universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Cui bono?"

Postedition von maschinellen Übersetzungen in der Literaturübersetzung

verfasst von / submitted by

Coni Cassius Zheng, BA

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

Master of Arts (MA)

Wien, 2021 / Vienna 2021

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. Waltraud Kolb

Inhaltsverzeichnis

Abkürzungsverzeichnis 1
1 Einleitung 2
2 Geschichte 4
2. 1 Ein holpriger Anfang 4
2. 2 Anfänglicher Optimismus bis zum ALPAC-Bericht 5
2. 3 Wiedergeburt nach dem ALPAC-Bericht 6
2. 4 Entwicklung in den 80er und 90er Jahren 7
2. 5 Aufstieg der neuronalen maschinellen Übersetzung 8
3 Was ist maschinelles Übersetzen? 9
3. 1 Definitionen der maschinellen Übersetzung 9
3. 2 Unterbegriffe 11
3. 2. 1 Machine-Aided Human Translation (MAHT) 11
3. 2. 2 Human-Aided Machine Translation (HAMT) 12
3. 2. 3 Fully Automatic High-Quality Translation (FAHQT) 12
4 Strategien und Ansätze der maschinellen Übersetzung 14
4. 1 Regelbasierte Übersetzungsstrategien 14
4. 1. 1 Direkte Systeme 14
4. 1. 2 Indirekte Systeme 15
4. 1. 2. 1 Interlinguabasierte Methode 15
4. 1. 2. 2 Transferbasierte Systeme 15
4. 2 Neuere Ansätze 16
4. 2. 1 Korpusbasierter Ansatz 16
4. 2. 2 Wissensbasierter Ansatz 17
4. 2. 3 Neuronale maschinelle Übersetzung 18
5 Probleme der maschinellen Übersetzung 19
5. 1 Morphologische Mehrdeutigkeit 20
5. 2 Lexikalische Mehrdeutigkeit 21
5. 3 Syntaktische Mehrdeutigkeit 23
6 Aktueller Forschungsstand der MÜ für den Bereich Literaturübersetzen 24
6. 1 Forschung ab 2010 24
6. 2 Forschung zur Lyrikübersetzung 25
6. 3 Weitere Studien 26
7 Translation Quality Assessment 29
7. 1 Begriffserklärung TQA – Translation Quality Assessment 29

7. 2 Evaluierungsansätze für die Bewertung von MÜ 30
7. 2. 1 Menschliche Evaluierung 30
7. 2. 1. 1 Angemessenheit und Flüssigkeit 31
7. 2. 1. 2 Lesbarkeit und Verständlichkeit 32
7. 2. 1. 3 Akzeptabilität 32
7. 2. 2 Automatisierte Evaluierung 33
7. 3 Fehlerklassifizierung 34
7. 3. 1 Fehlertypologien für die manuelle Klassifizierung 36
7. 3. 2 Herausforderungen der manuellen Fehlerklassifizierung 38
7. 3. 3 Automatische Fehlerklassifizierung 38
8 Postedition 40
8. 1 Was ist Postedition? 40
8. 2 Erforderliche Kompetenzen von PosteditorInnen 41
8. 3 Bisherige Forschung zu Postedition 42
9 Empirische Untersuchung 44
9. 1 Vorgehensweise und Vorstellung der ProbandInnen 44
9. 2 Der Versuchstext 46
9. 3 DeepL – Neuronale maschinelle Übersetzung online 46
10 Durchführung der Untersuchung 48
10. 1 Fragebögen der Gruppen A und B 48
10. 2 Fragebogen der Evaluatorin 63
10. 3 Die Evaluierungsbögen 64
10. 3. 1 Evaluierung von Text 1 64
10. 3. 2 Evaluierung von Text 2 65
10. 3. 3 Evaluierung von Text 3 66
10. 3. 4 Evaluierung von Text 4 67
10. 4 Diskussion der Daten 68
10. 4. 1 Der Zeitaufwand 68
10. 4. 2 Die Qualität der Zieltexte 70
10. 4. 3 Untersuchung der Messgrößen 71
10. 4. 4 Beziehungen und Ähnlichkeiten mit früheren Studien 79
10. 5 Fazit und Ausblick in die Zukunft 81
11 Conclusio 87
12 Literaturverzeichnis 89
13 Abbildungsverzeichnis 94
Anhang 97

Der Versuchstext (Originalfassung) 97
Übersetzung von DeepL 99
Fragebogen der ProbandInnen 103
Abstract 110

