin, dass Pre-Editing-Regel 3 (Explizieren) in beiden Fällen am wenigsten häufig negativ bewertet wurde, nämlich bei den Expert*innen gar nicht und bei den Lai*innen lediglich einmal. Wie in Tabelle 14 veranschaulicht spiegelt auch die Gesamtbewertung über alle Teilnehmer*innen hinweg diesen Trend wider.

Tabelle 14: Aggregierte Gesamtbewertung aller Teilnehmer*innen

	kP	1	2	3	4	5	6
Expert*innen	-79	39	18	58	66	23	-2
Lai*innen	-65	33	0	90	62	-38	9
Gesamtpunkte	-144	72	18	148	128	-15	7

Die deutlich positivsten Bewertungen erzielen auch über alle Teilnehmer*innen gerechnet die Regeln 1 (einfache Verben) mit 72 Punkten, Regel 3 (Explizieren) mit 148 Punkten und Regel 4 (präzise Wortwahl) mit 128 Punkten. Auch die Einigkeit zwischen den Gruppen in Bezug auf kP spiegelt sich in der Gesamtnegativbewertung von -144 Punkten wider. Bei den zwei konträr bewerteten Regeln 5 und 6 zeigt sich, dass die weniger deutlich gegensätzlich bewertete Regel 6 auf alle Teilnehmer*innen gerechnet noch ein schwach positives Ergebnis (7) erzielt, die Pre-Editing-Regel 5, bei der zwischen beiden Gruppen die meiste Uneinigkeit herrscht, jedoch von allen Teilnehmer*innen gesamt mit -15 Punkten negativ bewertet wurde, obwohl sie bei den Expert*innen ein positives Ergebnis erzielte (23).

5.2.3 Validierungstests

Zur Validierung der verwendeten BWS-Methode und um festzustellen, ob die Teilnehmer*innen idente Übersetzungen konsistent bewerten, wurden den Teilnehmer*innen im Verlauf der Umfrage Textpassage 1 als Textpassage 8 und Textpassage 4 als Textpassage 12 doppelt zur Bewertung vorgelegt. Tabelle 15 präsentiert die Ergebnisse dieser Validierungstests, getrennt nach Expert*innen (E) und Lai*innen (L). Die hellgrüne Markierung bedeutet, dass der Validierungstest bestanden wurde, also in beiden Textpassagen die gleichen Übersetzungen als beste und schlechteste gewählt wurden. Die dunkelgrüne Markierung bedeutet, dass der Validierungstest darüber hinaus perfekt bestanden wurde und nicht nur die gleichen Übersetzungen gewählt, sondern auch die gleiche Punktezahl in beiden Textpassagen vergeben wurde.

Tabelle 15: Auswertung der Validierungstests

	Textprobe 1				Textprobe 8			
	kP	1	4	2	kP	1	4	2
E1	-2		3			2	0	
E2		1	-1			1		-1
E3		2		-1	1	0		
E4			-1	3			-1	2
E5		2	-2			2	-1	
E6		3	-2			1	-1	
E7	-1	3			-1	2		
E8		3	-1			1	-1	
E9			3	-1			3	-1
E10			3	-1			3	0
E11		4	-1			2		-1
E12		4	-1			2	-1	

	Textprobe 4				Textprobe 12			
	kP	3	6	5	kP	3	6	5
E1	-4			1	-3			1
E2			-1	1	1		-1	
E3	-2			2		0		0
E4	-3	2			-3	2		
E5	-1			3	-2			3
E6	-4			1	-1			1
E7	-4	1			-2	1		
E8	-2			2	-1			3
E9		3	0		-1	3		
E10	-3	1				1		0
E11	-4			3	-1	3		
E12	-4	2			-2		1	

	Textprobe 1				Textprobe 8			
	kP	1	4	2	kP	1	4	2
L1		3	-2			0		0
L2	2	-1			2			-1
L3		2		-3		4		-4
L4		3		-1		4		-2
L5		4		0		4	-1	
L6	1			-1		1	-1	
L7		3	0			4	-3	
L8	-4	3			-4	3		
L9		-4	4				4	-2
L10			-2	3			-3	3
L11	-2	3				-1		4
L12		3		-1		0	-1	
L13		3		-2	-1	1		
L14		3	-1		2			-1
L15		4	-2			3		-2

	Textprobe 4				Textprobe 12			
	kP	3	6	5	kP	3	6	5
L1	-3	2			-2	1		
L2		3	-2		-2	2		
L3			-4	2	-4			4
L4	-4			2	-3			3
L5	0		2		-2		3	
L6		-1		0	-1		1	
L7	4	-3				0	3	
L8	-4		3		-4			2
L9	-4			3	-3		3	
L10	-3		3		-3		3	
L11		4	-3			-2		3
L12	-2	2			-2	3		
L13	-3		3		1			-1
L14	-3	3			-2	2		
L15	-3			3	-2			3

Bei den Expert*innen bestanden zwei Drittel, nämlich acht von 12 Personen, den ersten Validierungstest. Eine Person sogar perfekt. Den zweiten Validierungstest bestanden immerhin noch 50 % der Expert*innen, nämlich sechs von 12 und auch hier wurde der Test von einem Experten bzw. einer Expertin perfekt bestanden. Knapp die Hälfte, nämlich fünf der 12 Expert*innen, bestanden beide Validierungstests. Bei den Lai*innen ist die Quote etwas geringer, aber immerhin ein Drittel, nämlich 5 von 15 Personen, hat den ersten Validierungstest bestanden, darunter, wie bei den Expert*innen, eine Person perfekt. Immerhin knapp unter 50 % (7 von 15) der Lai*innen bestanden den zweiten Test, und auch hier hatte eine Person ein perfektes Ergebnis. Nur zwei der insgesamt 15 Lai*innen, also knapp 14 %, bestanden beide Validierungstests.

5.2.4 Auswirkungen der einzelnen Pre-Editing-Regeln auf die Qualität

Auf Grundlage der Bewertungen der Expert*innen und Lai*innen ergibt sich das in Tabelle 16 dargestellte Ranking der einzelnen Pre-Editing-Regeln, differenziert nach Gruppen sowie als Gesamtergebnis.

Tabelle 16: Ranking der Pre-Editing-Regeln nach Gruppen und Gesamt

	Expert*innen		Lai*innen		Gesamt	
	Regel	Punkte	Regel	Punkte	Regel	Punkte
1.	4	66	3	90	3	148
2.	3	58	4	62	4	128
3.	1	39	1	33	1	72
4.	5	23	6	9	2	18
5.	2	18	2	0	6	7
6.	6	-2	5	-38	5	-15
7.	kP	-79	kP	-65	kP	-144

Auffällig ist, dass kein Pre-Editing in den Bewertungen beider Gruppen den letzten Platz einnimmt. Da sämtliche mithilfe von Pre-Editing-Regeln überarbeiteten Textpassagen von beiden Gruppen durchwegs besser bewertet wurden als die entsprechenden Rohtextpassagen, zeigt sich, dass alle untersuchten Pre-Editing-Regeln zu einer Verbesserung der Übersetzungsqualität sowohl für Expert*innen als auch für Lai*innen führen.

Die deutlichsten Verbesserungen lassen sich bei den Pre-Editing-Regeln 3 (Explizieren), Regel 4 (präzise Wortwahl) und Regel 1 (einfache Verben) beobachten, die sowohl von Expert*innen als auch von Lai*innen besonders positiv bewertet wurden. Darüber hinaus zeigt sich eine ausgeprägte Differenz zwischen der am schlechtesten bewerteten Pre-Editing-Regel 5 (klare pronominale Bezüge) mit -15 Punkten und kein Pre-Editing, die mit -144 Punkten bewertet wurde. Damit erzielt kP ein um signifikante 129 Punkte schlechteres Ergebnis als die insgesamt am schlechtesten bewertete Pre-Editing-Regel 5.

5.2.5 Qualitative Analyse der optionalen Fragen zur Qualitätswahrnehmung

Von den 27 Teilnehmer*innen haben 10 Personen (3 Expert*innen und 7 Lai*innen) bei mindestens einer Textpassage eine Anmerkung zu den Übersetzungsvarianten bzw. eine Begründung für ihre Entscheidungen angegeben. Das Ziel dieser Frage ist, die Beweggründe der Umfrageteilnehmer*innen für die Auswahl der jeweils besten und schlechtesten Übersetzungsversionen zu analysieren und dadurch ein ganzheitlicheres Bild über die subjektive Qualitätswahrnehmung der Teilnehmer*innen zu erhalten. Die qualitative Analyse bietet somit einen

differenzierten Einblick in die sprachlichen, inhaltlichen und stilistischen Kriterien, anhand derer die Teilnehmer*innen die Qualität maschineller Übersetzungen bewerten.

Sowohl Translationsexpert*innen als auch Lai*innen legten großen Wert auf ähnliche Kerndimensionen der Übersetzungsqualität. Ein zentrales Kriterium für beide Gruppen ist die grammatikalische Korrektheit der Übersetzung. Fehler, wie falsche pronominale Bezüge oder fehlerhafte Kasuskonstruktionen wurden von Expert*innen und Lai*innen gleichermaßen als qualitätsvermindernd beschrieben, da sie einen unmittelbaren Einfluss auf die Klarheit, Verständlichkeit und Kohärenz der Texte haben. Ein weiteres stark qualitätsrelevantes Kernkriterium von beiden Gruppen betrifft die Idiomatizität und Natürlichkeit der Übersetzung sowie die Vermeidung lexikalischer Fehlgriffe. Expert*innen wie Lai*innen kritisierten Übersetzungsversionen, die aufgrund einer zu wörtlichen Übernahme aus dem Englischen im Deutschen unnatürlich klingen, darunter Formulierungen wie *windhungrig* oder *Nicht-Null-Auswirkungen* und unpassende Worte, z. B. Anglizismen wie *hostet*. Auch die strukturelle Lesbarkeit spielte für beide Gruppen eine wesentliche Rolle, so wurden lange, Schachtelsätze und ungünstige Gliedsatzkonstruktionen von Expert*innen und Lai*innen negativ bewertet, da sie den Lesefluss zum Teil erheblich störten. Insgesamt betrachtet ist die Qualitätsauffassung in beiden Gruppen multidimensional, geht also über die reine inhaltliche Richtigkeit hinaus und betrifft auch stilistische, strukturelle und sprachliche Aspekte.

Trotz vieler Gemeinsamkeiten zwischen Expert*innen und Lai*innen bei den generellen qualitätsrelevanten Kriterien, zeigen sich interessante Unterschiede in Akzentsetzungen und der Tiefe der Analysen. Die Translationsexpert*innen richteten ihren Fokus stärker auf linguistische Details und wiesen eine hohe Sensibilität für nuancierte sprachliche Probleme auf, insbesondere im Hinblick auf grammatikalische und referentielle Genauigkeit. Sie analysierten die Referenzen, Gliedsätze und Satz- und Pronominalbezüge detaillierter als Lai*innen und deduzierten, wie bestimmte Formulierungen und Gliedsatzkonstruktionen die Satzlogik beeinflussen. Zudem äußerten die Expert*innen insgesamt tiefergehende Reflexionen zur Verwendung von Metaphern und ihrer fachsprachlichen Angemessenheit. So wurde beispielsweise von einem Experten bzw. einer Expertin explizit erwähnt, dass die metaphorische Formulierung *das Gehirn* zwar auch angemessen sei und funktioniert, aber die Entscheidung schlussendlich auf die allgemeinere Beschreibung *zentrale Komponente* fiel, obwohl die Übersetzung weniger nah am englischen Original ist, aber im fachlichen Kontext die allgemeinere Formulierung ansprechender war. Ein Laie bzw. eine Laiin hat sich im kompletten Gegensatz dazu explizit für die Wahl einer Übersetzung mit *das Gehirn* ausgesprochen, weil es näher am Original ist.

