



universität
wien

MASTERARBEIT/ MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Wahrnehmung der Qualität maschineller
Übersetzungen im Vergleich zu Humanübersetzungen
durch Lai*innen und professionelle Translator*innen“

verfasst von / submitted by

Carina Dietl, BA

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

Wien, 2022 / Vienna 2022

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:

Ass.-Prof. Mag. Dr. Dagmar Gromann, BSc

Inhaltsverzeichnis

1. Einleitung	5
2. Qualitätsmessung in der Translationswissenschaft	8
2.1 Skopostheorie in der Translationswissenschaft	8
2.2 Funktionales Übersetzen nach Nord	10
2.3 Qualität nach Fehlerkategorien	11
2.4 Bewertung von Fachübersetzungen nach Ahrend.....	12
2.5 Verständlichkeitsmodell nach Göpferich.....	13
2.6 Normen in der Translationspraxis.....	16
2.6.1 ISO 17100	16
2.6.2 ISO 9001	16
2.7 Qualitätsabzüge bei professionellen Translator*innen	17
3. Maschinelle Übersetzungen	18
3.1 Grad der Bearbeitung der maschinellen Übersetzung.....	18
3.1.1 Pre-Editing	18
3.1.2 Rohübersetzung.....	19
3.1.3 Post-Editing.....	20
3.2 Arten und Entwicklung der maschinellen Übersetzung.....	21
3.2.1 Regelbasierte maschinelle Übersetzung.....	21
3.2.2 Statistische maschinelle Übersetzung	22
3.2.3 Neuronale maschinelle Übersetzung.....	23
3.3 Frei verfügbare benutzer*innenfreundliche maschinelle Übersetzungssysteme	24
3.3.1 SYSTRAN.....	24
3.3.2 Google Translate	25
3.3.3 Microsoft Bing	26
3.3.4 DeepL	26
4. Qualitätsbewertung maschineller Übersetzung	28
4.1 Maschinelle Evaluierung nach Metriken	28
4.1.1 ALPAC-Report.....	28
4.1.2 BLEU	28
4.1.3 METEOR	29
4.1.4 Die Qualitätsmetrik MQM	29
4.2 Humanauswertung maschineller Übersetzungen.....	31
4.2.1 Auswertung nach Fehlerkategorien.....	32
4.2.2 Bewertung nach dem Zweck (Skopos) der maschinellen Übersetzung	34
5. Aktueller Stand der Forschung	37

6. Methodik	40
6.1 Bewertungsbasis Accuracy, Fluency und Style	41
6.2 Zielgruppe der Umfrage	42
6.3 Auswahl des Übersetzungstools	43
6.4 Der Fragebogen	44
6.4.1 Demographische Daten	44
6.4.2 Textbewertung	45
6.4.3 Abschließende Fragen	45
6.5 Textauswahl und Begründung	45
6.5.1 Text 1: Humanübersetzung	46
6.5.2 Text 2: Maschinelle Übersetzung	46
6.5.3 Text 3: Maschinelle Übersetzung	46
6.5.4 Text 4: Humanübersetzung	47
6.5.5 Text 5: Humanübersetzung	48
6.5.6 Text 6: Maschinelle Übersetzung	48
6.5.7 Text 7: Humanübersetzung	49
6.5.8 Text 8: Maschinelle Übersetzung	49
6.5.9 Text 9: Maschinelle Übersetzung	50
6.6 Auswertung	50
7. Ergebnisse	51
7.1 Demographische Daten	51
7.2 Bewertung der Texte	53
7.2.1 Gesamtbewertung der Texte	54
7.2.2 Korrekte Erkennung der maschinellen Übersetzung	56
7.2.3 Gesamtübersicht über die Einschätzung von Lai*innen und professionellen Translator*innen	60
7.2.4 Schwierigkeitsgrad für die Proband*innen bei der Bewertung der Texte	61
7.2.5 Wichtigkeit der Aspekte Inhalt, Korrektheit und Stil für die Teilnehmer*innen...	62
7.3 Nutzung der maschinellen Übersetzungssysteme	63
7.4 Detailergebnisse nach Alter und Sprachkompetenz	64
7.4.1 Unterschiede in den Altersgruppen	64
7.4.2 Alter der professionellen Translator*innen	65
7.4.3 Alter der Lai*innen	66
7.4.4 Unterschiede in der Sprachkompetenz	67
7.5 Anmerkungen zur Qualität maschineller Übersetzungen	69
8. Diskussion	71
9. Fazit	75

10. Literaturverzeichnis	77
11. Anhang	81
Textauswahl	81
Der Fragebogen	90
Abstract (Deutsch).....	98
Abstract (English)	99

Abbildungsverzeichnis:

Abbildung 1: Verständlichkeitsmodell nach Göpferich (2006: 155)	14
Abbildung 2: Das magische Dreieck PMQS (2022)	17
Abbildung 3: DeepL Screenshot vom 8.2.2021	26
Abbildung 4: In welcher Branche sind Sie tätig?	52
Abbildung 5: Deutsch als Erstsprache.....	53
Abbildung 6: Englisch als Erstsprache.....	53
Abbildung 7: Schwierigkeitsgrad Texterkennung.....	62
Abbildung 8: Schwierigkeit Texterkennung	62

Tabellenverzeichnis:

Tabelle 1: Gesamtbewertung der Texte als Durchschnittswert auf Basis der drei Kategorien Accuracy, Fluency und Style und einer Skalenbewertung von 1-5	54
Tabelle 2: Durchschnittliche Gesamtbewertung der Humanübersetzungen	55
Tabelle 3: Durchschnittliche Gesamtbewertung der maschinellen Übersetzungen	55
Tabelle 4: Häufigkeit der erkannten maschinellen Übersetzungen.....	56
Tabelle 5: Häufigkeit der erkannten Humanübersetzungen	57
Tabelle 6: Gesamtergebnis MÜ und HÜ	58
Tabelle 7: Gesamtergebnis Lai*innen (MÜ)	58
Tabelle 8: Gesamtübersicht Lai*innen (HÜ)	58
Tabelle 9: Handelt es sich bei MÜ um HÜ?	59
Tabelle 10: Handelt es sich bei HÜ um HÜ	59
Tabelle 11: Gesamtübersicht Lai*innen.....	61
Tabelle 12: Einschätzung Inhalt, Korrektheit und Stil	63
Tabelle 13: Unterschiede Altersgruppen	64
Tabelle 14: Altersanteil professionelle Translator*innen	66
Tabelle 15: Altersanteil Lai*innen	67
Tabelle 16: Unterschiede Sprachkompetenz Gesamtteilnehmer*innen.....	68
Tabelle 17: Erkennung der MÜ nach Sprachkompetenz.....	69

1. Einleitung

Diese Arbeit widmet sich dem Qualitätsvergleich der neuronalen maschinellen Übersetzung (NMÜ) und der Humanübersetzung (HÜ). In den letzten Jahren ergaben sich rasante Entwicklungen im Bereich der maschinellen Übersetzung. Mittels künstlicher Intelligenz und neuronaler Netze können Maschinen übersetzen und auch lernen, sich zu verbessern (vgl. Schmalz 2019: 194ff.). Je nach Kontext erzielen frei verfügbare und benutzer*innenfreundliche NMÜ-Systeme von Google, DeepL etc. in einigen Fällen weniger gute Ergebnisse, während sie in anderen Situationen sehr gute liefern. Außerdem werden sie im Alltag von vielen Menschen verwendet, da etliche Tools online frei zugänglich sind.

Besonders Lai*innen, im Folgenden immer Personen, die nicht Translation studierten, verwenden diese tagtäglich, was bedeutet, dass maschinelle Übersetzungssysteme wie DeepL, SYSTRAN und Google Translate ihren Platz in der Gesellschaft gefunden haben. Dabei stellt sich allerdings die Frage, wie qualitativ hochwertig die ausgegebenen Übersetzungen dieser Systeme sind. Die Antwort darauf kann einerseits durch eine Einschätzung von Translator*innen und andererseits auch von Lai*innen gegeben werden. Bisher wurden zwar einige Studien zur Qualitätseinschätzung maschinell übersetzter Texte durchgeführt (vgl. Lusicky & Heinisch 2019 / Fiederer & O'Brien 2009). Die Teilnehmer*innen dieser Studien setzten sich allerdings hauptsächlich aus professionellen Translator*innen oder Student*innen der Translationswissenschaft zusammen. Da jedoch vor allem Lai*innen zu den Adressat*innen der Texte gehören, ist es in jedem Fall aufschlussreich, wie sie die Qualität maschineller Übersetzungen einschätzen.

Um dies tun zu können, muss zunächst ein für die Translationswissenschaft gültiger Qualitätsbegriff definiert werden, damit später auch festgesetzt werden kann, was ein gutes Translat ausmacht. In Kapitel 2 werden daher Qualitätsbegriffe aus der Translationswissenschaft vorgestellt und verglichen, ebenso wie Ansätze aus der Sprachindustrie. Dabei werden die aus der Sprachindustrie stammenden Normen ISO 17100 (Qualitätsstandard ISO 17100: 2020) und ISO 9001 (Qualitätsmanagement ISO 9001: 2020) näher beschrieben.

Im nächsten Kapitel wird die neuronale maschinelle Übersetzung (NMÜ) behandelt. Hier wird ein Überblick zum aktuellen Stand der Forschung geschaffen und anschließend auf die Verwendung von NMÜ durch Lai*innen und Translator*innen eingegangen. Dies ist nötig,

um die Qualität der Translate in weiterer Folge zu relativieren, denn je nach Kontext muss nicht jeder Text perfekt sein (vgl. Way 2018: 4ff.).

Weiters werden verschiedene Qualitätsbewertungsmethoden vorgestellt, die nur für Humanübersetzungen oder maschinelle Übersetzungen verwendet werden. Im empirischen Teil wird eine Umfrage erstellt, die sich auf die Parameter *Accuracy*, *Fluency* und *Style* nach Hutchins und Somers (1993) stützt. Dabei werden die Proband*innen die Qualität der maschinell erstellten Übersetzungen und der Humanübersetzungen anhand dieser genannten Parameter bewerten, ohne dabei zu wissen, um welche Art der Übersetzung es sich jeweils handelt. Die Teilnehmer*innen bestehen einerseits aus Lai*innen und andererseits aus professionellen Translator*innen.

Da Lai*innen, wie bereits erwähnt, den Großteil der von professionellen Translator*innen oder maschinellen Übersetzungen übersetzten Texte rezipieren, ist es sehr aufschlussreich, sie in die Forschung miteinzubeziehen. Einerseits kann so festgestellt werden, wie gut Lai*innen im Vergleich zu Translator*innen die einzelnen Texte bewerten können und welche Zielgruppe tatsächlich häufiger erkennt, ob es sich um eine maschinelle oder um eine Humanübersetzung handelt. Andererseits kann auch herausgefunden werden, auf welche Parameter Lai*innen im Vergleich zu professionellen Translator*innen besonders achten. Somit profitieren in weiterer Folge auch professionelle Translator*innen bei ihrer zukünftigen Arbeit davon, indem sie sich beim Übersetzen und der Revision auf die eruierten, den Lai*innen wichtige Parameter fokussieren.

Ziel der Arbeit ist in diesem Zusammenhang auch die Beantwortung der folgenden Forschungsfrage: „Wie nehmen professionelle Translator*innen und Lai*innen die Qualität maschineller Übersetzungssysteme im Vergleich zu Humanübersetzungen wahr?“ Dabei soll ein Qualitätsvergleich verschiedener Textsorten aufgestellt werden; und durch die Einschätzung der professionellen Translator*innen und Lai*innen wird analysiert, welche Textsorten sich für maschinelle Übersetzungen eignen – und welche eher weniger. Weiters soll herausgefunden werden, ob Lai*innen die Qualität der Translate anders bewerten als professionelle Translator*innen.

Die Ergebnisse sind für die Translationswissenschaft relevant, da Lai*innen bisher wenig Beachtung in der Forschung zu maschineller Übersetzung geschenkt wurde. Hierbei kann aufschlussreich sein, ob es einen Unterschied bei der Bewertung von Lai*innen und von professionellen Translator*innen gibt. der Lai*innen und professionellen Translator*innen

gibt. Wie bereits erwähnt, können auch professionelle Translator*innen von den Ergebnissen profitieren, da sie ein Bild zur Eignung von bestimmten Textsorten für maschinelle Übersetzung bekommen und im Gegensatz dazu auch erfahren, welche Textsorten eher einer Humanübersetzung bedürfen. Darüber hinaus ist es für sie – wie erwähnt – ebenso wertvoll zu wissen, was ihrem Zielpublikum, also den Lai*innen, an den Translaten bezüglich Qualität besonders wichtig ist.

Letztendlich können die Ergebnisse aber auch für Lai*innen relevant sein. Einerseits können sie durch die Teilnahme an durchgeführten Studien einen wesentlichen Beitrag zur Verbesserung von Übersetzungen leisten und andererseits kann es auch für sie aufschlussreich sein, was eine gute Übersetzung aus Sicht von professionellen Translator*innen ausmacht.

Dabei gilt es herauszufinden, welche Übersetzungen besser oder schlechter bewertet werden und ob bzw. welchen Unterschied es für Translator*innen und Lai*innen zwischen Humanübersetzungen und NMÜ gibt. Die Hypothese dabei ist, dass Lai*innen seltener erkennen, wann es sich um eine Humanübersetzung handelt, und sie die Qualität von maschinellen Übersetzungssystemen aufgrund von fehlendem Vorwissen und Sprachgefühl generell höher einstufen.

2. Qualitätsmessung in der Translationswissenschaft

Das folgende Kapitel umfasst die Qualitätsmessung von Humanübersetzungen und maschineller Übersetzung. In den folgenden Kapiteln werden einige wesentliche Theorien zur Qualitätsmessung näher beschrieben, die als Basis für die Umfrage im Methodikteil dienen. Dabei werden sowohl holistische Ansätze als auch Ansätze, die sich auf den Text und dessen Fehler beziehen, erläutert. Die hier beschriebenen Arten der Qualitätsmessung haben keinen Anspruch auf Vollständigkeit, da es noch einige weitere Möglichkeiten gibt, um Qualität zu messen.

Zunächst ist es notwendig, den Begriff Qualität für Übersetzungsdienstleistungen näher zu definieren. Eine Übersetzung kann nach dem Verständnis von Lai*innen *gut* oder *schlecht gelungen* sein. Für eine wissenschaftliche und fachlich kompetente Übersetzungsbewertung reicht es allerdings nicht aus, Qualität mit *gut* oder *schlecht* zu betiteln, da objektive Bewertungsrichtlinien fehlen. In den folgenden Kapiteln werden also verschiedene wissenschaftliche Ansätze vorgestellt, um den Begriff Qualität näher einzugrenzen. Weiters werden in den Kapiteln 2.6 bis 2.9 Methoden vorgestellt, wie die Qualität maschineller Übersetzungssysteme gemessen werden kann.

2.1 Skopostheorie in der Translationswissenschaft

Obwohl die Skopostheorie (Reiß & Vermeer 1984) auf das Jahr 1984 zurückgeht, ist sie auch heute noch von großer Bedeutung in der Translationswissenschaft. In ihrer Studie über die Qualität maschineller Übersetzung, die in Kapitel 3 vorgestellt wird, stützen sich zum Beispiel Lusicky und Heinisch (2019) auch heute noch auf die Skopostheorie.

Reiß und Vermeer (1984) entwickelten die Skopostheorie als Gegenströmung zur Äquivalenztheorie von Koller, die besagt, dass der Zieltext die gleichen Inhalte möglichst wortgetreu wiedergibt wie der Ausgangstext (vgl. Koller 2011: 218-277). Laut Koller (2011) müssen bei der Übersetzung bestimmte Forderungen nach Äquivalenz erfüllt werden, damit der Zieltext korrekt ist. Dabei gilt es, möglichst vielen Forderungen gerecht zu werden (vgl. Koller 2011: 218-277).

Reiß und Vermeer (1984) hingegen richten ihren Fokus auf den *Skopos*, also den Zweck eines Zieltextes. Sie vergleichen dabei nicht nur Ausgangstext und Zieltext inhaltlich-wörtlich miteinander, sondern betrachten den Text in einer bestimmten Situation (Reiß & Vermeer 1984:

18ff.). Sie sehen die Translation als *Handlung* in einem Kontext und nicht nur als bloße, möglichst wortgetreue Wiederherstellung des Ausgangstextes. Den Text bezeichnen sie dabei als ein *Informationsangebot* (Reiß & Vermeer 1984: 18-19). Dies bedeutet, dass es mehrere Möglichkeiten gibt, einen Text zu übersetzen und dass der/die Translator*in aus diesem sprachlichen Repertoire schöpfen kann (vgl. Reiß & Vermeer 1984: 18ff.).

Sie betrachten diese Handlung als in eine Situation eingebunden, weshalb freier als bei der Äquivalenzforderung übersetzt werden kann. Allerdings kommt es auch hier zu gewissen Einschränkungen bei der Übersetzung, denn nicht in jeder Situation darf alles gesagt, getan bzw. geschrieben werden (vgl. Reiß & Vermeer 1984: 18ff.).

Weiters gibt es laut Reiß und Vermeer nicht den einen Sach- oder Fachtext, sondern der Text wird in einer bestimmten Situation zu einem gewissen Zweck produziert, wodurch sich die translatorischen Entscheidungen ergeben (vgl. Reiß & Vermeer 1984: 29ff.) Es gibt demnach nicht die eine richtige Translation, die sich aus der textbezogenen Äquivalenztheorie ergäbe. Vielmehr kann die Struktur – eines Werbetextes beispielsweise – in der Zielsprache stark verändert werden und dennoch erfüllt der Text seine Funktion sehr gut, da er in einer bestimmten Situation wiedergegeben wird, in der der Inhalt adaptiert wird (vgl. Reiß & Vermeer 1984: 29 ff.). Dabei müssen die Translator*innen feststellen können, welchen Zweck die Translation haben und für welche Situation sie angefertigt werden soll. Dies ist in ihrem Berufsalltag jedoch nicht immer möglich, besonders dann nicht, wenn die Translator*innen einen Ausgangstext von einer Übersetzungsagentur erhalten, der kommentarlos – also ohne Angabe von Zielgruppe, Kontext etc. – verschickt wurde. Als höchstes Maß an Qualität gilt bei Reiß und Vermeer (1984) also die Erfüllung des Skopos. Er hat dabei immer Vorrang vor anderen Merkmalen (vgl. Reiß & Vermeer 1984: 95ff.).

Obwohl die Skopostheorie auf das Jahr 1984 zurückgeht, kann diese auch heute noch sowohl bei Humanübersetzungen als auch bei maschinellen Übersetzungen angewandt werden. Der Skopos gibt auch an, wie hoch die Qualität sein soll. Wenn zum Beispiel zwei Vertragspartner*innen vereinbaren, dass es nur darum geht, den Inhalt eines Vertrags verstehen zu können, dann muss dieser Vertrag nicht publizierfähig und makellos übersetzt werden. Ein solcher Zweck, also jener der bloßen Verständlichkeit in Bezug auf die Qualität der maschinellen Übersetzung, wird im Kapitel Grundlagen der maschinellen Übersetzung noch näher ausgeführt.

2.2 Funktionales Übersetzen nach Nord

Ein ähnlicher Ansatz wie der von Reiß und Vermeer (1984) wird von Nord (2006) verfolgt, die eine wesentliche Vertreterin des funktionalen Übersetzens ist. Sie betrachtet das funktionale Übersetzen als „eine Anwendung der Skopostheorie“ (Nord 2006: 15). Um eine qualitativ hochwertige Übersetzung zu liefern, benötigt es laut ihr nicht nur die Übersetzungskompetenz, sondern auch Sprach- und Sachkompetenz, also ein Grundwissen über das Thema (vgl. Nord 2006: 25 ff.).

Beim funktionalen Übersetzen ist der Zweck das oberste Ziel. Dieser ergibt sich auch hier aus der kommunikativen Situation. Der Zweck kann entweder im Übersetzungsauftrag beschrieben oder aber nicht explizit kommuniziert sein, beispielsweise wenn er in einer bestimmten Kultur Konvention ist. Die Funktion ist dabei zieltextorientiert (vgl. Nord 2006: 11-29).

Texte haben dabei jedoch selten nur eine Funktion. Darüber hinaus gilt zu beachten, dass bereits der Ausgangstext eine Funktion zu erfüllen hatte und seine Empfänger*innen eine Funktion erwarten. Der Text erhält dann erst in der Empfänger*innensituation die Funktion und die Sender*innen intendieren diese. Es liegt jedoch an den Übersetzer*innen, die Funktion zu schaffen (vgl. Nord 2006: 15ff.).

Der Begriff *Qualität* bedeutet bei Nord, wie sehr die angestrebte Intention der Autor*innen bzw. der Auftraggeber*innen erfüllt worden ist (vgl. Nord. 2006: 16ff.). Ein Translat ist also dann qualitativ hochwertig, wenn er die Funktion des Ausgangstextes erfüllt. Wichtig ist dabei, dass diese Funktion nicht erst nach der Übersetzung festgelegt wird, sondern sie bereits vor der Übersetzung existiert hat (vgl. Nord 2006: 11-29).

Nord (2006) wendet sich dabei auch von der Treue zum Ausgangstext ab und führt hingegen den Begriff *Loyalität* ein (vgl. Nord 2006: 16ff.). Dabei sind die Übersetzer*innen allen Beteiligten gegenüber, die mit der Übersetzung zu tun haben, zur Loyalität verpflichtet: dem/der Auftraggeber*in, dem Zielpublikum und dem/der Ausgangstextautor*in. Dabei verpflichtet sich der/die Übersetzer*in, den Auftrag gut zu erfüllen und diesen Personen gegenüber loyal zu sein (vgl. Nord 2006: 11-29). Ein Übersetzungsfehler ist bei Nord folglich nicht auf der sprachlichen Ebene zu finden:

Ein Übersetzungsfehler ist (aus funktionaler Perspektive) also eine Nicht-Erfüllung des Übersetzungsauftrags in bezug [sic] auf bestimmte funktionale Aspekte. Das bedeutet: Jede Übersetzungsleistung kann nur in bezug auf ein vorgegebenes (funktionales) Übersetzungsziel beurteilt werden. (Nord 2006: 17)

Dabei ist wesentlich, dass der/die Übersetzer*in die erwartete Funktion des Zieltextes kennt. Weiters führt Nord eine Gewichtung der Übersetzungsfehler an. Pragmatische Fehler, also Fehler, die die Funktion betreffen, sind dabei am schwerwiegendsten (vgl. Nord 2006: 17ff.). Darüber hinaus sind Abweichungen von kulturspezifischen Verhaltensnormen als Fehler zu rechnen. Sprachliche Fehler sind bei ihr keine Übersetzungsfehler, sondern sogenannte *Performanzfehler*, die daraus resultieren, dass die Zielsprache nicht ausreichend beherrscht wird (vgl. Nord 2006: 17ff.). Die Beherrschung der Zielsprache setzt sie jedoch bei professionellen Übersetzerinnen und Übersetzern voraus (vgl. Nord 2006: 11-29).

2.3 Qualität nach Fehlerkategorien

Eine andere Möglichkeit, Übersetzungen zu bewerten, ist der Einsatz von Fehlerkategorien, die einerseits textbezogen oder andererseits ganzheitlich sein können. Diesbezüglich wurden in den letzten Jahren bereits etliche unterschiedliche Kategorien erarbeitet (vgl. Mossop 2020, Hansen 2008: 313-327).

Mossop (2020) führt einen praktischen Ansatz für Translator*innen ein, um ihre Fähigkeiten zu verbessern. Sein Ansatz ist jedoch auch für Lektor*innen gedacht und enthält ein Bezugsraster. Für die Revision stellte er 14 Punkte auf, die es bei der Korrektur zu berücksichtigen gilt (Mossop 2020: 137ff.). Zu den Oberbegriffen zählt er Transferprobleme, inhaltliche Probleme, sprachliche Probleme, Probleme in Bezug auf die Präsentation des Textes (typographischer Aufbau und Übersichtlichkeit der Struktur des Textes) und Probleme der Spezifizierung (in Bezug auf Kund*innenwünsche) (vgl. Mossop 2020: 137 ff.).

Einen ähnlichen Ansatz bietet Hansen (2008). Sie erstellte ebenso ein Modell, das von Lektor*innen und Revisor*innen verwendet werden kann. Dabei beziehen sich die Fehlerkategorien nicht ausschließlich auf die sprachliche Ebene (vgl. Hansen 2008: 316ff.) Hansen führt sowohl sprachliche Fehler, wie z. B. Terminologie, Idiomatik, Stil und Orthographie, aber auch textexterne Faktoren, wie Situation und Zeit, Sender*in und Empfänger*in, an. Auch sie verfolgt das Ziel, bei der Optimierung von Translaten zu helfen und ein Raster für Revisor*innen anzubieten (vgl. Hansen 2008: 313-327.).

Delizée (2011) schlägt Fehlerkategorien für die Didaktik in der Translationswissenschaft vor. Diese entwarf sie speziell für Translationsstudent*innen des Schwerpunkts Fachübersetzen. Die Kategorien umfassen unter anderem Sprachkenntnisse des Ausgangs- und des Zieltextes, Translationskenntnisse und Fachkenntnisse (vgl. Delizée 2011:

11ff.). Zusätzlich zu den Fehlern, die zu Punkteabzügen führen, gibt es bei Delizée auch Pluspunkte für besonders gelungene Übersetzungen in diesen Bereichen. Dabei ist jedoch darauf zu achten, wie gravierend einzelne Fehler sind (vgl. Delizée 2011: 12ff.).

So kann ein bestimmter Fehler in einer Kategorie viel gravierender sein als zehn Fehler in einer anderen. Beispielsweise wäre es ein fataler Übersetzungsfehler, wenn in einem Beipackzettel stehen würde, dass ein Medikament zehnmal täglich anstelle von einmal täglich eingenommen werden müsse – auch wenn es sich höchstwahrscheinlich um einen Tippfehler und keinen translatorischen Fehler handeln würde (vgl. Delizée 2011: 9-25).

2.4 Bewertung von Fachübersetzungen nach Ahrend

Ahrend (2006) stellt zur Diskussion, wann eine Übersetzung als *richtig* oder *falsch* bzw. *gut* oder *schlecht* gilt. Akademisch betrachtet ist eine Übersetzung seiner Ansicht nach richtig, wenn sie vollständig und genau ist (vgl. Ahrend 2006: 31ff.). Er unterscheidet hierfür allgemeine und praxisorientierte Kriterien. Bei den allgemeinen Kriterien ist wichtig, dass die Übersetzer*innen die Ausgangs- und Zielsprache sehr gut beherrschen, eine verständliche Übersetzung liefern, sowie das Thema kennen und auch Mut zur freien Übersetzung haben (vgl. Ahrend 2006: 31).

Bei den praxisorientierten Kriterien geht auch er weg von den textuellen Merkmalen. Für ihn ist wichtig, was der/die Autor*in des Ausgangstexts tatsächlich meinte (vgl. Ahrend 2006: 32 ff.). Dies ist in der Praxis nicht immer möglich, da oft kein direkter Kontakt zum/zur Autor*in möglich ist. Dies gilt besonders dann, wenn man den zu übersetzenden Text als Übersetzer*in einer Agentur erhält. Allerdings besteht immer die Möglichkeit, nachzufragen bzw. in jedem Fall Kommentare zu vermerken (vgl. Ahrend 2006: 40 ff.).