Abkürzungsverzeichnis

AEM: Automatic Evaluation Metric

ALPAC: Automatic Language Processing Advisory Committee

BLEU: Bilingual Evaluation Understudy

EBMT: Example Based Machine Translation

FAHQT: Fully Automatic High-Quality Translation

HAMT: Human-Aided Machine Translation

HCI: Human-Computer Interaction

KI: Künstliche Intelligenz

MAHT: Machine-Aided Human Translation

MÜ: Maschinelle Übersetzung

MT: Machine Translation

NMÜ: Neuronale maschinelle Übersetzung

NMT: Neural Machine Translation

RBMT: Rule Based Machine Translation

SMT: Statistical Machine Translation

SYSTRAN: SYStem TRANslation

1 Einleitung

Seit der Geburtsstunde der maschinellen Übersetzung ist eine Vielfalt an Theorien, Ansätzen und Systemen der maschinellen Übersetzung entstanden, diskutiert und verbessert worden, wobei neue Ideen oftmals Begeisterung hervorriefen und große Hoffnungen weckten. Eine dieser Hoffnungen, wenn nicht gar die größte Hoffnung, waren vollautomatische Übersetzungssysteme, die ohne jegliche menschliche Hilfe in der Lage sind, Übersetzungen zu erzeugen, die qualitativ genauso hochwertig sind wie von HumanübersetzerInnen. Wird die maschinelle Übersetzung menschliche ÜbersetzerInnen in die Obsoleszenz drängen? Sind Google Übersetzer und ähnliche Maschinen aufgrund ihrer Geschwindigkeit und sprachlichen Möglichkeiten nicht bereits jetzt schon menschlichen ÜbersetzerInnen überlegen? Fragen, die vor allem für Personen, welche in der Sprachindustrie tätig sind, von größtem Interesse sind. Nicht ohne Grund verfolgen ÜbersetzerInnen auf der ganzen Welt wachsam Entwicklungen im Bereich der maschinellen Übersetzung, denn maschinelle Übersetzungssysteme haben die Arbeitsweisen und Methoden von Sprachdienstleistern bereits grundlegend verändert.

Während nach wie vor zwar kein „perfektes" MÜ-System existiert, nimmt der Gebrauch von Übersetzungssystemen Jahr für Jahr unweigerlich zu, was darauf hindeutet, dass auch die Qualität der erzeugten Übersetzungen steigt (vgl. Way 2018: 173). In der Tat werden maschinelle Übersetzungssysteme stetig besser. Mit dem Aufkommen des relativ neuen neuronalen maschinellen Übersetzungsparadigmas verbesserte sich die Qualität von MÜ beträchtlich. In der Forschung zur maschinellen Übersetzung wird der Fokus meist auf Fachtexte gelegt. Somit gibt es bloß eine sehr begrenzte Anzahl von Arbeiten, welche sich mit der Anwendung von MÜ bei literarischen Texten beschäftigten. Dies lässt sich womöglich dadurch erklären, dass eine zentrale Herausforderung bei der literarischen Übersetzung darin besteht, nicht nur die Bedeutung (wie in anderen Bereichen, z. B. der Übersetzung von technischen Texten), sondern auch das Leseerlebnis zu bewahren, sodass literarische ÜbersetzerInnen sorgfältig aus allen möglichen Übersetzungsoptionen auswählen müssen. Aufgrund dieser (und anderen) Herausforderungen, gelten literarische Texte nach wie vor als die größte Herausforderung für die maschinelle Übersetzung (vgl. Toral und Way 2018: 264-265).

Mit der verbesserten Qualität von maschinellen Übersetzungssystemen in den letzten Jahrzehnten hat auch die Postedition fortwährend an Bedeutung gewonnen. Angesichts der zunehmenden Qualität von maschinellen Übersetzungssystemen ist die Postedition eine übliche Tätigkeit in computergestützten Übersetzungsworkflows geworden. Eine Untersuchung beider Themen stellt den empirischen Teil der vorliegenden Arbeit dar. In der im Rahmen dieser Arbeit durchgeführten Studie werden Übersetzungen mit Posteditionen einer englischen Kurzgeschichte verglichen. Sowohl die Übersetzungen als auch die Posteditionen wurden von Studierenden des Masterstudiums der Translation mit dem Schwerpunkt „Übersetzen in

Literatur - Medien - Kunst“, ein Studiengang an der Universität Wien, verfasst. Die Posteditionen wurden an einer Rohübersetzung durchgeführt, welche zuvor vom Online-Übersetzungstool DeepL produziert wurde. Bewertet wurden die Zieltexte von einer professionellen Übersetzerin, welcher, um eine möglichst unparteiische Evaluierung zu erlauben, im Vorfeld nicht mitgeteilt wurde, bei welchen Texten es sich um Posteditionen handelte und bei welchen um Übersetzungen.

Der Aufbau der vorliegenden Arbeit ist wie folgt: Zuallererst wird die Geschichte der maschinellen Übersetzung dargelegt. Anschließend werden verschiedene Strategien und Ansätze von MÜ-Systemen erläutert. Probleme, mit denen die maschinelle Übersetzung zu kämpfen hat, werden geschildert. Das darauffolgende Kapitel widmet sich dem Translation Quality Assessment – der Begriff wird erklärt, verschiedene Evaluierungsansätze für die Bewertung von MÜ werden beschrieben und auch auf Fehlerklassifizierungen wird eingegangen. Das letzte theoretische Kapitel widmet sich der Postedition. Notwendige Kompetenzen von PosteditorInnen werden erwähnt und die bisherige Forschung zu Postedition wird veranschaulicht. Alle Inhalte ab Kapitel 9 beschäftigen sich mit der empirischen Untersuchung. Die Vorgehensweise, die ProbandInnen, der Versuchstext der Studie und das Programm DeepL werden vorgestellt. Nachfolgend wird die Durchführung der Untersuchung beschrieben. Die Arbeit schließt mit der Analyse und Diskussion der im Zuge der Studie gesammelten Daten.