Insgesamt spielte die Originaltreue für Lai*innen eine wichtigere Rolle bei der Beurteilung der Übersetzungsqualität als für Expert*innen.

Auch in der Tiefe der Ausführungen bzw. der Entscheidungsbegründungen unterscheiden sich die beiden Gruppen. Bis auf einen Laien bzw. eine Laiin, der*die ähnlich detaillierte und differenzierte Begründungen zur Wahl der besten und schlechtesten Übersetzungen angab, waren die Anmerkungen der Lai*innen im Allgemeinen weniger systematisch und analytisch, sondern spiegelten eher einen intuitiven Gesamteindruck in Form von *alle gleich schlecht/gut*, *liest sich holprig*, *klingt komisch* oder *macht keinen Sinn*. Auffällig ist, dass beide Gruppen erwähnten, dass zeitweise keine der Übersetzungen vollständig überzeugte und sie folglich in diesen Fällen einfach die am wenigsten schlechte Übersetzung als beste Variante wählten. Ein Laie bzw. eine Laiin gab in der abschließend Anmerkung zur Umfrage an, dass die Umfrage interessant war, eine andere Person erwähnte, dass es ihn bzw. sie gefreut hat. Ein Experte bzw. eine Expertin merkte an, dass er bzw. sie etwa bei der Hälfte der Umfrage die Konzentration für einzelne Begründungen verlor und es vor allem mit den Wiederholungen für die Validierungstests einen sehr hohen Zeit- und Konzentrationsaufwand darstellte so viele Textpassagen zu bewerten und detailliert zu begründen.

6 Diskussion

Ziel dieser Masterarbeit war es, durch die gezielte Anwendung von verschiedenen Pre-Editing-Regeln, die auf Basis einer Fehlertypologie spezifisch für den Fachbereich Astronautik sowie die Sprachkombination Englisch-Deutsch ausgewählt wurden, eine Qualitätsverbesserung der Ergebnisse neuronaler maschineller Übersetzung zu erzielen, indem eine Reduktion beziehungsweise Eliminierung identifizierter Übersetzungsfehler angestrebt wurde. Zu diesem Zweck wurde ein experimentelles Design entwickelt, das eine gleichmäßige und ausgewogene Verteilung von sechs ausgewählten Pre-Editing-Regeln sicherstellte. Jede Rohausgangstextpassage wurde dabei drei präeditierten Versionen gegenübergestellt, wobei unterschiedliche Gruppierungen aus je drei Regeln angewendet wurden. Die Regeln wurden systematisch und fair in verschiedenen Sets verglichen, um Verzerrungen zu vermeiden und die Vergleichbarkeit der Ergebnisse zu gewährleisten. Das Design wurde den Teilnehmer*innen im Rahmen einer Best-Worst-Scaling-Studie zur Bewertung vorgelegt. Zur Erhöhung der Validität sowie zur Überprüfung der Konsistenz der Entscheidungen wurden zwei Textpassagen in der Umfrage wiederholt.

Die Ergebnisse zeigen deutlich, dass Pre-Editing im Fachbereich Astronautik und der Sprachkombination Englisch-Deutsch sowohl für Expert*innen als auch für Lai*innen zu einer deutlich wahrnehmbaren Verbesserung der Qualität neuronaler maschineller Übersetzung führt. Das unterstreicht die Wirksamkeit des gewählten experimentellen Designs sowie dessen Umsetzung im Hinblick auf die Beantwortung der Forschungsfrage. Obwohl alle sechs Pre-Editing-Regeln im Fachbereich Astronautik und der Sprachkombination Englisch-Deutsch in beiden Gruppen zu einer wahrnehmbaren Qualitätsverbesserung gegenüber der uneditierten Textpassagen geführt haben, konnten insbesondere Regel 3 (Explizieren), Regel 4 (präzise Wortwahl) und Regel 1 (einfache Verben) konsistent hohe Bewertungen in beiden Gruppen erzielen. Obwohl eine gewisse Verbesserung der Übersetzungsqualität durch den Einsatz von Pre-Editing-Regeln erwartet wurde, ist das vergleichsweise schlechte Abschneiden sowie der erhebliche Abstand zwischen der am schlechtesten bewerteten Pre-Editing-Regel und kP überraschend. Dies deutet darauf hin, dass die unbearbeiteten Passagen von beiden Gruppen, insbesondere aber von den Expert*innen, die sie insgesamt tendenziell kritischer bewerteten als die Lai*innen, als qualitativ deutlich unzureichend wahrgenommen wurden.

Obwohl sich die zentralen qualitätsrelevanten Kriterien in beiden Gruppen weitestgehend überschnitten, unterschieden sich deren Begründungen insbesondere in der Akzentsetzung und der Tiefe der sprachlichen Analyse. Diese Unterschiede, sowohl in der allgemeinen

Bewertung als auch in der Argumentation, deuten darauf hin, dass ein translationswissenschaftlicher Ausbildungshintergrund möglicherweise mit einer bewussteren und kritischeren Wahrnehmung neuronale maschineller Übersetzungen einhergeht. Da die Differenzen jedoch insgesamt gering ausfielen, kann von einem grundsätzlich geteilten Qualitätsverständnis zwischen Expert*innen und Lai*innen, insbesondere in Bezug auf qualitätsrelevante Kriterien, ausgegangen werden, das sich auch in der qualitativen Analyse sowie in den quantitativen Ergebnissen größtenteils widerspiegelt. Besonders bemerkenswert ist, dass sowohl von Expert*innen als auch von Lai*innen in ihren Entscheidungsbegründungen explizit auf korrekte pronominale Bezüge als qualitätsrelevanten Aspekt hinweisen wurde, die dazugehörige Pre-Editing-Regel 5 (klare pronominale Bezüge) allerdings insgesamt über alle Teilnehmer*innen, und besonders in der Gruppe der Lai*innen, im Vergleich mit allen Pre-Editing-Regeln das schlechteste Ergebnis erzielte und eine Grundlage für eine weiterführende Erforschung dieses Widerspruchs bietet.

Die Ergebnisse diese Masterarbeit sind weitestgehend konsistent mit Resultaten früherer, vergleichbarer Studien, die ebenso eine Qualitätsverbesserung neuronaler maschineller Übersetzungen durch Pre-Editing feststellen konnten. Bemerkenswert ist dabei nicht nur die grundsätzliche Übereinstimmung hinsichtlich der positiven Auswirkung von Pre-Editing im Allgemeinen, sondern auch die inhaltliche Parallele zu Finetti (2020), die im Fachbereich Biolumineszenz und der Sprachkombination Englisch-Italienisch die Verwendung einfacher Verben als besonders qualitätsförderlich identifizierte. Eine vergleichbare Tendenz ließ sich mit der konstant hohen Bewertung beider Gruppen von Regel 1 (einfache Verben) in der vorliegenden Arbeit für den Fachbereich Astronautik und die Sprachkombination Englisch-Deutsch beobachten. Darüber hinaus zeigen sich hinsichtlich Regel 3 (Explizieren) und Regel 4 (präzise Wortwahl) ähnliche Effekte wie in der Studie von Göschl (2022), die im Fachbereich Elektromobilität und in der entgegengesetzten Sprachrichtung Deutsch-Englisch durchgeführt wurde. Beide Regeln trugen in beiden Arbeiten zur Verbesserung der Übersetzungsqualität bei. Eine signifikante Abweichung ergibt sich hingegen bei Regel 5 (klare pronominale Bezüge). Während diese im Bereich Elektromobilität zu einer deutlichen Qualitätssteigerung führte, erzielte sie im Bereich Astronautik auf alle Teilnehmer*innen bezogen das schwächste Ergebnis unter allen sechs getesteten Pre-Editing-Regeln.

Der Validierungstest zeigt, dass die Methode nicht vollständig zuverlässig ist, da die subjektiven Entscheidungen großteils nicht konsistent getroffen werden. Somit zeigen sich Schwierigkeiten der Wiederholbarkeit auf, wenn auch diese Abweichungen eventuell keine

Auswirkung auf die Gesamtergebnisse bei einer wiederholten Durchführung haben könnten. Somit ist die Validität der Methode nicht gänzlich eingeschränkt. Die besseren Ergebnisse der Translationsexpert*innen in den Validierungstests legen nahe, dass ein translationswissenschaftlicher Ausbildungshintergrund möglicherweise mit einer tiefergehenderen und systematischeren sprachlichen Analyse der Übersetzungen einhergeht und damit konsistenteren Bewertung unter gleichen Bedingungen ermöglicht. Diese Deduktion wird auch durch die Ergebnisse der qualitativen Analyse und den systematischeren und tieferen Analysen und Entscheidungsbegründungen der Translationsexpert*innen bestätigt. Das Ergebnis des Validierungstests unterstreicht die Bedeutung der Wahl einer adäquaten Evaluierungsmethode für einen so subjektiven Untersuchungsgegenstand wie maschinelle Übersetzungsqualität.

Die Anwendung von BWS bietet in diesem Zusammenhang einige wesentliche Vorteile. So erfordert das Verfahren beispielsweise keine Fehlerklassifikation, sondern ermöglicht eine intuitive Bewertung auf Basis der unmittelbaren Leseerfahrung, die insbesondere auch Personen ohne translationswissenschaftliche Ausbildung eine einfache Bewertung ermöglicht. Darüber hinaus zeichnet sich BWS durch eine einfache Datenerhebung und -auswertung aus. Die Kombination des Verfahrens mit einer Intensitätsskala von 0 bis +4 bzw. 0 bis -4 sowie der optionalen Möglichkeit zur Begründung der Entscheidungen erweitert die reine Präferenzentscheidung um zusätzliche quantitative und qualitative Analysedimensionen.

Es gibt jedoch auch einige Limitationen. So erwies sich die technische Umsetzung der BWS-Struktur mit gängigen Umfragetools als herausfordernd, da diese Methode nicht als integrierte Funktion zur Verfügung steht. Um ein Szenario zu realisieren, in dem Teilnehmer*innen keine Eingabefehler machen können, waren JavaScript-Kenntnisse erforderlich, um die BWS-Struktur entsprechend zu implementieren. Zudem wurde in dieser Studie keine Randomisierung der Präsentationsreihenfolge der Übersetzungsvarianten vorgenommen. Stattdessen wurde eine einmalig zufällig festgelegte Reihung allen Teilnehmenden in identischer Form vorgelegt. Dies könnte potenziell Reihungseffekte begünstigt und somit die Entscheidungen beeinflusst haben. Eine weitere Einschränkung stellt die relativ geringe Stichprobengröße von 27 Personen dar, auf deren Basis keine generalisierbaren Aussagen über die Wirksamkeit der Pre-Editing-Regeln getroffen werden können.

7 Fazit

In dieser Masterarbeit wurde die Auswirkung verschiedener Pre-Editing-Regeln auf die Qualität maschineller Übersetzung mithilfe von BWS im Fachbereich Astronautik und der Sprachkombination Englisch-Deutsch untersucht. Die Ergebnisse der Untersuchung zeigen deutlich, dass Pre-Editing ein wirksames Instrument zur Steigerung der Qualität maschineller Übersetzungen ist, insbesondere wenn die Änderungen auf Klarheit, Einfachheit und präzisen Sprachgebrauch abzielen. Besonders wirkungsvoll erwiesen sich dabei Maßnahmen wie das Explizieren von problematischen Ausdrücken, Metaphern und Ellipsen sowie die Verwendung präziser Fachbegriffe und einfacher Verben, die als konkrete Optimierungsstrategien im professionellen, wissenschaftlichen sowie im privaten Gebrauch dienen können. Die teilweise unterschiedlichen Bewertungen durch Translationsexpert*innen und Lai*innen insbesondere in Bezug auf pronominale Bezüge und die Tiefe der Entscheidungsbegründungen machen zudem deutlich, dass die Wahrnehmung von Übersetzungsqualität mitunter auch von der jeweiligen sprachlichen Vorbildung und Expertise beeinflusst wird. Insgesamt liefert die Masterarbeit nicht nur wertvolle Erkenntnisse zu effektiven Pre-Editing-Strategien, sondern auch zum Qualitätsverständnis von maschineller Übersetzung bei Translationsexpert*innen und Lai*innen im Fachbereich Astronautik und der Sprachkombination Englisch-Deutsch.