Ein weiterer Punkt ist, dass sich eine Übersetzung nicht wie eine Übersetzung lesen soll. Der Text soll wie ein Original klingen – sich also flüssig und ohne stilistische, kulturelle oder sprachliche Stolpersteine lesen (vgl. Ahrend 2006: 33). Als dritten Punkt nennt er, dass es wesentlich ist, dass es überhaupt eine Übersetzung gibt. Dies bedeutet, dass sich die Leser*innen des Zieltexts ein Bild davon machen können, worum es im Text geht (vgl. Ahrend 2006: 33ff.). Hierbei ist die oben angeführte akademische Qualität nicht notwendig, sondern es geht nur um ein erstes Erkennen des Inhalts (vgl. Ahrend 2006: 33ff.). Um dies zu erfüllen, ist keine professionelle Humanübersetzung nötig. Laut Ahrend wäre hier die maschinelle

Übersetzung ausreichend, um Texte zu erschaffen, die das Qualitätsmerkmal *gut genug* haben (vgl. Ahrend 2006: 31-42).

Als Beispiel einer besonderen Herausforderung für Übersetzer*innen führt er Rechtstexte der EU an, denn die Übersetzungen gelten in den jeweiligen Zielsprachen bzw. -staaten als eigenes Original. Der Zweck der Übersetzung ist hier also ein Original anzufertigen (vgl. Ahrend 2006: 36).

Als letzten Punkt fragt auch er nach dem Übersetzungszweck (vgl. Ahrend 2006: 34ff.). Als Übersetzer*in sollte man möglichst genau darüber Bescheid wissen, wozu die beauftragte Übersetzung genau benötigt wird. Es können sich hier große Unterschiede in der Gestaltung des Textes ergeben, wenn Übersetzungen veröffentlicht werden und die Namen der Übersetzer*innen aufscheinen sollen, oder aber wenn eine Übersetzung eine vergleichsweise sehr kurze Lebensdauer haben soll, weil sie in einem Unternehmen nur intern und für kurze Zeit benutzt wird (vgl. Ahrend 2006: 34-37).

Für eine sehr kurze Lebensdauer übersetzte Texte weisen oft die Qualität *gut genug* auf, d. h., sie ist nicht so hoch wie bei einer publizierfähigen Übersetzung (vgl. Ahrend 2006: 35ff.). Dennoch hat das im Vergleich zur Fachübersetzung *schlechtere* (maschinell) erstellte Translat seine Berechtigung und erfüllt seinen Zweck vollständig. Auch dieser Aspekt dient der Einschätzung der Qualität von maschinell erstellten Übersetzungen (vgl. Ahrend 2006: 31-42).

2.5 Verständlichkeitsmodell nach Göpferich

Zur Bewertung von Qualität hat Göpferich (2008: 291ff.) ein Verständlichkeitsmodell entwickelt. Dies ist angelehnt an das von Schulz von Thun (1981) erarbeitete Verständlichkeitskonzept der Instruktionspsychologie. Das Modell von Göpferich ist in Abb. 1 dargestellt und rückt nicht nur den Ausgangstext und den Zieltext in den Mittelpunkt, sondern vor allem auch die Zieltextrezipient*innen, d. h., die Personen, die den Text lesen. Dabei löst sich Göpferich vom Ausgangstext und bewegt sich zum Zieltext hin. Für sie zählt besonders der Faktor, wie gut eine Information aufgenommen wird (vgl. Göpferich 2008: 298ff.).

Verständlichkeit liegt im Zieltext also nicht per se vor, sondern wird erst von den Leserinnen durch deren Vorwissen und mentale Modelle konstruiert (vgl. Göpferich 2008: 298ff.). Es geht vorrangig nicht darum, ob einzelne Sätze im Zieltext grammatikalisch korrekt sind, sondern um die Frage, inwieweit der Sinn für die Leser*innen erfassbar ist (vgl. Göpferich 2008: 300ff.).

Autor*innen und Translator*innen sollen sich auf das Vorwissen der Zieltextrezipient*innen konzentrieren, um einen Text zu produzieren, der inhaltlich gut verstanden wird. Die Leser*innen wiederum erzeugen beim Rezipieren ein sogenanntes „mentales Denotationsmodell“ (Göpferich 2008: 298). Das bedeutet, dass sie aufgrund ihres Vorwissens ein bestimmtes Bild vor ihrem inneren Auge abrufen, wenn sie einen Text lesen – und dieses Bild ist im besten Fall dasselbe, das der/die Autor*in schaffen wollte (vgl. Göpferich 2008: 298ff.). Schulz von Thuns Verständlichkeitskonzept (1981) umfasst die vier Bereiche *Gliederung*, *Prägnanz*, *Motivation* und *Simplizität*, zu denen Göpferich zwei weitere Dimensionen zur Bewertung von Translaten hinzufügte, nämlich *Korrektheit* und *Perzipierbarkeit* (vgl. Göpferich 2008: 296ff.). Göpferich grenzt sich auch von den Termini im Sinne der Instruktionspsychologie ab, da diese laut ihr sehr vage definiert seien (vgl. Göpferich 2008: 295).

Abb. 1. stellt eine Übersicht der verschiedenen Verständlichkeitsebenen dar.

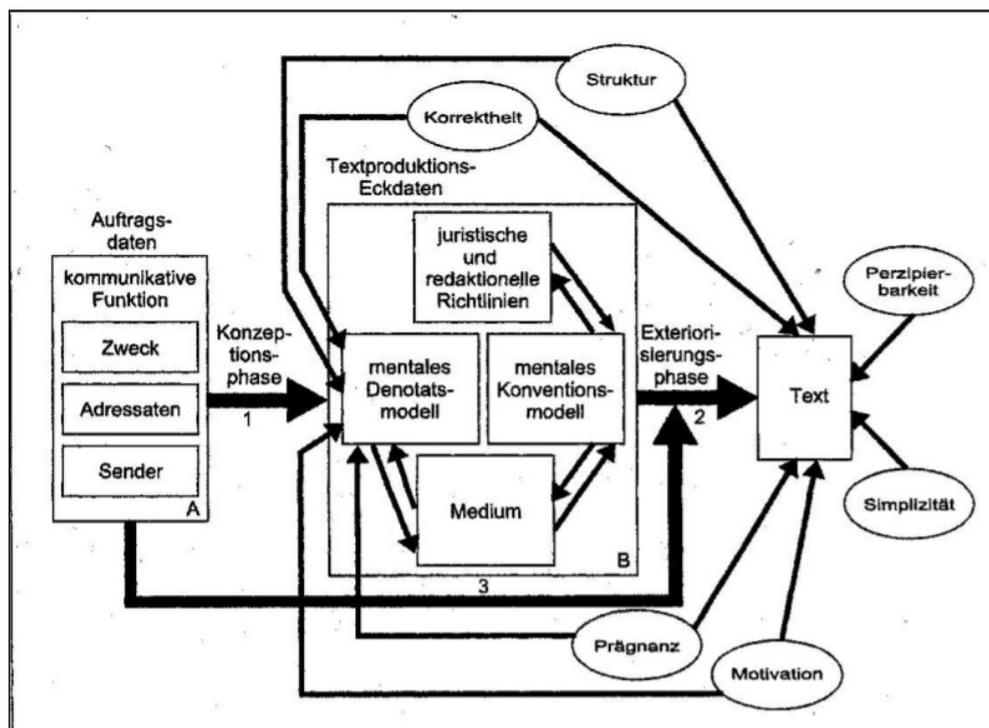


Abbildung 1: Verständlichkeitsmodell nach Göpferich (2006: 155)

Unter *Prägnanz* versteht Göpferich, dass das mentale Denotationsmodell bei den Zieltextrezipient*innen erfolgreich ist und die Kodierung im Text stimmt (vgl. Göpferich 2008: 303). Dabei muss der Kommunikationsrahmen stimmen und auch die weiteren fünf Bereiche erfüllt sein. Die *Prägnanz* ist außerdem dann erfüllt, wenn der Text mit den am wenigsten möglichen Wörtern auskommt, um das Denotationsmodell gut zu transportieren (vgl. Göpferich 2008: 303).

Göpferich unterscheidet zudem vier Verstöße gegen die *Prägnanz*, nämlich zu wenig oder zu viel Information für das Denotationsmodell, zu lange oder zu kurze Formulierungen, Tautologien (doppelte Wiedergabe des Sinns innerhalb einer Phrase) und Redundanzen (doppelte Sinnwiedergabe an anderer Stelle im Text, die überflüssig ist) (vgl. Göpferich 2008: 291-314).

Unter *Korrektheit* versteht sie die sprachliche Korrektheit, aber auch die Richtigkeit in Bezug auf die Einschätzung des Vorwissens der Leser*innen. Auch hier muss man also die Zielgruppe kennen, für die man einen Text formuliert (vgl. Göpferich 2008: 304).

Als nächsten Punkt nennt sie die *Motivation*. Hierunter fällt einerseits die Motivation, einen Text zu lesen, aber zeitgleich soll eine gewisse Motivation auch vom Text ausgehen. Er weckt einerseits das Interesse der Empfänger*innen und hält dieses Interesse während sie den Text lesen. Eine Möglichkeit, die Motivation zu wecken, besteht laut Göpferich darin, Beispiele anzuführen, die aus dem Alltag der Leser*innen stammen (vgl. Göpferich 2008: 304ff.).

Unter *Struktur* versteht Göpferich nicht die optische Gliederung, sondern allein die inhaltliche Strukturierung. Sie unterscheidet dabei in Makro- und Mikrostruktur. Unter Makrostruktur versteht sie die inhaltliche Gliederung in Absätzen und größere Einheiten und unter Mikroebene versteht sie den Sinn, der nicht über zwei Sätze hinausgeht (vgl. Göpferich 2008: 305-306).

Simplizität bedeutet, dass die Wortwahl und die Syntax so einfach wie möglich sind. Dabei ist eine gewisse Konsistenz in Lexik und Syntax wichtig, und es gilt, Fachsprache, Komposita und Nominalstil zu vermeiden (vgl. Göpferich 2008: 307ff.).

Unter *Perzipierbarkeit* versteht Göpferich ausschließlich optische Signale, die die Leser*innen ansprechen. Dabei nennt sie unter anderem Aufzählungszeichen, die in einigen Fällen zur Leserlichkeit des Textes beitragen (vgl. Göpferich 2008: 308ff.).

Durch den Zusammenhang der angeführten Punkte mit dem Denotationsmodell schafft sie eine Verbindung von textuellen und außertextuellen Faktoren sowie Denkmodellen. Dabei bleiben Ausgangs- und Zieltext zwar im Fokus, allerdings richtet sich ihr Ansatz auch auf die situationsbedingten Faktoren, wie etwa das Vorwissen der Leser*innen. Da es hauptsächlich darum geht, den Zieltext richtig zu verstehen, wird ihm mehr Bedeutung als dem Ausgangstext beigemessen. Das Modell kann sowohl bei einsprachigen Texten als auch beim Übersetzen angewendet werden und ist auch für die Bewertung von maschinellen Übersetzungen geeignet (vgl. Göpferich 2008: 296-312).

2.6 Normen in der Translationspraxis

Im folgenden Kapitel werden die Qualitätsnormen ISO 17100 und ISO 9001 und deren wesentliche Bedeutung für die Bewertung der Qualität von Translaten näher beschrieben.

2.6.1 ISO 17100

In der Translationspraxis findet die ISO 17100: 2015 als jene Norm Anwendung, die von zahlreichen Übersetzungsbüros verwendet wird (vgl. Qualitätsstandard ISO 17100: 2015). Diese Norm umfasst Prozesse, die eingehalten werden müssen, damit die Qualität einer Übersetzung gewährleistet werden kann. Dabei werden Begriffe wie *Übersetzen* (vgl. Qualitätsstandard ISO 17100: 2015: 16) klar definiert, die nötigen Kompetenzen festgelegt und die Arbeitsschritte, die die Qualität sicherstellen sollen, vorgegeben. Beim Übersetzen müssen Kategorien eingehalten werden, wie zum Beispiel Syntax, Orthographie und semantische Korrektheit, aber auch die Formatierung. Zusätzlich muss auf die Zielgruppe und den Zweck der Übersetzung geachtet werden (vgl. Qualitätsstandard ISO 17100: 2015: 16).

2.6.2 ISO 9001

Zusätzlich existiert die Qualitätsnorm ISO 9001, anhand derer die Qualität einer Übersetzung gemessen werden kann. Diese Norm wurde ursprünglich hinsichtlich der Qualitätsanforderungen innerhalb von Unternehmen erstellt. Durch ihre allgemein formulierten Inhalte kann sie jedoch auch für die Übersetzungspraxis verwendet werden. Bei der Definition von Qualität handelt es sich nicht um eine klar abgegrenzte Definition, die am Übersetzungsmarkt zu sehr einschränken würde, sondern die Qualität wird definiert als „Grad, in dem ein Satz inhärenter Merkmale eines Objekts [...] *Anforderungen* [...] erfüllt“ (Qualitätsmanagement ISO 9001:2015: 11). Diese Anforderungen können je nach Übersetzungsauftrag mit den Kund*innen vereinbart werden. Aufgrund der schriftlichen Festlegung dieser Anforderungen ist es später leichter festzustellen, ob der Faktor *Qualität* erfüllt worden ist oder nicht. Die nachstehenden Kapitel beziehen sich auf den Qualitätsbegriff der ISO 9001:2015, wenn über Qualität geschrieben wird (vgl. Qualitätsmanagement ISO 9001).

2.7 Qualitätsabzüge bei professionellen Translator*innen

Professionelle Übersetzer*innen sind oft geprägt von Fehlerkategorien, auf die sie beim Erstellen von Texten in ihrer Ausbildung achten mussten, und legen daher besonderen Wert darauf, in dieser Hinsicht qualitativ hochwertige Texte zu liefern. Besonders im Unterricht wird der Qualitätsstandard „gut genug“ (Bowker 2020: 462), der bei Texten mit kurzer Lebensdauer oft ausreichend ist, nicht vergeben. Auch die Berufsverbände – beispielsweise UNIVERSITAS in Österreich – zeichnen sich dadurch aus, dass ihre Mitglieder sehr hochwertige Translate liefern, also solche Übersetzungen, die weit über „gut genug“ hinausgehen (vgl. UNIVERSITAS).

Dennoch gibt es Umstände, unter denen von Translator*innen erwartet wird, eine Übersetzung zu liefern, die qualitativ nicht hochwertig sein und/oder maschinell erstellt werden soll. In diesem Zusammenhang spiegelt das aus dem Projektmanagement stammende magische Dreieck die Faktoren *Zeit*, *Kosten* und *Qualität* wider, wie nachstehend in Abb. 2. dargestellt.



Abbildung 2: Das magische Dreieck PMQS (2022)

Das Dreieck ist so zu verstehen, dass der/die Translator*in dann eine qualitativ hochwertige Übersetzung liefern kann, wenn er/sie mehr Zeit zur Verfügung hat. Diese hat dann allerdings auch ihren Preis. Falls Kund*innen eine günstigere Übersetzung wünschen, leidet in den meisten Fällen die Qualität darunter (vgl. Bowker 2020: 453ff.). Das bedeutet, wenn die Übersetzungspreise von selbständigen Translator*innen den Kund*innen zu hoch sind, dann wenden sie sich oft an eine günstigere Agentur. Doch dabei muss ihnen bewusst sein, dass das Translat eine niedrigere Qualität haben kann. Es besteht darüber hinaus die Möglichkeit, dass

Translator*innen entsprechend den Preisanforderungen von Kund*innen selbst Übersetzungen anfertigen, die ihrer Meinung nach qualitativ nicht hochwertig sind (vgl. Bowker 2020: 459ff.).

Sollte letzteres mit den Kund*innen explizit vereinbart sein, gilt auch in diesem Fall die Erfüllung des Prädikats *gut genug* (vgl. Bowker 2020: 462) und beide Vertragspartner*innen sind zufrieden. Wenn sehr wenig Zeit zur Verfügung steht, können oder müssen Translator*innen auf maschinelle Übersetzungssysteme zurückgreifen. Dies ist in der Regel mit ihren Kund*innen vereinbart, ebenso, dass der knappe Liefertermin der beauftragten Übersetzung kein qualitativ hochwertiges Translat ermöglicht (vgl. Bowker 2020: 459-463).

3. Maschinelle Übersetzungen

Durch die sehr hohe Datenmenge, die derzeit verfügbar ist, ist es praktisch nicht möglich, alle existierenden Texte von (professionellen) Translator*innen übersetzen zu lassen. Im Jahr 2012 wurde Google Translate von mehr als 200 Millionen Internetnutzerinnen und -nutzern in Anspruch genommen, was einer Menge von einer Million Büchern und ca. 75 Milliarden Wörtern pro Tag entspricht (vgl. Way 2018: 3). Die maschinelle Übersetzung hat somit neben der Humanübersetzung ihren fixen Platz in der Übersetzungsbranche gefunden (vgl. Way 2018: 1-22). Im Folgenden soll ein kurzer Überblick über die Entwicklung der maschinellen Übersetzung bis hin zu ihrer Qualitätsbewertung gegeben werden.

3.1 Grad der Bearbeitung der maschinellen Übersetzung

Texte, die mithilfe der maschinellen Übersetzung übersetzt werden, werden häufig einer zusätzlichen Bearbeitung unterzogen. Way (2013: 2ff.) unterteilt diese Bearbeitungsformen grundsätzlich in: Rohübersetzung, leichtes *Post-Editing* und *Full-Post-Editing*. Weiters findet oft auch das Pre-Editing Anwendung, bei dem ein Text vor seiner Übersetzung bearbeitet wird (vgl. Way 2013: 2ff.).

3.1.1 Pre-Editing

Unter *Pre-Editing* wird ein Bearbeitungsschritt verstanden, der noch vor der eigentlichen, maschinellen Übersetzung eines Textes erfolgt. *Pre-Editing* umfasst Veränderungen des Ausgangstextes, damit dieser leichter und korrekter übersetzt werden kann. Die Vorbereitung

kann von monolingualen Personen durchgeführt werden (vgl. Popović 2018: 131ff.), wobei es laut Hutchins (2004: 7ff.) auch sinnvoll sei, für diese Aufgabe bilinguale Personen zu beauftragen, da sie Änderungen (zum Beispiel bei Komposita) bereits mit Hinblick auf den entsprechenden Zieltext vornehmen und die Übersetzung damit vorab optimieren können (vgl. Popović 2018: 131ff., Hutchins 2004: 7ff.).

Pre-Editing kann auf verschiedenen Ebenen erfolgen. Hutchins (2004: 8ff) unterscheidet dabei zwischen:

- *Wortebene*: Hier wird bestimmt, um welche Wortart es sich handelt.
- *Eigennamen*: Bei diesen kann festgelegt werden, ob sie unübersetzt bleiben.
- *Komposita*: Auf dieser Ebene wird entschieden, wie ein zusammengesetztes Wort übersetzt werden soll.
- *Zeichensetzung*: Hierbei wird definiert, ob eine bestimmte Zeichensetzung übernommen wird oder nicht.

Weiters können laut Hutchins (vgl. 2004: 3ff.) Sätze entsprechend gekürzt werden, damit diese von der Maschine leichter übersetzt werden. Zusätzlich kann auch mit kontrollierten Sprachen gearbeitet werden. Dabei wird das Vokabular eingegrenzt und die Terminologie vorab festgelegt, was zu einer verbesserten Leistung der maschinellen Übersetzung führt (vgl. Hutchins 2004: 3ff.).

3.1.2 Rohübersetzung

Unter Rohübersetzung im Zusammenhang mit maschineller Translation wird verstanden, dass ein maschinell übersetzter Zieltext nicht vom Menschen bearbeitet wird (vgl. Way 2018: 4). Bei der Rohübersetzung, die für manche Zwecke ausreichend ist, kann davon ausgegangen werden, dass Grammatikfehler enthalten sind und das Translat vor allem stilistische Fehler aufweist (vgl. Way 2018: 4ff.).

Sie wird unter anderem für das sogenannte *Gisting* verwendet, wobei es darum geht, lediglich den groben Sinn des Ausgangstextes zu erfassen. Nach dem *Gisting* kann dann entschieden werden, ob es sich lohnt, den Ausgangstext professionell übersetzen zu lassen. *Gisting* wird laut Way (2018) auch bei der Polizei und bei Überwachungsdiensten eingesetzt, wenn Dokumente nur in einer Fremdsprache vorliegen (vgl. Way 2018: 6). Weiters kann die Rohübersetzung laut Way (vgl. 2018: 4ff.) sehr von Nutzen sein, wenn es darum geht, den Sinn

von kurzen Texten zu erfassen, die zudem keine lange Lebensdauer haben. Way führt hierfür unter anderem E-Mails und Inhalte von Online-Chats an (vgl. Way 2018: 4 ff.).

Im Methodikteil werden ebenso Rohübersetzungen, also Zieltexte, die nicht bearbeitet worden sind, verwendet. Ein Beispiel für Rohübersetzungen, die unbearbeitet an Kund*innen geliefert wurden, sind die kanadischen Wetterberichte, die mit dem System METEO MT in den 1990er Jahren übersetzt wurden (vgl. Bowker 2020: 454 ff.). Diese Übersetzungen waren deshalb sehr erfolgreich, weil sich das Vokabular und die Fachausdrücke bei Wetterberichten auf einen kleinen Teil einer Sprache beschränken. Wie das Beispiel der Wetterberichte zeigt, ist es also sinnvoll, für maschinelle Übersetzungssysteme, die Rohübersetzungen liefern sollen, eine Nische zu finden, da hier die Termini begrenzt sind (vgl. Bowker 2020: 454 ff.).

3.1.3 Post-Editing

Unter *Post-Editing* wird in diesem Zusammenhang die Revision von maschinell übersetzten Texten, also Zieltexten, verstanden. Aufgrund des Pre-Editings von Ausgangstexten und der Verwendung von kontrollierter Sprache wird der Output der maschinellen Übersetzung verbessert und somit auch das *Post-Editing* erleichtert (vgl. Allen 2003: 298). Beim *Post-Editing* kann zwischen unterschiedlichen Graden der Bearbeitung unterschieden werden. Diese reichen vom minimalen *Post-Editing* bis hin zum *Full-Post-Editing* (vgl. Allen 2003: 304-306). Auch hier ist sehr subjektiv, was unter *minimal* zu verstehen ist, und so wird der Grad des *Post-Editings* in der Praxis unterschiedlich realisiert. Dabei gilt, dass (individuelle) Richtlinien eingehalten werden sollen; darüber hinaus braucht man für diese Art der Revision eine Lernphase (vgl. Allen 2003: 304), auch weil schwer zu definieren ist, welche Fehler korrigiert werden sollen.

Professionelle Übersetzer*innen und Revisor*innen tendieren laut Allen (2003: 305) dazu, beim minimalen *Post-Editing* zu viel zu korrigieren. Weiters kann es vorkommen, dass zu wenig korrigiert wird oder auch zu viele kleine Fehler korrigiert werden, aber grobe Fehler im Zieltext bestehen bleiben (vgl. Allen 2003: 305ff.). Beim *Full-Post-Editing* sollen alle Fehler korrigiert werden, damit ein Dokument vollständig überarbeitet beispielsweise an Kund*innen gesendet werden kann. Das *Full-Post-Editing* ist aufgrund der vollständigen Bearbeitung umstritten, da diese sehr viel zeitlichen Aufwand beansprucht, der ebenso gut schon für eine Humanübersetzung verwendet werden könnte. Doch laut Allen ist *Full-Post-*

Editing nach wie vor schneller das Übersetzer*innen, die ohne CAT-Tool arbeiten (vgl. Allen 2003: 306ff.).

Die unterschiedlichen Formen des *Post-Editing* haben auch verschiedene Einsatzgebiete. Rohübersetzungen werden – wie bereits erläutert – für *Gisting* verwendet, bei dem der übersetzte Text nur einen ersten Eindruck vom Thema vermitteln muss (vgl. Way 2013: 2ff.). *Light-Post-Editing* wird für Online-Hilfen, Tutorials und Wissensforen verwendet (vgl. Way 2013: 4ff.). Hierbei liegt der Fokus auf der einfachen grammatikalischen Korrektur (vgl. Way 2013), während das *Full-Post-Editing* für sensible Dokumente und die externe Kommunikation verwendet wird. In diesem Fall muss neben der Grammatik auch der Stil verbessert werden (vgl. Way 2013 4ff., Allen 2003: 297-316).

3.2 Arten und Entwicklung der maschinellen Übersetzung

Was die maschinelle Übersetzung betrifft, so wird im Allgemeinen zwischen drei Arten unterschieden, nämlich der regelbasierten, der statistischen und der neuronalen Übersetzung, die im Folgenden vorgestellt werden.

3.2.1 Regelbasierte maschinelle Übersetzung

Bei der regelbasierten maschinellen Übersetzung wird der Ausgangstext auf mehreren Ebenen analysiert und anschließend der Zieltext unter Einhaltung von lexikalischen Regeln erstellt (vgl. Ruiz 2018: 948). Das entsprechende Regelwerk muss vorab von Linguist*innen erstellt werden und sehr ausführlich sein. Dies bedeutet, dass Linguist*innen sehr viel Zeit und Arbeit dafür aufwenden (vgl. Ruiz 2018: 948ff.).

Zu den Regeln gehören: das Aufteilen des Satzes in Wörter, das Nachschlagen in Wörterbüchern, die syntaktische Analyse, deren Transfer sowie die richtige Satzstellung und ebenso Regeln, um den Zieltext zu produzieren (vgl. Melby 2010: 419ff., Dorothy 2020: 434 ff.). Die regelbasierte maschinelle Übersetzung war in den 1950ern bis 1980ern populär und wurde von der statistischen maschinellen Übersetzung abgelöst (vgl. Melby 2010: 419ff.). Sie existiert allerdings noch immer für Sprachenpaare, in denen es wenig Trainingsmaterial gibt (vgl. Melby 2010: 419-436, Dorothy 2020: 434ff.).

Die regelbasierte maschinelle Übersetzung kann laut Dorothy in drei Arten unterschieden werden (vgl. Dorothy 2020: 434-436). Es gibt die direkte Übersetzung, den

Transfer und die Übersetzung mittels *Interlingua*. Bei der direkten Übersetzung wird Wort für Wort übersetzt, was keine idiomatischen Ergebnisse liefert (vgl. Dorothy 2020: 434ff.). Beim Transfer wird der Ausgangstext zunächst analysiert und anschließend in die Zielsprache transferiert. Dabei werden Zieltextkandidaten mittels der festgelegten Sprachregeln eruiert und erzeugt. Was die Übertragung durch *Interlingua* betrifft, so wird eine Zwischensprache erzeugt (vgl. Dorothy 2020: 435). Hier geht es um die Repräsentation des Inhalts, weshalb der Zieltext keine Spuren des Ausgangstexts mehr enthält. Dieser Ansatz wurde in seinen Anfängen zwar erforscht, aber nie weiterverfolgt (vgl. Dorothy 2020: 428-445).