Die konkreten Fragestellungen der vorliegenden Arbeit lauten somit:

Ist es mittels Postedition einer vom Online-Tool DeepL erzeugten Übersetzung möglich, gleichwertige oder sogar bessere deutsche Übersetzungen einer englischen Kurzgeschichte zu erzeugen im Vergleich zu Übersetzungen von HumanübersetzerInnen?

In welchen Aspekten schneiden HumanübersetzerInnen besser und in welchen schlechter ab?

Es wird angenommen, dass MÜ-Systeme Übersetzungen von literarischen Texten von geringerer Qualität produzieren als HumanübersetzerInnen. Ferner wird davon ausgegangen, dass selbst nach dem Posteditions-Prozess eines solchen maschinell generierten Textes zwar keine so hohe Qualität erlangt wird, wie sie von menschlichen ÜbersetzerInnen erreicht wird, dafür aber ein signifikanter Zeitgewinn erzielt wird.

2 Geschichte

In diesem Kapitel wird ein konziser Überblick über die abwechslungsreiche Geschichte des maschinellen Übersetzens gegeben. Das Kapitel beginnt zwar mit den Entwicklungen im 17. Jahrhundert, der Fokus liegt jedoch vor allem auf den zahlreichen Entwicklungen im 20. Jahrhundert. Vorgänger von maschinellen Übersetzungssystemen werden vorgestellt, die plötzlich erreichte Bekanntheit durch das Weaver-Memorandum wird erläutert, die Folgen des ALPAC-Berichts und der Weg zur heutigen Kommerzialisierung von MÜ-Systemen werden geschildert. Folgende Grafik stellt eine Auswahl der wichtigsten Meilensteine in der Geschichte der maschinellen Übersetzung dar:

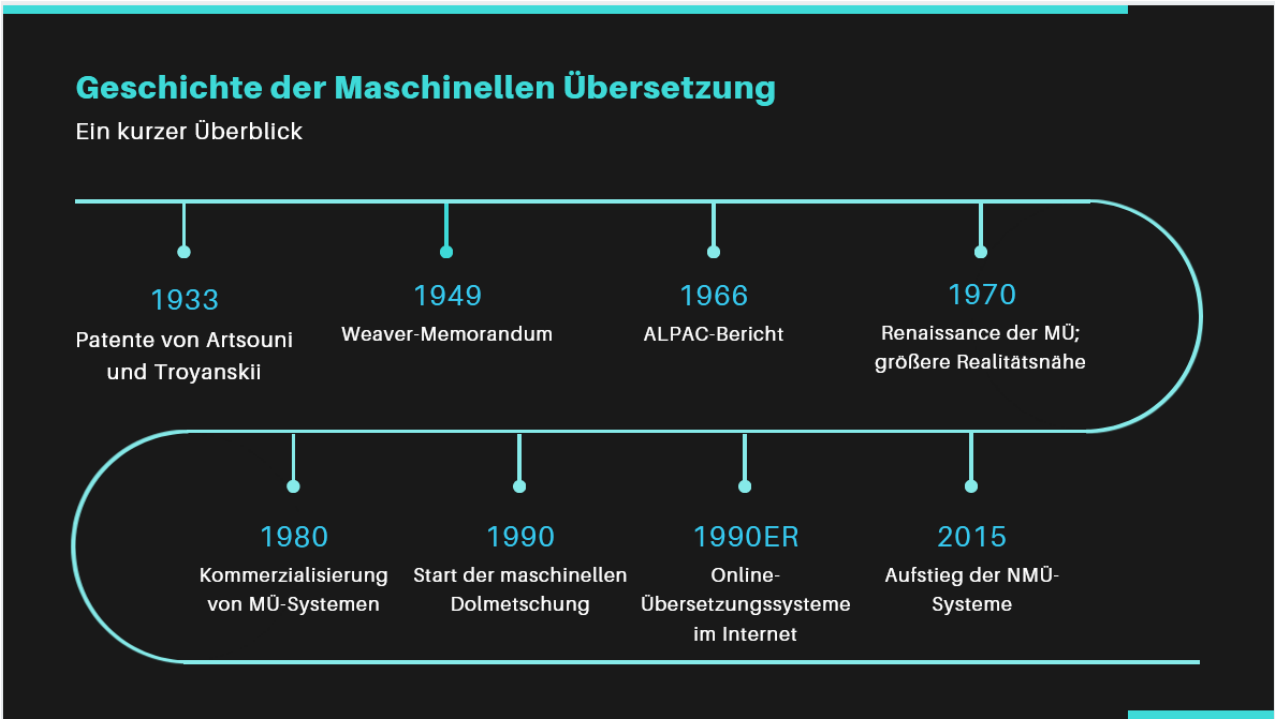

Abbildung 1: Die Geschichte der maschinellen Übersetzung (eigene Grafik, erstellt am 13.08.2020)