Die Ergebnisse dieser Masterarbeit liefern eine Reihe an Möglichkeiten für weitere Untersuchungen in diesem Bereich. Der Widerspruch in Bezug auf pronominale Bezüge, sowohl zwischen den Gruppen von Expert*innen und Lai*innen im Zuge dieser Arbeit, als auch fach- und sprachübergreifend mit vergleichbaren Arbeiten, legt nahe, dass die Rolle pronominaler Bezüge im Pre-Editing-Prozess in zukünftigen Studien vertieft untersucht werden sollte. Obwohl die vorliegende Masterarbeit, insbesondere im Zusammenspiel mit früheren, vergleichbaren Studien, wertvolle Einblicke in potenzielle Überschneidungen qualitätsfördernder Pre-Editing-Regeln liefert, erscheint es erforderlich, weiterführende Untersuchungen in unterschiedlichen Fachbereichen und Sprachkombinationen durchzuführen. Ziel solcher Studien sollte es sein, die Generalisierbarkeit der identifizierten Regeln zu überprüfen, um mögliche sprach- oder fachbereichsunabhängige Muster zu identifizieren.

Darüber hinaus könnte eine Untersuchung der Wahrnehmung von Übersetzungsqualität innerhalb eines spezifischen Fachbereichs durch Translationsexpert*innen einerseits und Fachexpert*innen dieses Fachbereichs andererseits weiterführende Erkenntnisse über divergierende Qualitätskriterien und Bewertungsperspektiven liefern. Ergänzend wäre eine groß angelegte Studie von Interesse, in der Pre-Editing-Regeln entweder als Plug-in oder im Rahmen des

Trainings maschineller Übersetzungsmodelle in ein öffentlich zugängliches NMÜ-System integriert werden. Durch breit angelegtes, sprach-, fach- und kulturunabhängiges Feedback aus der allgemeinen Bevölkerung ließen sich daraus weiterführende Aussagen über die Wirksamkeit der implementierten Regeln im Hinblick auf die wahrgenommene Übersetzungsqualität ableiten.

Bibliografie

Agarwal, Abhaya & Lavie, Alon (2008). METEOR, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In: *Proceedings of the Third Workshop on Statistical Machine Translation*. 115-118.

Aggarwal, Charu C. (2018). *Neural Networks and Deep Learning: A Textbook*. New York: Springer.

Almár, Iván & Both, Elöd (2002). THE DELIMITATION PROBLEM - SELECTING THE BASIC LIST OF TERMS FOR AN ASTRONAUTICAL DICTIONARY. *Acta Astronautica*, 50 (2), 83-87.

ALPAC (1966). *Languages and Machines: Computers in Translation and Linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington D.C.: National Academy of Sciences, National Research Council (1416).

Angelo, Joseph A. (2006). *Encyclopedia of space and astronomy*. New York: Facts on File.

Arif, Samee; Khan, Aamina Jamal; Abbas, Mustafa; Raza, Agha Ali & Athar, Awais (2025). WER We Stand: Benchmarking Urdu ASR Models. In: *Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE: Association for Computational Linguistics*. 5952-5961.

Arnold, Doug (1994). *Machine Translation: An Introductory Guide*. 1. Aufl. Manchester: NCC Blackwell.

Avramidis, Eleftherios (2019). *Comparative Quality Estimation for Machine Translation*. Dissertation. Universität des Saarlandes.

Bahdanau, Dzmitry; Cho, Kyunghyun & Bengio, Yoshua (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In: *International Conference on Learning Representations (ICLR 2015)*. 1-15.

Banerjee, Satanjeev & Lavie, Alon (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: *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.

Bañón, Marta; Chen, Pinzhen; Haddow, Barry; Heafield, Kenneth; Hoang, Hieu; Esplà-Gomis, Miquel; Forcada, Mikel L.; Kamran, Amir; Kirefu, Faheem; Koehn, Philipp; Ortiz

Rojas, Sergio; Pla Sempere, Leopoldo; Ramírez-Sánchez, Gema; Sarrías, Elsa; Strelec, Marek; Thompson, Brian; Waites, William; Wiggins, Dion & Zaragoza, Jaume (2020). ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics. 4555-4567.

Bennett, Scott & Gerber, Laurie (2003). Inside commercial machine translation. In: Somers, Harold (Hrsg.) *Computers and Translation. A translator's guide*. Amsterdam/Philadelphia: John Benjamins Publishing Company. 175-190.

Bernth, Arendse & Gdaniec, Claudia (2001). MTranslatability. *Machine Translation*, 16 (3), 175-218.

Bruhn, Manfred (2020). *Qualitätsmanagement für Dienstleistungen: Handbuch für ein erfolgreiches Qualitätsmanagement. Grundlagen - Konzepte - Methoden*. 12. Aufl. Berlin Heidelberg: Springer.

Burchardt, Aljoscha & Porsiel, Jörg (2017). Vorwort: Was kann maschinelle Übersetzung und was nicht? In: Porsiel, Jörg (Hrsg.) *Maschinelle Übersetzung: Grundlagen für den professionellen Einsatz*. Berlin: BDÜ Weiterbildungs- und Fachverlags GmbH. 11-18.

Callison-Burch, Chris; Osborne, Miles & Koehn, Philipp (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy: Association for Computational Linguistics. 249-256.

Campbell, Paul A. (1962). History and Background of Astronautics. In: *Lectures in Aerospace Medicine*. 7-20.

Carstensen, Kai-Uwe; Ebert, Christian; Ebert, Cornelia; Jekat, Susanne; Langer, Hagen & Klabunde, Ralf (2010). *Computerlinguistik und Sprachtechnologie: Eine Einführung*. 3. Aufl. Heidelberg: Spektrum Akademischer Verlag.

Castilho, Sheila; Doherty, Stephen; Gaspari, Federico & Moorkens, Joss (2018). Approaches to Human and Machine Translation Quality Assessment. In: Moorkens, Joss; Castilho, Sheila; Gaspari, Federico & Doherty, Stephen (Hrsg.) *Translation Quality Assessment: From Principles to Practice*. Cham: Springer International Publishing. 9-38.

Castilho, Sheila; Moorkens, Joss; Gaspari, Federico; Sennrich, Rico; Sosoni, Vilelmini; Georgakopoulou, Panayota; Lohar, Pintu; Way, Andy; Miceli-Barone, Antonio Valerio & Gialama, Maria (2017). A comparative quality evaluation of PBSMT and NMT using

professional translators. In: *Proceedings of Machine Translation Summit XVI: Research Track*. Nagoya, Japan. 116-131.

Caswell, Isaac & Bowen, Liang (2020). Recent Advances in Google Translate. https://research.google/blog/recent-advances-in-google-translate/ (Stand: 10.05.2025).

Chan, Sin-wai (2018). *The Human Factor in Machine Translation*. London: Routledge.

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

Clifford, Andrew (2001). Discourse Theory and Performance-Based Assessment: Two Tools for Professional Interpreting. *Meta*, 46 (2), 365-378.

DeepL (2021). How does DeepL work? https://www.deepl.com/en/blog/how-does-deepl-work (Stand: 16.05.2025).

DeepL SE (2025). About DeepL. https://www.deepl.com/en/publisher (Stand: 16.05.2025).

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton & Toutanova, Kristina (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics. 4171-4186.

Doherty, Stephen (2017). Issues in Human and Automatic Translation Quality Assessment. In: Kenny, Dorothy (Hrsg.) *Human Issues in Translation Technology*. London: Routledge. 131-148.

Dorr, Bonnie; Olive, Joseph; McCary, John & Christianson, Caitlin (2011). Machine Translation Evaluation and Optimization. In: Olive, Joseph; Christianson, Caitlin & McCary, John (Hrsg.) *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. New York: Springer. 745-843.

Drugan, Joanna (2013). *Quality In Professional Translation: Assessment and Improvement*. London: Bloomsbury Academic.

Dunne, Keiran J. (2006). Putting the cart behind the horse: Rethinking localization quality management. In: Dunne, Keiran J. (Hrsg.) *Perspectives on Localization*. Philadelphia: John Benjamins Publishing Company. 95-117.

Europarat (2025). Common European Framework of Reference for Languages (CEFR). https://www.coe.int/en/web/common-european-framework-reference-languages (Stand: 04.05.2025).

European Space Agency (2019). ESA Highlights 2019. https://esamultimedia.esa.int/docs/corporate/ESA_Highlights_2019_LR.pdf (Stand: 12.05.2025).

European Space Agency (2023). ESA Highlights 2023. https://esamultimedia.esa.int/docs/corporate/ESA_Highlights_2023.pdf (Stand: 12.05.2025).

European Space Agency (2024). ESA Highlights 2024. https://esamultimedia.esa.int/docs/corporate/ESA_Highlights_2024.pdf (Stand: 27.04.2025).

European Space Agency (2025). ESA Highlights. https://www.esa.int/About_Us/ESA_Publications/ESA_Highlights (Stand: 01.05.2025).

Farrús, Mireia; Ruiz Costa-Jussà, Marta; Mariño Acebal, José Bernardo & Rodríguez Fonollosa, José Adrián (2010). Linguistic-based Evaluation Criteria to identify Statistical Machine Translation Errors. In: *14th Annual Conference of the European Association for Machine Translation*. Saint Raphaël, France: European Association for Machine Translation.

Finetti, Fiorenza (2020). *Pre-Editing als Vorschlag zur Verbesserung Neuronaler Maschineller Übersetzung*. Masterarbeit. Universität Wien.

Flanagan, Mary (1994). Error Classification for MT Evaluation. In: *Proceedings of the First Conference of the Association for Machine Translation in the Americas*. Columbia, Maryland, USA. 65-72.

Forcada, Mikel L. (2010). Machine Translation Today. In: Gambier, Yves & van Doorslaer, Luc (Hrsg.) *Handbook of Translation Studies*. John Benjamins Publishing Company. 215-223.

Forcada, Mikel L. (2017). Making sense of neural machine translation. *Translation Spaces*, 6 (2), 291-309.

Gabriel, Peter (2019). Einleitung: KI ohne Grenzen? In: Wittpahl, Volker (Hrsg.) *Künstliche Intelligenz: Technologie / Anwendung / Gesellschaft*. Berlin, Heidelberg: Springer. 95-98.

Gingerich, Owen (1986). Islamic astronomy. *Scientific American*, 254 (4), 74-83.

Görz, Günther; Schmid, Ute & Braun, Tanya (2020). *Handbuch der Künstlichen Intelligenz*. Berlin, Boston: De Gruyter Oldenbourg.

Göschl, Janina (2022). *Die Auswirkung von Pre-Editing-Methoden auf die Translationsqualität von maschinell übersetzten Texten und deren Einschätzung durch Sprachexpert*innen im Vergleich zu Lai*innen*. Masterarbeit. Universität Wien.

Guerberof Arenas, Ana (2019). Pre-editing and post-editing. In: Angelone, Erik; Ehrensberger-Dow, Maureen & Massey, Gary (Hrsg.) *The Bloomsbury Companion to Language Industry Studies*. London: Bloomsbury Academic. 333-360.

Guo, Junliang; Zhang, Zhirui; Xu, Linli; Chen, Boxing & Chen, Enhong (2021). Adaptive Adapters: An Efficient Way to Incorporate BERT Into Neural Machine Translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 1740-1751.

Hanslmeier, Arnold (2020). *Einführung in Astronomie und Astrophysik*. 4. Aufl. Berlin, Heidelberg: Springer.