3.2.2 Statistische maschinelle Übersetzung

Bei der statistischen maschinellen Übersetzung (SMÜ) werden große Korpora verwendet und daraus die Vorschläge, die statistisch am häufigsten vorkommen, als Übersetzungslösung verwendet (vgl. Schmalz 2019: 198). Das bedeutet, es wird ermittelt, wie häufig ein Wort bzw. eine Phrase auf dieselbe Weise übersetzt wurde. Die statistische maschinelle Übersetzung braucht sehr viel Trainingsmaterial in Form von Korpora und Paralleltexten sowohl der Ausgangs- als auch der Zielsprache, das auf Satzebene aligniert wird (vgl. Schmalz 2019: 198). Für eine korrekte Übersetzung kommt es dabei vor allem auch auf den Kontext der zur Verfügung stehenden Texte an. Das bedeutet, wenn Trainingsmaterial zum Beispiel nur im Bereich Politik vorliegt, dann kann das System einzelne Phrasen in anderen Bereichen nicht übersetzen oder die Kollokationen sind nicht korrekt (vgl. Schmalz 2019: 198ff.). Das System ist dafür umso besser und genauer, je mehr Trainingsmaterial beider Sprachen in einem Fachbereich vorliegt.

Weiters bedarf es monolinguale Texte, damit ein möglichst flüssiger Text in der Zielsprache erstellt werden kann (vgl. Schmalz 2019: 203ff.). Probleme ergeben sich auch bei Sprachenpaaren, bei denen sich die Wortreihenfolge unterscheidet (vgl. Ruiz 2018: 956).

Zusammengefasst bedeutet das: Die statistische maschinelle Übersetzung hat eine dreiteilige Architektur. Im ersten Schritt wird die maschinelle Übersetzung mittels Parallelkorpora trainiert. Was den Zieltext angeht, so werden, wie erläutert, monolinguale Korpora verwendet, um ein Zieltextmodell zu erschaffen (vgl. Dorothy 2020: 435-436).

Eine zweite Phase wird als *Tuning* bezeichnet, wobei das Material von Linguist*innen geändert wird (vgl. Dorothy 2020: 435-436). In der dritten Phase, dem *Decoding*, wird daraus ein Zieltext geschaffen (vgl. Dorothy 2020: 436). Dabei wird nicht nur ein Zieltextsatz für einen

Ausgangstext erstellt, sondern tausende Sätze. Der Satz, der statistisch am korrektesten ist, wird dann für den Zieltext ausgewählt.

Die statistische maschinelle Übersetzung eignet sich eher schlechter für Sprachen mit stark flektierten Verben (vgl. Ruiz 2018: 4, Dorothy 2020: 435ff.). Außerdem kann es zu Auslassungen von einzelnen Wörtern oder zu inkonsistenten Übersetzungen (auch innerhalb eines Satzes) kommen (vgl. Dorothy 2020: 435 ff.). Das heißt, dass ein Wort innerhalb eines Satzes oder eines Textes unterschiedlich übersetzt wird. Besonders bei Fachübersetzungen, die eine konsistente Terminologie erfordern, ist die SMÜ daher eher weniger geeignet, außer sie wurde speziell für eine bestimmte Domäne trainiert.

Ein weiterer Nachteil ist, dass die Übersetzungen der SMÜ im Vergleich zur neuronalen maschinellen Übersetzung nicht immer natürlich klingen (vgl. Schmalz 2019: 188ff.). Die statistische maschinelle Übersetzung dominierte den Markt bis in die 2010er Jahre und wurde anschließend von der neuronalen maschinellen Übersetzung als Marktführerin abgelöst (vgl. Schmalz 2019: 198ff.).

3.2.3 Neuronale maschinelle Übersetzung

Die neuronale maschinelle Übersetzung (NMÜ) basiert auf künstlicher Intelligenz (KI). Um zu lernen, verwendet der Computer neuronale Netze, die nach dem Vorbild des menschlichen Gehirns modelliert sind. Beim sogenannten *Deep Learning*, dem Training mit neuronalen Netzen, handelt es sich um ein Verfahren, bei dem die Neuronen des Computers auf zahlreichen Ebenen angeordnet sind, damit der Computer auch schwierigere Verbindungen erkennen kann (vgl. Matthiesen 2017: 27-36). Mittels KI werden tausende Neuronen erstellt, die mit anderen Neuronen verbunden sind (vgl. Dorothy 2020: 436 ff.). Der *gold standard*, also die Messlatte, für eine qualitativ hochwertige Übersetzung, ist dabei die Humanübersetzung. Diese wird als Trainingsmaterial verwendet (vgl. Dorothy 2020: 436ff.).

Bei der neuronalen maschinellen Übersetzung wird immer ein vollständiger Satz analysiert und übersetzt (vgl. Schmalz 2019: 198ff.). Ein Vorteil, der sich dadurch ergibt, ist, dass die Sätze natürlicher klingen als bei der wort- oder phrasenbasierten statistischen Übersetzung. Auch kann das System Wortverbindungen, die sehr weit voneinander entfernt sind, gut erkennen – so zum Beispiel, wenn der zweite Teil eines mehrteiligen Verbs in einer Sprache erst am Ende des Satzes steht (vgl. Schmalz 2019: 198ff.).

Damit die neuronale maschinelle Übersetzung gute Übersetzungen liefern kann, ist auch hier sehr wichtig, dass möglichst viel Trainingsmaterial zur Verfügung steht (vgl. Schmalz 2019: 198ff.). Ein Nachteil der neuronalen Übersetzung ist, dass für einen Fachterminus der Ausgangssprache nicht immer ein und derselbe Zielsprachenterminus verwendet wird. Daher kommen auch hybride Verfahren zum Einsatz, mit denen man zum Beispiel Datenbanken einbinden kann, womit fixe Phrasen bzw. bestimmte Wörter immer gleich übersetzt werden (vgl. Schmalz 2019: 198ff.).

Ein Nachteil der NMÜ ist, dass das Training länger dauert und das Modell zwar die gewünschten flüssigen Passagen liefert, diese allerdings inhaltlich nicht akkurat sind. Weiters werden teilweise sogenannte *non-words* gebildet, d. h., das Modell lässt bestimmte Sätze einfach aus (vgl. Dorothy 2020: 437). Ein weiteres Problem für Anwender*innen ist, dass das Training der neuronalen maschinellen Übersetzung nicht sehr transparent ist. Somit werden die Verbesserung des Workflows und das Vermeiden von Fehlern erschwert (vgl. Dorothy 2020: 438). Auch das Tool, das in der vorliegenden Arbeit für die Analyse bzw. die Umfrage verwendet wird, lässt im Namen DeepL bereits erkennen, dass es auf Deep Learning beruht (vgl. DeepL 2022).

3.3 Frei verfügbare benutzer*innenfreundliche maschinelle Übersetzungssysteme

3.3.1 SYSTRAN

SYSTRAN (1968) wurde in den 1970ern als hochwertiges Übersetzungssystem vermarktet. Dieses Übersetzungssystem hat als erstes die regelbasierte Übersetzung angewandt (vgl. Dorothy 2020: 432). Die Besonderheit bei SYSTRAN war, dass es nicht zwischen zwei Sprachen übersetzen sollte, sondern die Notwendigkeit bestand, dass es zwischen mehreren Sprachen übersetzen kann, damit es die Bedürfnisse der europäischen Bürger*innen und somit die der Mehrsprachigkeit erfüllen konnte (vgl. Dorothy 2020: 432ff.) SYSTRAN wurde auch von der CEC (Kommission der Europäischen Gemeinschaft) für die Sprachenkombination Englisch-Deutsch verwendet. Bis 2010 wurde es ebenso von der Europäischen Kommission benutzt und anschließend von einem internen Programm abgelöst (vgl. Dorothy 2020: 432ff., Schmalz 2019: 194 ff.).

Die Bearbeitung mithilfe von SYSTRAN funktionierte nicht so, dass ein vollständiger Satz gelesen und übersetzt wurde, sondern das Programm zerlegte einen Satz in einzelne Teile und übersetzte diese dann mithilfe von Wörterbüchern (vgl. Schmalz 2019: 135ff.). Wie bereits beschrieben basierte SYSTRAN anfangs auf regelbasierter maschineller Übersetzung. Seine Software wurde auch durch Subsysteme unterstützt, die die eingebundenen Wörterbücher erschufen und diese auch updateten. Es dauerte etliche Jahre, bis die Software entwickelt werden konnte, wobei vor allem Mehrdeutigkeiten in der Sprache zu Beginn nicht richtig übersetzt wurden (vgl. Schmalz 2019: 195). Mittlerweile bietet SYSTRAN jedoch mit *Pure Neural Machine Translation* (PNMT) ein Übersetzungssystem an, das mit neuronaler Übersetzung arbeitet. Das Unternehmen bietet Übersetzungen in 140 Sprachen an, wobei auf der frei verfügbaren Webseite derzeit maximal 200 Zeichen übersetzt werden können. Bei der Übersetzung scheinen zudem zusätzlich zum Übersetzungsvorschlag ein Wörterbuch und mit ihm Übersetzungsalternativen auf, was beispielsweise bei Google Translate nicht der Fall ist (vgl. SYSTRAN 2022).

3.3.2 Google Translate

Google Translate wurde im Jahr 2007 gegründet und unterstützte zu Beginn nur die Sprachen Arabisch und Englisch. Mittlerweile werden Übersetzungen in 103 Sprachen angeboten. Die am häufigsten verwendeten Sprachen sind Englisch, Arabisch, Spanisch, Portugiesisch, Russisch und Indonesisch (vgl. Sen 2016). Die anfängliche Architektur von *Google Translate* war eine statistische maschinelle Übersetzungssoftware, bei der zu Beginn eine Datenbank automatisch ausgewertet und danach die statistisch am häufigsten vorkommende Übersetzung ausgegeben wurde. In diesem Kontext ist es wichtig zu beachten, in welchem Themengebiet das Tool trainiert wurde, da es ansonsten besonders bei mehrdeutigen Wörtern zu Fehlübersetzungen kommen konnte (vgl. Schmalz 2019: 198 ff.).

Mittlerweile arbeitet *Google Translate* mit einem neuronalen Übersetzungssystem. Das Unternehmen brachte dieses im Jahr 2016 auf den Markt (vgl. Schmalz 2019: 198 ff.). Zu diesem Zeitpunkt waren neuronale maschinelle Übersetzungssysteme laut Schmalz (2019) allerdings noch nicht so ausgereift wie gut trainierte statistische Übersetzungssysteme (vgl. Schmalz 2019: 198 ff.).

3.3.3 Microsoft Bing

Das Tool *Bing* wurde von Microsoft gegründet und war von Anfang an online verfügbar. Es sollte dabei helfen, das Internet zu erkunden. Dabei konzentrierte sich Bing auf vier Bereiche, nämlich das Einkaufen, das Planen eines Urlaubs, das Nachlesen über eine Krankheit und die Unterstützung beim Auffinden eines Geschäfts (vgl. Microsoft Bing Blogs 2011). Mittlerweile heißt das Tool *Microsoft Bing* (vgl. Microsoft Bing Blogs 2011). Mit ihm kann in 109 Sprachen maschinell übersetzt werden. Bing bietet auch ein Service an, mit dessen Hilfe Benutzer*innen Fragen stellen können und eine Antwort erhalten, die in über 100 Sprachen verfügbar ist (vgl. Microsoft Bing Blogs 2011).

3.3.4 DeepL

DeepL GmbH ist ein Unternehmen, das sich auf maschinelle Übersetzungen durch künstliche Intelligenz spezialisiert hat. Der Korpora des Tools DeepL basiert auf den angezeigten Ergebnissen von *linguee.de*. Auf dieser Webseite werden den Besucher*innen der Webseite mit Hilfe von *Webcrawling* an einem Ort gesammelte Paralleltexte von verschiedenen Webseiten angezeigt (vgl. DeepL 2022). Das Ergebnis der gesuchten Übersetzung für den Begriff *Masterarbeit* ist nachstehend in Abb. 3 zu sehen:

[...] Unterrichtsfächern, wobei eines der Praktika ein Forschungspraktikum sein kann, 6 C für das Mastermodul und 20 C für die Masterarbeit . ↳ uni-goettingen.de	[...] whereby one of the internships may be a research internship, 6 C for the master's module, and 20 C for the master's thesis . ↳ uni-goettingen.de
Die Masterarbeit ist eine Arbeit, die dem Nachweis der Befähigung dient, wissenschaftliche Themen selbstständig sowie inhaltlich [...] ↳ jointdegree.eu	The Master's thesis serves to prove that the student is qualified to work independently on academic topics in a methodologically [...] ↳ jointdegree.eu
(6) Die Masterarbeit kann durch einen gesonderten Antrag auch in Form einer Gruppen- oder Projektarbeit ausgegeben werden, [...] ↳ mastersprogram.de	(6) Upon request, the Master's thesis may be assigned as a group thesis or a group project if the candidate's individual [...] ↳ mastersprogram.de
Bei der Abgabe der Masterarbeit hat die Kandidatin oder der Kandidat schriftlich zu versichern, dass sie bzw. er die Arbeit [...] ↳ mastersprogram.de	In submitting the Master's thesis, the candidate must certify in writing that he or she has independently composed the thesis [...] ↳ mastersprogram.de

Abbildung 3: DeepL Screenshot vom 8.2.2021

Gibt man in der Suchleiste den gewünschten Begriff, in unserem Fall das Wort *Masterarbeit* ein, so wird das Wort auf der linken Seite in allen gefundenen deutschsprachigen Texten gelb angezeigt, während auf der rechten Seite die englischen Pendant dargestellt und ebenso hervorgehoben werden. Linguee hat zum Zeitpunkt der vorliegenden Recherche bereits zehn Milliarden Anfragen beantwortet (vgl. DeepL 2022).

Dadurch verfügt DeepL mittlerweile über eine sehr hohe Datenmenge, die für das weitere Training der maschinellen Übersetzung eingesetzt werden kann und Übersetzungen immer besser macht. Der DeepL-Übersetzer wurde 2007 vorgestellt und ist kostenlos verfügbar (vgl. DeepL 2022). Er bietet nicht nur bei der Übersetzung von einzelnen Wörtern oder Wortverbindungen Hilfestellung, sondern man kann bei der kostenpflichtigen Version Deep Pro ganze Dokumente hochladen und mit nur einem Mausklick übersetzen lassen. Seit Beginn arbeitet das Übersetzungsprogramm DeepL mit neuronaler Übersetzung (vgl. DeepL 2022).

4. Qualitätsbewertung maschineller Übersetzung

Hinsichtlich der maschinellen Übersetzung wurden bisher einige Ansätze erforscht, um den Zieltext qualitativ bewerten zu können. Einige von ihnen werden im folgenden Unterkapitel vorgestellt. Beim Einsatz von maschineller Übersetzung gibt es sowohl holistische Ansätze als auch Ansätze, die mit Fehlerkategorien arbeiten. Bei der Bewertung maschineller Übersetzungen kann zusätzlich zwischen rein maschineller Evaluierung und Humanauswertung unterschieden werden.

4.1 Maschinelle Evaluierung nach Metriken

Im folgenden Kapitel werden jene Ansätze vorgestellt, die maschinelle Übersetzungen anhand von Metriken bewerten: der ALPAC-Report, BLEU, METEOR und die Qualitätsmetrik MQM.

4.1.1 ALPAC-Report

Ein erster Ansatz zur Bewertung von maschineller Übersetzung war der ALPAC-Report (ALPAC 1966). Das Akronym setzt sich aus den Anfangsbuchstaben des Automatic Language Processing Advisory Committee zusammen. Beim ALPAC-Report wurden die Parameter *fidelity* und *intelligibility* eingesetzt (vgl. Way 2013: 6ff.). Mit *fidelity* ist gemeint, dass der Zieltext die gleiche Information wie der Ausgangstext beinhaltet. Unter *intelligibility* wird verstanden, dass der Zieltext der Zielsprache entspricht, also idiomatisch und verständlich ist (vgl. Way 2013: 6ff.). Der ALPAC-Report führte seinerseits zur Unterbrechung der Forschung in den USA, da die maschinelle Übersetzungsqualität als *nicht gut genug* eingestuft wurde (vgl. Way 2013: 6ff.).

4.1.2 BLEU

Eine weitere Metrik ist die *Bilingual Evaluation Study* (BLEU)-Metrik, die bei jedem Sprachenpaar die Qualität des Zieltextes bewerten soll (vgl. Papineni et al. 2002: 311ff.). Der *gold standard* ist dabei immer die menschliche Übersetzung (vgl. Way 2018: 7-8). Dies bedeutet, je näher eine maschinelle Übersetzung einer Humanübersetzung ist, desto besser wird sie bewertet. Bei BLEU handelt es sich um eine Metrik, bei der diese Nähe gemessen wird. Dafür werden Korpora mit Referenztexten von Übersetzungen guter Qualität herangezogen

(vgl. Papineni et al. 2002: 311ff.). Es gilt, dass gute maschinelle Übersetzungen mehr Übereinstimmungen mit den Referenztexten enthalten als schlechtere Übersetzungen.

Diese Übereinstimmungen werden in Form von *n-grams* gemessen (vgl. Jurafsky & Martin 2021: 2ff.). Die Variable *n* gibt dabei die Anzahl der Wörter mit Übereinstimmungen aus dem Referenztext an. So handelt es sich zum Beispiel bei „please turn“ um ein 2-gram, da dies aus zwei Wörtern besteht (vgl. Jurafsky & Martin 2021: 2ff.).

Bei der Evaluierung nach dem *gold standard* der menschlichen Übersetzung wird jedoch der Faktor ausgelassen, dass auch Humanübersetzungen Fehler enthalten können. Da es heutzutage bessere Evaluierungsformen gibt, wird BLEU kaum noch eingesetzt (vgl. Chatzikoumi 2020: 140ff.).

4.1.3 METEOR

Die Metric for Evaluation of Translation with Explicit ORdering (METEOR) (vgl. Lavie & Denkowski 2009) wurde 2004 entwickelt, um die Schwächen von BLEU auszugleichen. Anders als bei BLEU bewertet METEOR die Übersetzung nicht mithilfe von n-grams, sondern es wird Wort für Wort verglichen (vgl. Lavie & Denkowski 2009: 105ff.). Zudem analysiert METEOR nicht wie BLEU mehrere Sätze gleichzeitig, sondern eine Referenz-Übersetzung des Ausgangstextes. Die Metrik analysiert gleiche Wörter (exakte Übereinstimmungen), gleiche Wortstämme, aber auch Synonyme (vgl. Lavie & Denkowski 2009: 106ff.). Um die Synonyme analysieren zu können bedarf es allerdings einer Datenbank. METEOR wurde ursprünglich für das Sprachenpaar Arabisch-Englisch und die Sprachen Spanisch, Französisch und Englisch entwickelt (vgl. Lavie & Denkowski 2009: 106ff.). Ein Training in anderen Sprachen ist laut den Entwickler*innen allerdings ebenso möglich. Die Parameter, die bei METEOR gemessen werden, sind *Accuracy* und *Fluency*, die beispielsweise auch bei der MQM-Metrik vorkommen, die im nächsten Kapitel vorgestellt wird (vgl. Lavie & Denkowski 2009: 105-115).

4.1.4 Die Qualitätsmetrik MQM

Die Multidimensional Quality Metric (MQM) (vgl. Lommel et al. 2014: 455ff.) gilt als eine Metrik, die nicht prozess- oder projektorientiert ist, sondern die sich rein auf die Bewertung von Translaten fokussiert (vgl. Lommel et al. 2014: 455ff.) und aus mehreren

Qualitätsparametern besteht, die Ausgangstext und Zieltext getrennt und auch in Verbindung miteinander bewerten. Dabei handelt es sich um folgende Punkte:

- *Accuracy* (dt. Präzision): Hierbei wird Bezug darauf genommen, wie präzise die Beziehung zwischen Ausgangstext (AT) und Zieltext (ZT) ist.
- *Verity* (dt. Wahrheit): Fehler bestehen hier darin, dass der Zieltext nicht an die entsprechenden kulturellen Gegebenheiten angepasst worden ist.
- *Fluency* (dt. Flüssigkeit): Der Text ist sprachlich nicht idiomatisch gestaltet. Diese Fehler können von Personen bemerkt werden, die nur Zugang zum Zieltext haben.
- *Style* (dt. Stil): Es treten stilistische Fehler auf.
- *Terminology* (dt. Terminologie): Termini wurden falsch übersetzt¹ (vgl. Lommel et al. 2014:455ff.).
- *Design* (dt. Gestaltung): Dieser Parameter richtet sich nicht auf die Sprache, sondern auf die richtige Darstellung von Tabellen und die Formatierung.

Weitere Parameter gelten für die Lokalisierungsbranche. Diese sind *Locale Conventions* (dt. Konventionen der Gebietsschema) und *Internationalization* (dt. Internationalisierung).

Die Metrik ist dabei nicht statisch, sondern bietet eine Liste von Fehlerkategorien (s. o.), die individuell an die Übersetzung oder das maschinelle Übersetzungssystem angepasst werden können. Dabei haben die Fehlerkategorien noch Unterkategorien, die verwendet oder ausgelassen werden können. Der Punkt *Design* hat beispielsweise 33 Unterpunkte, die für die Bewertung verwendet werden können (vgl. Lommel 2018: 120ff.).

Dabei lässt sich die Metrik für die Bewertung von Humanübersetzungen und maschinellen Übersetzungen gleichermaßen verwenden. Laut Lommel (2018) geht es nicht bloß darum, Fehler zu sammeln und dann festzustellen, dass eine Übersetzung mit vielen Fehlerpunkten schlecht und eine Übersetzung mit nur einem Fehler gut sein muss, sondern es ist auch von Bedeutung, wie schwerwiegend diese Fehler sind (vgl. Lommel 2018: 120ff.).

Daher wurde zusätzlich zur MQM-Metrik eine Liste von Fehlerkategorien eingeführt. Es gibt kritische Fehler (*critical*), schwere Fehler (*major*), kleine Fehler (*minor*) und Änderungen, die keine Fehler sind (*null*). Die kritischen Fehler sind so schwerwiegend, dass eine Übersetzung ihren Zweck nicht erfüllt. Dabei reicht oft ein einziger Fehler in einer Anleitung, durch den die

¹Die Definitionen wurden direkt vom Deutschen Forschungszentrum für Künstliche Intelligenz (2015) übernommen und übersetzt.

Übersetzung nicht mehr brauchbar ist (vgl. Lommel 2018: 120ff.) Lommel (2018) führt dabei das Beispiel der Gewichtsobergrenze in einer Zentrifuge von „2 pounds“ an, das mit „2 kg“ anstelle von 0,9 kg übersetzt wurde. Besonders schlimm sind solche nur scheinbar kleinen Fehler, weil sie vorerst nicht auffallen. Wie bereits in Kapitel 1.3 erwähnt, können sie allerdings verheerende Folgen haben, wenn beispielsweise ein Medikament eingenommen werden muss und die Mengenangabe falsch angegeben ist (vgl. Lommel 2018: 120ff.).

Unter schweren Fehlern (*major mistakes*) versteht man bei der MQM-Metrik, dass die Bedeutung des Textes negativ verändert worden ist und zwar so stark, dass der Zieltext an dieser Stelle keinen Sinn ergibt. Das heißt, der/die Leser*in versteht einzelne Passagen gar nicht oder falsch (vgl. Lommel 2018: 120ff.). Kleine Fehler (*minor mistakes*) (vgl. Lommel 2018: 120) sind Fehler, die den Text nicht gänzlich unbrauchbar machen. Dabei kann es sich um vereinzelte Grammatik- oder Rechtschreibfehler handeln. Solche kleinen Fehler müssen vor der Auslieferung nicht zwangsläufig korrigiert werden. Fehler, die keine Fehler darstellen (*null*) (vgl. Lommel 2018: 120) sind zum Beispiel nachträgliche Änderungen in der Terminologie, die mit Kund*innen ausgemacht wurden und vom System dennoch als Fehler markiert werden (vgl. Lommel 2018: 120ff.).

Die MQM-Metrik verfolgt somit einen holistischen Ansatz und kann darüber hinaus von den Benutzer*innen individuell angepasst werden, da nicht alle Unterkategorien verwendet werden müssen. Dabei kann der/die Benutzer*in ebenso anpassen, wie wichtig ihm/ihr bestimmte Fehler sind (vgl. Lommel 2018: 180).

4.2 Humanauswertung maschineller Übersetzungen

Eine große Herausforderung, die sich bei der Bewertung durch Menschen präsentiert, ist zunächst das Profil der Annotator*innen, also jener Personen, die die Qualität der Übersetzung bewerten und den Text korrigieren. Ein wesentlicher Nachteil bei der Einschätzung von guten Übersetzungen ist die Subjektivität. Das heißt, wenn die zuständigen Personen zwei Sätze für das Training miteinander vergleichen, wird nicht immer gleich deutlich, warum ein Satz die bessere und der andere die schlechtere Übersetzung darstellt (vgl. Popović 2018: 129 ff.). White (2003) führt dabei Faktoren an, die die Auswertung beeinflussen können. Ein Problem kann hierbei sein, dass nicht bekannt ist, auf welchem Ausgangstext die Übersetzung basiert, denn meist haben die Annotator*innen nur den Zieltext vorliegen. Dies kann besonders dann

problematisch sein, wenn die Ausgangstexte bereits Fehler enthalten haben und diese dann übernommen werden (White 2003: 211ff., Popović 2018: 129ff.).

Weiters sollten auch Faktoren beachtet werden, die das Umfeld der Person betreffen, die aber meist ausgelassen werden, wie der mentale Zustand der Person (zum Beispiel nach einem langen Tag) und die Motivation (vgl. Popović 2018: 129ff.). Meist ist die Person beim ersten Satz noch konzentrierter als beim letzten Satz, der bewertet werden soll. Bei der Auswahl der Annotator*innen werden zusätzlich entweder Erstsprachler*innen oder bilinguale Personen eingesetzt (vgl. Popović 2018: 130). Erstsprachler*innen analysieren dabei die Qualität des Zieltexts und bilinguale Personen werden eingesetzt, um Ausgangs- und Zieltext zu bewerten (vgl. White 2003: 129ff., Popović 2018: 129ff.).

Zudem kann unterschieden werden, ob professionelle Annotator*innen, wie Linguist*innen, oder Lai*innen eingesetzt werden. Dabei sind in der Forschung der maschinellen Übersetzung Lai*innen eher die Ausnahme. Es existieren allerdings einige wenige Studien, bei denen Lai*innen und Student*innen eingesetzt werden (vgl. Castilho et al. 2018: 9-39). Ein Nachteil hiervon ist, dass Lai*innen nicht immer sprachlich ausreichend gebildet sind, um alle Fehler zu erkennen. Auch die Menge der Annotator*innen kann stark voneinander abweichen (vgl. Castilho et al. 2018: 9-39, Chatzikoumi 2020: 137-158).

Wenn Lai*innen eingesetzt werden, können mit dem sogenannten *Crowdsourcing* tausende Personen Übersetzungen bewerten. *Crowdsourcing* bedeutet, dass Personen im Internet dazu aufgerufen werden, an einem Projekt mitzuwirken, und anschließend eine *Crowd* (die Menge) eine Aufgabe übernimmt (vgl. Chatzikoumi 2020: 146). Wie auch bei der Humanübersetzung kann die maschinelle Übersetzung von der ganzheitlichen Analyse bis hin zur Bewertung von Fehlerkategorien ausgewertet werden.

4.2.1 Auswertung nach Fehlerkategorien

Popović (2018) schlägt für die Bewertung maschineller Übersetzungen einen Katalog an Fehlerkategorien vor. Das Ziel dabei ist es, ein Fehlerprofil für den Zieltext zu erhalten. Die Fehlerkategorien können manuell bestimmt werden. Weiters ist es auch wesentlich, eine Entscheidung bezüglich der Menge der Fehlerkategorien zu treffen. Dabei zählt nicht nur die Anzahl an Kategorien, sondern auch, wie genau diese formuliert sind (vgl. Popović 2018: 129ff.).