2. 1 Ein holpriger Anfang

Den exakten Anfangszeitpunkt der maschinellen Übersetzung festzumachen ist schwierig. Bereits Descartes und Leibniz spekulierten, dass die Erfindung von Wörterbüchern, welche auf universellen nummerischen Codes basieren, Sprachbarrieren überbrücken könnten (vgl. Hutchins und Somers 1992: 5). Mechanische Wörterbücher wurden unter anderem von Autoren wie Johann Joachim Becher (1661) und Athanasius Kircher (1663) veröffentlicht. Diese Ideen könnte man zwar als erste Anfänge der maschinellen Übersetzung klassifizieren, allerdings waren sie für die Entstehung und Entwicklung von sogenannten internationalen „Plansprachen“

(darunter ist die bekannteste gewiss Esperanto) von weitaus größerer Relevanz (vgl. Ramlow 2009: 54). Tatsächlich musste sich die Menschheit fast bis zur Mitte des 20. Jahrhunderts gedulden, bis die ersten konkreten Schritte zur Mechanisierung von Translation gemacht wurden. 1933 wurden, unabhängig voneinander, zwei Patente für maschinengestützte Übersetzungssysteme angemeldet – eines in Frankreich von George Artsouni und eines in Russland von Petr Troyanskii. Der Prototyp von ersterem wurde 1937 veröffentlicht und stieß anfangs auf großes Interesse, doch durch den Ausbruch des Ersten Weltkrieges geriet diese Erfindung rasch in Vergessenheit. Troyanskiis Übersetzungssystem war komplexer (der Übersetzungsprozess war bei ihm in drei Phasen gegliedert) und er stellte 1939 auch noch eine verbesserte Version seiner Übersetzungsmaschine vor, allerdings stießen seine Bemühungen in der Sowjetunion auf wenig Anerkennung (vgl. Krenz/Ramlow 2008: 28-29).

2. 2 Anfänglicher Optimismus bis zum ALPAC-Bericht

Dem Thema der maschinellen Übersetzung wurde erst 1949 ein gewisser öffentlicher Bekanntheitsgrad zuteil; vorherige Entwicklungen waren bloß einem äußerst kleinen Expertenkreis vertraut. Warren Weaver veröffentlichte in jenem Jahr ein Memorandum, in welchem er behauptete, langjährig bestehende Probleme von Übersetzungen mit der damals aufkommenden Computertechnologie lösen zu können (vgl. Arnold et al. 1994: 13). Das MÜ-System der „ersten Generation“ war jedoch, trotz des anfänglichen Optimismus, alles andere als perfekt. Damals herrschte noch die Prämisse, dass es für jegliche Wörter der Ausgangssprache immer genau eine Entsprechung in der Zielsprache gebe. Somit würde die konsequente Zuordnung dieser „gleichen“ Wörter ausreichen, um eine adäquate Übersetzung zu schaffen (vgl. Blatt et al. 1985: 46-47). Ferner kamen der syntaktischen und vor allem semantischen Analyse des Ausgangstextes bestenfalls untergeordnete Rollen zu (vgl. Austermühl 2001: 155). Trotz augenscheinlicher Mängel kam es 1954 zu einer Demonstration an der Georgetown University: 49 sorgfältig ausgewählte russische Sätze wurden, bei einem sehr eingeschränkten Wortschatz von 250 Vokabeln und nur sechs Grammatikregeln, ins Englische übersetzt. Trotz des äußerst geringen wissenschaftlichen Mehrwerts war diese Vorführung beeindruckend genug, um großzügige Förderungen für die Forschung von maschineller Übersetzung in den Vereinigten Staaten sicherzustellen. Außerdem war dies auch der Startschuss für ähnliche Projekte in anderen Ländern – allen voran in der Sowjetunion (vgl. Hutchins/Somers 1992: 5-6).

Die darauffolgenden, forschungsintensiven Jahre brachten viele neue, bedeutende Erkenntnisse nicht nur für die maschinelle Übersetzung, sondern auch für die Bereiche Künstliche Intelligenz, Computerlinguistik und Linguistik. Das primäre Ziel dieser Forschung – vollautomatische, qualitativ hochwertige Übersetzungen durch Übersetzungssysteme

(FAHQT), war jedoch nach wie vor unerreicht. Ernüchterung ersetzte die anfängliche Euphorie. Je mehr geforscht wurde, desto offenkundiger wurde die Tatsache, dass die Komplexität der linguistischen Probleme unterschätzt worden war (vgl. Hutchins/Somers 1992: 6-7). Bar Hillels Bericht, welcher 1960 in der Zeitschrift *Advances in Computers* erschien, war sehr kritisch; vollständig automatisierte Übersetzungen in erstklassiger Qualität seien laut ihm nicht nur in den nächsten Jahren, sondern grundsätzlich unmöglich. Vielmehr plädierte er für die Entwicklung maschineller Übersetzungssysteme, in denen der Mensch verstärkt in den Prozess eingreifen kann (vgl. Krenz/Ramlow 2008: 30).