Hess, Stephane; Daly, Andrew & Batley, Richard (2018). Revisiting consistency with random utility maximisation: theory and implications for practical work. *Theory and Decision,* 84 (2), 181-204.

Hiebl, Bettina & Gromann, Dagmar (2024). Comparative Quality Assessment of Human and Machine Translation with Best-Worst Scaling. In: *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.

Hiraoka, Yusuke & Yamada, Masaru (2019). Pre-editing Plus Neural Machine Translation for Subtitling: Effective Pre-editing Rules for Subtitling of TED Talks. In: *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*. Dublin, Ireland: European Association for Machine Translation. 64-72.

House, Juliane (2014). *Translation quality assessment: Past and present*. New York: Routledge.

Hutchins, John W. (1995). Machine translation: A brief history. In: Koerner, E. F. K. & Asher, R. E. (Hrsg.) *Concise history of the language sciences*. Amsterdam: Pergamon. 431-444.

Hutchins, John W. (2000). The first decades of machine translation: overview, chronology, sources. In: Hutchins, John W. (Hrsg.) *Early years in machine translation: memoirs and biographies of pioneers*. Amsterdam: John Benjamins Publishing Co. 1-15.

Hutchins, John W. & Somers, Harold L. (1992). *An Introduction to Machine Translation*. London: Academic Press.

Jaiswal, Ashish; Babu, Ashwin Ramesh; Zadeh, Mohammad Zaki; Banerjee, Debapriya & Makedon, Fillia (2021). A Survey on Contrastive Self-Supervised Learning. *Technologies*, 9 (1).

Jiménez-Crespo, Miguel A. (2015). The Evaluation of Pragmatic and Functionalist Aspects in Localization: Towards a Holistic Approach to Quality Assurance. *The Journal of Internationalization and Localization*, 1 (1) 60-93.

Jiménez-Crespo, Miguel A. (2018). Crowdsourcing and Translation Quality: Novel Approaches in the Language Industry and Translation Studies. In: Moorkens, Joss; Castilho, Sheila; Gaspari, Federico & Doherty, Stephen (Hrsg.) *Translation Quality Assessment: From Principles to Practice*. Cham: Springer International Publishing. 69-93.

Jordan, Michael P. (1999). Using acronyms in technical writing. *Discourse and Writing/Rédactologie*, 15 (1), 54-60.

Jurafsky, Dan & Martin, James H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3. Aufl. https://web.stanford.edu/~jurafsky/slp3/.

Kardakis, Spyridon; Perikos, Isidoros; Grivokostopoulou, Foteini & Hatzilygeroudis, Ioannis (2021). Examining Attention Mechanisms in Deep Learning Models for Sentiment Snalysis. *Applied Sciences*, 11 (9).

Kelleher, John D. (2019). *Deep Learning*. Cambridge, Massachusetts: MIT press.

Kenny, Dorothy (2019). Machine Translation. In: Wilson, Philip & Rawling, Piers (Hrsg.) *The Routledge Handbook of Translation and Philosophy*. London: Routledge. 428-445.

Kenny, Dorothy (2020). Machine Translation. In: Baker, Mona & Saldanha, Gabriela (Hrsg.) *The Routledge Encyclopedia of Translation Studies*. 3. Aufl. London, New York: Routledge. 305-310.

Kiritchenko, Svetlana & Mohammad, Saif (2017). Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics. 465-470.

Koby, Geoffrey S.; Fields, Paul; Hague, Daryl R.; Lommel, Arle & Melby, Alan (2014). Defining Translation Quality. *Tradumàtica*, 12, 413-420.

Koehn, Philipp (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In: *Proceedings of Machine Translation Summit X: Papers*. Phuket, Thailand. 79-86.

Koehn, Philipp (2010). *Statistical Machine Translation*. Cambridge, New York: Cambridge University Press.

Koehn, Philipp (2020). *Neural Machine Translation*. Cambridge: Cambridge University Press.

Krenz, Michael & Ramlow, Markus (2008). *Maschinelle Übersetzung und XML im Übersetzungsprozess: Prozesse der Translation und Lokalisierung im Wandel*. Berlin: Frank & Timme.

Krucien, Nicolas; Watson, Verity & Ryan, Mandy (2017). Is Best-Worst Scaling Suitable for Health State Valuation? A Comparison with Discrete Choice Experiments. *Health Economics*, 26 (12), e1-e16.

Kudo, Taku & Richardson, John (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*: *System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics. 66-71.

Lavie, Alon; Sagae, Kenji & Jayaraman, Shyamsundar (2004). The Significance of Recall in Automatic Metrics for MT Evaluation. In: Frederking, Robert E. & Taylor, Kathryn B. (Hrsg.) *Machine Translation: From Real Users to Research*. Berlin, Heidelberg: Springer. 134-143.

Levenshtein, Vladimir I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10 (8), 707-710.

Lewis-Kraus, Gideon (2016). The great AI awakening. *The New York Times Magazine*, 14 (12).

Li, Xiuhua & Li, Li (2015). Characteristics of English for Science and Technology. In: *2015 International Conference on Humanities and Social Science Research*. Atlantis Press. 161-165.

LimeSurvey GmbH (2025). LimeSurvey - Free Online Survey Tool. https://www.limesurvey.org/ (Stand: 04.05.2025).

Limkonchotiwat, Peerat; Ponwitayarat, Wuttikorn; Udomcharoenchaikit, Can; Chuangsuwanich, Ekapol & Nutanong, Sarana (2022). CL-ReLKT: Cross-lingual Language Knowledge Transfer for Multilingual Retrieval Question Answering. In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics. 2141-2155.

Lommel, Arle (2018). Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In: Moorkens, Joss; Castilho, Sheila; Gaspari, Federico & Doherty, Stephen (Hrsg.) *Translation Quality Assessment: From Principles to Practice*. Cham: Springer International Publishing. 109-127.

Lommel, Arle; Burchardt, Aljoscha & Uszkoreit, Hans (2015). *Multidimensional Quality Metrics (MQM) Definition*. Deutsches Forschungszentrum für Künstliche Intelligenz

GmbH. https://web.archive.org/web/20211216145215/https:/www.qt21.eu/mqm-definition/issues-list-2015-12-30.html (Stand: 27.04.2025).

Lommel, Arle; Uszkoreit, Hans & Burchardt, Aljoscha (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, 12, 455-463.

Louviere, Jordan J.; Flynn, Terry N. & Marley, A. A. J. (2015). *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge: Cambridge University Press.

Louviere, Jordan J. & Woodworth, George G. (1990). *Best worst scaling: A model for largest difference judgments*. Working Paper. Faculty of Business, University of Alberta.

Luong, Minh-Thang (2016). *Neural Machine Translation*. Dissertation. Stanford University.

Mateo, Roberto Martínez (2014). A deeper look into metrics for translation quality assessment (TQA): A case study. *Miscelánea - Departamento de Filología Inglesa y Alemana, Universidad de Zaragoza*, 49 (49), 73-93.

Mathangi, Ragmadura (2020). Was sind NLP und NLG und wie funktionieren sie? In: Kabel, Peter (Hrsg.) *Dialog zwischen Mensch und Maschine: Conversational User Interfaces, intelligente Assistenten und Voice-Systeme*. Wiesbaden: Springer Fachmedien. 39-63.

Maučec, Mirjam Sepesy & Donaj, Gregor (2020). Machine translation and the Evaluation of its Quality. In: Sadollah, Ali & Shishir Sinha, Tilendra (Hrsg.) *Recent Trends in Computational Intelligence*. London: IntechOpen. 143-162.

Mayer, Martina (2013). *Sprachpflege und Sprachnormierung in Frankreich am Beispiel der Fachsprachen vom 16. Jahrhundert bis in die Gegenwart*. Innsbruck University Press.

Mercader-Alarcón, Julia & Sánchez-Matínez, Felipe (2016). Analysis of translation errors and evaluation of pre-editing rules for the translation of English news texts into Spanish with Lucy LT. *Tradumàtica*, 14, 172-186.

Miyata, Rei & Fujita, Atsushi (2017). Dissecting human pre-editing toward better use of off-the-shelf machine translation systems. In: *Proceedings of the 20th Annual Conference of the European Association for Machine Translation (EAMT)*. 54-59.

Mügge, Uwe (2002). Möglichkeiten für das Realisieren einer einfachen Kontrollierten Sprache. *Lebende Sprachen*, 47 (3), 110-114.

Munkova, Dasa; Hajek, Petr; Munk, Michal & Skalka, Jan (2020). Evaluation of Machine Translation Quality through the Metrics of Error Rate and Accuracy. *Procedia Computer Science*, 171, 1327-1336.

Nießen, Sonja; Och, Franz Josef; Leusch, Gregor & Ney, Hermann (2000). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC'00)*. Athen: European Language Resources Association (ELRA).

Nord, Christiane (2006). Translationsqualität aus funktionaler Sicht. In: Schippel, Larisa (Hrsg.) *Übersetzungsqualität: Kritik, Kriterien, Bewertungshandeln*. Berlin: Frank & Timme. 11-30.

Österreichisches Normungsinstitut, Institut, & Austrian Standards Institute (2016). Übersetzungsdienstleistungen - Anforderungen an Übersetzungsdienstleistungen: (ISO/17100: 2015). Wien: Austrian Standards Institute. Ausgabe 2016-04-01.

Pannekoek, Anton (1989). *A history of astronomy*. New York: Dover Publications, Inc.

Papineni, Kishore; Roukos, Salim; Ward, Todd & Zhu, Wei-Jing (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. 311-318.

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

Popović, Maja (2018). Error Classification and Analysis for Machine Translation Quality Assessment. In: Moorkens, Joss; Castilho, Sheila; Gaspari, Federico & Doherty, Stephen (Hrsg.) *Translation Quality Assessment: From Principles to Practice*. Cham: Springer International Publishing. 129-158.

Porsiel, Jörg (2020). Welcome to the Machine! Zwischen Goldgräberstimmung und der Suche nach dem Heiligen Gral. In: Porsiel, Jörg (Hrsg.) *Maschinelle Übersetzung für Übersetzungsprofis*. Berlin: BDÜ Weiterbildungs- und Fachverlags GmbH. 2-10.

Quoc, Le V. & Schuster, Mike (2016). A Neural Network for Machine Translation, at Production Scale. https://research.google/blog/a-neural-network-for-machine-translation-at-production-scale/ (Stand: 10.05.2025).

Reiß, Katharina & Vermeer, Hans (1984). *Grundlegung einer allgemeinen Translationstheorie*. Tübingen: Max Niemeyer Verlag.

Roturier, Johann (2004). Assessing a set of Controlled Language rules: Can they improve the performance of commercial Machine Translation systems. In: *Proceedings of Translating and the Computer 26*. London, UK: Aslib. 1-14.

Rudinger, Rachel; Naradowsky, Jason; Leonard, Brian & Van Durme, Benjamin (2018). Gender Bias in Coreference Resolution. In: *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, Louisiana: Association for Computational Linguistics. 8-14.

Rumelhart, David E.; Hinton, Geoffrey E. & Williams, Ronald J. (1986). Learning Internal Representations by Error Propagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1, 318-362.

SAE International J2450 (2001). Translation Quality Metric J2450_200112. Warrendale, USA: SAE International (Vehicle E E System Diagnostic Standards Committee).

Sager, Juan C. (1989). Quality and Standards: The Evaluation of Translations. In: Picken, Catriona (Hrsg.) *The Translator's Handbook*. London: Aslib. 91-102.

Sánchez-Gijón, Pilar & Kenny, Dorothy (2022). Selecting and preparing texts for machine translation: Pre-editing and writing for a global audience. In: Kenny, Dorothy (Hrsg.) *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Berlin: Language Science Press. 81-103.

Sarkhel, Juran K. & Tripathi, Sneha (2010). Approaches to Machine Translation. *Annals of Library and Information Studies*, 57, 388-393.