Dabei sollen laut Popović (2018) nicht nur linguistische Aspekte, sondern auch Kategorien, die die Übersetzungsstrategie betreffen, miteinbezogen werden. Manche Fehlerkategorien sind auch nur für eine Sprache wichtig, zum Beispiel was die Grammatik betrifft, und sollten daher nicht bei der Übersetzung angewandt werden (vgl. Popović 2018: 139ff.). Je nach Tool und Aufgabe bzw. Genauigkeit der Übersetzung kann deshalb mit unterschiedlich genauen Fehlerkategorien gearbeitet werden (vgl. Popović 2018: 139ff.).

Kirchhoff et al. (2012) verwenden zum Beispiel folgende Kategorien:

- *Untranslated Word*: Hier wurde ein Wort nicht übersetzt, sondern in der Ausgangssprache gelassen.
- *Missing Word*: Ein Wort des AT wurde ausgelassen.
- *Word sense error*: Der Sinn eines Wortes wurde nicht richtig wiedergegeben.
- *Morphologie*: Die morphologischen Eigenschaften sind falsch.
- *Word order error*: Die Satzstellung ist falsch.
- *Spelling*: Die Rechtschreibung ist inkorrekt.
- *Superfluous Word*: Ein Wort ist redundant oder überflüssig.
- *Diacritics*: Die diakritischen Zeichen wie Umlaute oder Accents sind falsch gesetzt oder nicht übernommen worden.
- *Punctuation*: Die Satzzeichen sind fehlerhaft oder überflüssig.
- *Capitalization*: Die Groß-/Kleinschreibung stimmt nicht.
- *Pragmatic/Cultural Error*: Die Übersetzung ist für die Zielkultur nicht angebracht.
- *Other*: Hierzu gehört alles, was noch nicht genannt wurde (vgl. Kirchhoff et al, 122).

Weiters ist es sinnvoll, darauf zu achten, wie schwerwiegend Fehler für die Zielgruppe sind. So stufen Lai*innen kleine Grammatikfehler als weniger schlimm ein als Fehler den Sinn eines Satzes betreffend (vgl. Kirchhoff et al. 2012: 120). Was beim *Post-Editing* noch zu beachten ist, ist die Tatsache, dass für diesen Bearbeitungsschritt nicht zwingend ein Ausgangstext benötigt wird. Dieser ist laut Popović allerdings sehr wohl bei der Bewertung von Übersetzungen von Bedeutung (vgl. Popović 2018: 129-158).

Eine Herausforderung für die menschliche Bewertung maschineller Übersetzungen kann laut Popović (2018) das Profil der Evaluator*innen sein (vgl. Popović 2018: 139). Letztere können bilingual oder einsprachig arbeiten. Zudem kann die von zwei geschulten Personen

durchgeführte Bewertung eines Textes ein Problem die Konsistenz betreffend herbeiführen, da beide nicht zwangsläufig dieselbe Ansicht zu einer guten/schlechten Übersetzung teilen (vgl. Popović 2018: 139).

Weiters ist es schwierig, festzustellen, wie viele Fehlerkategorien verwendet werden sollten und wie genau die Kategorie ist. Hierbei gilt: Je genauer die Fehlerkategorie festgelegt ist, desto besser werden die Ergebnisse der Auswertung sein (vgl. Popović 2018: 139). Ein letzter Aspekt, der laut Popović (2018) zu beachten ist, ist die Definition der Fehlerklassen. Die Fehlerklasse *Terminologie* ist zum Beispiel bei Fachtexten wichtig und kann bei allgemeinsprachlichen Texten ausgelassen werden (vgl. Popović 2018: 139 ff.).

Weiters ist bei der Fehlerkategorie auch zu beachten, was für die Benutzer*innen ein leichter oder schwerer Fehler ist. Darüber hinaus ist von Bedeutung, wie gut das Tool ist und nicht nur die Übersetzung ist. Laut White (2003) kann ein Tool auch zufällig gute Ergebnisse generieren. Deshalb schlägt er vor, auch das Tool zu analysieren. Er unterscheidet dabei in Black-Box- und Glass-Box-Analyse. Bei der viel häufiger vorkommenden Black-Box-Evaluierung wird nur der Zieltext bewertet. Hingegen ist bei der Glass-Box-Evaluierung auch ein Einblick in das System möglich und eine Evaluierung der Gründe und der Art der Übersetzung des Systems (vgl. White 2003: 217ff.). Dies kann sehr sinnvoll sein, allerdings ist es die Umsetzung nicht immer möglich, besonders wenn für die Forschung mit frei zugänglichen Tools gearbeitet wird und zwar der translatorische Hintergrund gegeben ist, der technische Hintergrund jedoch fehlt.

4.2.2 Bewertung nach dem Zweck (Skopos) der maschinellen Übersetzung

Die in Kapitel 1.1. erwähnte Skopostheorie kann auch bei maschinellen Übersetzungen angewandt werden. Der Zweck von Übersetzungen allgemein reicht vom Attribut *publizierfähig* bis hin zum ersten Erfassen des Sinnes. Ebenso kann die Lebensdauer eines Translats sehr unterschiedlich aussehen. Texte können weniger als eine Minute online oder jahrelang als Publikation verfügbar sein.

Maschinelle Übersetzungen lassen sich nicht immer 1:1 mit Humanübersetzungen vergleichen bzw. kann die Qualität einer maschinellen Übersetzung im klassischen Sinn viel schlechter sein (mit Grammatikfehlern, Auslassungen, stilistischen Fehler) und dennoch erfüllt die Übersetzung ihren Zweck (z. B. Erfassen des Sinns) gemäß Reiß und Vermeer (1984). Eine Übersetzung kann 20 Grammatikfehler enthalten, deshalb bei einem System mit

Fehlerkategorien durchfallen und dennoch für die Leser*innen brauchbar sein, da sich der Inhalt trotzdem erschließen lässt (vgl. Way 2013: 4ff.).

Oft ist, wie bereits in Kapitel 2.1. erwähnt, der Zweck der maschinellen (Roh)Übersetzung ein erstes Erfassen des Sinnes. Danach kann zum Beispiel entschieden werden, ob es sich lohnt, den Ausgangstext professionell übersetzen zu lassen. Way (2013) bezeichnet die unterschiedlichen Übersetzungsqualitäten als „*fitness for purpose*“ (Way 2013: 2) und er geht weg vom „*one size fits it all*“ (Way 2013: 9). Dabei ist es auch notwendig, *fitness for purpose* zu definieren, da auch dies eine sehr subjektive Bedeutung haben kann (vgl. Way 2013: 9).

Laut Bowker gilt *fitness for purpose* als das, was mit den Kund*innen vereinbart wurde, damit der Text seinen Zweck erfüllt, der sehr individuell sein kann (vgl. Bowker 2020: 461). Von der anfänglichen Erwartung an die maschinelle Übersetzung hochwertige Translate zu generieren (*fully-automatic high-quality translation*, auch FAHQT), was das Ziel der ersten Entwickler*innen maschineller Übersetzung (MÜ) war, ist man bereits abgewichen (vgl. Bowker 2020: 453ff.).

Die Skopostheorie kann auch im Speziellen beim Qualitätsbegriff von maschinellen Übersetzungen angewandt werden. Diese Texte haben ihre Daseinsberechtigung und können als qualitativ *gut genug* eingestuft werden.

Jedoch ist auch die Lebensdauer eines Translats ausschlaggebend für dessen Qualität. Hierbei ist es sehr wichtig zu wissen, ob ein Text publiziert werden soll und wie lange für das Internet übersetzte Texte online sind (vgl. Lusicky & Heinisch 2019: 42ff.). Viele Online-Shops verwenden rein maschinelle Übersetzungen, da die Texte offensichtlich ihren Zweck erfüllen und nicht allzu lange online sind (vgl. Way 2013: 1ff.).

Auch das funktionale Übersetzen nach Nord (2006) kann hier angewandt werden. Dabei wird der Zweck beim *Gisting* von allen Akteur*innen erfüllt, wenn dies die Erwartung (die Funktion) des Textes war. Die vereinbarte Qualität ist hierbei *gut genug* und alle Beteiligten sind damit zufrieden. In diesem Fall handelt es sich bei den Beteiligten um die Produzent*innen der maschinellen Übersetzung und die Leser*innen.

Bei der *Gisting*-Übersetzung können die im ersten Kapitel erwähnten Ansätze gut eingebracht werden. Wenn man allerdings zum Beispiel nur die Fehlerkategorien nach Delizée (2011) übernimmt, und das Ergebnis von Fachübersetzer*innen mit dem einer Rohübersetzung

vergleicht, so schneidet die maschinelle Übersetzung jedenfalls schlechter ab, als wenn man den Zweck beider Übersetzungen berücksichtigt.

5. Aktueller Stand der Forschung

Wie in den Grundlagenkapiteln erwähnt, existieren bereits etliche Studien, in denen die Qualität der maschinellen Übersetzung alleine oder im Vergleich zur Humanübersetzung bewertet wurde. In diesen Studien haben Translator*innen bzw. Translationswissenschaftler*innen die Übersetzungen bewertet (vgl. Jia et al. 2019: 60-86, Koehn et al. 2008: 139-142, Lusicky & Heinisch 2019: 42-48, Fiederer & O'Brien 2009: 52-74, Koponen 2010: 1-12). Allerdings wurde bislang noch nicht verglichen, wie Lai*innen und Übersetzer*innen die Qualität der maschinellen Übersetzung und der Humanübersetzung bewerten. Dennoch existieren einige Studien, die dem Forschungsziel der vorliegenden Masterarbeit ähnlich sind und an dieser Stelle näher ausgeführt werden (vgl. Fiederer & O'Brien 2009: 52-74, Lusicky & Heinisch 2019: 42-48).

Fiederer und O'Brien (2009) verglichen die Qualität maschineller Übersetzung mit post-editierten Texten und Humanübersetzungen. Dabei verwendeten sie Englisch als Ausgangssprache und Deutsch als Zielsprache. Als MÜ-Tool verwendeten sie IBM WebSphere (vgl. IBM 2020). Die Proband*innen waren Masterstudent*innen an der Dublin City University, deren Bildungs- oder Erstsprache Deutsch war. Unter den Texten waren 30 Ausgangstexte, drei humanübersetzte Texte und drei post-editierte Versionen.

Für die Bewertung der Qualität verwendeten sie basierend auf Hutchins und Somers (1993: 163) die angepassten Parameter *Accuracy*, *Clarity* und *Style*, die wie folgt definiert werden:

- *Accuracy* (dt. Präzision) bedeutet, wie genau Informationen aus dem Original in den Zieltext übersetzt werden.
- *Clarity* (dt. Klarheit) bezieht sich auf die leichte/schwere Verständlichkeit einer Übersetzung.
- *Style* (dt. Stil) bedeutet, dass die sprachlichen – stilistischen – Feinheiten zum Inhalt und zur Intention passen.

Für den Parameter *Style* entschieden sie sich, da dieser bisher kaum in Bezug auf maschinelle Übersetzung bewertet wurde (vgl. Fiederer & O'Brien 2009: 57). Weiters verwiesen sie auf die Diskussionen in der Translationswissenschaft und Sprachindustrie zur Frage, welche Parameter die besten für eine Evaluierung seien (vgl. Fiederer & O'Brien 2009: 56). Dies zeigt auf, dass bestimmte Parameter je nach Kontext anders anzuwenden sind.

Um zu eruieren, ob die Humanübersetzung immer die bessere Wahl ist, wurden die Proband*innen neben der Bewertung der Übersetzungen auch gefragt, welche die beste Übersetzung sei, um feststellen zu können, ob sie die Humanübersetzung immer bevorzugen würden (vgl. Fiederer & O'Brien 2009: 54).

Die Ergebnisse der Studie waren, dass die Humanübersetzungen besser bei *Clarity* und *Style* abschnitten, während die maschinelle Übersetzung besser beim Parameter *Accuracy* bewertet wurde (vgl. Fiederer & O'Brien 2009: 64). Weiters bieten die Autor*innen einen Überblick über die diversen Qualitätsparameter, die für Studien eingesetzt werden können (vgl. Fiederer & O'Brien: 63ff.).

Eine ähnliche Studie wurde von Lusicky und Heinisch (2019) durchgeführt. Dabei haben 47 Teilnehmer*innen des Masterstudiums Translation der Universität Wien Texte auf ihre Qualität hin bewertet. Zur Bewertung wurde die MQM-Metrik (vgl. Lommel et al. 2014: 455ff.) und der TAUS Dynamic Quality Framework (DQF) (vgl. TAUS 2020) herangezogen, die beide aus der Sprachindustrie stammen. Die Student*innen mussten zuerst angeben, welche Fehlerkategorien sie am ehesten erwarten würden und anschließend mussten sie die Kategorien bewerten.

Die häufigsten Fehlerkategorien waren dabei in den Parametern *Accuracy* und *Fluency* sowohl bei den Erwartungen als auch beim tatsächlichen Resultat zu finden. Weiters waren die anfänglichen Einschätzungen zur MÜ höher als das ausgegebene Resultat. Neben der zu bewertenden Qualität wurden die Studienteilnehmer*innen zusätzlich befragt, welche Erwartungen sie generell an die maschinelle Übersetzung hatten. Dabei ging der Großteil der Teilnehmer*innen davon aus, dass die MÜ eine *Gisting-Translation* (siehe Kapitel 3.1.2) produzieren würde, während nur 11 % der Teilnehmer*innen ein publizierfähiges Translat erwarteten.

Auch wenn in den beiden Studien der Fokus nicht zusätzlich auf Lai*innen lag, bieten sie eine gute Basis für weitere Arbeiten in diesem Bereich. Bei Fiederer und O'Brien (2019) wurden die Fehlerkategorien genauestens erarbeitet und bei Lusicky und Heinisch (2019) wurde der Fokus zusätzlich zur Bewertung von MÜ auch auf die Erwartungen gelegt, was aufzeigt, dass die gleichzeitige Existenz von MÜ und Humanübersetzung für unterschiedliche Zwecke durchaus sinnvoll ist.

Eine weitere Studie zum Vergleich von maschineller und Humanübersetzung wurde von Koponen (2010) durchgeführt. Dabei wurden drei verschiedene Ausgangstexte zu

verschiedenen Themen mittels regelbasierter maschineller Übersetzung, statistischer maschineller Übersetzung und Übersetzung durch Translator*innen durchgeführt und danach evaluiert. Koponen (2010) verwendete dabei keine Metrik, sondern führte fünf Fehlerkategorien ein:

- *Omitted concept* (dt. ausgelassener Begriff),
- *Added concept* (dt. hinzugefügter Begriff),
- *Untranslated concept* (dt. nicht übersetzter Begriff),
- *Mistranslated concept* (dt. falsch übersetzter Begriff) und
- *Substituted concept* (dt. ersetzter Begriff; d. h., Begriff wurde im Zieltext durch anderen ersetzt) (Koponen 2010:4).

Um die Ergebnisse auszuwerten, führte sie Punkte ein und eine höhere Summe bedeutete mehr Fehler. Was die Fehlerkategorien der einzelnen Begriffe betrifft, so stellte sich heraus, dass die statistische maschinelle Übersetzung die meisten Fehler aufwies. Sie schnitt auch bei den Fehlern zur Beziehung der einzelnen Begriffe am schlechtesten ab. Weiters wurde bei der Untersuchung die Fehlerkategorie *substituted concept* tabellarisch dargestellt. Hierbei ließ sich feststellen, dass die Humanübersetzung die meisten Fehler ergab: 215 Punkte, im Vergleich zu 60 Punkten für SMÜ und 12 Punkten für RBMÜ. Diese hohe Punkteanzahl ergab sich deshalb, weil der/die Translator*in Ergänzungen und Erklärungen hinzufügte (vgl. Koponen 2010: 1-12).

Dies bedeutet, dass textuelle – inhaltliche – Veränderungen, die vermeintliche Fehler gemäß einem Schema darstellen, nicht immer tatsächliche Fehler sind, die die Qualität eines Translates vermindern. Im Gegenteil, ein solcher Eingriff in den Text erhöht sie. Denn durch hinzugefügte Erklärungen werden den Rezipient*innen des Zieltextes ausgangssprachliche Inhalte erklärt, die in der Zielkultur so nicht vorkommen, was schließlich zu einem besseren Textverständnis führt.

6. Methodik

Im folgenden Kapitel wird die Forschungsmethode, einschließlich der Bewertungsbasis, der Zielgruppen und der Erstellung und Auswertung des Fragebogens, vorgestellt. Um die Forschungsfrage, ob professionelle Translator*innen die Qualität maschineller Übersetzung im Vergleich zur Humanübersetzung anders einschätzen als Lai*innen, beantworten zu können, wurde eine quantitative Untersuchungsmethode anhand eines Online-Fragebogens gewählt. Mit dieser Methode können möglichst viele Proband*innen aus verschiedenen Fachbereichen erreicht werden. Ausgehend von früheren Forschungen von Lusicky und Heinisch (2019) und Fiederer und O'Brien (2019) zum Thema Qualitätsbewertung maschineller Übersetzungen, in denen Student*innen zur Qualität von maschinell übersetzten Texten befragt wurden, ist das Ziel der vorliegende Masterarbeit, herauszufinden, ob und wie sich die Bewertung der Qualität maschineller Übersetzung bei Lai*innen und professionellen Translator*innen unterscheidet. Dabei wird von der Hypothese ausgegangen, dass Lai*innen die Qualität der maschinellen Übersetzung höher als professionelle Translator*innen einstufen und aufgrund fehlender translatorischer Kenntnisse seltener erkennen, wann es sich um eine maschinelle Übersetzung handelt.

Um dies beantworten zu können, wurden den Proband*innen zehn englische Ausgangstexte und zehn deutsche Zieltex te gleichzeitig in einer Tabelle zur Bewertung vorgelegt. Dabei galt es also nicht nur, das zielsprachige Translat auf seine Qualität hin zu bewerten, sondern auch anzugeben, wie gut die Übersetzung des ausgangssprachlichen Textes gelungen war. Weiters wurden die Proband*innen bei jedem Text gefragt, ob es sich ihrer Ansicht nach um eine maschinelle Übersetzung handelt. So sollte herausgefunden werden, wie häufig die Proband*innen maschinelle Übersetzungen tatsächlich erkennen. In der vorliegenden Studie wurden den Proband*innen Texte aus verschiedenen Themengebieten vorgelegt, um auch hier eruieren zu können, ob Unterschiede in der Einschätzung der Übersetzungsqualität bestehen. Diese umfassten wirtschaftliche Texte, literarische Texte sowie einen Sportartikel.

Durch die Online-Umfrage ließen sich in sehr kurzer Zeit viele Ergebnisse generieren. Ein nicht zu unterschätzender Nachteil bestand jedoch darin, dass Rückfragen nicht möglich waren. Dies ist bei Unklarheiten oder gar Fehlern bei der Erstellung des Fragebogens kontraproduktiv. Es lässt sich deshalb nur eingeschränkt beurteilen, ob die Texte beispielsweise

schwer verständlich waren oder Unklarheiten enthielten. Um diesen Faktor abzumildern, konnten die Proband*innen in einem Kommentarfeld – zusätzlich zur Bewertung – Anmerkungen zu den Texten und ihren Anliegen machen und ebenso wurde vor Aussendung des Fragebogens eine Pilotstudie durchgeführt. Dafür wurde eine kleine Anzahl an Personen vor der finalen Freigabe gebeten, an der Umfrage teilzunehmen und anschließend Feedback zur Klarheit der Aufgabenstellung und der Antwortmöglichkeiten zu geben.

Zunächst kamen für die Erstellung des Fragebogens unter anderem die Tools Lamapoll (vgl. Lamapoll 2020) und Soscisurvey (vgl. Soscisurvey 2020) in Frage, da man bei diesen Tools auch Textfelder einfügen kann. Letztendlich wurde jedoch LimeSurvey ausgewählt, da man mit diesem Tool die Felder frei bearbeiten kann und auch Bedingungen gesetzt werden konnten. Zum Beispiel verlangte eine positive Antwort auf die Frage „Sind Sie als professionelle*r Translator*in tätig?“ eine Bedingung, damit die Frage „In welcher Branche sind Sie tätig?“ nicht mehr angezeigt wurde. Dies ermöglichte die Erstellung von einem einzigen Fragebogen für alle Befragten – also Lai*innen und gleichzeitig professionelle Translator*innen.

Neben der Übersetzungsbewertung wurden weitere aufschlussreiche Punkte in die Online-Umfrage mit aufgenommen, nämlich das Alter, der Beruf und die Sprachkenntnisse der jeweiligen Person. Zudem wurde das Level der Sprachkenntnisse der Proband*innen abgefragt, um dessen Einfluss auf ihre Revisionskenntnisse abschätzen zu können.

Ziel war es, mithilfe der Parameter numerisch darstellen zu können, ob es einerseits einen Unterschied zwischen maschineller Übersetzung und Humanübersetzung gibt und andererseits, wie gut diese Zieltexte von Lai*innen und Translator*innen eingestuft werden. Um die Ergebnisse präsentieren zu können, wurde mit dem Tabellenkalkulationsprogramm Microsoft Excel gearbeitet, worauf im Kapitel Auswertung noch näher eingegangen wird. Die Proband*innen bewerteten die Übersetzungen auf einer Skala von 1-5. Somit konnte auch sichergestellt werden, dass es einerseits keine Mittelwerte gibt, die in großer Anzahl nicht aussagekräftig gewesen wären und andererseits die Ergebnisse übersichtlich dargestellt werden konnten.

6.1 Bewertungsbasis Accuracy, Fluency und Style

Um die Übersetzungen bewerten zu können, werden als Basis die Parameter *Accuracy*, *Fluency* und *Style* nach Hutchins und Somers (1993) verwendet. Als Forschungsgrundlage gelten die

bereits erwähnten Studien von Lusicky und Heinisch (2019) und Fiederer und O'Brien (2009). Lusicky und Heinisch (2019) wählten Bewertungen auf Basis der MQM-Metrik und Fiederer und O'Brien (2009) wählten die Parameter *Clarity*, *Style* und *Accuracy*.

Zunächst war für den Vergleich von Humanübersetzungen mit maschinellen Übersetzungen wichtig, Parameter zu finden, die für beide Arten der Übersetzungen angewendet werden können. Die Wahl fiel auf die Qualitätsparameter *Accuracy*, *Fluency* und *Style*. Mithilfe dieser Parameter kann ein ganzheitliches Bild der Qualität sowohl für Humanübersetzungen als auch für die maschinellen Übersetzungen geschaffen werden. Da maschinelle Übersetzungen und Humanübersetzungen meist einen unterschiedlichen Skopos haben, wäre eine Einteilung in Fehlerkategorien (vgl. Mossop 2020, Hansen 2008) nicht sinnvoll gewesen, da anzunehmen war, dass maschinelle Übersetzungen im Vergleich zu Humanübersetzungen schlechter abschneiden würden, wodurch ein holistischer Vergleich nicht möglich gewesen wäre.

Zunächst kam auch die Qualitäts-Metrik MQM in Frage, die allerdings für die Auswertung zu viele Parameter beinhaltet hätte. Außerdem hätte mittels einer Metrik ausgewertet werden müssen, was den Rahmen der vorliegenden Masterarbeit gesprengt hätte. Die gewählten Parameter *Accuracy*, *Fluency* und *Style* konnten hingegen mithilfe des Tabellenkalkulationsprogramms Microsoft Excel ausgewertet werden.

Für die Bewertung wurde eine Notenskala von 1-5 erstellt, womit die Ergebnisse in den genannten drei Qualitätsbereichen numerisch dargestellt werden konnten. Dies ermöglichte sowohl, dass die drei Parameter für maschinelle Übersetzungen und Humanübersetzungen unabhängig voneinander analysiert werden konnten, als auch, dass ihr Gesamtdurchschnitt der einzelnen Texte berechnet werden konnte, um ihre Ergebnisse zu präsentieren.

Wie bereits erwähnt, ergab die Studie von Fiederer und O'Brien (vgl. 2009), dass Humanübersetzungen in den Bereichen *Style* und *Clarity* besser abgeschnitten hatten, maschinelle Übersetzungen jedoch im Bereich *Accuracy*. In der vorliegenden Arbeit galt es herauszufinden, ob es einen ähnlichen Unterschied hinsichtlich der Parameter bei den jeweiligen Übersetzungen und den unterschiedlichen Proband*innengruppen gibt.

6.2 Zielgruppe der Umfrage

Wie erwähnt, umfasst die Zielgruppe der Umfrage einerseits Lai*innen und andererseits auch professionelle Translator*innen. Die Gruppe der Lai*innen bezieht sich anders als bei

bisherigen Studien nicht ausschließlich auf Studierende, sondern auf eine möglichst breite Gruppe und umfasst grundsätzlich alle, die nicht als professionelle Translator*innen tätig sind.

Die Gruppe der Lai*innen wurde einerseits aus Personen zusammengestellt, die dem Aufruf in den folgenden drei Umfragegruppen auf Facebook nachkamen: *UMFRAGEN! – Studienarbeiten – Meinungsforschung, Umfragen & Umfrageteilnehmer finden* und *Umfragen für Studienarbeiten*. Andererseits wurde die Gruppe der Lai*innen durch das Teilen der Umfrage via *WhatsApp* und auf *LinkedIn* mithilfe einiger Kontakte erweitert.

Die Gruppe der professionellen Translator*innen ist ebenso breit aufgestellt und umfasst Dolmetscher*innen, literarische Übersetzer*innen, Fachübersetzer*innen und andere Sprachdienstleister*innen. Um professionelle Translator*innen zu erreichen, wurde ein Mail an den Berufsverband *UNIVERSITAS* mit der Bitte um Weiterleitung an diese geschrieben und ein Posting in der Facebook-Gruppe *Master Dolmetschen und Übersetzen* veröffentlicht.

6.3 Auswahl des Übersetzungstools

Für die maschinellen Übersetzungen wurde das Übersetzungstool DeepL verwendet. DeepL ist seit 2017 online verfügbar und wird von der gleichnamigen Firma DeepL GmbH betrieben (siehe Kapitel 3.3.4.). Das Tool DeepL entstand aus dem Unternehmen Linguee GmbH. Linguee GmbH bietet als eines der ersten Unternehmen im deutschsprachigen Raum die Suche nach Übersetzungsvorschlägen an. Dabei werden Ausgangstext und Zieltext einander gegenübergestellt und die entsprechenden Stellen markiert.

Bei DeepL wurde mittels *Webcrawling* bereits eine sehr hohe Datenmenge gesammelt, da diese für Linguee generiert wurden. *Webcrawling* bedeutet, dass wie im Falle von Linguee, zweisprachige Paralleltexte von Webseiten gesammelt und automatisch aligniert werden (vgl. DeepL 2020, Big Data Insider 2020). DeepL verfügte für das Training über die Daten von Linguee. Die Entscheidung für *DeepL* im Gegensatz zu anderen frei verfügbaren Übersetzungsprogrammen fußt primär darauf, dass es momentan sehr populär ist und für die gute Qualität von Übersetzungen sowohl von Lai*innen im privaten Umfeld der Autorin als auch von einigen ihrer Professor*innen gelobt wird. Weiters wurde im Rahmen des Masterstudiums Fachübersetzen in mehreren Lehrveranstaltungen *Post-Editing* von mit *DeepL* übersetzten Texten gelehrt. Ein zu berücksichtigender Nachteil von Linguee ist jedoch, dass die Ergebnisse nicht intern im Unternehmen geprüft werden. So muss angemerkt werden, dass es häufig zu unterschiedlichen Übersetzungsvorschlägen kommt und diese kritisch betrachtet

werden müssen. Das liegt daran, dass auch von Lai*innen oder Maschinen übersetzte Webseiten als Übersetzungsvorschläge zur Basis des *Webcrawling* dienen.