Der ALPAC-Bericht (Automatic Language Processing Advisory Committee) kam 1966 zu einem ähnlich aussichtslosen Fazit für den Einsatz von MÜ-Systemen. Auch dieser Bericht zweifelte an der Realisierbarkeit von preisgünstigen, schnell verfügbaren Übersetzungen in einer brauchbaren Qualität, welche maschinell angefertigt worden sind. Der ALPAC-Bericht wurde nicht einfach so hingenommen – zahleiche Kritiker warfen der Kommission unter anderem Kurzsichtigkeit vor und, dass eine Beurteilung der MÜ nach einer so kurzen Forschungszeit verfrüht sei (vgl. Blatt et al. 1985: 52-53). Ungeachtet der Kritik hatte der ALPAC-Bericht fatale Auswirkungen auf die MÜ-Forschung. Die staatlichen Förderungen der USA wurden drastisch reduziert und jeglicher Optimismus verschwand (vgl. Arnold et al. 1994: 14). Außerdem führten die vernichtenden Schlussfolgerungen des Berichts dazu, dass US-amerikanische WissenschaftlerInnen nun versuchten, sich vom Stigma der maschinellen Übersetzung zu distanzieren (vgl. Hutchins 2000: 11). Infolgedessen stand die MÜ-Forschung in der USA praktisch über ein Jahrzehnt lang still – der Schaden an der öffentlichen Wahrnehmung von MÜ, der durch den Bericht angerichtet worden war, hatte noch langfristigere Auswirkungen. Außerhalb der USA wurde jedoch weitergeforscht, etwa in Westeuropa und Kanada. Denn in den Staaten der Europäischen Gemeinschaft stieg die Nachfrage nach Übersetzungsdienstleistungen stetig (vgl. Hutchins/Somers 1992: 7). Es sollte allerdings bis in die 1970er Jahre dauern, bis die maschinelle Übersetzung eine Renaissance erlebte (vgl. Arnold et al. 1994: 15).

2. 3 Wiedergeburt nach dem ALPAC-Bericht

Die MÜ-Forschung hatte sich gewandelt – infolge des ALPAC-Berichts wurde eine größere Realitätsnähe praktiziert. Nun wurde versucht, maschinelle Übersetzung in konkreten Arbeitsbereichen und Sachgebieten anzuwenden und nicht mehr wahllos Texte zu übersetzen. Ferner lag der Fokus der MÜ-Forschung auf folgenden drei Bereichen: Entwicklung von transferbasierten Systemen, Entwicklung neuartiger interlinguabasierter Systeme und die Einbeziehung der Erforschung von Künstlicher Intelligenz. Ab Mitte der 1970er kam außerdem dazu, dass eine rege Nachfrage nach und auch Interesse an MÜ-Systemen entstand (vgl.

Krenz/Ramlow 2008: 33). Die Europäischen Gemeinschaften erwarben die englisch-französische Version des SYSTRAN-Systems, ein deutlich verbesserter „Nachfahre“ der ersten Systeme, welche an der Georgetown University entwickelt worden waren. Eine russisch-englische Version desselben Systems wurde von der NASA und der US-amerikanischen Luftwaffe verwendet. In diesen Jahren kam es auch zu den ersten Erfolgen in Japan (vgl. Arnold et al. 1994: 15). Die MÜ-Forschung hatte es nun endlich geschafft, sich von dem Image eines suspekten, rein theoretischen Forschungsfeldes zu distanzieren und der Welt einen praktischen Mehrwert bieten zu können (vgl. Hutchins 2000: 12). Die maschinelle Übersetzung war wiedergeboren.

2. 4 Entwicklung in den 80er und 90er Jahren

In den 80er Jahren war eine zunehmende Kommerzialisierung der maschinellen Übersetzung bemerkbar. Eine Vielzahl von kommerziellen Übersetzungssystem wurde in Japan entwickelt – diese waren oft computergestützte Systeme für das Sprachenpaar Japanisch-Englisch. Doch auch in Nordamerika und Europa mehrte sich die Zahl der kommerziellen MÜ-Systeme. Ebenso entstanden nun auch in Ländern wie der Sowjetunion, China, Korea und Taiwan kommerzielle Systeme, wenn auch in geringerer Anzahl (vgl. Krenz/Ramlow 2008: 34).

Ende der 80er Jahre wurde im maschinellen Übersetzungsprozess auch erstmals auf korpusbasierte Verfahren zurückgegriffen. Hierbei wird zwischen dem statistischen und dem beispielbasierten Ansatz unterschieden. Davor waren MÜ-Systeme von einem regelbasierten Ansatz geprägt, bei welchem unter anderem Regeln für die Syntaxanalyse und den lexikalischen Transfer angewandt wurden (vgl. Ramlow 2009: 67). Ein weiterer neuer Ansatz war eine sich an Künstlicher Intelligenz orientierende, wissensbasierte Methode (vgl. Krenz/Ramlow 2008: 35).

Die 90er Jahre läuteten das Zeitalter der Computerisierung ein, mit welcher eine weiter zunehmende Kommerzialisierung der MÜ-Systeme einherging. MÜ-Systeme wurden nun nicht nur von internationalen Organisationen genutzt, sondern auch von Unternehmen und professionellen ÜbersetzerInnen (vgl. Krenz/Ramlow 2008: 28-29). Im Zuge der Globalisierung wurde es zu einer Notwendigkeit für multinationale Unternehmen, Dokumente wie Gebrauchsanweisungen, Produktbeschreibungen und dergleichen rasch und in einer zufriedenstellenden Qualität zu übersetzen, um keine internationalen Marktanteile zu verlieren.