Schmalz, Antonia (2019). Maschinelle Übersetzung. In: Wittpahl, Volker (Hrsg.) *Künstliche Intelligenz: Technologie / Anwendung / Gesellschaf*t. Berlin, Heidelberg: Springer. 194-211.

Secară, Alina (2005). Translation evaluation: A state of the art survey. In: *Proceedings of the eCoLoRe/MeLLANGE Workshop*. Leeds, United Kingdom: Centre for Translation Studies, University of Leeds. 39-44.

Seretan, Violeta; Bouillon, Pierrette & Gerlach, Johanna (2014). A Large-Scale Evaluation of Pre-editing Strategies for Improving User-Generated Content Translation. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland: European Language Resources Association (ELRA) 1793-1799.

Sheshadri, Akshay Krishna; Rao Vijjini, Anvesh & Kharbanda, Sukhdeep (2021). WER-BERT: Automatic WER Estimation with BERT in a Balanced Ordinal Classification Paradigm. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.* Online: Association for Computational Linguistics. 3661-3672.

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.

Specia, Lucia; Raj, Dhwaj & Turchi, Marco (2010). Machine translation evaluation versus quality estimation. *Machine Translation*, 24 (1), 39-50.

Stankevičiūtė, Gilvilė; Kasperavičienė, Ramunė & Horbačauskienė, Jolita (2017). Issues in Machine Translation: A Case of Mobile Apps in the Lithuanian and English Language Pair. *International Journal on Language, Literature and Culture in Education*, 4, 75-88.

Stein, Daniel (2009). Maschinelle Übersetzung - ein Überblick. *Journal for Language Technology and Computational Linguistics*, 24 (3), 5-18.

Stymne, Sara & Ahrenberg, Lars (2012). On the practice of error analysis for machine translation evaluation. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA). 1785-1790.

Suojanen, Tytti; Koskinen, Kaisa & Tuominen, Tiina (2015). *User-Centered Translation*. London: Routledge.

Sutskever, Ilya; Vinyals, Oriol & Le, Quoc V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27, 1-9.

SYSTRAN (2025a) SYSTRAN products to translate for all your needs. https://www.systransoft.com/de/ubersetzungsprodukt/ (Stand: 24.02.2025).

SYSTRAN (2025b) SYSTRAN translate Private Cloud. https://www.systransoft.com/de/ubersetzungsprodukt/translate-private-cloud/ (Stand: 24.02.2025).

Tenney, Merle D. (1983). Machine translation, machine-aided translation, and machine-impeded translation. In: *Proceedings of Translating and the Computer 5: Tools for the trade*. London, UK: Aslib. 105-113.

Thurstone, Louis L. (1927). A law of comparative judgment. *Psychological Review*, 34 (4), 273-286.

Toma, Peter (1977). Systran as a multilingual machine translation system. In: *Proceedings of the Third European Congress on Information Systems and Networks, Overcoming the language barrier*. München: Verlag Dokumentation, 569-581.

Toma, Peter (1986). Systran's contribution to mankind. *Terminologie et Traduction*, 1.

Turian, Joseph P.; Shen, Luke & Melamed, I. Dan (2003). Evaluation of machine translation and its evaluation. In: *Proceedings of Machine Translation Summit IX: Papers*. New Orleans, USA.

van Dalen, Benno (2013). Ptolemäus in der islamischen Welt. *Akademie Aktuell*, 46, 28-32.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukasz & Polosukhin, Illia (2017). Attention is All you Need. In: *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA, USA.

Volkova, Irina D. (2018). Text Globalization and Machine Translation (Case of English-Russian Translation of US Travel Site). *SHS Web of Conferences*, 50, 1-6.

Way, Andy (2018). Quality Expectations of Machine Translation. In: Moorkens, Joss; Castilho, Sheila; Gaspari, Federico & Doherty, Stephen (Hrsg.) *Translation Quality Assessment: From Principles to Practice*. Cham: Springer International Publishing. 159-178.

Weaver, Warren (1949). Memorandum on Translation. https://www.mt-archive.net /50/Weaver-1949.pdf (Stand: 10.05.2025).

Werthmann, Antonina & Witt, Andreas (2014). Maschinelle Übersetzung - Gegenwart und Perspektiven. In: Stickel, Gerhard (Hrsg.) *Translation and Interpretation in Europe. Contributions to the Annual Conference 2013 of EFNIL in Vilnius*. Frankfurt am Main/Berlin/Bern/Brüssel/New York/Oxford/Wien: Lang. 79-103.

Woolard, Edgar W. (1949). The evolution of fundamental astronomical concepts as reflected in the terminology of astronomy. *Popular Astronomy*, 57, 482-488.

Woyde, Rick (2001). Introduction to the SAE J2450 translation quality metric. *Language International*, 13 (2), 37-39.

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 (2019). The impact of Google neural machine translation on post-editing by student translators. *The Journal of Specialised Translation*, 31 (1), 87-106.

Zhang, JiaJun & Zong, Chengqing (2020). Neural Machine Translation: Challenges, Progress and Future. *Science China Technological Sciences*, 63 (10), 2028-2050.

Zhao, Jieyu; Wang, Tianlu; Yatskar, Mark; Ordonez, Vicente & Chang, Kai-Wei (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics. 2979-2989.

Anhang 1

TEXTPASSAGE 1	What makes an inaugural flight successful is the immense preparation and attention to detail leading up to it. “An impressive number of tests and fault reports are produced over years of development,” says Michel. “The attention to detail was extreme – even the discovery of a hair or a small mark on a part could prompt a fault report.” “For Ariane 6, there were two years of combined tests, followed by six months of meticulous preparation leading up to this inaugural mission,” Tony recalls (European Space Agency 2024: 22).
TEXTPASSAGE 2	Keeping an eye out for lightning. The first of six Meteosat Third Generation weather satellites is placed within Ariane 5’s fairing for protection during launch. From 36 000 km above the equator, it will observe Earth’s atmosphere in improved detail. Its novel Lightning Imager will monitor lightning events, even in daylight, allowing severe storms to be detected early and warnings issued (European Space Agency 2023: 16).
TEXTPASSAGE 3	ENHANCING METEOROLOGY. Two MetOp Second Generation satellites have been brought together to stand side-by-side for testing at Airbus’s facilities in Toulouse, France. The new MetOp-SG mission, a collaboration between Eumetsat and ESA, consists of six satellites in total. It is based on a pair of satellites, type-A and type-B, each carrying different instruments. The series is set to deliver a wealth of meteorological information for over 20 years (European Space Agency 2024: 12).
TEXTPASSAGE 4	HERA IS COMPLETE. ESA’s asteroid mission ready to be tested. Hera’s Core Module can be thought of as the brains of the mission, hosting its onboard computer, mission systems and instruments (European Space Agency 2023: 68).
TEXTPASSAGE 5	Two satellites will be launched together into orbit to maintain formation relative to each other down to a few millimetres, creating an artificial solar eclipse in space. This mission is being put together for ESA by a consortium led by Spain’s Sener, with participation by more than 29 companies from 14 countries (European Space Agency 2024: 30).
TEXTPASSAGE 6	An object greater than 15 m across can survive intact as it enters Earth’s atmosphere and can cause considerable damage when it impacts or burns up mid-air. Sixty-six million years ago, the Chicxulub impactor in Mexico, at more than 11 km in diameter, was deadly for three quarters of life on planet Earth (including, of course, the dinosaurs). In 1908, the forest-flattening Tunguska event in Russia was most likely caused by an asteroid measuring about 50 m across disintegrating in mid-air (European Space Agency 2019: 24).
TEXTPASSAGE 7	The NEOCC keeps track of known NEAs, which currently number more than 21 300, and publishes a risk list that tracks all objects for which a non-zero impact probability has been detected. The NEOCC risk list shows that the asteroid Apophis – approximately 350 m in diameter – could potentially collide with Earth in 2068. The same

	asteroid is due to pass near to Earth in 2029, enabling astronomers to make very precise observations (European Space Agency 2019: 29).
TEXTPASSAGE 9	BURIED WATER ICE AT MARS' EQUATOR? Windswept piles of dust, or layers of ice? ESA's Mars Express has revisited one of Mars' most mysterious features – the Medusae Fossae Formation (MFF) – to clarify its composition. Its findings suggest layers of water ice stretching several kilometres below ground – the most water ever found in this part of the planet (European Space Agency 2024: 13).
TEXTPASSAGE 10	"With DART and Hera, we will have a fully validated asteroid deflection technique enabling humankind to avoid future impacts." Ian points out that working closely with NASA on the AIDA collaboration is an important aspect of the project: "The collaboration with another agency and its investigation team is important as this is what would happen in a real impact scenario (European Space Agency 2019: 26).
TEXTPASSAGE 11	READY FOR MOON DELIVERY. Engineers and technicians keep a close eye on the third European Service Module (ESM-3), an essential component of NASA's Orion spacecraft leaving the integration hall at Airbus Defence and Space in Bremen, Germany (European Space Agency 2024: 72).

Anhang 2

Textpassage 1	
AT 1.1 kein Pre-Editing	**ZT 1.1**
What makes an inaugural flight successful is the immense preparation and attention to detail leading up to it. “An impressive number of tests and fault reports are produced over years of development,” says Michel. “The attention to detail was extreme – even the discovery of a hair or a small mark on a part could prompt a fault report.” “For Ariane 6, there were two years of combined tests, followed by six months of meticulous preparation leading up to this inaugural mission,” Tony recalls.	Was einen Erstflug erfolgreich macht, ist die immense Vorbereitung und Liebe zum Detail, die zu ihm führt. „In jahrelanger Entwicklung werden eine beeindruckende Anzahl von Tests und Fehlerberichten erstellt“, sagt Michel. „Die Liebe zum Detail war extrem – selbst die Entdeckung eines Haares oder einer kleinen Markierung an einem Teil konnte eine Fehlermeldung auslösen.“ „Für Ariane 6 gab es zwei Jahre kombinierter Tests, gefolgt von sechs Monaten akribischer Vorbereitung, die zu dieser Eröffnungsmission führten“, erinnert sich Tony.
AT 1.2 Pre-Editing-Regel 1	**ZT 1.2**
What makes an inaugural flight successful is the immense preparation and attention to detail preceding it. “An impressive number of tests and fault reports are produced over years of development,” says Michel. “The attention to detail was extreme – even the discovery of a hair or a small mark on a part could prompt a fault report.” “For Ariane 6, there were two years of combined tests, followed by six months of meticulous preparation preceding this inaugural mission,” Tony recalls.	Was einen Erstflug erfolgreich macht, ist die immense Vorbereitung und Liebe zum Detail, die ihm vorausgeht. „In jahrelanger Entwicklung werden eine beeindruckende Anzahl von Tests und Fehlerberichten erstellt“, sagt Michel. „Die Liebe zum Detail war extrem – selbst die Entdeckung eines Haares oder einer kleinen Markierung an einem Teil konnte eine Fehlermeldung auslösen.“ „Für Ariane 6 gab es zwei Jahre kombinierter Tests, gefolgt von sechs Monaten akribischer Vorbereitung vor dieser Eröffnungsmission“, erinnert sich Tony.
AT 1.3 Pre-Editing-Regel 4	**ZT 1.3**
What makes an inaugural flight successful is the immense preparation and attention to detail leading up to it. “An impressive number of tests and fault reports are produced over years of development,” says Michel. “The attention to detail was extreme – even the discovery of a hair or a small mark on a part could prompt a fault report.” “For Ariane 6, there were two years of combined tests, followed by six months of meticulous preparation leading up to this inaugural flight,” Tony recalls.	Was einen Erstflug erfolgreich macht, ist die immense Vorbereitung und Liebe zum Detail, die zu ihm führt. „In jahrelanger Entwicklung werden eine beeindruckende Anzahl von Tests und Fehlerberichten erstellt“, sagt Michel. „Die Liebe zum Detail war extrem – selbst die Entdeckung eines Haares oder einer kleinen Markierung an einem Teil konnte eine Fehlermeldung auslösen.“ „Für Ariane 6 gab es zwei Jahre kombinierter Tests, gefolgt von sechs Monaten akribischer Vorbereitung, die zu diesem Eröffnungsflug führten“, erinnert sich Tony.
AT 1.4 Pre-Editing-Regel 2	**ZT 1.4**
What makes an inaugural flight successful is the immense preparation and attention to detail leading up to it. “Over years of development, we produced an impressive number of tests and fault reports,” says Michel. “The attention to detail was extreme – even the	Was einen Erstflug erfolgreich macht, ist die immense Vorbereitung und Liebe zum Detail, die zu ihm führt. „In jahrelanger Entwicklung haben wir eine beeindruckende Anzahl an Tests und Fehlerberichten erstellt“, sagt Michel. „Die Liebe zum Detail