6.4 Der Fragebogen

Für die Befragung der Proband*innen wurde der Online-Fragebogen (siehe Anhang) ausgewählt, damit eine möglichst große Zielgruppe erreicht werden konnte.

6.4.1 Demographische Daten

Im demographischen Teil des Fragebogens wurden die Teilnehmer*innen zunächst zu ihrem Alter und Geschlecht befragt. Weiters wurden sie gebeten, ihren höchsten Bildungsabschluss und ihre Arbeitsstelle anzugeben.

Letzteres wurde in die Branchen Hotel- und Gastgewerbe, Handel, Lehr- und Kulturbereich und Technik und Sozialberufe unterteilt, um die häufigsten Berufsarten abzudecken. Eine weitere Frage sollte Aufschluss darüber geben, ob die Teilnehmer*innen zum Zeitpunkt der Umfrage als professionelle Translator*innen tätig waren, um die beiden Gruppen der Lai*innen und Translator*innen einander gegenüberzustellen.

Weiters wurden die Teilnehmer*innen nach ihrer höchsten abgeschlossenen Ausbildung gefragt. Zu den Auswahlmöglichkeiten zählten Lehre, berufsbildende höhere Schule, Studium und Doktorat.

Um die Textbewertung besser einschätzen zu können, wurden die Teilnehmer*innen auch nach ihrer Erstsprache und ihrem Sprachniveau befragt. Die erste Frage lautete „Ist Deutsch Ihre Erstsprache?“; die zweite lautete „Wie hoch schätzen Sie Ihr Sprachniveau auf Deutsch ein?“ – sofern die erste verneint wurde. Die gleichen Fragen wurden auch für gestellt. Damit das Sprachniveau einheitlich angegeben werden konnte, wurden die Niveaustufen des europäischen Referenzrahmens im Fragebogen kurz erklärt. So konnten die Befragten anschließend zwischen den Niveaus A1, A2, B1, B2, C1 und C2 auswählen. Um die Erfahrung der Teilnehmer*innen mit Übersetzungsprogrammen einschätzen zu können, wurden sie auch gefragt, ob sie frei verfügbare Übersetzungstools nutzen. Für deren Verständnis wurde hinzugefügt, dass es sich bei diesen Tools um Programme wie DeepL, Google Translate oder SYSTRAN handelt.

6.4.2 Textbewertung

Im zweiten Teil der Umfrage wurden den Proband*innen neun englische Ausgangstexte und Zieltexte vorgelegt, wobei sie die Translate auf einer Notenskala von 1-5 bewerteten. Für diese Bewertung standen drei Fragen zur Verfügung: Wurde der Inhalt korrekt übersetzt? (1-5), Wie flüssig wurde der Text übersetzt? (1-5) und Wie gut ist der Stil des Zieltextes? (1-5). Im nächsten Schritt wurden die Teilnehmer*innen gefragt, ob es sich ihrer Einschätzung nach beim vorliegenden Text um eine maschinelle Übersetzung handelt.

6.4.3 Abschließende Fragen

Zum Abschluss des Fragebogens wurden die Teilnehmer*innen gebeten, offene Fragen zu beantworten. Die erste Frage lautete: „Wie leicht ist es Ihnen nach Ihrem Gefühl gegangen, menschliche Übersetzungen von maschineller Übersetzung zu unterscheiden“. Weiters wurden sie auch noch gefragt, wie wichtig Ihnen die Aspekte Inhalt, Korrektheit und Stil bei Übersetzungen generell sind. Am Ende konnten sie noch Anmerkungen zur Qualität und dem Einsatz maschineller Übersetzungen machen.

6.5 Textauswahl und Begründung

Für die Humanübersetzungen wurden Texte aus verschiedenen Themengebieten ausgewählt, um eruieren zu können, ob es einen Qualitätsunterschied hinsichtlich verschiedener Themen gibt. Dabei wurden verschiedene Themengebiete, von Politik bis Literatur, abgedeckt, womit für die Humanübersetzungen ein breites Spektrum an Texten zur Verfügung stand. Für die maschinell übersetzten Texte wurden verschiedene Webseiten zu unterschiedlichen Themen ausgewählt, um auch hier mehrere Domänen abzudecken und feststellen zu können, wie hoch die Qualität der Texte von DeepL gemäß der Parameter *Accuracy*, *Fluency* und *Style* ist (siehe Anhang Textauswahl). Zudem wurde ein Text der Modulprüfung Fachübersetzen des Wiener Zentrums für Translationswissenschaft miteinbezogen, der maschinell übersetzt wurde (siehe Anhang Textauswahl, Text 3).

6.5.1 Text 1: Humanübersetzung

Beim ersten Text der Umfrage handelte es sich um einen Auszug aus einem alltagssprachlichen Zeitungsbericht über einen festgenommenen Privatdetektiv. Aus Sicht der Autorin wurde der Inhalt korrekt übersetzt, es wurde lediglich in der deutschen Übersetzung des Satzes „he left her loud voicemails“ das Wort „laut“ weggelassen. Da es sich um einen sehr standardsprachlichen Text ohne Eingriffe in den Zieltext handelt, könnte auch angenommen werden, dass die Übersetzung maschinell erstellt wurde. Tatsächlich wurde die Übersetzung aber von einem Menschen angefertigt.

6.5.2 Text 2: Maschinelle Übersetzung

Um mehrere Themengebiete abzudecken und am Ende feststellen zu können, wie gut die unterschiedlichen Textsorten (maschinell) übersetzt wurden, wurde ein an Lai*innen gerichteter informativer Text aus dem Fachbereich Medizin ausgewählt. Text 2 stammt aus der Online-Enzyklopädie Britannica (vgl. Britannica 2022) und beschreibt den Begriff „in vitro fertilization (IVF)“, also künstliche Befruchtung.

Text 2 wurde deswegen gewählt, da er sehr wörtlich übertragen wurde, so zum Beispiel der Begriff „medical procedure“ im folgenden Satz:

“In vitro fertilization (IVF), also called test-tube conception, **medical procedure** in which mature egg cells are removed from a woman”

Das „medical procedure“ wurde ins Deutsche anhand von einer Einschlebung übernommen:

„...*Reagenzglasbefruchtung* genannt, **medizinisches Verfahren**, bei dem...“.

Weiters wurde „fertilized eggs (ova)“ mit „befruchtete Eizellen (Eizellen)“ übersetzt, da *fertilized egg* und *ova* im Englischen synonym verwendet werden können. Auch Lai*innen sollte hier auffallen, dass beide Begriffe mit „befruchtete Eizellen (Eizellen)“ gleich übersetzt worden sind. Daher ist bei dieser Übersetzung anzunehmen, dass die Mehrheit der befragten Personen von einer maschinell gefertigten ausgeht.

6.5.3 Text 3: Maschinelle Übersetzung

Bei Text 3 handelte es sich um einen Prüfungstext der Modulprüfung „Technik und Naturwissenschaft“, die am Zentrum für Translationswissenschaft der Universität Wien als eine der schriftlichen Abschlussprüfungen für den Schwerpunkt „Fachübersetzen“ gilt“. Der Text

stammt aus der medizinischen Fachzeitschrift *Research Reports in Clinical Cardiology* und beschreibt, wie sich die westliche Ernährung auf unsere Gesundheit auswirkt (vgl. Carrera et al. 2011).

Text 3 wurde ausgewählt, da er eine Art Qualitätsprüfung für angehende Übersetzer*innen darstellt und somit für dieses Experiment auch als Prüfung für ein maschinelles Übersetzungssystem herangezogen wird. Aus Sicht der Autorin hat DeepL alle Termini korrekt übersetzt, weshalb man darauf schließen könnte, dass es sich bei diesem Text um eine Humanübersetzung handelt. Der Text wurde sehr ausgangstextnahe übersetzt, was für diese Textsorte durchaus angemessen ist; ebenso wurde inhaltlich alles korrekt übertragen. Somit ist auch deshalb anzunehmen, dass sowohl Lai*innen als auch professionelle Translator*innen den Text als vom Menschen übersetzt einstufen werden.

6.5.4 Text 4: Humanübersetzung

Bei Text 4 (vgl. Twain) handelt es sich um einen Auszug aus dem Vorwort der Übersetzung des Literaturklassikers *Die Abenteuer des Tom Sawyers* von Mark Twain, die 1900 auf Deutsch erschienen ist und von H. Hellwag übersetzt wurde (vgl. Twain 1876, Twain/Hellwag, 1900).

Dieser Text wurde ausgewählt, da an einigen Stellen die Struktur des Ausgangstextes nicht übernommen wurde. So wurde „men and women“ nicht mit „Männer und Frauen“, sondern im Kontext ebenfalls korrekt mit „Erwachsene“ übersetzt. Weiters wurde bei der Textstelle „what queer enterprises they engaged in“ abgewichen von einer wortwörtlichen Übersetzung des Inhalts, sondern der Übersetzer entschied sich für „welcher Art ihr Ehrgeiz und ihre Unternehmungen waren“. Dies ist aus Sicht der Autorin korrekt übersetzt.

An anderen Stellen wurde allerdings die Satzstruktur einschließlich der Satzzeichen übernommen. So wurde der Satz „(...) I hope it will not be shunned by men and women on that account; for part of my plan has been to try to pleasantly remind adults“ übersetzt mit „so darf ich doch wohl hoffen, dass es auch von Erwachsenen nicht ganz unbeachtet gelassen werde, habe ich doch darin versucht, ihnen auf angenehme Weise zu zeigen (...)“.

Aufgrund der nicht wortwörtlichen Übersetzung ist anzunehmen, dass die Umfrageteilnehmer*innen den Text als eine Humanübersetzung einstufen, wobei die sehr ausgangstextnahe Satzstruktur auch auf eine maschinelle Übersetzung schließen lassen könnte.

6.5.5 Text 5: Humanübersetzung

Bei Text 5 handelt es sich um einen wirtschaftlichen Text, der die Folgen der globalen Finanzkrise zwischen 2007 und 2009 behandelt (siehe Anhang, Text 5). Im englischen Ausgangstext sind Akronyme wie *ZIRP* enthalten, die von der/dem Humanübersetzer*in mithilfe von Klammersausdrücken erklärt werden, wie etwa bei „ZIRP (Nullzinspolitik)“.

Aufgrund dieser zusätzlichen Erklärungen von Akronymen in Klammern, die von einem maschinellen Übersetzungssystem nicht zu erwarten sind, sollte für die Lai*innen und Übersetzer*innen leicht zu erkennen sein, dass es sich hierbei nicht um eine maschinelle Übersetzung von *DeepL* handelt. Auch das folgende Beispiel macht deutlich, dass der Text nicht maschinell übersetzt wurde. Englischer Ausgangstext:

„No central bank had considered any of these measures (zero interest rate policy, quantitative easing, credit easing, forward guidance, negative deposit rate, and unlimited foreign exchange intervention, respectively) before 2008.“

Deutsche Übersetzung: „Vor 2008 hätte keine Zentralbank irgendeine dieser Maßnahmen in Erwägung gezogen. [...].“ Der ursprüngliche Satz wurde im Deutschen mit zwei Sätzen gelöst und die Jahreszahl im ersten Satz an den Anfang gestellt. Diese Anpassung der Satzstruktur, also der neue Satzbau lassen eine Humanübersetzung gut erkennen.

Die Erklärung der Abkürzungen durch Klammersausdrücke, die Umstellung der Jahreszahl und die Trennung über Satzgrenzen hinweg sind von einer maschinellen Übersetzung nicht zu erwarten, weshalb angenommen wird, dass hier vor allem die professionellen Translator*innen erkennen, dass es sich um eine Humanübersetzung handelt.

6.5.6 Text 6: Maschinelle Übersetzung

Bei diesem Text handelt es sich um die maschinelle Übersetzung eines Sportberichts zum Thema Football (vgl. Miller, Kerry 2021). Darin wird beschrieben, wie das Basketballteam Houston Cougars ein Spiel gewonnen hat.

Diese Textsorte wurde ausgesucht, da, wie für einen Sportbericht üblich, bestimmte Kollokationen enthalten sind, wie etwa „claw all the way back to tie the game“, was mit „den ganzen Weg zurückkam“ übersetzt wurde und idiomatisch nicht einwandfrei klingt. Hierbei handelt es sich allerdings eher um einen Stilfehler als um einen inhaltlichen Fehler. Es ist

anzunehmen, dass professionelle Translator*innen den Text als maschinell übersetzt einschätzen und Lai*innen gegebenenfalls als Humanübersetzung.

6.5.7 Text 7: Humanübersetzung

Der Text stammt aus dem Themenbereich Politik und wurde sehr idiomatisch von einem Menschen übersetzt (vgl. Annan, Kofi 2015). Er wurde 2015 verfasst und beschreibt die damalige Krise Syriens, Iraks und der Ukraine und die Tatsache, dass sie eine Bedrohung für die Charta der Vereinten Nationen und eine Gefahr für die Demokratie darstellt.

Der Text wurde gewählt, da er sehr idiomatisch übersetzt wurde, wie beispielsweise die Ausgangs- und Zieltextstellen „that **have embraced** democratic ideals“ und „die sich zu demokratischen Idealen **bekennen**“ zeigen. Daher ist anzunehmen, dass die Proband*innen korrekterweise erkennen, dass es sich um eine Humanübersetzung handelt. Weiters wurde zu Beginn des Textes in der Übersetzung eine kleine Änderung vorgenommen, die sinngemäß nur minimal vom Ausgangstext abweicht. Es wurde nämlich „**as 2015 begins**“ auf „**Das Jahresende 2014**“ geändert. Da diese Änderung sofort ins Auge sticht, ist anzunehmen, dass professionelle Translator*innen und auch Lai*innen diese Abweichung erkennen und den Text daher als Humanübersetzung einstufen.

6.5.8 Text 8: Maschinelle Übersetzung

Bei Text 8 handelt es sich um eine Textstelle aus dem Klassiker „Eine Weihnachtsgeschichte“ von Charles Dickens und seine Übersetzung von Leo Feld aus dem Jahr 1917 (vgl. Charles Dickens 1917).

Der Text wurde viel zu wörtlich übersetzt – so wurde zu Beginn beispielsweise „he was dead: to begin with“ mit „Marley war tot: zu Beginn“ übertragen. Ein interessanter und wohl auch für Lai*innen auffälliger Fehler ist, dass *DeepL* den Namen *Scrooge* an einer einzigen Textstelle auf *Dagobert* abänderte. Dabei ist anzunehmen, dass dies daher rührt, dass Onkel Dagobert aus der Disney-Reihe in der englischen Originalversion den Namen Scrooge Mc Duck trägt (vgl. Dolle-Weinkauff 2012). Dieser kommt im Korpus von *DeepL* wohl häufiger vor. Da *DeepL* den Namen nur an einer Stelle änderte, was sehr auffällig ist, kann davon ausgegangen werden, dass beide Gruppen die maschinell übersetzte Version erkennen.

6.5.9 Text 9: Maschinelle Übersetzung

Für Text 9 wurde eine zweite Textpassage von Mark Twains Die Abenteuer des Tom Sawyers ausgewählt, um sie mit der Humanübersetzung zu vergleichen (vgl. Twain 1876). In der kurzen Textpassage geht es um die alltäglichen Sorgen, die sich Tom macht, und darum, dass diese zwischenzeitlich weniger geworden waren.

Hervorzuheben ist der Teil „his troubles were one whit less heavy“, wobei “whit” mit “keinen Deut“ und somit eher wörtlich übersetzt wurde. Da der Inhalt an den anderen Stellen korrekt und auch idiomatisch übertragen wurde, wird bei diesem Text davon ausgegangen, dass Lai*innen nicht erkennen, dass es sich um eine maschinelle Übersetzung handelt.

6.6 Auswertung

Um die in LimeSurvey generierten Ergebnisse der Umfrage auszuwerten, wurden die Umfrageergebnisse zunächst als Excel-Datei heruntergeladen. Die restliche Auswertung erfolgte ausschließlich in Excel, wobei Tabellen der einzelnen Ergebnisse generiert werden. In einem ersten Schritt wurden alle Ergebnisse jener Teilnehmer*innen, die die Umfrage nicht vollständig ausgefüllt hatten, aussortiert.

Anschließend wurde die Spalte der professionellen Translator*innen grün markiert, um mit dem Filter „nach Farbe filtern“ arbeiten zu können. Somit ließen sich in einem nächsten Schritt die beiden Hauptgruppen der professionellen Translator*innen und der Lai*innen einander gut gegenüberstellen. Danach wurde eine Gesamttabelle mithilfe der Summenfunktion von Excel ein Gesamtergebnis der Parameter *Accuracy*, *Fluency* und *Style* in einem neuen Excel-Arbeitsblatt generiert, um aufzuzeigen, wie gut die Texte im Durchschnitt von Lai*innen und professionellen Translator*innen bewertet wurden.

Schlussendlich wurde ausgewertet, ob die Texte von Lai*innen und/oder professionellen Translator*innen korrekt als maschinelle Übersetzung ermittelt worden waren. Hierzu wurden in der obersten Zeile Filter gesetzt und in Prozent berechnet, wie viele Lai*innen und anschließend Translator*innen bei den Texten eine richtige Einschätzung getroffen hatten. Weiters galt es herauszufinden, wie bestimmte Altersgruppen die Qualität wahrgenommen hatten und ob es Unterschiede in den Berufsgruppen gab. Durch die numerische Auswertung in Excel konnte übersichtlich dargestellt werden, wie Lai*innen und professionelle

Translator*innen gesamt und gesondert und in verschiedenen Alters- und Berufsgruppen die Texte bewertet hatten.

7. Ergebnisse

Wie erwähnt, wurden für die Generierung der Ergebnisse die beiden Gruppen der professionellen Translator*innen und Lai*innen gebeten, den in LimeSurvey erstellten Online-Fragebogen auszufüllen. Alle Teilnehmer*innen erhielten denselben Fragebogen. Es war zur Unterscheidung anzugeben, ob man als professionelle*r Übersetzer*in arbeitet. Bei „Ja“ wurde mittels der Funktion *Bedingung* die weitere Frage nach der Branche der Befragten nicht mehr angezeigt. Neben der Unterscheidung in Lai*innen und professionelle Translator*innen wurde auch bei englischen und deutschen Erstsprachler*innen mit dieser Funktion gearbeitet.

Bei dieser Frage ging es darum, welche Erstsprache die Teilnehmer*innen haben. Wurde „Ist Deutsch Ihre Erstsprache?“ mit „Ja“ beantwortet, so wurde die Frage nach dem Sprachenlevel gar nicht erst angezeigt. Dadurch war es möglich, mehrere Personengruppen mit nur einem Fragebogen zu erreichen. Insgesamt nahmen 118 Personen an der Umfrage teil, wobei der Fragebogen von 44 Personen vollständig ausgefüllt wurde. Die Hauptgruppen der Lai*innen und professionellen Translator*innen wurden beide in ähnlicher Proportion erreicht, nämlich 21 Lai*innen und 23 professionelle Translator*innen.

7.1 Demographische Daten

Der Großteil der Teilnehmer*innen war weiblich. Es nahmen 34 weibliche Personen (77 %) und 10 männliche Personen (23 %) an der Umfrage teil. Die Auswahl eines dritten Geschlechts wurde dementsprechend 0-mal (0 %) ausgewählt.

Was das Alter betrifft, so umfasste die Gruppe der 19 bis 25-Jährigen 14 Personen (31,8 %) und die Gruppe der 26 bis 35-Jährigen 12 Personen (27 %). Weiters war die Gruppe der 36 bis 55-Jährigen mit nur sechs Personen (also 13,6 %) vertreten; und nur zwei Personen (4,55 %) waren in einem Alter zwischen 56 und 70 Jahren. Eine Person (2,27 %) der Gesamtteilnehmer*innen war über 70 Jahre alt. Somit war die jüngste Altersgruppe zwischen 19 und 25 Jahre am größten.

Bezogen auf die Ausbildung hat keine*r der Teilnehmer*innen, die die Umfrage vollständig ausfüllten, angegeben, eine Lehre als höchsten Bildungsabschluss absolviert zu haben. Fünf

Personen (11,36 %) beendeten eine höhere berufliche Ausbildung. Mit 86,36 % hatte der Großteil der Teilnehmer*innen (38 Personen) ein Studium absolviert; und zwei Personen hatten darüber hinaus ein Doktoratsstudium abgeschlossen.

Der Anteil der professionellen Translator*innen und Lai*innen war – wie bereits erläutert – beinahe gleich hoch: 23 Teilnehmer*innen (52,27 %) waren als professionelle Translator*innen tätig, während 21 Teilnehmer*innen (47,73 %) angaben, in einer anderen Berufsgruppe zu arbeiten. Somit konnten die beiden Gruppen gut miteinander verglichen werden.

Was die Branche der Lai*innen betrifft, so waren drei Teilnehmer*innen (6,82 %) in der Wissenschaft tätig und weitere drei (6,82 %) arbeiteten in der Administration oder führten eine Büroarbeit aus, wie in Abb. 4 dargestellt. In den weiteren Berufssparten, wie Handel, Lehr- und Kulturbereich, Metall- und Elektrobereich, Sozialberufe und Technik war jeweils eine Person tätig (2,27 % pro Sparte). Der Großteil, nämlich neun Personen, arbeiteten in sonstigen Berufsgruppen (43 %).

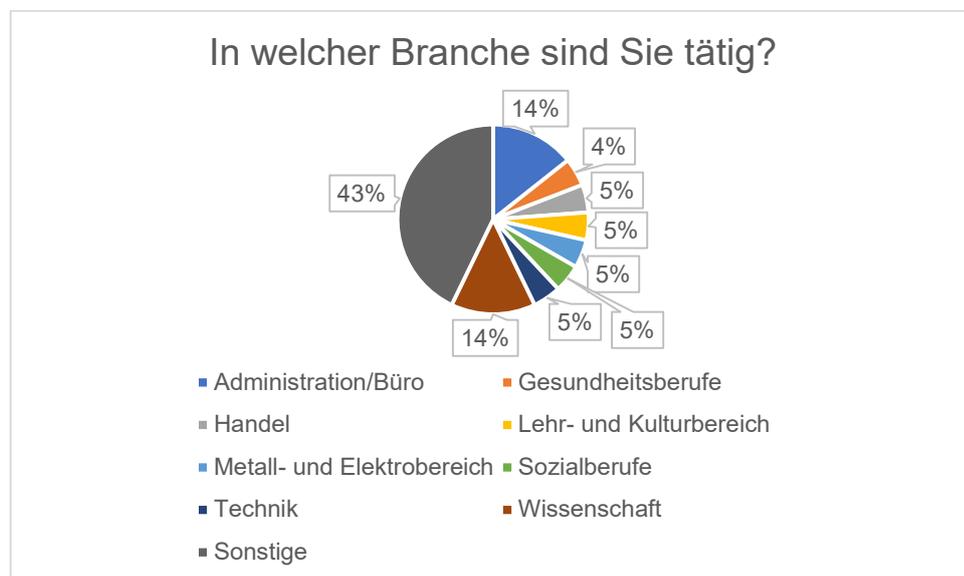


Abbildung 4: In welcher Branche sind Sie tätig?

Was die Erstsprache betrifft, haben 42 Personen (95,46 %) angegeben, die Sprache Deutsch als Erstsprache zu sprechen, wie in Abb. 5 ersichtlich. Die zwei Personen (4,55 %) mit Deutsch als Zweitsprache gaben an, dass sie Deutsch auf C2-Niveau beherrschen. Drei Personen (6,82 %) gaben zusätzlich an, dass neben Deutsch auch Englisch ihre Erstsprache ist; diese Funktion bzw. Angabemöglichkeit ließ der Fragebogen zu und ist in Abb. 6 dargestellt.

Bei der Frage nach der Sprachkompetenz haben zwei Teilnehmer*innen (4,55 %) angegeben, dass sie Englisch auf dem Niveau A2 beherrschen. Drei Proband*innen (6,82 %) gaben an, Englisch auf B1-Niveau zu beherrschen und weitere sieben Personen (15,90 %) gaben das Niveau B2 an. Weitere neun Personen (20,45 %) gaben an, dass sie Englisch auf C1-Niveau beherrschen. Der Großteil der Teilnehmer*innen, nämlich 23 Personen (52,27 %), gab an, Englisch auf C2-Niveau zu beherrschen. Die zwei Teilnehmer*innen (5 %), deren Erstsprache nicht Deutsch war, gaben an, dass sie Deutsch auf C2-Niveau sprechen.

Bei der Frage nach der Verwendung von frei verfügbaren maschinellen Übersetzungssystemen wie DeepL, SYSTRAN und Google Translate gaben lediglich fünf Personen (11,36 %) an, dass sie noch nie ein maschinelles Übersetzungssystem genutzt hatten.

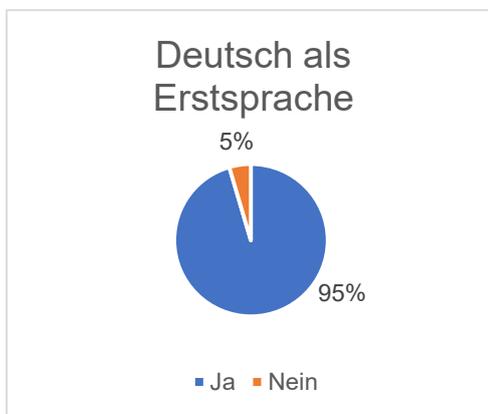


Abbildung 5: Deutsch als Erstsprache

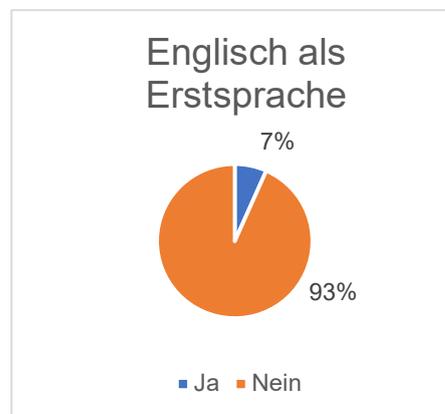


Abbildung 6: Englisch als Erstsprache

Hierbei ist aufschlussreich, wie viele professionelle Translator*innen und Lai*innen maschinelle Übersetzungssysteme nutzten. In der Gruppe der Translator*innen nutzten 21 Personen (91,30 %) maschinelle Übersetzungssysteme, wohingegen nur zwei Personen (8,70 %) die Systeme nicht nutzten. Die Frage, ob maschinelle Übersetzungssysteme auch zu beruflichen Zwecken genutzt werden, wurde nicht gestellt und bleibt somit offen. Auch der Großteil der Lai*innen gab an, bereits mit maschinellen Übersetzungssystemen gearbeitet zu haben, nämlich insgesamt 18 Personen (85,71 %) Lediglich drei Personen (14,29 %) der Lai*innen hatten noch nicht damit gearbeitet.