Des Weiteren begann in diesem Jahrzehnt auch die Forschung des maschinellen Dolmetschens. Hierbei wurden zwei Verfahren kombiniert – die der akustischen Erkennung und die der Interpretation sprachlicher Äußerungen. Die erste erfolgreiche Demonstration eines

Dolmetschsystems in und aus den Sprachen Deutsch, Englisch und Japanisch fand 1993 statt. Ferner wurden nun auch maschinelle Hilfsmittel wie elektronische Glossare, Wörterbücher und Terminologiedatenbanken eingesetzt. Von diesen Neuerungen profitierten jedoch nicht nur professionelle ÜbersetzerInnen; Übersetzungssysteme wurden nun auch von Privatpersonen benutzt, die vor allem E-Mails und Homepages aus dem Internet übersetzten. Sowohl Angebot als auch Nachfrage für Übersetzungssysteme für den privaten Bereich verzeichneten ein starkes und stetiges Wachstum (vgl. Ramlow 2009: 68-70).

Eine weitere bemerkenswerte Tendenz war die territoriale Verschiebung der MÜ-Forschung – während Impulse in der Vergangenheit stets aus Nordamerika, Westeuropa und Japan kamen, stieg die Relevanz der Forschung in Ländern wie China, Taiwan, Korea und Indien stetig. Im Zuge dieser Entwicklung lag der Fokus nun auch nicht mehr so stark wie früher auf einigen ausgewählten, international „wichtigen" westeuropäischen Sprachen; ab Mitte der 90er Jahre wurde verstärkt versucht, MÜ-Systeme mit einer wachsenden Zahl von Sprachenpaaren zu entwickeln. Laut Ramlow (2009: 36) ist insgesamt feststellbar, dass die maschinelle Übersetzung über Jahre hinweg zwar primär ein forschungsorientiertes Projekt war, im Laufe der Jahrzehnte jedoch nicht nur in der Übersetzungspraxis an Relevanz gewonnen hat, sondern auch zu einem kommerziellen Produkt geworden ist.

Mit dem Aufkommen des Internets ab der Mitte der 90er Jahre werden nun auch Online-Übersetzungssysteme im Internet angeboten. Die Qualität der Übersetzungen, die von solchen Systemen erzeugt wurden, ließ jedoch vor allem anfangs sehr zu wünschen übrig (vgl. Ramlow 2009: 70). Trotzdem wurde ab den 90er Jahren eine immense Menge an Wörtern von Onlinesystemen übersetzt. 2012 wurden circa 75 Milliarden Wörter vom bekannten Programm Google Übersetzer übersetzt – und das täglich. Im Mai 2016 waren es in etwa 143 Milliarden Wörter in über 100 verschiedenen Sprachkombinationen (vgl. Way 2018: 161).

Ebenfalls in den 90er Jahren begann der Aufstieg von statistischen Ansätzen bei MÜ-Systemen. Einer davon, der phrasenbasierte Ansatz, dominierte den Bereich der maschinellen Übersetzung für Jahre – bis zur Mitte der 2000er Jahre, als den neuronalen Systemen der Durchbruch gelang (vgl. Koehn 2017: 6).

2. 5 Aufstieg der neuronalen maschinellen Übersetzung

Bereits in den 90er Jahren präsentierten Forcada und Castaño et al. Modelle für die MÜ, welche mit neuronalen Netzwerken arbeiteten. Allerdings überstieg die Komplexität der Berechnungen, welche für neuronale maschinelle Übersetzung (NMÜ) notwendig ist, die rechnerischen Kapazitäten, über die die Computer der damaligen Zeit verfügten. Aus diesem

Grund wurde dem neuronalen Ansatz für die MÜ für mehr als zwei Jahrzehnte nur wenig Beachtung geschenkt. In diesem Zeitraum traten korpusbasierte Ansätze wie die der statistischen MÜ in den Vordergrund. Erst durch die Pionierarbeit von Schwenk im Jahr 2007 wurden neuronale Sprachmodelle in „traditionellen" statistischen MÜ-Systemen integriert. Diese neuen Ideen wurden anfangs allerdings bloß langsam aufgenommen, hauptsächlich aufgrund computertechnischer Bedenken (vgl. Koehn 2017: 5).

In den 2010er Jahren folgte ein kometenhafter Aufstieg der NMÜ-Systeme. 2015 wurde bei der Conference on Machine Translation (WMT) nur ein einziges, rein neuronales MÜ-System vorgelegt. Dieses System war zwar durchaus wettbewerbsfähig, wurde jedoch von traditionelleren Systemen übertroffen. Bereits bei der darauffolgenden Conference on Machine Translation 2016 gewann ein neuronales maschinelles Übersetzungssystem aber in fast allen Sprachenpaaren. 2017 wurden fast ausschließlich NMÜ-Systeme eingereicht (vgl. Koehn 2017: 5-6). Zahlreiche Studien belegen inzwischen, dass neuronale Systeme in der Lage sind, signifikant bessere Ergebnisse zu erzielen als phrasenbasierte Systeme. Dennoch weisen neuronale Systeme auch Probleme auf – insbesondere im Umgang mit seltenen Wörtern und langen Sätzen (Bentivogli et al 2016: 9).