discovery of a hair or a small mark on a part could prompt a fault report." "For Ariane 6, we conducted two years of combined tests, followed by six months of meticulous preparation leading up to this inaugural mission," Tony recalls.	war extrem – sogar die Entdeckung eines Haares oder einer kleinen Markierung an einem Teil konnte eine Fehlermeldung auslösen." „Für Ariane 6 führten wir zwei Jahre kombinierte Tests durch, gefolgt von sechs Monaten akribischer Vorbereitung, die zu dieser Eröffnungsmission führten", erinnert sich Tony.
Textpassage 2	
AT 2.1 kein Pre-Editing	**ZT 2.1**
Keeping an eye out for lightning. The first of six Meteosat Third Generation weather satellites is placed within Ariane 5's fairing for protection during launch. From 36 000 km above the equator, it will observe Earth's atmosphere in improved detail. Its novel Lightning Imager will monitor lightning events, even in daylight, allowing severe storms to be detected early and warnings issued.	Blitze im Auge. Der erste von sechs Meteosat Wettersatelliten der dritten Generation wird während des Starts zum Schutz in der Verkleidung der Ariane 5 platziert. Von 36 000 km über dem Äquator aus wird es die Erdatmosphäre genauer beobachten. Der neuartige Lightning Imager überwacht Blitzereignisse auch bei Tageslicht, sodass schwere Stürme frühzeitig erkannt und Warnungen ausgegeben werden können.
AT 2.2 Pre-Editing-Regel 5	**ZT 2.2**
Keeping an eye out for lightning. The first of six Meteosat Third Generation weather satellites is placed within Ariane 5's fairing for protection during launch. From 36 000 km above the equator, the satellite will observe Earth's atmosphere in improved detail. The novel Lightning Imager onboard the satellite will monitor lightning events, even in daylight, allowing severe storms to be detected early and warnings to be issued.	Blitze im Auge. Der erste von sechs Meteosat Wettersatelliten der dritten Generation wird während des Starts zum Schutz in der Verkleidung der Ariane 5 platziert. Von 36 000 km über dem Äquator aus wird der Satellit die Erdatmosphäre detaillierter beobachten. Der neuartige Lightning Imager an Bord des Satelliten überwacht Blitzereignisse auch bei Tageslicht, sodass schwere Stürme frühzeitig erkannt und Warnungen ausgegeben werden können.
AT 2.3 Pre-Editing-Regel 3	**ZT 2.3**
Observing lightning activity. The first of six Meteosat Third Generation weather satellites is placed within Ariane 5's fairing for protection during launch. From an altitude of 36 000 km above the equator, it will observe Earth's atmosphere in improved detail. Its novel Lightning Imager will monitor lightning events, even in daylight, allowing severe storms to be detected early and warnings issued.	Beobachtung der Blitzaktivität. Der erste von sechs Meteosat Wettersatelliten der dritten Generation wird während des Starts zum Schutz in der Verkleidung der Ariane 5 platziert. Aus einer Höhe von 36 000 km über dem Äquator wird es die Erdatmosphäre genauer beobachten. Der neuartige Lightning Imager überwacht Blitzereignisse auch bei Tageslicht, sodass schwere Stürme frühzeitig erkannt und Warnungen ausgegeben werden können.
AT 2.4 Pre-Editing-Regel 2	**ZT 2.4**
Keeping an eye out for lightning. We placed the first of six Meteosat Third Generation weather satellites within Ariane 5's fairing to protect it during launch. From 36 000 km above the equator, it will observe Earth's atmosphere in improved detail. Its novel Lightning Imager will monitor lightning events, even in daylight,	Blitze im Auge. Wir platzierten den ersten von sechs Meteosat-Wettersatelliten der dritten Generation innerhalb der Verkleidung von Ariane 5, um ihn während des Starts zu schützen. Von 36 000 km über dem Äquator aus wird es die Erdatmosphäre genauer beobachten. Der neuartige Lightning Imager überwacht Blitzereignisse, auch bei Tageslicht, sodass wir

allowing us to detect severe storms early and issue warnings.	schwere Stürme frühzeitig erkennen und Warnungen ausgeben können.
Textpassage 3	
AT 3.1 kein Pre-Editing	**ZT 3.1**
ENHANCING METEOROLOGY. Two MetOp Second Generation satellites have been brought together to stand side-by-side for testing at Airbus's facilities in Toulouse, France. The new MetOp-SG mission, a collaboration between Eumetsat and ESA, consists of six satellites in total. It is based on a pair of satellites, type-A and type-B, each carrying different instruments. The series is set to deliver a wealth of meteorological information for over 20 years.	VERBESSERUNG DER METEOROLOGIE. Zwei MetOp-Satelliten der zweiten Generation wurden zusammengeführt, um in den Airbus-Werken in Toulouse, Frankreich, parallel zu testen. Die neue MetOp-SG-Mission, eine Kooperation zwischen Eumetsat und ESA, besteht aus insgesamt sechs Satelliten. Es basiert auf einem Paar von Satelliten, Typ A und Typ B, die jeweils unterschiedliche Instrumente tragen. Die Serie soll über 20 Jahre lang eine Fülle meteorologischer Informationen liefern.
AT 3.2 Pre-Editing-Regel 5	**ZT 3.2**
ENHANCING METEOROLOGY. Two MetOp Second Generation satellites have been brought together to stand side-by-side for testing at Airbus's facilities in Toulouse, France. The new MetOp-SG mission, a collaboration between Eumetsat and ESA, consists of six satellites in total. The basis is a pair of satellites, type-A and type-B, each carrying different instruments. The series is set to deliver a wealth of meteorological information for over 20 years.	VERBESSERUNG DER METEOROLOGIE. Zwei MetOp-Satelliten der zweiten Generation wurden zusammengeführt, um in den Airbus-Werken in Toulouse, Frankreich, parallel zu testen. Die neue MetOp-SG-Mission, eine Kooperation zwischen Eumetsat und ESA, besteht aus insgesamt sechs Satelliten. Die Basis ist ein Paar Satelliten, Typ A und Typ B, die jeweils unterschiedliche Instrumente tragen. Die Serie soll über 20 Jahre lang eine Fülle meteorologischer Informationen liefern.
AT 3.3 Pre-Editing-Regel 1	**ZT 3.3**
ENHANCING METEOROLOGY. Two MetOp Second Generation satellites have been joined side-by-side for testing at Airbus's facilities in Toulouse, France. The new MetOp-SG mission, a collaboration between Eumetsat and ESA, consists of six satellites in total. It is based on a pair of satellites, type-A and type-B, each carrying different instruments. The series is set to deliver a wealth of meteorological information for over 20 years.	VERBESSERUNG DER METEOROLOGIE. Zwei MetOp-Satelliten der zweiten Generation wurden in den Airbus-Werken in Toulouse, Frankreich, parallel getestet. Die neue MetOp-SG-Mission, eine Kooperation zwischen Eumetsat und ESA, besteht aus insgesamt sechs Satelliten. Es basiert auf einem Paar von Satelliten, Typ A und Typ B, die jeweils unterschiedliche Instrumente tragen. Die Serie soll über 20 Jahre lang eine Fülle meteorologischer Informationen liefern.
AT 3.4 Pre-Editing-Regel 4	**ZT 3.4**
ADVANCING METEOROLOGY. Two MetOp Second Generation satellites have been brought together to stand side-by-side for testing at Airbus's facilities in Toulouse, France. The new MetOp-SG mission, a collaboration between Eumetsat and ESA, consists of six satellites in total. It is based on satellite pairs, type-A and type-B, each equipped with different instruments.	FORTSCHRITT IN DER METEOROLOGIE. Zwei MetOp-Satelliten der zweiten Generation wurden zusammengeführt, um in den Airbus-Werken in Toulouse, Frankreich, parallel zu testen. Die neue MetOp-SG-Mission, eine Kooperation zwischen Eumetsat und ESA, besteht aus insgesamt sechs Satelliten. Es basiert auf Satellitenpaaren, Typ A und Typ B, die jeweils mit unterschiedlichen

The series is set to deliver a wealth of meteorological information for over 20 years.	Instrumenten ausgestattet sind. Die Serie soll über 20 Jahre lang eine Fülle meteorologischer Informationen liefern.
Textpassage 4	
AT 4.1 kein Pre-Editing	**ZT 4.1**
HERA IS COMPLETE. ESA's asteroid mission ready to be tested. Hera's Core Module can be thought of as the brains of the mission, hosting its onboard computer, mission systems and instruments.	HERA IST KOMPLETT. Die Asteroidenmission der ESA ist bereit, getestet zu werden. Heras Kernmodul kann als das Gehirn der Mission betrachtet werden, die ihre Bordcomputer, Missionssysteme und Instrumente hostet.
AT 4.2 Pre-Editing-Regel 3	**ZT 4.2**
HERA IS COMPLETE. ESA's asteroid mission ready to be tested. Hera's Core Module can be thought of as the central component of the mission, hosting its onboard computer, mission systems and instruments.	HERA IST KOMPLETT. Die Asteroidenmission der ESA ist bereit, getestet zu werden. Heras Kernmodul kann als zentrale Komponente der Mission betrachtet werden, die ihren Bordcomputer, ihre Missionssysteme und Instrumente hostet.
AT 4.3 Pre-Editing-Regel 6	**ZT 4.3**
HERA IS COMPLETE. ESA's asteroid mission ready to be tested. Hera's Core Module can be thought of as the brains of the mission, that hosts its onboard computer, mission systems, and instruments.	HERA IST KOMPLETT. Die Asteroidenmission der ESA ist bereit, getestet zu werden. Heras Kernmodul kann als das Gehirn der Mission betrachtet werden, das seinen Bordcomputer, Missionssysteme und Instrumente hostet.
AT 4.4 Pre-Editing-Regel 5	**ZT 4.4**
HERA IS COMPLETE. ESA's asteroid mission ready to be tested. Hera's Core Module can be thought of as the brains of the mission, hosting the onboard computer, mission systems and instruments.	HERA IST KOMPLETT. Die Asteroidenmission der ESA ist bereit, getestet zu werden. Heras Kernmodul kann als Gehirn der Mission betrachtet werden, das den Bordcomputer, Missionssysteme und Instrumente hostet.
Textpassage 5	
AT 5.1 kein Pre-Editing	**ZT 5.1**
Two satellites will be launched together into orbit to maintain formation relative to each other down to a few millimetres, creating an artificial solar eclipse in space. This mission is being put together for ESA by a consortium led by Spain's Sener, with participation by more than 29 companies from 14 countries.	Zwei Satelliten werden gemeinsam in die Umlaufbahn gebracht, um die Formation relativ zueinander auf wenige Millimeter zu halten und eine künstliche Sonnenfinsternis im Weltraum zu erzeugen. Diese Mission wird für die ESA von einem Konsortium unter der Leitung des spanischen Senders Sener unter Beteiligung von mehr als 29 Unternehmen aus 14 Ländern zusammengestellt.
AT 5.2 Pre-Editing-Regel 6	**ZT 5.2**
Two satellites will be launched together into orbit to maintain formation relative to each other down to a few millimetres, which will create an artificial solar eclipse in space. This mission is being put together for ESA by	Zwei Satelliten werden gemeinsam in die Umlaufbahn gebracht, um die Formation relativ zueinander auf wenige Millimeter zu halten, was eine künstliche Sonnenfinsternis im Weltraum erzeugen wird. Diese Mission wird für die ESA von einem Konsortium