7.2 Bewertung der Texte

Im folgenden Kapitel werden die Ergebnisse der Textbewertung vorgestellt. Zunächst wird eine Gesamtbewertung der Texte abgegeben und anschließend werden die Unterschiede zwischen Lai*innen und professionellen Translator*innen präsentiert. Um eruieren zu können, wie gut

die maschinellen Texte im Gegensatz zu den Humanübersetzungen eingestuft worden sind, wurde zunächst eine Gesamtauswertung der Texte erstellt, bei der in Excel mit der Funktion Metrik alle Bewertungen (in den Bereichen *Accuracy*, *Fluency* und *Style*) zusammengezählt und dividiert.

7.2.1 Gesamtbewertung der Texte

	Gesamtauswertung aller Proband*innen	Accuracy	Fluency	Style
Text 1 (HÜ)	2,65	2,60	2,50	2,90
Text 2 (MÜ)	2,11	1,80	2,20	2,40
Text 3 (MÜ)	1,96	1,80	1,90	2,20
Text 4 (HÜ)	2,53	2,40	2,40	2,70
Text 5 (HÜ)	1,88	1,80	1,90	1,90
Text 6 (MÜ)	2,21	2,00	2,20	2,40
Text 7 (HÜ)	1,95	2,00	1,80	2,10
Text 8 (MÜ)	3,50	3,60	3,40	3,70
Text 9 (MÜ)	2,35	2,30	2,40	2,50

Tabelle 1: Gesamtbewertung der Texte als Durchschnittswert auf Basis der drei Kategorien Accuracy, Fluency und Style und einer Skalenbewertung von 1-5

Wie in Tabelle 1 ersichtlich, erhielt keine der Übersetzungen von den Proband*innen (44 Personen, 100 %) eine schlechtere Durchschnittsnote als 3,50 (bei Text 8, einer maschinellen Übersetzung) auf einem Spektrum von der Bestnote 1 bis zur schlechtesten Bewertung mit der Note 5. Die Proband*innen bewerteten Text 5, eine Humanübersetzung, mit der Note 1,88 am besten. Die maschinellen Übersetzungen erhielten im Durchschnitt die Note 2,42 und die Humanübersetzungen die Note 2,30. Im Bereich *Accuracy* wurden sowohl zwei maschinelle Übersetzungen als auch eine Humanübersetzung mit der Bestnote 1,80 bewertet. Am schlechtesten wurde Text 8, eine maschinelle Übersetzung, bewertet. Diese erhielt die Note 3,60. Was den Parameter *Fluency* betrifft, so erhielt eine Humanübersetzung die Note 1,80. Am schlechtesten wurde auch hier eine maschinelle Übersetzung mit 3,40 bewertet. Im Bereich *Style* wurde Text 5, eine Humanübersetzung, mit 1,90 am besten bewertet; die schlechteste Note erhielt wiederum Text 8, der maschinell übersetzt wurde.

	Gesamtbewertung Humanübersetzungen	Accuracy	Fluency	Style
Text 1	2,65	2,60	2,50	2,90
Text 4	2,53	2,40	2,40	2,70
Text 5	1,88	1,80	1,90	1,90
Text 7	1,95	2,00	1,80	2,10
Gesamtdurchschnitt	2,25	2,20	2,15	2,40

Tabelle 2: Durchschnittliche Gesamtbewertung der Humanübersetzungen

Wie in Tabelle 2 ersichtlich, erhielten die Humanübersetzungen von den 44 Proband*innen (100 %) Noten zwischen 1,88 und 2,65. Am besten wurde der Parameter *Fluency* gesamtdurchschnittlich mit der Note 2,15 bewertet, gefolgt von *Accuracy* mit der Note 2,20. Der Bereich *Style* schnitt im Durchschnitt mit der Note 2,20 ab.

	Gesamtbewertung Maschinelle ÜS	Accuracy	Fluency	Style
Text 2	2,11	1,80	2,20	2,40
Text 3	1,96	1,80	1,90	2,20
Text 6	2,21	2,00	2,20	2,40
Text 8	3,50	3,60	3,40	3,70
Text 9	2,55	2,30	2,40	2,50
Gesamtdurchschnitt	2,47	2,30	2,42	2,64

Tabelle 3: Durchschnittliche Gesamtbewertung der maschinellen Übersetzungen

Im Gegensatz dazu erhielten die maschinellen Übersetzungen eine minimal schlechtere Gesamtdurchschnittsnote von 2,47, wie in Tabelle 3 ersichtlich. Am besten wurden die maschinellen Übersetzungen im Bereich *Accuracy* bewertet: Hier wurde die Note 2,30 vergeben. Der Parameter *Fluency* wurde mit 2,42 bewertet und im Bereich *Style* erhielten die maschinellen Übersetzungen eine Durchschnittsnote von 2,64.

Die maschinellen Übersetzungen wurden somit etwas schlechter bewertet als die Humanübersetzungen. Die Bestnote der maschinellen Übersetzungen lag bei 1,96 (Text 3) im Vergleich zur Bestnote der Humanübersetzungen bei 1,88 (Text 5). Beim am schlechtesten bewerteten Text handelt es sich um eine maschinelle Übersetzung mit der Note 3,5, wobei die schlechteste Humanübersetzung mit 2,65 benotet wurde.

7.2.2 Korrekte Erkennung der maschinellen Übersetzung

Im folgenden Kapitel werden die Antworten zur Frage dargestellt, ob die Texte maschinell übersetzt wurden oder nicht. Zunächst wird die Gesamtauswertung aller Teilnehmer*innen präsentiert und anschließend werden die Unterschiede in den Gruppen der Lai*innen und der professionellen Translator*innen dargestellt.

MÜ: Wurde der Text maschinell erstellt?	Gesamte Teilnehmer*innenanzahl (44 P., 100 %)	
Text 2	Ja (26 P./59,09 %)	Nein (18 P./40,91 %)
Text 3	Ja (16 P./36,36 %)	Nein (28 P./63,64 %)
Text 6	Ja (23 P./52,27 %)	Nein (21 P./47,43 %)
Text 8	Ja (41 P./93,18 %)	Nein (3 P./6,82 %)
Text 9	Ja (23 P./52,27 %)	Nein (21 P./47,73 %)

Tabelle 4: Häufigkeit der erkannten maschinellen Übersetzungen

Wie in Tabelle 4 dargestellt, erkannten die Proband*innen viermal mehrheitlich (>50 %) richtig, dass es sich bei einem der Texte um eine maschinelle Übersetzung handelte. Bei Text 8, dem Ausschnitt aus der Weihnachtsgeschichte von Charles Dickens, waren die Ergebnisse

sehr aussagekräftig, denn 93,18 % erkannten korrekt, dass es sich hierbei um eine maschinelle Übersetzung handelte. Dies liegt vermutlich daran, dass, wie bereits erwähnt, die Syntax nicht idiomatisch übertragen wurde und der Name *Scrooge* in diesem Kontext falsch mit *Dagobert* übersetzt worden war.

Bei den anderen Texten waren die Antworten jedoch nicht eindeutig, sondern beinahe ausgeglichen, wie etwa bei Text 6, wo 52,27 % für eine maschinelle Übersetzung stimmten und 47,43 % dagegen. Bei Text 3 schätzte die Mehrheit der Proband*innen (63,64 %) die Übersetzung sogar als eine Humanübersetzung ein. Dabei handelte es sich hierbei um das maschinell übersetzte Modulprüfungsbeispiel. Es ist anzunehmen, dass der Text mehrheitlich als Humanübersetzung eingestuft wurde, da die Termini korrekt übertragen wurden und die Satzstruktur beibehalten wurde.

HÜ: Wurde der Text maschinell erstellt?	Gesamte Teilnehmer*innenanzahl (44 P.), 100 %	
Text 1	Ja (38 P./86,36 %)	Nein (6 P./13,64 %)
Text 4	Ja (12 P./50 %)	Nein (12 P./50 %)
Text 5	Ja (8 P./18,18 %)	Nein (36 P./81,82 %)
Text 7	Ja (3P./13,64 %)	Nein (41 P./93,18 %)

Tabelle 5: Häufigkeit der erkannten Humanübersetzungen

Was die Humanübersetzungen betrifft, so stimmte die Einschätzung der Proband*innen dreimal mehrheitlich, wie in Tabelle 5 dargestellt. Im Gegensatz zu den maschinellen Übersetzungen waren die Ergebnisse hierbei sogar aussagekräftiger, wie anhand der Texte 1, 5 und 7 ersichtlich ist. Bei Text 1, der mit 86,36% mehrheitlich als maschinelle Übersetzung eingestuft wurde, handelt es sich um einen Zeitungsbericht und Text 5, der von 81,82% der Proband*innen als Humanübersetzung bewertet wurde, ist ein Zeitungsartikel zu einem

wirtschaftlichen Thema. Da in Text 5 Abkürzungen für das Zielpublikum erklärt wurden, ist anzunehmen, dass dieser deshalb als Humanübersetzung eingestuft worden ist.

Gesamtergebnis	Der Text ist eine Humanübersetzung	Der Text ist eine maschinelle Übersetzung
Gesamtübersicht Humanübersetzungen	105 Antworten/59,66 %	71 Antworten/40,34 %
Gesamtübersicht maschinelle Übersetzungen	91 Antworten/41,36 %	129 Antworten/58,64 %

Tabelle 6: Gesamtergebnis MÜ und HÜ

In Tabelle 6 ist das prozentuelle Ergebnis aller Teilnehmer*innen für die Humanübersetzungen und maschinellen Übersetzungen dargestellt. Es wurden die Gesamtergebnisse addiert und davon wurde das prozentuelle Ergebnis der Gesamtantworten berechnet.

Wie in Tabelle 6 auch ersichtlich, stuften 60,91 % der Teilnehmer*innen die Humanübersetzungen korrekt als solche ein. Was die maschinellen Übersetzungen betrifft, so stuften sie 58,64 % der Proband*innen korrekt als solche ein. Dies zeigt, dass das Gesamtergebnis der Proband*innen bei beiden Arten der Übersetzung sehr nahe beieinander liegt.

In den folgenden Tabellen 7 und 8 wird dargestellt, wie häufig Lai*innen erkannten, ob es sich um eine maschinelle Übersetzung handelt oder nicht.

Maschinelle Übersetzung: Handelt es sich um eine maschinelle Übersetzung? (Lai*innen, 21 P., 100 %)		
Text	Ja	Nein
Text 2	7 P./33 %	14 P./ 66,66 %
Text 3	9 P./42,86 %	12 P./57,14 %
Text 6	13 P./61,90 %	8 P./38,10 %
Text 8	20 P./95,24 %	1 P./4,76 %
Text 9	10 P./47,62 %	11 P./52,38 %

Tabelle 7: Gesamtergebnis Lai*innen (MÜ)

Humanübersetzungen: Handelt es sich um eine maschinelle Übersetzung? (Lai*innen, 21 P., 100 %)		
Text	Ja	Nein
Text 1	17 P./80,95 %	4 P./19,07 %
Text 4	12 P./57,14 %	9 P./42,86 %
Text 5	4 P./19,05 %	17 P./80,95 %
Text 7	3 P./14,29 %	18 P./85,71 %

Tabelle 8: Gesamtübersicht Lai*innen (HÜ)

Wie aus Tabelle 7 und 8 ersichtlich ist, schätzte die Mehrheit der Lai*innen (>50 %) insgesamt nur bei drei Texten richtig ein, ob es sich um eine maschinelle Übersetzung oder eine Humanübersetzung handelt.

Die Prozentzahl der Ergebnisse der Gruppe waren nicht immer eindeutig. Diese waren teilweise sehr knapp, wie aus den Texten 4 (Zeitungsbericht), 5 (Zeitungsbericht) und 9 (Literaturübersetzung) hervorgeht. Bei Text 3 (Prüfungstext) lagen neun Personen (42,86 %) richtig und 12 Personen (57,14 %) falsch. Auch bei Text 4 (Literaturübersetzung) lagen 12 Personen (57,14 %) richtig und neun Personen (42,86 %) falsch. Ebenso bei Text 9 (Literaturübersetzung) waren die Anteile der richtigen und falschen Antworten beinahe gleich hoch. Der Text wurde von elf Personen (52,38 %) falsch als Humanübersetzung eingestuft und von zehn Personen (47,62 %) richtig als maschinelle Übersetzung. Insgesamt lagen die Lai*innen bei den maschinellen Übersetzungen viermal richtig und viermal falsch; bei den Humanübersetzungen erkannten sie zweimal korrekt, dass es sich um eine Humanübersetzung handelt; dreimal lagen sie falsch.

Maschinelle Übersetzung: Handelt es sich um eine maschinelle Übersetzung? (prof. T., 21 P., 100 %)		
Text	Ja	Nein
Text 2	19 P./82,61 %	4 P./17,39 %
Text 3	7 P./30,43 %	16 P./69,57 %
Text 6	10 P./43,48 %	13 P./56,52 %
Text 8	21 P./91,30 %	2 P./8,70 %
Text 9	13 P./56,52 %	10 P./43,48 %

Tabelle 9: Handelt es sich bei MÜ um HÜ?
(prof. Translator*innen)?

Humanübersetzungen: Handelt es sich um eine maschinelle Übersetzung? (prof. T., 21 P., 100 %)		
Text	Ja	Nein
Text 1	18 P./78,26 %	5 P./21,74 %
Text 4	10 P./43,48 %	13 P./56,52 %
Text 5	4 P./17,39 %	19 P./82,61 %
Text 7	3 P./13,04 %	20 P./86,96 %

Tabelle 10: Handelt es sich bei HÜ um HÜ?
(prof. Translator*innen)?

Wie in Tabelle 9 und 10 ersichtlich, erkannte die Gruppe der professionellen Translator*innen häufiger, ob es sich um eine maschinelle Übersetzung handelt oder nicht. Insgesamt lag die Mehrheit der Gruppe (>50 %) sechsmal richtig und nur dreimal falsch.

Was die maschinellen Übersetzungen betrifft, so erkannten die professionellen Translator*innen bei drei Texten korrekt, dass es sich dabei um eine solche Übersetzung handelt. Bei den Humanübersetzungen lagen sie ebenso dreimal als Mehrheit (> 50 %) richtig.

Im Vergleich zu den Lai*innen waren die Ergebnisse meist eindeutiger und prozentual nicht knapp beieinander, wie etwa bei Text 8, wo 21 Personen (91,30 %) mit „Ja“ und nur zwei Personen (8,70 %) mit „Nein“ abstimmten. Lediglich bei Text 4, Text 6 und Text 9 waren die Ergebnisse denen der Lai*innen ähnlich. Bei Text 4 und Text 6 stimmten zehn Personen für „Ja“ (43,48 %) und 13 Personen (56,52 %) für „Nein“; bei Text 9 stimmten 13 Personen (56,52 %) für „Ja“, während zehn Personen (43,48 %) für „Nein“ stimmten.

7.2.3 Gesamtübersicht über die Einschätzung von Lai*innen und professionellen Translator*innen

Wie in Tabelle 11 dargestellt, erkannte die Mehrheit der professionellen Translator*innen (über 50 %) häufiger, um welche Art der Übersetzung es sich bei den Texten handelt. Die Gruppe der Lai*innen erkannte bei vier von neun Texten korrekt, ob es sich um eine maschinelle oder eine Humanübersetzung handelt, wohingegen die Gruppe der professionelle Translator*innen bei sechs von neun Texten korrekt angegeben hat, ob der Text maschinell oder von einem Menschen übersetzt wurde. Wie bereits zu den vorherigen Tabellen erwähnt und auch in der Tabelle 11 angeführt, liegen die Ergebnisse jedoch häufig sehr nahe beieinander, wie etwa bei Text 4, wo 57,14 % der Lai*innen falsch und 56,52 % der professionellen Translator*innen richtig erkannten, um welche Art der Übersetzung es sich handelt.

Gesamtübersicht der Einschätzung von Lai*innen und professionellen Translator*innen über die Art der Übersetzung		
Text	Lai*innen gesamt mehrheitlich richtig gestimmt	Professionelle Translator*inne n gesamt mehrheitlich richtig gestimmt
Text 1 (HÜ)	Nein (80,95 % stimmten falsch)	Nein (78,26 % stimmten falsch)

Text 2 (MÜ)	Nein (66,66 % stimmten falsch)	Ja (82,61 % stimmten richtig)
Text 3 (MÜ)	Nein (57,14 % stimmten falsch)	Nein (69,57 % stimmten falsch)
Text 4 (HÜ)	Nein (57,14% stimmten falsch)	Ja (56,52% stimmten richtig)
Text 5 (HÜ)	Ja (80,95 % stimmten richtig)	Ja (82,61 % stimmten richtig)
Text 6 (MÜ)	Ja (61,90 % stimmten richtig)	Nein (56,52 % stimmten falsch)
Text 7 (MÜ)	Ja (85,71 % stimmten richtig)	Ja (86,96 % stimmten richtig)
Text 8 (MÜ)	Ja (95,24 % stimmten richtig)	Ja (91,30 % stimmten richtig)
Text 9 (MÜ)	Nein (52,38% stimmten falsch)	Ja (56,52% stimmten falsch)

Tabelle 11: Gesamtübersicht Lai*innen
und prof. Translator*innen

7.2.4 Schwierigkeitsgrad für die Proband*innen bei der Bewertung der Texte

Nachdem sie die Texte vorgelegt bekamen, wurden die Proband*innen befragt, wie leicht es ihnen gefallen war, maschinelle Übersetzungen zu erkennen (Abb. 8 und 9). Mit 22 Personen (50 %) gab die Hälfte der Proband*innen zum Thema Erkennen der MÜ „mittelschwer

gefallen“ an. Nur neun Personen (20,45 %) gaben an, dass es ihnen leichtgefallen war, dies zu erkennen. Eine MÜ zu erkennen ist sechs Personen schwergefallen (13,63 %) und eine Person (4,76 %) fand die Bewertung sehr schwer. Für lediglich zwei Personen (4,54 %) war es sehr leicht, die Bewertung durchzuführen.

Ähnlich zu den Ergebnissen der Lai*innen (Abb. 7) empfand auch der Großteil der professionellen Translator*innen (13 Personen, 59 %) als mittelschwer, die Texte richtig als MÜ oder HÜ einzustufen. Dies ist aus Abb. 8 ersichtlich. In dieser Gruppe anzumerken ist jedoch, dass die Zuordnung nur eine Person schwer fand (4,35 %), wohingegen fünf Lai*innen (25,80 %) sie schwer fanden. Sechs Translator*innen fiel die Aufgabe (27,08 %) leicht; im Vergleich zu drei Lai*innen (15,29 %). Eine Person in der Gruppe der Lai*innen (4,76 %) empfand es als sehr leicht; ebenso zwei professionelle Translator*innen (8,70 %).

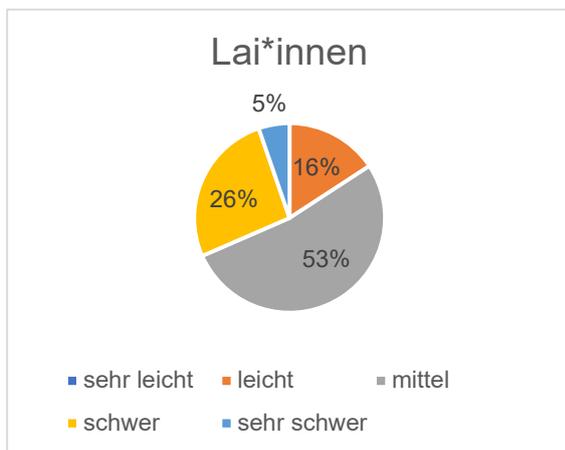


Abbildung 7: Schwierigkeitsgrad Texterkennung Lai*innen

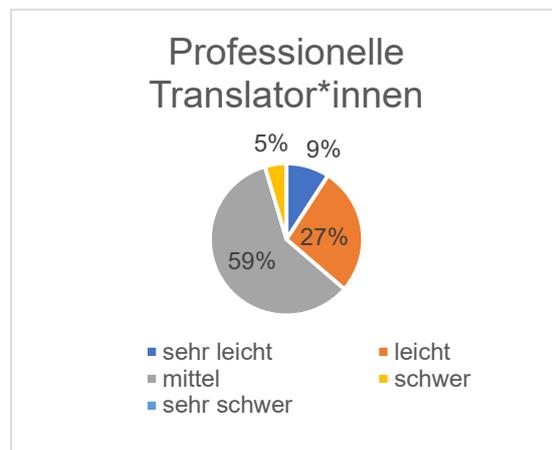


Abbildung 8: Schwierigkeit Texterkennung professionelle Translator*innen

7.2.5 Wichtigkeit der Aspekte Inhalt, Korrektheit und Stil für die Teilnehmer*innen

Um zu erfahren, wie wichtig den Teilnehmer*innen die Bereiche Inhalt, Korrektheit und Stil sind, wurden sie gebeten, diese auf einer Notenskala von 1 bis 5 zu bewerten, wie in der folgenden Tabelle 12 präsentiert.

	Inhalt (Skala von 1-5)	Korrektheit (Skala von 1-5)	Stil (Skala von 1-5)
Lai*innen	1,61	1,61	2,04

Professionelle Translator*innen	1,21	1,13	1,70
Gesamtdurchschnitt	1,41	1,37	1,87

Tabelle 12: Einschätzung Inhalt, Korrektheit und Stil

Wie aus Tabelle 12 hervorgeht, ist für den Großteil der Lai*innen der Inhalt sehr wichtig: 61,90 % vergaben die Note 1. Weitere vier Personen (19,05 %) vergaben die Note 2 im Bereich Inhalt. Lediglich drei Personen (14,29 %) vergaben die Note 3 und nur eine Person (4,76 %) vergab die Note 4.

Was die Korrektheit betrifft, so hat auch hier die Mehrheit die Note 1 vergeben, nämlich 12 Personen (57,14 %); weitere sieben Personen (33,33 %) vergaben die Note 2. Die Note 3 wurde von einer Person (4,76 %) vergeben; ebenso die Note 5 (4,76 %). Die Note 4 wurde im Bereich Korrektheit nicht vergeben. Was den Stil betrifft, so vergaben vier Personen (19,05 %) die Note 1 und 12 Personen (57,14 %) die Note 2. Weitere fünf Personen (23,81 %) vergaben die Note 3. Die Noten 4 und 5 wurden in diesem Bereich nicht vergeben.

Auch die Mehrheit der professionellen Translator*innen (18 Personen, 78,26 %) bewertete den Inhalt mit der Note 1 als besonders wichtig. Weitere fünf Personen (21,74 %) vergaben die Note 2. Die Noten 3-5 wurden von den professionellen Translator*innen nicht vergeben. Was die Korrektheit der Texte betrifft, so vergaben 20 Personen (86,96 %) die Note 1 und zwei Personen (8,70 %) die Note 3. Die Noten 2, 4 und 5 wurden nicht vergeben. Was den Stil betrifft, so vergaben neun Personen (39,13 %) die Note 1. Die Mehrheit vergab beim Stil die Note 2, nämlich 12 Personen (52,17 %). Weitere zwei Personen (8,70 %) vergaben in diesem Bereich die Note 3.

7.3 Nutzung der maschinellen Übersetzungssysteme

Der Großteil der Teilnehmer*innen gab an, bereits mit einem maschinellen Übersetzungstool wie DeepL, SYSTRAN oder Google Translate gearbeitet zu haben (insgesamt 39 Personen, 88,64 %). Nur fünf Personen (11,36 %) gaben hingegen an, noch nicht mit einem der Tools gearbeitet zu haben. In der Gruppe der Lai*innen gaben drei Personen (14,28 %) an, bislang nicht mit Übersetzungstools gearbeitet zu haben; in der Gruppe der professionellen

Translator*innen waren es zwei Personen (8,70 %). Die Personen, die noch nicht mit Übersetzungstools gearbeitet hatten, waren – im Gegensatz zur Erwartungshaltung der Autorin – nicht älter als 26-35 Jahre. Eine Person (2,27 %) war im Alter von 19-25 Jahren und drei Personen (6,82 %) waren zwischen 26-35 Jahren.

7.4 Detailergebnisse nach Alter und Sprachkompetenz

Im folgenden Kapitel werden die Unterschiede nach Altersgruppen und Sprachkompetenz dargestellt.

7.4.1 Unterschiede in den Altersgruppen

Wie bereits in Kapitel 6.6.1 erwähnt, war die Gruppe der 26 bis 35-Jährigen (21 Personen, 47,73 %) am häufigsten vertreten. Nachstehend wird dargestellt, welchen Einfluss das Alter auf die Bewertung hatte.

Alter (Profis und Lai*innen, 44 P., 100 %)	Gesamt	MÜ	HÜ
19-25 Jahre (14 P., 31,82 %)	2,88	2,87	2,89
26-35 Jahre (21 P., 47,73 %)	2,06	2,12	2,00
36-55 Jahre (6 P., 13,64 %)	2,28	2,53	2,02
56-70 Jahre (2 P., 4,55 %)	2,28	2,60	1,96
Über 70 Jahre (1 P., 2,27 %)	1,69	1,87	1,50

Tabelle 13: Unterschiede Altersgruppen

Am besten wurden die Translate, wie in Tabelle 13 ersichtlich, von der Gruppe der über 70-Jährigen bewertet (eine Person, 2,27 %). Diese vergab eine Gesamtdurchschnittsnote von 1,50 und bewertete die maschinellen Übersetzungen und Humanübersetzungen insgesamt am

besten. Sie zählte zur Gruppe der Lai*innen. Die maschinellen Übersetzungen wurden mit 1,87 bewertet und die Humanübersetzungen mit 1,50. Da es sich um eine einzige Bewertung handelt, ist diese jedoch für diese Altersgruppe nicht repräsentativ.

Nach der Gruppe der 70-Jährigen vergab die Gruppe der 26 bis 35-Jährigen (21 Personen, 47,73 %) mit einer Gesamtdurchschnittsnote von 2,00 die besten Noten insgesamt. Die Gruppe der 36 bis 55-Jährigen (13,64 %) bewertete die Übersetzungen am drittbesten mit einem Gesamtdurchschnitt von 2,02. Diese Gruppe vergab für die Humanübersetzungen in allen drei Bereichen bessere Noten. Die Gruppe der 56 bis 70-Jährigen (4,55 %) vergab mit einem Gesamtdurchschnitt von 2,28 ähnliche Noten. Sie bewertete die maschinellen Übersetzungen mit 2,60 und die Humanübersetzungen mit 1,96. Die Gruppe der 19 bis 25-Jährigen (31,82 %) bewertete die Texte mit einem Gesamtdurchschnitt von 2,88 in allen Bereichen am schlechtesten. Die maschinellen Übersetzungen erhielten einen Gesamtdurchschnitt von 2,87 und die Humanübersetzungen eine Note von 2,89. Interessant ist, dass die jüngste Gruppe auch die kritischste Bewertung abgegeben hat, wohingegen die zweitjüngste Gruppe die beste Bewertung abgegeben hat. Beide Gruppen zusammen stellen die Mehrheit der Teilnehmer*innen dar.

7.4.2 Alter der professionellen Translator*innen

Für die Studie ebenso aufschlussreich war die Beantwortung der Frage, ob ältere Translator*innen die maschinellen Übersetzungen aufgrund ihrer Berufserfahrung strenger bewerten als jüngere Kolleg*innen. Dies konnte nicht bestätigt werden, wie in Tabelle 14 ersichtlich ist. Am strengsten wurden die Translate von der Gruppe der 19 bis 25-Jährigen Translator*innen mit einem Gesamtdurchschnitt von 2,66 bewertet. Sie bewerteten die maschinellen Übersetzungen mit 2,75 und die Humanübersetzungen mit 2,60.