a consortium led by Spain's Sener, with participation by more than 29 companies from 14 countries.	unter der Leitung des spanischen Senders Sener unter Beteiligung von mehr als 29 Unternehmen aus 14 Ländern zusammengestellt.
AT 5.3 Pre-Editing-Regel 2	**ZT 5.3**
Two satellites will be launched together into orbit to maintain formation relative to each other down to a few millimetres, creating an artificial solar eclipse in space. A consortium led by Spain's Sener, with participation by more than 29 companies from 14 countries, is putting together this mission for ESA.	Zwei Satelliten werden gemeinsam in die Umlaufbahn gebracht, um die Formation relativ zueinander auf wenige Millimeter zu halten und eine künstliche Sonnenfinsternis im Weltraum zu erzeugen. Ein Konsortium unter der Leitung des spanischen Senders Sener, an dem mehr als 29 Unternehmen aus 14 Ländern beteiligt sind, stellt diese Mission für die ESA zusammen.
AT 5.4 Pre-Editing-Regel 4	**ZT 5.4**
Two satellites will be launched into orbit together to maintain formation relative to each other accurate to the millimetre, creating an artificial solar eclipse in space. This mission is being put together for ESA by a consortium led by Spain's Sener, with participation from more than 29 companies from 14 countries.	Zwei Satelliten werden gemeinsam in die Umlaufbahn gebracht, um die Formation relativ zueinander millimetergenau zu halten und so eine künstliche Sonnenfinsternis im Weltraum zu erzeugen. Diese Mission wird für die ESA von einem Konsortium unter der Leitung des spanischen Senders Sener unter Beteiligung von mehr als 29 Unternehmen aus 14 Ländern zusammengestellt.
Textpassage 6	
AT 6.1 kein Pre-Editing	**ZT 6.1**
An object greater than 15 m across can survive intact as it enters Earth's atmosphere and can cause considerable damage when it impacts or burns up mid-air. Sixty-six million years ago, the Chicxulub impactor in Mexico, at more than 11 km in diameter, was deadly for three quarters of life on planet Earth (including, of course, the dinosaurs). In 1908, the forest-flattening Tunguska event in Russia was most likely caused by an asteroid measuring about 50 m across disintegrating in mid-air.	Ein Objekt mit einem Durchmesser von mehr als 15 m kann intakt überleben, wenn es in die Erdatmosphäre eintritt, und kann erheblichen Schaden anrichten, wenn es mitten in der Luft auftrifft oder verbrennt. Vor 66 Millionen Jahren war der Chicxulub-Impaktor in Mexiko mit einem Durchmesser von mehr als 11 km für drei Viertel des Lebens auf dem Planeten Erde tödlich (einschließlich natürlich der Dinosaurier). Im Jahr 1908 wurde das waldverflachende Tunguska-Ereignis in Russland höchstwahrscheinlich durch einen Asteroiden verursacht, der etwa 50 m groß war und sich in der Luft auflöste.
AT 6.2 Pre-Editing-Regel 2	**ZT 6.2**
An object greater than 15 m across can survive intact as it enters Earth's atmosphere and can cause considerable damage when it impacts or burns up mid-air. Sixty-six million years ago, the Chicxulub impactor in Mexico, at more than 11 km in diameter, was deadly for three quarters of life on planet Earth (including, of course, the dinosaurs). In 1908, an asteroid measuring about 50 m across disintegrating in mid-air most likely caused the forest-flattening Tunguska event in Russia.	Ein Objekt mit einem Durchmesser von mehr als 15 m kann intakt überleben, wenn es in die Erdatmosphäre eintritt, und kann erheblichen Schaden anrichten, wenn es mitten in der Luft auftrifft oder verbrennt. Vor 66 Millionen Jahren war der Chicxulub-Impaktor in Mexiko mit einem Durchmesser von mehr als 11 km für drei Viertel des Lebens auf dem Planeten Erde tödlich (einschließlich natürlich der Dinosaurier). Im Jahr 1908 verursachte ein Asteroid

	mit einer Größe von etwa 50 Metern, der sich in der Luft auflöste, höchstwahrscheinlich das waldverflachende Tunguska-Ereignis in Russland.
AT 6.3 Pre-Editing-Regel 1	**ZT 6.3**
An object greater than 15 m across can survive intact as it enters Earth's atmosphere and can cause considerable damage when it impacts or explodes mid-air. Sixty-six million years ago, the Chicxulub impactor in Mexico, at more than 11 km in diameter, was deadly for three quarters of life on planet Earth (including, of course, the dinosaurs). In 1908, the forest-flattening Tunguska event in Russia was most likely caused by an asteroid measuring about 50 m across disintegrating in mid-air	Ein Objekt mit einem Durchmesser von mehr als 15 m kann intakt überleben, wenn es in die Erdatmosphäre eintritt, und kann erheblichen Schaden verursachen, wenn es mitten in der Luft auftrifft oder explodiert. Vor 66 Millionen Jahren war der Chicxulub-Impaktor in Mexiko mit einem Durchmesser von mehr als 11 km für drei Viertel des Lebens auf dem Planeten Erde tödlich (einschließlich natürlich der Dinosaurier). Im Jahr 1908 wurde das waldverflachende Tunguska-Ereignis in Russland höchstwahrscheinlich durch einen Asteroiden verursacht, der etwa 50 m groß war und sich in der Luft auflöste.
AT 6.4 Pre-Editing-Regel 3	**ZT 6.4**
An object greater than 15 m across can survive intact as it enters Earth's atmosphere and can cause considerable damage when it impacts the surface or burns up mid-air. Sixty-six million years ago, the Chicxulub impactor in Mexico, at more than 11 km in diameter, was deadly for three quarters of life on planet Earth (including, of course, the dinosaurs). In 1908, the devastating Tunguska event in Russia was most likely caused by an asteroid measuring about 50 m across disintegrating in mid-air.	Ein Objekt mit einem Durchmesser von mehr als 15 m kann intakt überleben, wenn es in die Erdatmosphäre eintritt, und kann erheblichen Schaden verursachen, wenn es auf die Oberfläche auftrifft oder in der Luft verbrennt. Vor 66 Millionen Jahren war der Chicxulub-Impaktor in Mexiko mit einem Durchmesser von mehr als 11 km für drei Viertel des Lebens auf dem Planeten Erde tödlich (einschließlich natürlich der Dinosaurier). Im Jahr 1908 wurde das verheerende Tunguska-Ereignis in Russland höchstwahrscheinlich durch einen Asteroiden verursacht, der etwa 50 m groß war und sich in der Luft auflöste.
Textpassage 7	
AT 7.1 kein Pre-Editing	**ZT 7.1**
The NEOCC keeps track of known NEAs, which currently number more than 21 300, and publishes a risk list that tracks all objects for which a non-zero impact probability has been detected. The NEOCC risk list shows that the asteroid Apophis – approximately 350 m in diameter – could potentially collide with Earth in 2068. The same asteroid is due to pass near to Earth in 2029, enabling astronomers to make very precise observations.	Das NEOCC verfolgt bekannte NEAs, die derzeit mehr als 21 300 umfassen, und veröffentlicht eine Risikoliste, die alle Objekte verfolgt, für die eine Wahrscheinlichkeit von Nicht-Null-Auswirkungen festgestellt wurde. Die NEOCC-Risikoliste zeigt, dass der Asteroid Apophis – ca. 350 m Durchmesser – 2068 möglicherweise mit der Erde kollidieren könnte. Derselbe Asteroid wird 2029 in der Nähe der Erde vorbeiziehen, sodass Astronomen sehr präzise Beobachtungen machen können.
AT 7.2 Pre-Editing-Regel 1	**ZT 7.2**
The NEOCC monitors known NEAs, which currently number more than 21 300, and publishes a risk list that tracks all objects for which a non-zero impact	Das NEOCC überwacht bekannte NEAs, die derzeit mehr als 21 300 umfassen, und veröffentlicht eine Risikoliste, die alle Objekte verfolgt, für die eine

probability has been detected. The NEOCC risk list shows that the asteroid Apophis – approximately 350 m in diameter – could potentially collide with Earth in 2068. The same asteroid is due to pass near to Earth in 2029, enabling astronomers to make very precise observations.	Wahrscheinlichkeit von Nicht-Null-Auswirkungen festgestellt wurde. Die NEOCC-Risikoliste zeigt, dass der Asteroid Apophis – ca. 350 m Durchmesser – 2068 möglicherweise mit der Erde kollidieren könnte. Derselbe Asteroid wird 2029 in der Nähe der Erde vorbeiziehen, sodass Astronomen sehr präzise Beobachtungen machen können.
AT 7.3 Pre-Editing-Regel 6	**ZT 7.3**
The NEOCC keeps track of known NEAs, which currently number more than 21 300, and publishes a risk list that tracks all objects for which a non-zero impact probability has been detected. The NEOCC risk list shows that the asteroid Apophis – approximately 350 m in diameter – could potentially collide with Earth in 2068. The same asteroid is due to pass near to Earth in 2029, which enables astronomers to make very precise observations.	Das NEOCC verfolgt bekannte NEAs, die derzeit mehr als 21 300 umfassen, und veröffentlicht eine Risikoliste, die alle Objekte verfolgt, für die eine Wahrscheinlichkeit von Nicht-Null-Auswirkungen festgestellt wurde. Die NEOCC-Risikoliste zeigt, dass der Asteroid Apophis – ca. 350 m Durchmesser – 2068 möglicherweise mit der Erde kollidieren könnte. Derselbe Asteroid soll 2029 in der Nähe der Erde vorbeiziehen, was es Astronomen ermöglicht, sehr präzise Beobachtungen vorzunehmen.
AT 7.4 Pre-Editing-Regel 2	**ZT 7.4**
The NEOCC keeps track of known NEAs, which currently number more than 21 300, and publishes a risk list that tracks all objects for which NEOCC has detected a non-zero impact probability. The NEOCC risk list shows that the asteroid Apophis – approximately 350 m in diameter – could potentially collide with Earth in 2068. The same asteroid is due to pass near to Earth in 2029, enabling astronomers to make very precise observations.	Das NEOCC verfolgt die bekannten NEAs, die derzeit mehr als 21 300 umfassen, und veröffentlicht eine Risikoliste, die alle Objekte verfolgt, für die das NEOCC eine Aufprallwahrscheinlichkeit ungleich null erkannt hat. Die NEOCC-Risikoliste zeigt, dass der Asteroid Apophis – ca. 350 m Durchmesser – 2068 möglicherweise mit der Erde kollidieren könnte. Derselbe Asteroid wird 2029 in der Nähe der Erde vorbeiziehen, sodass Astronomen sehr präzise Beobachtungen machen können.
Textpassage 9	
AT 9.1 kein Pre-Editing	**ZT 9.1**
BURIED WATER ICE AT MARS' EQUATOR? Windswept piles of dust, or layers of ice? ESA's Mars Express has revisited one of Mars' most mysterious features – the Medusae Fossae Formation (MFF) – to clarify its composition. Its findings suggest layers of water ice stretching several kilometres below ground – the most water ever found in this part of the planet.	VERGRABENES WASSEREIS AM MARSÄQUATOR? Windhungrige Staubberge oder Eisschichten? Der Mars-Express der ESA hat eines der geheimnisvollsten Merkmale des Mars – die Medusae-Fossae-Formation (MFF) – überarbeitet, um seine Zusammensetzung zu klären. Seine Funde deuten auf mehrere Kilometer unter der Erde verlaufende Wassereisschichten hin – das meiste Wasser, das jemals in diesem Teil des Planeten gefunden wurde.
AT 9.2 Pre-Editing-Regel 4	**ZT 9.2**
SUBSURFACE WATER ICE AT MARS' EQUATOR? Wind-formed piles of dust, or layers of ice? ESA's Mars Express has reinvestigated one of Mars' most mysterious regions – the Medusae Fossae	UNTERIRDISCHES WASSEREIS AM MARSÄQUATOR? Durch Wind geformte Staubhaufen oder Eisschichten? Der Mars-Express der ESA hat eine der geheimnisvollsten Regionen des Mars – die

Formation (MFF) – to clarify its composition. Its findings suggest layers of water ice stretching several kilometres below ground – the most water ever found in this part of the planet.	Medusae-Fossae-Formation (MFF) – erneut untersucht, um ihre Zusammensetzung zu klären. Seine Funde deuten auf mehrere Kilometer unter der Erde verlaufende Wassereisschichten hin – das meiste Wasser, das jemals in diesem Teil des Planeten gefunden wurde.
AT 9.3 Pre-Editing-Regel 5	**ZT 9.3**
BURIED WATER ICE AT MARS' EQUATOR? Windswept piles of dust, or layers of ice? ESA's Mars Express has revisited one of Mars' most mysterious features – the Medusae Fossae Formation (MFF) – to clarify the composition of this formation. The findings suggest layers of water ice stretching several kilometres below ground – the most water ever found in this part of the planet.	VERGRABENES WASSEREIS AM MARSÄQUATOR? Windhungrige Staubberge oder Eisschichten? Der Mars-Express der ESA hat eines der geheimnisvollsten Merkmale des Mars – die Medusae-Fossae-Formation (MFF) – überarbeitet, um die Zusammensetzung dieser Formation zu klären. Die Funde legen Wassereisschichten nahe, die sich mehrere Kilometer unter der Erde erstrecken – das meiste Wasser, das jemals in diesem Teil des Planeten gefunden wurde.
AT 9.4 Pre-Editing Regel 3	**ZT 9.4**
BURIED WATER ICE AT MARS' EQUATOR? Are they windswept piles of dust, or layers of ice? ESA's Mars Express has revisited the Medusae Fossae Formation (MFF), one of Mars' most mysterious features, to clarify its composition. Its findings suggest there are layers of water ice stretching several kilometres below ground – the most water ever found in this part of the planet.	VERGRABENES WASSEREIS AM MARSÄQUATOR? Sind es windgepeitschte Staubberge oder Eisschichten? Der Mars-Express der ESA hat die Medusae-Fossae-Formation (MFF), eines der geheimnisvollsten Merkmale des Mars, erneut untersucht, um ihre Zusammensetzung zu klären. Die Funde deuten darauf hin, dass sich mehrere Kilometer unter der Erde Wassereisschichten erstrecken – das meiste Wasser, das jemals in diesem Teil des Planeten gefunden wurde.
Textpassage 10	
AT 10.1 kein Pre-Editing	**ZT 10.1**
"With DART and Hera, we will have a fully validated asteroid deflection technique enabling humankind to avoid future impacts." Ian points out that working closely with NASA on the AIDA collaboration is an important aspect of the project: "The collaboration with another agency and its investigation team is important as this is what would happen in a real impact scenario.	„Mit DART und Hera verfügen wir über eine vollständig validierte Asteroidenablenktechnik, mit der die Menschheit zukünftige Einschläge vermeiden kann." Ian weist darauf hin, dass die enge Zusammenarbeit mit der NASA bei der AIDA-Zusammenarbeit ein wichtiger Aspekt des Projekts ist: „Die Zusammenarbeit mit einer anderen Agentur und ihrem Untersuchungsteam ist wichtig, da dies in einem echten Wirkungsszenario passieren würde.
AT 10.2 Pre-Editing-Regel 6	**ZT 10.2**
"With DART and Hera, we will have a fully validated asteroid deflection technique, that enables humankind to avoid future impacts." Ian points out that working closely with NASA on the AIDA collaboration is an important aspect of the project: "The collaboration with	„Mit DART und Hera verfügen wir über eine vollständig validierte Asteroidenablenktechnik, die es der Menschheit ermöglicht, zukünftige Einschläge zu vermeiden." Ian weist darauf hin, dass die enge Zusammenarbeit mit der NASA bei der AIDA-Zusammenarbeit ein wichtiger Aspekt des Projekts ist: „Die

another agency and its investigation team is important as this is what would happen in a real impact scenario.	Zusammenarbeit mit einer anderen Agentur und ihrem Untersuchungsteam ist wichtig, da dies in einem echten Wirkungsszenario passieren würde.
AT 10.3 Pre-Editing-Regel 1	**ZT 10.3**
"With DART and Hera, we will have a fully validated asteroid deflection technique enabling humankind to avoid future impacts." Ian indicates that working closely with NASA on the AIDA collaboration is an important aspect of the project: "The collaboration with another agency and its investigation team is important as this is what would happen in a real impact scenario.	„Mit DART und Hera verfügen wir über eine vollständig validierte Asteroidenablenktechnik, mit der die Menschheit zukünftige Einschläge vermeiden kann." Ian erklärt, dass die enge Zusammenarbeit mit der NASA bei der AIDA-Zusammenarbeit ein wichtiger Aspekt des Projekts ist: „Die Zusammenarbeit mit einer anderen Agentur und ihrem Untersuchungsteam ist wichtig, da dies in einem echten Wirkungsszenario passieren würde.
AT 10.4 Pre-Editing-Regel 5	**ZT 10.4**
"With DART and Hera, we will have a fully validated asteroid deflection technique enabling humankind to avoid future impacts." Ian points out that working closely with NASA on the AIDA collaboration is an important aspect of the project: "The collaboration with another agency and that agency's investigation team is important as this is what would happen in a real impact scenario.	„Mit DART und Hera verfügen wir über eine vollständig validierte Asteroidenablenktechnik, mit der die Menschheit zukünftige Einschläge vermeiden kann." Ian weist darauf hin, dass die enge Zusammenarbeit mit der NASA bei der AIDA-Zusammenarbeit ein wichtiger Aspekt des Projekts ist: „Die Zusammenarbeit mit einer anderen Agentur und dem Untersuchungsteam dieser Agentur ist wichtig, da dies in einem echten Wirkungsszenario passieren würde.
Textpassage 11	
AT 11.1 kein Pre-Editing	**ZT 11.1**
READY FOR MOON DELIVERY. Engineers and technicians keep a close eye on the third European Service Module (ESM-3), an essential component of NASA's Orion spacecraft leaving the integration hall at Airbus Defence and Space in Bremen, Germany.	BEREIT FÜR MONDLIEFERUNG. Ingenieure und Techniker haben das dritte europäische Servicemodul (ESM-3) im Auge, ein wesentlicher Bestandteil des Orion-Raumschiffs der NASA, das die Integrationshalle bei Airbus Defence and Space in Bremen, Deutschland, verlässt.
AT 11.2 Pre-Editing-Regel 6	**ZT 11.2**
READY FOR MOON DELIVERY. Engineers and technicians keep a close eye on the third European Service Module (ESM-3), an essential component of NASA's Orion spacecraft, which is leaving the integration hall at Airbus Defence and Space in Bremen, Germany.	BEREIT FÜR MONDLIEFERUNG. Ingenieure und Techniker haben das dritte europäische Servicemodul (ESM-3), ein wesentlicher Bestandteil der Orion-Raumsonde der NASA, das die Integrationshalle bei Airbus Defence and Space in Bremen, Deutschland, verlässt, genau im Blick.
AT 11.3 Pre-Editing-Regel 4	**ZT 11.3**
READY FOR LUNAR MISSION DEPLOYMENT. Engineers and technicians keep a close eye on the third European Service Module (ESM-3), a critical component of NASA's Orion spacecraft, exiting the integration hall at Airbus Defence and Space in Bremen, Germany.	BEREIT FÜR DIE MONDMISSION. Ingenieure und Techniker haben das dritte europäische Servicemodul (ESM-3), eine wichtige Komponente des Orion-Raumschiffs der NASA, beim Verlassen der Integrationshalle bei Airbus Defence and Space in Bremen, Deutschland, genau im Blick.

AT 11.4 Pre-Editing-Regel 3	ZT 11.4
READY TO BE SENT TO THE MOON. Engineers and technicians closely watch the third European Service Module (ESM-3), an essential component of NASA's Orion spacecraft leaving the integration hall at Airbus Defence and Space in Bremen, Germany.	BEREIT, ZUM MOND GESCHICKT ZU WERDEN. Ingenieure und Techniker beobachten aufmerksam das dritte europäische Servicemodul (ESM-3), eine wesentliche Komponente des Orion-Raumschiffs der NASA, das die Integrationshalle bei Airbus Defence and Space in Bremen, Deutschland, verlässt.

Zusammenfassung

Die zunehmende Verbreitung maschineller Übersetzung in professionellen und privaten Kontexten erfordert neue Strategien zur Sicherung der Übersetzungsqualität. Diese Masterarbeit untersucht, welche gezielt ausgewählte Pre-Editing-Regeln, also kleine sprachliche Anpassungen des Ausgangstextes vor der maschinellen Übersetzung, die Übersetzungsqualität im Fachbereich Astronautik und der Sprachkombination Englisch-Deutsch verbessern können. Auf der Grundlage einer Fehlertypologie und bestehender Erkenntnisse zu MTranslatability wird analysiert, welche spezifischen Pre-Editing-Regeln zu einer Qualitätsverbesserung gegenüber dem unbearbeiteten Rohtext führen. Die Bewertung der Ergebnisse erfolgt durch Translationsexpert*innen und Lai*innen mit Best-Worst-Scaling, um unterschiedliche Perspektiven auf Übersetzungsqualität und Nutzungsanforderungen zu berücksichtigen. Ziel ist die Identifizierung wirkungsvoller Pre-Editing-Regeln, die den Post-Editing-Aufwand verringern, die MÜ-Qualität steigern und somit zur Effizienzsteigerung professioneller Übersetzungsprozesse sowie zur besseren Nutzbarkeit von MÜ für nicht-professionelle Anwender*innen beitragen. Die Ergebnisse zeigen, dass alle untersuchten Pre-Editing-Regeln zu einer Qualitätsverbesserung führen, insbesondere solche, die auf Klarheit, Einfachheit und präzisen Sprachgebrauch abzielen. Als besonders wirkungsvoll erwiesen sich im Fachgebiet Astronautik und der Sprachkombination Englisch-Deutsch das Explizieren von problematischen Ausdrücken, Metaphern und Ellipsen sowie die Verwendung präziser Fachbegriffe und einfacher Verben, die als konkrete Optimierungsstrategien im professionellen, wissenschaftlichen und privaten Gebrauch dienen können.

Abstract

The growing use of machine translation in both professional and private contexts, calls for new strategies to ensure translation quality. This master's thesis explores how targeted pre-editing rules, i.e., small adjustments made to the source text before machine translation, can improve translation quality in the field of astronautics for the English-German language pair. Based on an error typology and existing research on MTranslatability, the study analyzes which specific pre-editing rules lead to quality improvements compared to unedited source texts. The results are evaluated by both translation experts and non-experts using a Best-Worst Scaling method to reflect different perspectives on translation quality and user needs. The goal is to identify effective pre-editing rules that reduce post-editing effort, enhance machine translation quality, and thus increase the efficiency of professional translation processes while also improving usability for non-professional users. The findings show that all tested pre-editing rules improve translation quality, especially those focused on clarity, simplicity, and precise language use. Particularly effective in the context of astronautics and the English-German language combination were strategies such as clarifying ambiguous expressions, metaphors, and ellipses, as well as using precise technical terms and simple verbs. These can serve as effective optimization strategies for professional, academic, and personal use.