Die Gruppe der 26 bis 35-Jährigen (43,48 %) der professionellen Translator*innen hingegen vergab die besten Noten – ein Ergebnis, das deckungsgleich mit der Gruppe der Lai*innen ist (Durchschnitt von 2,04). Sie bewerteten die MÜ mit 2,24 und die HÜ mit 1,91.

Alter professionelle Translator*innen (23 P./100 %)	Gesamtdurchschnitt	MÜ	HÜ

19-25 Jahre (5 P./21,74 %)	2,66	2,75	2,60
26-35 Jahre (10 P./43,48 %)	2,04	2,24	1,91
36-55 Jahre (5 P./21,74 %)	2,36	2,57	2,02
56-70 Jahre (2 P./8,70 %)	2,45	2,60	1,96
Über 70 Jahre (1 P./4,35 %)	1,69	1,87	1,50

Tabelle 14: Altersanteil professionelle Translator*innen

7.4.3 Alter der Lai*innen

In Tabelle 15 wird dargestellt, wie die Proband*innen die Texte nach Altersgruppe bewertet haben. Auch hier bewertete die Gruppe der 19 bis 25-Jährigen die Übersetzungen gesamtdurchschnittlich am strengsten mit einer Note von 2,76. Im Gegensatz dazu wurden die Translate von der Gruppe der 26 bis 35-Jährigen am besten mit einer Gesamtdurchschnittsnote von 1,68 bewertet. Diese Gruppe bewertete gemeinsam mit der Gruppe der 36 bis 55-Jährigen die maschinellen Übersetzungen am besten. Beide vergaben eine Durchschnittsnote von 2,00. Die Humanübersetzungen wurden von den Lai*innen generell strenger bewertet als die maschinellen Übersetzungen. Die Gruppe der 26 bis 35-Jährigen vergab die beste Durchschnittsnote mit 2,12, gefolgt von der Gruppe der 19 bis 25-Jährigen, die eine Durchschnittsnote von 2,87 vergab. Die Gruppe der 36 bis 55-Jährigen hat in diesem Bereich die schlechteste Durchschnittsnote von 2,92 vergeben.

Alter der Lai*innen (21 P./100 %)	Gesamtdurchschnitt	MÜ	HÜ

19-25 Jahre (9 P./42,86 %)	2,76	2,67	2,87
26-35 Jahre (11 P./52,38 %)	1,68	2,00	2,12
36-55 Jahre (1 P./4,35 %)	2,19	2,00	2,92

Tabelle 15: Altersanteil Lai*innen

7.4.4 Unterschiede in der Sprachkompetenz

Wie im Kapitel zu den demographischen Angaben der Umfrageteilnehmer*innen ersichtlich, gaben drei Personen an, dass auch Englisch ihre Erstsprache sei. Im folgenden Kapitel wird dargestellt, wie sich die Sprachkompetenz auf die Bewertung der Translate auswirkt und wie häufig Personen mit deutscher und englischer Sprachkompetenz erkennen, ob es sich um eine maschinelle Übersetzung handelt.

Sprachlevel Englisch (Erstsprache Deutsch)	Gesamtdurchschnitt	MÜ	HÜ
Niveau A2 (2 P./4,88 %)	2,09	1,87	2,30
Niveau B1 (3 P./14,29 %)	2,18	2,38	1,98
Niveau B2 (7 P./33,33 %)	2,64	2,58	2,69
Niveau C1 (9 P./42,86 %)	2,56	2,50	2,62

Niveau C2 (15 P./71,43 %)	2,18	2,34	2,02
--	------	------	------

Tabelle 16: Unterschiede Sprachkompetenz Gesamtteilnehmer*innen

Entgegen der Erwartung, dass die Strenge der Bewertungen von – insbesondere maschinellen – Übersetzungen mit höherem Sprachniveau steigen würde, lieferten die diesbezüglichen Antworten der Umfrage ein anderes Ergebnis (siehe Tabelle 16). Am besten wurden die Translate zwar von den Personen mit Sprachniveau A2 mit einem Gesamtdurchschnitt von 2,09 bewertet. Sie bewerteten auch die maschinellen Übersetzungen am besten mit einem Durchschnitt von 1,87. Dennoch vergaben aber auch jene Personen mit Sprachniveau C2 – fünf Lai*innen und zehn professionelle Translator*innen – eine gute Gesamtdurchschnittsnote von 2,18 für alle Translate. Sie bewerteten die maschinellen Übersetzungen mit 2,34 und die Humanübersetzungen mit 2,02.

Ganz knapp dahinter liegt die gesamtdurchschnittliche Bewertung der Übersetzungen der Personen mit dem Sprachniveau B1. Sie vergaben dieselbe Gesamtdurchschnittsnote von 2,18 und unterscheiden sich sowohl bei den Humanübersetzungen als auch bei den maschinellen Übersetzungen nur minimalst von der Sprachniveau-Gruppe C2.

Die Teilnehmer*innen mit dem Niveau C1 (vier Lai*innen und fünf professionelle Übersetzer*innen) bewerteten die Übersetzungen etwas strenger mit einer Gesamtdurchschnittsnote von 2,56.

Am strengsten bewerteten jedoch die Teilnehmer*innen mit dem Sprachniveau B2 die Translate. Anzumerken hierbei ist, dass diese sieben Personen alle La*innen waren. Sie vergaben eine Gesamtdurchschnittsnote von 2,64 und bewerteten die maschinellen Übersetzungen mit der Note 2,58 und die Humanübersetzungen mit 2,69.

Richtige Einschätzung der Übersetzungen nach Sprachkompetenz	Maschinelle Übersetzungen korrekt erkannt	Humanübersetzungen korrekt erkannt
Niveau A2	50 %	87,7 %
Niveau B1	53,53 %	75 %

Niveau B2	42,86 %	65 %
Niveau C1	57,78 %	58,33 %
Niveau C2	61 %	56,25 %

Tabelle 17: Erkennung der MÜ nach Sprachkompetenz

Obwohl bei der Bewertung der maschinellen und Humanübersetzungen keine eindeutigen Ergebnisse bezüglich der Sprachkompetenz ermittelt werden konnten, waren die Unterschiede bei der korrekten Einstufung von maschineller und Humanübersetzung sehr wohl ersichtlich, wie in Tabelle 17 dargestellt. Die Teilnehmer*innen mit Sprachniveau C2 erkannten am häufigsten, ob es sich beim jeweiligen zu bewertenden Text um eine maschinelle Übersetzung handelt oder nicht. Sie erkannten weiters zu 56,25 % korrekt, ob es sich um eine Humanübersetzung handelt. Die Teilnehmer*innen mit dem Sprachniveau A1 erkannten mit 87,7 % am häufigsten richtigerweise die Humanübersetzungen, gefolgt von B1 und B2, was bemerkenswert ist. Die A1-Gruppe bestand jedoch nur aus zwei Personen und gilt daher als nicht repräsentativ für Personen für dieses Sprachniveau.

7.5 Anmerkungen zur Qualität maschineller Übersetzungen

Am Ende der Bewertungsfragen wurden die Teilnehmer*innen noch gebeten, Anmerkungen zur Qualität und zum Einsatz der maschinellen Übersetzungen zu machen. Zwölf Personen fügten einen Kommentar ein, wie beispielsweise:

Die Qualität von maschineller Übersetzung ist in den letzten Jahren unglaublich gestiegen. Besonders stilistisch wissen maschinelle Übersetzungen oft überraschenderweise zu überzeugen.

In den Kommentaren wurde von einigen auch erklärt, wie sie bei den Bewertungen vorgegangen waren. So gab eine Person an, sie habe zunächst analysiert, ob zu wörtlich übersetzt worden war, und anschließend die Kollokationen und Redewendungen kontrolliert.

Weiters wurde angemerkt, dass die maschinelle Übersetzung nur für bestimmte Textsorten passe und generell für eine gute Übersetzung bereits der Ausgangstext „gut“ sein müsse. Ein*e professionelle*r Translator*in gab zudem auch an, dass DeepL erstaunlich gut sei:

MT wie DeepL erstaunlich gut, meist korrekte Ü die wenig Nachbearbeitung erfordern – nicht bei allen Textsorten! Kommerziell im normalen Leben eigentlich durchaus ausreichend.

Wie im Theorieteil der Masterarbeit ausgeführt, kommentierte eine Person hinsichtlich der Skopostheorie und schlug vor, Qualität anhand dieser zu bewerten. Sie verglich darüber hinaus die maschinellen Rohübersetzungen mit publizierten Humanübersetzungen und schlug vor, dass letztere, wenn sie maschinell übersetzt seien, von einem Menschen post-editiert werden sollten. Eine weitere Teilnehmerin gab an, dass schon der Ausgangstext „gut“ sein und die Übersetzung an das Zielpublikum angepasst werden muss, um eine „gute“ Übersetzung zu erhalten. Daher sollte im Idealfall ein*e professionelle*r Humanübersetzer*in dafür beauftragt werden, der/die angemessen bezahlt wird, um eine qualitativ hochwertige Übersetzung zu erhalten. Eine weitere Probandin gab an, dass DeepL für standardisierte Fachsprache brauchbar sei, bei literarischen Texten allerdings noch scheitere.

8. Diskussion

Lusicky und Heinisch (2019) führten eine Studie durch, bei der sie die Parameter der MQM-Metrik verwendeten. Wie bereits erwähnt, beschränkten sich die Studienteilnehmer*innen dabei auf Studierende des ZTW – keine Lai*innen wurden befragt. Basierend auf den MQM-Parametern *Accuracy*, *Fluency* und *Style* ergab die Studie von Lusicky und Heinisch (2019), dass die häufigsten Fehlerkategorien der maschinellen Übersetzungen in den Bereichen *Accuracy* und *Fluency* zu finden waren.

Auch Fiederer und O'Brien (2009) führten eine ähnliche Studie zum Vergleich von maschineller Übersetzung und Humanübersetzung durch, in der sich die Proband*innen auf Masterstudierende beschränkten. Sie verwendeten nach Hutchins (2004) die Qualitätsparameter *Accuracy*, *Fluency* und *Style*. Dabei ging hervor, dass die maschinelle Übersetzung beim Punkt *Accuracy* bessere Ergebnisse lieferte als die Humanübersetzung. Hingegen schnitt die Humanübersetzung bei den Punkten *Fluency* und *Style* besser ab. In der Umfrage der vorliegenden Masterarbeit wurden ebenso die Qualitätsparameter *Accuracy*, *Fluency* und *Style* nach Hutchins (2004) für die Bewertung der Translate verwendet. Die Zielgruppe der vorliegenden Masterarbeit beschränkte sich jedoch nicht nur auf professionelle Translator*innen bzw. Student*innen, sondern auch auf Lai*innen.

Neben der Bewertung der Texte und deren Einstufung in Humanübersetzung oder maschinelle Übersetzung wurden diese auch nach den Kriterien *Accuracy*, *Fluency* und *Style* bewertet. Die Umfrage ergab in diesen Bereichen, dass die Lai*innen die maschinelle Übersetzungen im Bereich *Accuracy* besser bewerteten als die Humanübersetzungen. Ein ähnliches Ergebnis zeigte sich, wie bereits angeführt, bei Fiederer und O'Brien (2009), denn auch in dieser Studie lieferte die maschinelle Übersetzung in dieser Kategorie die besten Ergebnisse. In den Bereichen *Fluency* und *Style* bewerteten die Lai*innen die maschinellen Übersetzungen etwas besser als die Humanübersetzungen. Hingegen bewerteten die professionellen Translator*innen die Humanübersetzungen in allen Bereichen besser als die maschinellen Übersetzungen. Dies ergibt ein anderes Ergebnis als bei Fiederer und O'Brien (2009), wo die maschinellen Übersetzungen von allen Proband*innen im Bereich *Accuracy* am besten bewertet wurden. Auch unterscheidet sich diese Bewertung von der Studie von Lusicky und Heinisch (2019), in der der Bereich *Accuracy* neben *Fluency* die meisten Fehler enthielt.

Weiters wurden die Proband*innen auch gefragt, wie wichtig ihnen die Parameter *Inhalt*, *Korrektheit* und *Stil* generell für Übersetzungen sind. So zeichnete sich das Bild ab, dass professionellen Translator*innen Korrektheit am wichtigsten ist und Lai*innen Korrektheit und Inhalt am wichtigsten sind. Beide Gruppen gaben außerdem an, dass der Stil einer Übersetzung von Bedeutung ist. Bei der Übersetzung soll also besonders auf eine korrekte Übertragung des Inhalts und auf Genauigkeit geachtet werden. Weiters erkannten die professionellen Translator*innen häufiger, um welche Art der Übersetzung es sich handelt.

Ziel der Masterarbeit war es, herauszufinden, ob Lai*innen die Qualität der maschinellen Übersetzungen in den Bereichen *Accuracy*, *Fluency* und *Style* höher einstufen als professionelle Translator*innen und welcher Unterschied sich in diesen Bereichen zwischen Humanübersetzungen und maschinellen Übersetzungen feststellen lassen. Das Ergebnis der Fragen zeigte, wie bereits angenommen, dass Lai*innen die Qualität der maschinellen Übersetzungen besser einstufen als professionelle Translator*innen. Mittels eines Online-Fragebogens im Tool LimeSurvey wurde den Teilnehmer*innen gleichzeitig der englische Ausgangstext und der deutsche Zieltext vorgelegt. Die Teilnehmer*innen wurden gebeten, fünf maschinelle Übersetzungen und vier Humanübersetzungen auf Ihre Qualität hin zu bewerten.

Insgesamt nahmen 44 Personen an der Umfrage teil, wobei der Anteil an professionellen Translator*innen und Lai*innen sehr ausgewogen war und die beiden Gruppen daher gut miteinander verglichen werden konnten. Obwohl Deutsch als Erstsprache keine Voraussetzung für die Umfrage war, gab dennoch der Großteil der Befragten an, Deutsch als Erstsprache zu haben. Dies ergibt sich daraus, dass die Umfrage im deutschsprachigen Raum (online) veröffentlicht wurde.

Für die Auswertung wurde zunächst eine Gesamtbewertung aller Teilnehmer*innen vorgenommen, um darstellen zu können, wie gut die maschinellen Übersetzungen und Humanübersetzungen von den Proband*innen bewertet werden. Diese ergab, dass die Humanübersetzungen etwas besser bewertet worden sind als die maschinellen Übersetzungen. Die Werte sind jedoch nicht sehr aussagekräftig, da die Ergebnisse sehr knapp beieinander liegen. Insgesamt schätzte die Gruppe der Lai*innen weniger oft richtig ein, um welche Art der Übersetzung es sich handelt.

Was die Textsorten betrifft, so ergab sich aus der Umfrage, dass der am besten bewertete Text eine literarische Humanübersetzung war und der am schlechtesten bewertete Text eine maschinelle Übersetzung. Auch die MÜ war eine Übersetzung eines literarischen Textes. Aus

der besten Bewertung lässt sich ableiten, dass Literaturübersetzungen Kreativität erfordern und dabei über Satzgrenzen hinweg übersetzt werden sollte. Bei der maschinell übersetzten Literaturübersetzung wurde der Text von DeepL hingegen sehr wörtlich übersetzt, woraus man schließen kann, dass die Maschine für diese Art von Texten nach heutigem Stand noch nicht geeignet ist. Bei jener maschinellen Übersetzung, die am häufigsten als Humanübersetzung eingestuft wurde, handelte es sich um einen verständlichen Fachtext, der sich auch an Lai*innen richtet. Da Fachtexte per se keine kreativen Sprachelemente, Metaphern oder andere rhetorische Stilmittel enthalten, mit denen Übersetzungssysteme Schwierigkeiten hätten, eignete sich DeepL für diese Art von Text gut – weshalb das System bei der Bewertung so gut abschnitt.

Die Annahme, dass professionelle Translator*innen maschinelle Übersetzungen häufiger erkennen, wurde bestätigt. Das bedeutet, es ist wahrscheinlich, dass professionelle Translator*innen aufgrund ihrer Ausbildung, Sprachkompetenz und Übersetzungserfahrung (mit und ohne Übersetzungstools) häufiger erkennen, um welche Art von Übersetzung es sich handelt. Weiter gibt dies die Bestätigung dafür, dass Translator*innen wissen, worauf es bei einer Humanübersetzung ankommt, um ein qualitativ hochwertiges Ergebnis zu erzielen; nämlich darauf, den Inhalt nicht Wort für Wort zu übersetzen, sondern ihn an das Zielpublikum und die Zielsprache anzupassen. Ebenso ist dabei die idiomatische Übersetzung über Satzgrenzen hinweg wichtig. Aus diesen Gründen werden maschinelle Übersetzungen von professionellen Translator*innen häufiger als solche erkannt.

Allerdings konnte jene Annahme, dass es den professionellen Translator*innen leichter fallen würde, zu erkennen, ob es sich um eine maschinelle Übersetzung handelt, nicht bestätigt werden. Bei der Einstufung des Schwierigkeitsgrades gaben in beiden Gruppen nämlich mehr als die Hälfte der Teilnehmer*innen an, dass das Erkennen der maschinellen Übersetzung mittelschwer für sie gewesen sei. Dabei ist anzunehmen, dass professionelle Übersetzer*innen von Grund auf kritischer bewerten, auch etwaige Humanübersetzungen – und die Einstufung in maschinell oder nicht ihnen deshalb schwerer fiel.

Auch in den Altersgruppen konnten Unterschiede festgestellt werden. Hier war ersichtlich, dass die Gruppe der 19 bis 25-Jährigen die Texte am strengsten bewertete, was sowohl auf die Lai*innen als auch auf die professionellen Translator*innen zutrifft.

Die anfängliche Erwartung, dass die Teilnehmer*innen mit besseren Englischkenntnissen eher erkennen würden, um welche Art der Übersetzung es sich handelt, konnte nicht bestätigt werden. Die Teilnehmer*innen mit Sprachkenntnissen C2 erkannten die maschinellen

Übersetzungen am häufigsten, allerdings erkannten die Teilnehmer*innen mit Sprachniveau B1 die Humanübersetzungen am häufigsten korrekt. Hierbei lässt sich vermuten, dass Personen mit C2-Niveau vor allem (von der Maschine) nicht idiomatische Kollokationen und Wendungen häufiger erkennen als Personen mit weniger guten Sprachkenntnissen, die zwar den Inhalt verstehen, aber weniger mit der Idiomatik vertraut sind. Weiters kann auch angenommen werden, dass generell Personen mit sehr guten Sprachkenntnissen (nicht nur Translator*innen) auch Humanübersetzungen strenger auf Fehler hin bewerten; vor allem, wenn diese bei einer Bewertungsaufgabe gesucht werden sollen. So kann es dazu kommen, dass auch Humanübersetzungen als maschinell übersetzte Texte eingestuft werden.

Die Teilnehmer*innen hatten abschließend die Möglichkeit, Feedback zur Umfrage zu geben. Dabei hat sich unter anderem herausgestellt, dass für eine aussagekräftigere Bewertung mehr Kontext nötig wäre. Für den Zweck der Studie wurde auf die Angabe des Kontexts und den Zweck der Übersetzungen jedoch absichtlich verzichtet, da die Proband*innen beispielsweise bei der Angabe „für *Gisting*-Zwecke verwendet“ dann eher erkannt hätten, woher die Texte stammen bzw. ob es sich um eine maschinelle Übersetzung handelt. Um diese Vorab-Beeinflussung zu verhindern, sollte sich die Bewertung rein auf sprachlich-idiomatischer Ebene der Texte abspielen.

Für die Ausweitung der Umfrage hätte die Frage nach der Revision der Humanübersetzungen anhand von professionellen Translator*innen miteinbezogen werden können. Da die humanübersetzten Texte zum Zeitpunkt der Erstellung des Fragebogens frei zur Verfügung standen, wurde jedoch nicht geprüft, ob es sich bei ihnen bereits um redigierte Version handelte. Speziell bei der Literaturübersetzung kann mit großer Wahrscheinlichkeit davon ausgegangen werden, dass diese, wie üblich, schon lektoriert wurde.

Bei den maschinell übersetzten Texten handelte sich um Rohfassungen, die mit DeepL erstellt wurden. Durch die Umfrage hat sich herauskristallisiert, dass sich die Qualität von nicht-posteditierten maschinellen Übersetzungen für die Übertragung bestimmter Textsorten geeignet ist, nämlich dann, wenn eine wörtliche Formulierung von einfachen, kurzen Sätzen im Ausgangstext gewährleistet ist und keine Fachterminologie enthalten ist. Denn diese würde Recherche und Überprüfung professioneller Translator*innen benötigen. Weiterhin ist auch auf den Zweck der Texte zu achten. Maschinell erstellte Texte werden von der Zielgruppe der Lai*innen bereits tagtäglich zumeist online genutzt, was bedeutet, dass der Zweck dieser Übersetzungen ein anderer ist als jener einer anspruchsvollen literarischen Übersetzung, die in

zigfacher Ausführung gedruckt und publiziert wird. Da Lai*innen darüber hinaus meist auch die Zielgruppe humanübersetzter Texte darstellen, war es der Autorin und den Forschungsfortschritt wichtig, sie in die Umfrage miteinzubeziehen.

9. Fazit

Die vorliegende Masterarbeit widmete sich der Wahrnehmung von Qualität maschineller Übersetzungen und Humanübersetzungen durch Lai*innen und professionelle Translator*innen. Um Qualität zu definieren, wurden der Begriff aus Sicht der Translationswissenschaft sowie der maschinellen Übersetzung im Grundlagenkapitel erläutert und auch der unterschiedliche Einsatz von maschineller Übersetzung und Humanübersetzung beschrieben. Für die vorliegende Masterarbeit wurde damit zusammenhängend eine Studie zur Qualitätserfassung von maschineller Übersetzung und Humanübersetzung durchgeführt, an der neben professionellen Translator*innen auch Lai*innen teilgenommen haben. Ziel der Arbeit war es also, die folgende Forschungsfrage zu beantworten: Stufen professionelle Translator*innen die Qualität von maschinellen Übersetzungen im Vergleich zu Humanübersetzungen anders ein als Lai*innen?

Die Online-Umfrage zur Beantwortung der Forschungsfrage ergab, dass Lai*innen die maschinellen Übersetzungen in den Bereichen Accuracy, Fluency und Style besser als die Humanübersetzungen einstufen und die professionellen Translator*innen bei den Humanübersetzungen bessere Noten vergaben. Zwar konnte bei der Gesamtbewertung der Texte aller Proband*innen kein wesentlicher Unterschied zwischen Humanübersetzungen und maschinellen Übersetzungen gezeigt werden; und der Unterschied zwischen richtig erkannten maschinellen Übersetzungen und Humanübersetzungen war nicht aussagekräftig.

Allerdings konnte ebenso erforscht werden, dass professionelle Translator*innen häufiger als Lai*innen erkennen, um welche Art der Übersetzung es sich handelt, womit die Forschungsfrage bestätigt werden konnte.

Zunächst wurde zudem angenommen, dass Teilnehmer*innen mit besseren Sprachkenntnissen häufiger erkennen würden, ob es sich um eine maschinelle Übersetzung handelt oder nicht, was nicht bestätigt werden konnte. Daraus lässt sich schließen, dass Personen mit höheren Sprachkenntnissen alle Arten der Translate strenger bewerten und dadurch Humanübersetzungen auch häufiger als maschinelle Übersetzung einstufen.

Da die Translate meist für Lai*innen angefertigt werden, ist es sinnvoll, diese Zielgruppe auch in Zukunft bei der Qualitätsbewertung von (maschinell erstellten) Translaten einzubinden. Dies ist wichtig, um feststellen zu können, wie gut sie Texte bewerten, wie häufig sie Translate als maschinell übersetzt und als Humanübersetzung einstufen und jedenfalls auch in Erfahrung zu bringen, was sie sich von (maschinell) übersetzten Texten erwarten. Die Studie hat gezeigt, dass auch Lai*innen die Qualität von sowohl maschinell übersetzten Texten als auch Humanübersetzungen bewerten und hilfreiche Kommentare und Anmerkungen zur Qualität von Übersetzung liefern können.

In der vorliegenden Umfrage ließ sich zudem feststellen, dass bei der Qualitätsbewertung von Übersetzungen der Unterschied von einer Textsorte zur nächsten im Zusammenhang mit der maschinellen Übersetzung beachtet werden sollte. So erhielt ein maschinell übersetzter literarischer Text die schlechtesten Noten und ein maschinell übersetzter wissenschaftlicher Fachtext die Bestnoten. Da dies anhand von zwei Textpassagen wissenschaftlich weniger aussagekräftig ist, müssen in diesem Bereich noch weiterführende, umfangreichere Studien durchgeführt werden – vor allem, wenn in Zukunft auch literarischer Texte und ähnliche Textsorten erfolgreich maschinell übersetzt werden sollen. Da die Qualität der neuronalen maschinellen Übersetzungssysteme in den letzten Jahren immer besser wurde, bleibt offen, inwieweit sich die Qualität von maschinellen Übersetzungen in den nächsten Jahren verbessert. Eine Wiederholung der Umfrage in erweitertem Ausmaß wäre demnach sinnvoll, um den Forschungsstand zur Qualität der Übersetzungen in ein paar Jahren erörtern zu können.

10. Literaturverzeichnis

Ahrend, Klaus (2006). Kriterien für die Bewertung von Fachübersetzungen. In: Schippel, Larisa (Hg.) *Übersetzungsqualität: Kritik – Kriterien – Bewertungshandeln*. Berlin: Frank & Timme, 31-42.

Allen, Jeff (2003). Post-editing. In: Somers, Harold (Hg.) *Computers and translation: a translator's guide*. Amsterdam: John Benjamins, 297-317.

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.

Big Data Insider (2020). Was ist ein Webcrawler? <https://www.bigdata-insider.de/was-ist-ein-webcrawler-a-704217> (Stand: 15.11.2022).

Bowker, Lynne (2020). Fit-for-purpose translation. In: O'Hagan, Minako (Hg.) *The Routledge Handbook of Translation and Technology*, Routledge: London, 453-468.

Castilho, Sheila. Doherty, Stephen. Gaspori, Federico. Moorkens, Joss (2018). Approaches to Human and Machine Translation Quality Assessment. In: Castilho, Sheila. Doherty, Stephens. Gaspori, Federico. Moorkens, Joss (Hg.) *Translation Quality Assessment*. Springer: Cham, 9-38.

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

DeepL (2020). Über Uns. <https://www.deepl.com/jobs.html> (Stand: 15.11.2022).

Delizée, Anne (2011). A global rating scale for the summative assessment of pragmatic translation at Master's level: an attempt to combine academic and professional criteria. In: Depraetere, Ilse (Hg.) *Perspectives on Translation Quality*. Berlin: De Gruyter Mouton, 9-23.

Dolle-Weinkauff, Bernd (2012). Dagobert Duck. Vom Griesgram zum Fantastilliardär. *Forschung Frankfurt*. (2), 124-127.

Dorothy, Kenny (2020). Machine Translation. In: O'Hagan, Minako (Hg.) *The Routledge Handbook of Translation and Technology*, Routledge: London, 453-468.

Fiederer, Rebecca. O'Brien, Sharon (2009). Quality and Machine Translation: A realistic objective? *The Journal of Specialised Translation* (11), 52-74.

Hutchins, John. Somers, Harold L. (1992) *An introduction to machine translation*. London: Academic Press limited, 161-174.

Göpferich, Susanne (2008). Textverstehen und Textverständnis. In: Janich, Nina. (Hg.) *Textlinguistik: 15 Einführungen*. Tübingen: Gunter Narr Verlag, 291-314.

Göpferich, Susanne (2006). *Textproduktion im Zeitalter der Globalisierung*. Tübingen: Stauffenburg-Verlag, 155.

Hansen, Gyde (2008). A Classification of Errors in Translation. In: Forstner, Martin. Lee-Jahnke, Hannelore. Schmitt, Peter A. (Hg.) *CIUTI-Forum 2019. Enhancing Translation Quality: Ways, Means, Methods*. Berlin: Lang, 313-327.

Hutchins, John (2004). Current commercial machine translation systems and computer-based translation tools: system types and their uses. *International journal of translation* 17(1-2), 5-38.

IBM (2020). IBM Websphere Application Server. <https://www.ibm.com/at-de/marketplace/java-ee-runtime> (Stand: 15.11.2022).

Jia, Yanfang. Carl, Michael. Wang, Xiangling (2019). How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. *The Journal of Specialised Translation* 31. 60-86.

Jurafsky, Daniel. Martin, James (2021). *Speech and Language Processing*. New Jersey: Prentice Hall, 1-27.

Kirchhoff, Katrin. Capurro, Daniel. Turner, Anne (2012). Evaluating user preferences in machine translation using conjoint analysis. *The Prague Bulletin of Mathematical Linguistics*. 108(1): 121-132.

Koehn, Philipp. Arun, Abhishek. Hoang, Hieu (2008). Towards better Machine Translation Quality for the German-English Language Pairs. In: Callison-Burch, Chris. Koehn, Philipp. Monz, Christof. Schroeder, Josh. Fordyce, Cameron Shaw (Hg.) *Proceedings of the Third Workshop on Statistical Machine Translation*, 139-142.

Koller, Werner (2011). *Einführung in die Übersetzungswissenschaft*. Stuttgart: Francke Attempto Verlag, 218-277.

Koponen, Maarit (2010). Assessing machine translation quality with error analysis. *Electronic proceeding of the KäTu symposium on translation and interpreting studies* 4, 1-12.

Lamapoll (2020). LamaPoll. <https://www.lamapoll.de> (Stand: 15.11.2022).

Lavie, Alon. Denkowski, Michael (2009). The METEOR metric for automatic evaluation of machine translation. *Machine Translation* (23), 105-115.

Lommel, Arle. Uszkoreit, Hans. Burchardt, Aljoscha (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció* (12), 455-463.

Lommel, Arle (2018). Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies In: Moorkens, Joss. Castilho, Sheila. Gaspari, Federico. Doherty, Stephen (Hg.) *Translation Quality Assessment*, 109-127.

Lusicky, Vesna, Heinisch, Barbara (2019). User expectations towards machine translation: A case study. *Proceeding of MT Summit XVII* (2), 42-48.

Matthiesen, Aron (2017). *Maschinelle Übersetzung im Wandel. Die Auswirkungen von künstlicher Intelligenz auf maschinelle Übersetzungssysteme*. Epubli.

Melby, Alan K. (2010). Future of machine Translation. Musings on Weaver's memo. In: O'Hagan, Minako (Hg.) *The Routledge Handbook of Translation and Technology*, 419-436.

Microsoft Bing Blogs (2011). Microsoft's New Search at Bing.com Helps People Make Better Decisions. <https://news.microsoft.com/2009/05/28/microsoft-new-search-at-bing-com-helps-people-make-better-decisions> (Stand: 15.06.2022).

Mossop, Brian (2020). *Revising and Editing for Translators*. Web: Taylor and Francis.

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

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

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

Qualitätsmanagement ISO 9001 (2020). Was ist Qualitätsmanagement nach ISO 9001? https://www.qualitaetsmanagement.me/qualitaetsmanagement_iso_9001/ (Stand 11.06.2020).

Qualitätsstandard ISO 17100 (2020). ISO 17100 – ANFORDERUNGEN AN ÜBERSETZUNGSDIENSTLEISTUNGEN. <https://qualitätsstandard.iso17100.com> (Stand: 15.11.2022).

Reiß, Katharina. Vermeer, Hans J. (1984). *Grundlegung einer allgemeinen Translationstheorie. Linguistische Arbeiten: Vol. 147*. Tübingen: Max Niemeyer.

Ruiz, Marta (2018). From Feature to Paradigm: Deep Learning. *Machine Translation. Journal of Artificial Intelligence Research* (61), 947-974.

Schmalz, Antonia (2019). Maschinelle Übersetzung. In: Wittenpahl, Volker (Hg.) *Künstliche Intelligenz*. Berlin: Springer, 194-208.

Schulz von Thun, Friedemann (1981). *Miteinander reden: Störungen und Klärungen*. Hamburg: Rowohlt.

Sen, Diya (2016). Google Translate Celebrates 10th Birthday: Here are some Interesting Facts About The Digital Linguist.

<https://www.techtimes.com/articles/155051/20160501/google-translate-celebrates-10th-birthday-here-are-some-interesting-facts-about-the-digital-linguist.htm> (Stand: 15.11.2022)

Soscisurvey (2020). Sosci. Der onlineFragebogen. <https://www.soscisurvey.de> (Stand: 15.11.2022).

TAUS (2020). The Language Data Network. <https://www.taus.net/> (Stand: 15.11.2022).

UNIVERSITAS (2022). Mit Sprachprofis zum Erfolg. <https://www.universitas.org/> (Stand: 15.11.2022).

Way, Andy (2013) Traditional and Emerging Use-Cases for MT. *Proceedings of Translating and the Computer*, 1-22.

Way, Andy (2018). Quality expectations of machine translation. In: Moorkens, Joss. Castilho, Sheila. Gaspari, Federico. Doherty, Stephen (Hg.) *Translation quality assessment*. Cham: Springer, 159-178.

White, John S (2003). How to evaluate machine translation. In: Somers, Harold (Hg.) *Computer and Translations: A translator's guide*. Amsterdam/Philadelphia: John Benjamins, 211-244.

SYSTRAN (2022). SYSTRAN Translate. <https://translate.SYSTRAN.net/> (Stand: 15.11.2022).

11. Anhang

Textauswahl

Text 1: Humanübersetzung

<p>The newspaper hired private detective Glenn Mulcaire to hack and monitor Pawlby's voicemails, but Coulson also called the court and left her loud voicemails. According to research, the cabin experience also depends on more than just the seat width.</p> <p>'The city definitely forms part of the comedy DNA of all of Larry's work', Steinberg said.</p>	<p>Die Zeitung beauftragte den Privatdetektiv Glenn Mulcaire, Pawlbys Voicemails zu hacken und sie zu überwachen, aber Coulson rief laut dem Gericht auch an und hinterließ ihr Voicemails. Untersuchungen zufolge hängt die Kabinenerfahrung auch von mehr als der Sitzbreite ab. „Die Stadt ist definitiv in der Comedy-DNA aller Arbeiten von Larry enthalten“, sagte Steinberg.</p>
---	---

Quelle Ausgangstext: Facebook Research (2014).

<https://github.com/facebookresearch/evaluation-of-nmt-bt/blob/main/data/newstest2014-ende.r1.r1.en> (Stand 15.11.22).

Quelle Zieltext: Sergey Edunov, Ott, Myle, Ranzota, Marc Aurelio, Auli, Michael. (2020) On the evaluation of machine translation systems trained with back-translations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2836-2846.

Text 2: Maschinelle Übersetzung (DeepL)

<p>In vitro fertilization (IVF), also called test-tube conception, medical procedure in which mature egg cells are removed from a woman, fertilized with male sperm outside the body, and inserted into the uterus of the same or another woman for normal gestation. Although IVF with reimplantation of fertilized eggs (ova) has long been widely used in animal breeding, the first successful birth of a human child from IVF, carried out by British gynecologist Patrick Steptoe and British medical researcher Robert Edwards, did not take place until 1978.</p>	<p>In-vitro-Fertilisation (IVF), auch Reagenzglasbefruchtung genannt, medizinisches Verfahren, bei dem einer Frau reife Eizellen entnommen, mit männlichen Spermien außerhalb des Körpers befruchtet und in die Gebärmutter derselben oder einer anderen Frau zur normalen Schwangerschaft eingesetzt werden. Obwohl die IVF mit Reimplantation der befruchteten Eizellen (Eizellen) in der Tierzucht schon lange angewendet wird, fand die erste erfolgreiche Geburt eines menschlichen Kindes durch IVF, durchgeführt vom britischen Gynäkologen Patrick Steptoe und dem britischen Medizinforscher Robert Edwards, erst 1978 statt.</p>
---	--

Quelle Ausgangstext: Britannica (2022) in vitro fertilization.

<https://www.britannica.com/science/in-vitro-fertilization> (Stand 15.11.2022).

Text 3: Maschinelle Übersetzung (DeepL)

<p>This mismatch between our ancient physiology and the Western diet and lifestyle underlies many so-called diseases of civilization, including coronary heart disease, obesity, hypertension, type 2 diabetes, epithelial cell cancers, autoimmune disease, and osteoporosis, which are rare or virtually absent in hunter-gatherers and other non-westernized populations.</p>	<p>Diese Diskrepanz zwischen unserer alten Physiologie und der westlichen Ernährung und Lebensweise liegt vielen sogenannten Zivilisationskrankheiten zugrunde, darunter koronare Herzkrankheit, Fettleibigkeit, Bluthochdruck, Typ-2-Diabetes, Epithelzellkrebs, Autoimmunerkrankungen und Osteoporose, die bei Jägern und Sammlern und anderen nicht verwestlichten Populationen selten oder praktisch nicht vorhanden sind.</p>
--	--

Quelle Ausgangstext: Carrera-Bastos, Pedro, Fontes-Villalba, Maelan. O’Keefe, James H. Lindeberg, Staffan. Cordain, Loren (2011). The Western diet and lifestyle and diseases of civilization. *Research Reports in Clinical Cardiology* 2 (2), 2-15.

Text 4: Humanübersetzung

<p>Although my book is intended mainly for the entertainment of boys and girls, I hope it will not be shunned by men and women on that account; for part of my plan has been to try to pleasantly remind adults of what they once were themselves, and of how they felt and thought and talked, and what queer enterprises they sometimes engaged in.</p>	<p>Obwohl dies Buch vor allem zur Unterhaltung der kleinen Welt geschrieben wurde, so darf ich doch wohl hoffen, dass es auch von Erwachsenen nicht ganz unbeachtet gelassen werde, habe ich doch darin versucht, ihnen auf angenehme Weise zu zeigen, was sie einst selbst waren, wie sie fühlten, dachten, sprachen, und welcher Art ihr Ehrgeiz und ihre Unternehmungen waren.</p>
---	---

Quelle Ausgangstext: Twain, Mark (1876). The Adventures of Tom Sawyer. Hartford: American Publishing Company.

Quelle Zieltext: Twain, Mark/Hellwag, H (1900). Die Abenteuer des Tom Sawyers. Halle a. d. Saale: Otto Hendel.

Text 5: Humanübersetzung

<p>NEW YORK – Who would have thought that six years after the global financial crisis, most advanced economies would still be swimming in an alphabet soup – ZIRP, QE, CE, FG, NDR, and U-FX Int – of unconventional monetary policies? No central bank had considered any of these measures (zero interest rate policy, quantitative easing, credit easing, forward guidance, negative deposit rate, and unlimited foreign exchange intervention, respectively) before 2008.</p>	<p>NEW YORK – Wer hätte gedacht, dass die meisten entwickelten Ökonomien sechs Jahre nach der globalen Finanzkrise immer noch in einer Buchstabensuppe unkonventioneller geldpolitischer Strategien – ZIRP (Nullzinspolitik), QE (quantitative Lockerung), CE (Kreditlockerung), FG (Forward Guidance), NDR (negativer Einlagenzins) und U-FX Int (unbeschränkte Deviseninterventionen) – schwimmen würden? Vor 2008 hätte keine Zentralbank irgendeine dieser Maßnahmen in Erwägung gezogen.</p>
---	---

Quelle Ausgangstext: Roubine, Nouriel (2015) <https://www.stern.nyu.edu/experience-stern/faculty-research/roubini-unconventional-truth> (Stand: 15.11.2022)

Quelle Zieltext: Capital (2015). <https://www.capital.de/wirtschaft-politik/eine-unkonventionelle-wahrheit-3912> (Stand: 15.11.2022)

Text 6: Maschinelle Übersetzung (DeepL)

<p>The Houston Cougars are headed to the Final Four for the first time since 1984 following their 67-61 Monday night victory over the Oregon State Beavers. After leading 34-17 at halftime, Houston allowed Oregon State to claw all the way back to tie the game at 55-55, but the Cougars' prowess both on defense and on the offensive glass was just too much in the end. Houston also made too many threes against an Oregon State team that had excelled on the defensive perimeter for the past few weeks. Marcus Sasser led all scorers with 20 points and five made triples.</p>	<p>Die Houston Cougars stehen nach ihrem 67:61-Sieg am Montagabend gegen die Oregon State Beavers zum ersten Mal seit 1984 wieder im Final Four. Nach einer 34:17-Führung zur Halbzeit ließ Houston zu, dass Oregon State den ganzen Weg zurückkam und das Spiel mit 55:55 ausglich, aber die Fähigkeiten der Cougars in der Defensive und in der Offensive waren am Ende einfach zu viel. Houston traf auch zu viele Dreier gegen ein Team von Oregon State, das in den letzten Wochen vor allem in der Defensive brilliert hatte. Marcus Sasser führte alle Scorer mit 20 Punkten und fünf getroffenen Dreiern an.</p>
--	--

Quelle Ausgangstext: Miller, Kerry (2021). Bleacher Report.

<https://bleacherreport.com/articles/2938784-houston-will-finally-face-challenges-after-historically-easy-final-four-path> (Stand 15.11.2022)

Text 7: Humanübersetzung

<p>GENEVA – As 2015 begins, the values enshrined in the Charter of the United Nations and the Universal Declaration of Human Rights are under threat. Around the world, personal liberty, human rights, and democracy are at risk – even in countries that have embraced democratic ideals. The international community is deeply divided, blocking progress on a host of global challenges, ranging from the crises in Syria, Iraq, and Ukraine to climate change and international trade.</p>	<p>GENF – Das Jahresende 2014 ist durch eine Bedrohung der in der Charta der Vereinten Nationen und der Allgemeinen Erklärung der Menschenrechte verkörperten Werte gekennzeichnet. Weltweit sind die persönliche Freiheit, die Menschenrechte und die Demokratie in Gefahr – selbst in Ländern, die sich zu demokratischen Idealen bekennen. Die internationale Gemeinschaft ist zutiefst zerstritten, was Fortschritte bei einer Vielzahl globaler Herausforderungen blockiert – von den Krisen in Syrien, dem Irak und der Ukraine bis hin zum Klimawandel und zum Welthandel.</p>
---	---

Quelle Ausgangstext: Annan, Kofi (2015). World Economic Forum. Three trends destabilising the world in 2015. <https://www.weforum.org/agenda/2015/01/how-to-save-global-order#:~:text=As%202015%20begins%2C%20the%20values%20enshrined%20in%20the,eve n%20in%20countries%20that%20have%20embraced%20democratic%20ideals> (Stand: 15.11.2022)

Quelle Zieltext: Annan, Kofi/ Doolan, Jan (2015). Die Rettung der Weltordnung. <https://www.project-syndicate.org/onpoint/saving-global-order-and-international-system-by-kofi-a-annan-2015-01/german> (Stand: 15.11.2022).

Text 8: Maschinelle Übersetzung

<p>Marley was dead: to begin with.</p> <p>There is no doubt whatever about that.</p> <p>The register of his burial was signed by the clergyman, the clerk, the undertaker, and the chief mourner. Scrooge signed it. And Scrooge 's name was good upon 'Change, for anything he chose to put his hand to. Old Marley was as dead as a door-nail. Mind! I don't mean to say that I know, of my own knowledge, what there is particularly dead about a door-nail.</p> <p>I might have been inclined, myself, to regard a coffin-nail as the deadest piece of ironmongery in the trade.</p>	<p>Marley war tot: zu Beginn.</p> <p>Daran gibt es überhaupt keinen Zweifel.</p> <p>Das Register seiner Beerdigung wurde von dem Geistlichen, dem Schreiber, dem Bestatter und dem Haupttrauernden unterschrieben.</p> <p>Dagobert hat es unterschrieben. Und Scrooge's Name war gut auf ändern, für alles, was er wählte, um seine Hand zu setzen. Der alte Marley war so tot wie ein Türnagel.</p> <p>Achtung! Ich will nicht sagen, daß ich aus eigener Kenntnis weiß, was an einem Türnagel besonders tot ist. Ich selbst wäre geneigt gewesen, einen Sargnagel als das toteste Stück Eisenwaren zu betrachten, das es gibt.</p>
--	--

Quelle Ausgangstext: Dickens, Charles (1843). A Christmas Carol. London: Chapman & Hall.

Text 9: Maschinelle Übersetzung (DeepL)

Not because his troubles were one whit less heavy and bitter to him than a man 's are to a man but because a new and powerful interest bore them down and drove them out of his mind for the time – just as men 's misfortunes are forgotten in the excitement of new enterprises.	Nicht, weil seine Sorgen für ihn einen Deut weniger schwer und bitter waren als für einen Mann, sondern weil ein neues und mächtiges Interesse sie niederdrückte und sie für die Zeit aus seinem Gedächtnis vertrieb – so wie das Unglück eines Menschen in der Aufregung eines neuen Unternehmens vergessen wird.
--	--

Quelle Ausgangstext: Twain, Mark (1876). Tom Sawyer. Hartford: American Publishing Company.

Fragebogen

Demographische Daten:

Wie alt sind Sie?

- unter 18
- 19-25
- 26-35
- 36-55
- 56-70
- über 70 Jahre

Was ist Ihr Geschlecht?

- männlich
- weiblich
- Sonstiges:

Was ist Ihre höchste abgeschlossene Schule bzw. Ausbildung?

- Lehre
- Berufsbildende höhere Schule (Matura)
- Studium
- Doktorat

Sind Sie als professionelle*r Translator*in tätig?

- Ja
- Nein

In welcher Branche sind Sie tätig?

- Hotel- und Gastgewerbe
- Administration/Büro
- Handel
- Metall- und Elektroberufe

- Lehr- und Kulturbereich
- Gesundheitsberufe
- Technik
- Wissenschaft
- Sozialberufe

Ist Deutsch Ihre Erstsprache?

- Ja
- Nein

Wie hoch schätzen Sie Ihr Sprachniveau auf Deutsch ein?

- A1
- A2
- B1
- B2
- C1
- C2

Wie hoch schätzen Sie Ihr Sprachniveau auf Englisch ein?

- A1
- A2
- B1
- B2
- C1
- C2

Nutzen Sie (frei verfügbare) maschinelle Übersetzungssysteme? *

- Ja
- Nein

*Zu den frei verfügbaren maschinellen Übersetzungssystemen zählen unter anderem Google Translate, DeepL und SYSTRAN.

Textbewertung:

Im Folgenden werden zehn Texte vorgestellt. Bei fünf davon handelt es sich um Humanübersetzungen und fünf sind maschinelle Übersetzungen. Bitte bewerten Sie die Texte auf ihre Qualität.

Bitte beachten Sie: Maschinelle Übersetzung wird meist für andere Zwecke als die professionelle Übersetzung verwendet. So gibt es zum Beispiel maschinell übersetzte Texte, bei denen es nur darum geht, den Sinn des Ausgangstexts zu verstehen. Diese Texte können zum Beispiel viele Grammatikfehler enthalten, erfüllen aber dennoch ihren Zweck.

Text 1: Bitte bewerten Sie die folgende Übersetzung. 1 ist dabei die Bestnote und 5 die schlechteste Note.

	1	2	3	4	5
Der Inhalt ist korrekt übersetzt.					
Der Text ist flüssig übersetzt.					
Der Stil ist gut.					

Der Text ist maschinell übersetzt

- Ja
- nein

Haben Sie Anmerkungen zum Ausgangstext oder der Übersetzung?

Text 2: Bitte bewerten Sie die folgende Übersetzung. 1 ist dabei die Bestnote und 5 die schlechteste Note.

	1	2	3	4	5
Der Inhalt ist korrekt übersetzt.					
Der Text ist flüssig übersetzt.					
Der Stil ist gut.					

Der Text ist maschinell übersetzt

- Ja
- nein

Haben Sie Anmerkungen zum Ausgangstext oder der Übersetzung?

--

Text 3: Bitte bewerten Sie die folgende Übersetzung. 1 ist dabei die Bestnote und 5 die schlechteste Note.

	1	2	3	4	5
Der Inhalt ist korrekt übersetzt.					
Der Text ist flüssig übersetzt.					
Der Stil ist gut.					

Der Text ist maschinell übersetzt

- Ja
- nein

Haben Sie Anmerkungen zum Ausgangstext oder der Übersetzung?

--

Text 4: Bitte bewerten Sie die folgende Übersetzung. 1 ist dabei die Bestnote und 5 die schlechteste Note.

	1	2	3	4	5
Der Inhalt ist korrekt übersetzt.					

Der Text ist flüssig übersetzt.					
Der Stil ist gut.					

Der Text ist maschinell übersetzt

- Ja
- nein

Haben Sie Anmerkungen zum Ausgangstext oder der Übersetzung?

Text 5: Bitte bewerten Sie die folgende Übersetzung. 1 ist dabei die Bestnote und 5 die schlechteste Note.

	1	2	3	4	5
Der Inhalt ist korrekt übersetzt.					
Der Text ist flüssig übersetzt.					
Der Stil ist gut.					

Der Text ist maschinell übersetzt

- Ja
- Nein

Haben Sie Anmerkungen zum Ausgangstext oder der Übersetzung?

Text 6: Bitte bewerten Sie die folgende Übersetzung. 1 ist dabei die Bestnote und 5 die schlechteste Note.

	1	2	3	4	5

Der Inhalt ist korrekt übersetzt.					
Der Text ist flüssig übersetzt.					
Der Stil ist gut.					

Der Text ist maschinell übersetzt

- Ja
- Nein

Haben Sie Anmerkungen zum Ausgangstext oder der Übersetzung?

Text 7: Bitte bewerten Sie die folgende Übersetzung. 1 ist dabei die Bestnote und 5 die schlechteste Note.

	1	2	3	4	5
Der Inhalt ist korrekt übersetzt					
Der Text ist flüssig übersetzt					
Der Stil ist gut					

Der Text ist maschinell übersetzt

- Ja
- Nein

Haben Sie Anmerkungen zum Ausgangstext oder der Übersetzung?

Text 8: Bitte bewerten Sie die folgende Übersetzung. 1 ist dabei die Bestnote und 5 die schlechteste Note.

	1	2	3	4	5
Der Inhalt ist korrekt übersetzt.					
Der Text ist flüssig übersetzt.					
Der Stil ist gut.					

Der Text ist maschinell übersetzt

- Ja
- nein

Haben Sie Anmerkungen zum Ausgangstext oder der Übersetzung?

Text 9: Bitte bewerten Sie die folgende Übersetzung. 1 ist dabei die Bestnote und 5 die schlechteste Note.

	1	2	3	4	5
Der Inhalt ist korrekt übersetzt.					
Der Text ist flüssig übersetzt.					
Der Stil ist gut.					

Der Text ist maschinell übersetzt

- Ja
- Nein

Haben Sie Anmerkungen zum Ausgangstext oder der Übersetzung?

Wie leicht ist es Ihnen nach Ihrem Gefühl gegangen, menschliche Übersetzung von maschineller Übersetzung zu unterscheiden?

- sehr leicht
- leicht
- mittel
- schwer
- sehr schwer

Welcher der hier beurteilten Aspekte ist für Sie bei Übersetzungen generell am wichtigsten? Bitte bewerten Sie dies auf einer Schulnotenskala von 1-5.

	1	2	3	4	5
Inhalt					
Korrektheit					
Stil					

Haben Sie noch Anmerkungen zur Qualität und dem Einsatz der maschinellen Übersetzung?

Haben Sie weitere Anmerkungen zum Thema dieser Studie und maschineller Übersetzung generell?

Abstract (Deutsch)

Die vorliegende Masterarbeit beschäftigt sich mit der Evaluierung der Qualität von Humanübersetzungen im Vergleich zu maschinellen Übersetzungen durch Lai*innen und professionelle Translator*innen. Im theoretischen Teil wurden zunächst der Begriff Qualität dargelegt und verschiedene Arten zur Bewertung der Qualität von Humanübersetzungen und von maschinellen Übersetzungen präsentiert.

Ziel der Arbeit war es, herauszufinden, ob professionelle Translator*innen die Qualität von Humanübersetzungen anders einstufen als Lai*innen. Dafür wurde eine Online-Umfrage erstellt, die im Methodikteil thematisiert wurde. Translator*innen und Lai*innen bewerteten die Qualität von insgesamt neun Texten, wobei vier maschinell und fünf von Humanübersetzer*innen übersetzt wurden. Als Bewertungsbasis wurden die Parameter *Accuracy*, *Fluency* und *Style* nach Hutchins und Somers (1993) herangezogen. Die Studie ergab, dass professionelle Translator*innen häufiger erkennen, ob es sich um eine maschinelle Übersetzung handelt und sie Humanübersetzungen jedoch auch strenger bewerten als Lai*innen dies tun. Aus diesem Grund stuften sie diese daher teilweise als maschinelle Übersetzungen ein. Unterschiede konnten zudem in Bezug auf Sprachkompetenz und Altersgruppen festgestellt werden, wobei jüngere Teilnehmer*innen die maschinell übersetzten Texte häufiger als solche erkannten.

Darüber hinaus ergab die praktische Untersuchung, dass verschiedene Textsorten unterschiedlich bewertet wurden und sich bestimmte Textsorten (z. B. der Fachbericht) gemäß der Bewertung der Teilnehmer*innen bereits für maschinelle Übersetzungssysteme eignen – während vor allem literarische Texte besser von Humanübersetzer*innen übersetzt werden sollten.

Bisherige Studien umfassten meist professionelle Translator*innen als Zielgruppe, wobei die Einbeziehung von Lai*innen ebenso aufschlussreiche Ergebnisse und Input für die Forschung lieferte, da diese meist die Zielgruppe der übersetzten Texte darstellen.

Abstract (English)

This Master's thesis is about evaluating the quality of human translation and machine translation by examine translations done by professional translators and lay persons. The term *quality* for translations is described in the theoretical part, as well as different means to evaluate the quality of human and machine translation.

This study aimed to find out whether professional translators evaluate the quality of human and machine translations differently from lay persons. In an online survey, professional translators and lay persons had to evaluate the quality of nine texts in total; some of them were translated by DeepL, the other ones by human translators.

Besides answering the question whether these texts are machine translations or human translations, the participants were also asked to evaluate the translations on the basis of *accuracy, fluency* and *style* by Hutchins and Somers (1993). The study shows that professional translators properly identified more often whether it had been a machine translation or a human translation. However, professional translators are stricter when it comes to evaluating human translations, which is why they sometimes identified them as machine translations. Differences could also be found with regards to language competency and age group, whereby younger participants correctly identified machine translations more often than other age groups.

Since different text types were used for the survey, the study also gives a glimpse at the applicability and proper use of some text types in machine translation. In the study, a machine-translated literature translation showed bad results, while a machine-translated technical text was regarded as a human translation by the majority of the participants because of its accuracy, fluency and style. However, more research would be necessary since the text sample was too small and resources were limited.