Während es nach wie vor noch kein „perfektes" MÜ-System gibt, nimmt der Gebrauch von Übersetzungssystemen Jahr für Jahr zu, was darauf hindeutet, dass auch die Qualität der erzeugten Übersetzungen steigt (vgl. Way 2018: 173) Die Geschichte der maschinellen Übersetzung ist noch lange nicht zu Ende erzählt.

3 Was ist maschinelles Übersetzen?

In diesem Kapitel werden die wichtigsten Begriffe für den Bereich der maschinellen Übersetzung erläutert. Allen voran der Begriff der maschinellen Übersetzung an sich und die unterschiedlichen Assoziationen dazu. Zusätzlich werden gebräuchliche Unterbegriffe erklärt. Da Englisch die Lingua Franca in der MÜ-Forschung ist, werden in der Folge auch vermehrt englische Begriffe und Abkürzungen verwendet.

3. 1 Definitionen der maschinellen Übersetzung

Zuallererst sei hier die zentrale Definition von Hutchins/Somers (1992: 3) angeführt:

„The term Machine Translation (MT) is the now traditional and standard name for computerised systems responsible for the production of translations from one natural language into another, with or without human assistance."

Außerdem führen Hutchins/Somers (1992: 3) aus, dass computerbasierte, ÜbersetzerInnen unterstützende Werkzeuge wie Terminologiedatenbanken, Wörterbücher und Textbearbeitungsprogramme in dieser Definition nicht inkludiert sind. Dafür sind auch Systeme gemeint, in welchen BenutzerInnen den Computer in der Erstellung von Übersetzungen unterstützen, was die Vorbereitung von Texten und Revisionen des Outputs impliziert. Der zentrale Punkt der maschinellen Übersetzung ist jedoch, laut den beiden Autoren, die Automatisierung des *gesamten* Translationsprozesses. Bereits jene Definition, die circa 30 Jahre zuvor im ALPAC-Bericht verwendet wurde, um die MÜ zu beschreiben, betont, dass der ganze Übersetzungsprozess ohne menschliche Hilfe vonstattengehen sollte:

"'Machine Translation' presumably means going by algorithm from machine-readable source text to useful target text, without recourse to human translation or editing." (vgl. ALPAC 1966: 19)

Zweifellos sind beide angeführten Definitionen sehr alt. Deswegen sei hier eine rezentere Begriffsbestimmung angeführt:

> Die maschinelle Übersetzung ist nicht eine durch den Computer unterstützte Übersetzung, hier muss man deutlich zwischen der maschinellen Übersetzung und der durch den Computer unterstützten unterscheiden. In der maschinellen Übersetzung wird der Übersetzungsvorgag ausschließlich durch den Computer ohne menschliche Intervention durchgeführt (obwohl die menschliche Intervention manchmal am Anfang oder zum Schluss der Übersetzung in der Form der Präeditation oder Posteditation geeignet ist), d. h. der Computer liefert selbst die Übersetzung. In der durch den Computer unterstützten Übersetzung ist diese durch den Übersetzer – Menschen realisiert, der beim Übersetzungsvorgang verschiedene Übersetzungsmittel benutzt, d. h. CAT-Instrumente unterstützen nur die Übersetzung, die der Übersetzer durchführt. (vgl. Bánik et al. 2019: 21)

Wie auch bei den anderen Definitionen wird hier hervorgehoben, dass der Übersetzungsvorgang „ausschließlich durch den Computer" und „ohne menschlicher Intervention" stattfindet. Unterschiede lassen sich jedoch beim Gebrauch einer Prä- oder Postedition ausmachen. Es ist also feststellbar, dass sich der Kerngedanke der maschinellen Übersetzung, trotz der zahllosen, bedeutenden Fortschritte in diesem Bereich, bemerkenswerterweise nicht (oder kaum) verändert hat. Diese Definition von Bánik et al. wird auch jene sein, nach der sich diese Arbeit im späteren Verlauf orientieren wird.

3. 2 Unterbegriffe

Abhängig davon, ob und wie sehr die/der ÜbersetzerIn in den maschinellen Übersetzungsprozess eingreift, wird grundsätzlich zwischen Machine-Aided Human Translation (MAHT) und Human-Aided Machine Translation (HAMT) unterschieden. Eine eindeutige Grenzziehung zwischen den ersten beiden erweist sich jedoch als problematisch – sogenannte CAT-Tools (Computer-Aided Translation beziehungsweise Computer-Assisted Translation) sind oftmals sogar grenzübergreifend. Vollautomatische Übersetzungssysteme, welche ohne menschlichen Einfluss Übersetzungen produzieren, nennt man Fully Automatic High-Quality Translation (FAHQT) (vgl. Bánik et al. 2019: 20-21; Hutchins/Somers 1992: 3). Diese Begriffe werden heute noch verwendet. In den folgenden Unterkapiteln werden diese Begriffe genauer erklärt. Eine gute Übersicht über die vielen Unterbegriffe bietet folgende Grafik:

<table>
<tr><td>Keine (-)</td><td colspan="2">Intervention durch den Menschen</td><td>Ganze (+)</td></tr>
<tr><td>Ganze (+)</td><td colspan="2">Automatisierung</td><td>Keine (-)</td></tr>
<tr><td>FAHQMT

voll automatische maschinelle Übersetzung in hoher Qualität</td><td>HAMT

durch den Menschen unterstützte maschinelle Übersetzung</td><td>MAHT

durch den Computer unterstützte menschliche Übersetzung</td><td>HT

menschliche Übersetzung / humane Übersetzung</td></tr>
<tr><td colspan="2">Maschinelle Übersetzung</td><td colspan="2">Menschliche Übersetzung</td></tr>
<tr><td></td><td colspan="2">durch den Computer unterstützte Übersetzung</td><td></td></tr>
</table>

Abbildung 2: Beziehung zwischen der Intervention des Menschen und der Automatisierung der Übersetzung (vgl. Bánik et al. 2019: 21).

3. 2. 1 Machine-Aided Human Translation (MAHT)

Machine-Aided Human Translation oder maschinengestützte Humanübersetzung steht ÜbersetzerInnen bereits seit geraumer Zeit zur Verfügung. Hierbei handelt es sich in erster Linie um Computerprogramme, welche übersetzerische Arbeitsprozesse unterstützen und somit optimieren sollen. Dazu zählen unter anderem Textverarbeitungsprogramme (mit integrierten

Rechtschreib- und Grammatiküberprüfungstechniken), elektronische Wörterbücher (ein- und mehrsprachig), elektronische Terminologiedatenbanken, Glossare oder Textkorpora (vgl. Ramlow 2009: 113). Diese Programme haben die Art und Weise, wie ÜbersetzerInnen auf der ganzen Welt arbeiten, von Grund auf verändert. Die rasante Entwicklung der Technologie, die weit verbreitete Nutzung des Internets und die Globalisierung haben zu einer Zunahme des zu übersetzenden Materials geführt. Durch diese Faktoren wurde eine Effizienzsteigerung im Übersetzungsprozess notwendig. Der Einsatz von MAHT-Technologie ist eine Möglichkeit, die Produktivität zu steigern (vgl. Temizöz 2016: 646).

3. 2. 2 Human-Aided Machine Translation (HAMT)

Bei HAMT-Systemen wird die Übersetzung primär von einem Übersetzungssystem, also von einer Maschine/einem Programm und nicht von menschlichen ÜbersetzerInnen, durchgeführt. Ein/e HumanübersetzerIn muss den Übersetzungsprozess allerdings unterstützen – wenn der Eingriff vor dem Übersetzungsprozess stattfindet, wird der Begriff „Präedition“ verwendet, falls er danach von statten geht der Begriff „Postedition“. Bei ersterem können komplexe syntaktische Strukturen vereinfacht und Mehrdeutigkeiten beseitigt werden. Beim Postediting wird die maschinell erzeugte Übersetzung von menschlichen ÜbersetzerInnen überarbeitet und Fehler wie Sinnentstellungen, Syntaxfehler und nicht idiomatische Übersetzungen werden korrigiert beziehungsweise verbessert (vgl. Ramlow 2009: 115). Wird während des Übersetzungsprozesses eingegriffen, so nennt man dies Interaktion. Hierbei kann ein Mensch Ambiguitäten auflösen, dem Programm bei der Interpretation von Strukturen oder bei der Auswahl von der korrekten Lexik unterstützen (vgl. Hutchins/Somers 1992: 151).

3. 2. 3 Fully Automatic High-Quality Translation (FAHQT)

“Vollautomatische” maschinelle Übersetzungen bezeichnen vollständige Übersetzungsprozesse, bei denen keine menschliche ÜbersetzerInnen eingreifen. Der Begriff wurde von Yehoshua Bar-Hillel geprägt, welcher konsequent betonte, dass eine solche vollautomatische Übersetzungen nicht nur ein unrealistisches Ziel für die Forschung sei, sondern grundsätzlich unmöglich. Seine Kritik beeinflusste auch den späteren ALPAC-Bericht (vgl. Hutchins/Somers 1992: 148). Die Notwendigkeit einer menschlichen Revision, ähnlich wie bei der HAMT, wird auch bei folgenden Stellen in der Fachliteratur evident:

„In FAMT [sc. Fully Automatic Machine Translation] there is no human intervention between the input of the original text and the final raw machine output of the translated text. Of course, revision of the raw output may be required." (vgl. Lehrberger/Bourbeau 1988:8)

Bei diesem Zitat von Lehrberger/Bourbeau ist eine anschließende Revision durch einen Menschen eine Selbstverständlichkeit. Ähnlich sehen es Goshawke et al. (1987: 6):

> Machine Translation (MT) is the transfer of meaning from one natural (human) language to another with the aid of a computer. There are few systems that are, or even attempt to be, complete machine translation systems in themselves – nearly all systems are Machine Aided Translation (MAT), involving human help either at the input stage (pre-editing) or the output stage (post-editing) or both.

Trotz des beträchtlichen Alters beider Zitate hat deren Inhalt noch bis heute Bestand; es gibt nach wie vor keine vollautomatischen Übersetzungssysteme, welche ohne Prä- beziehungsweise Postedition makellose Übersetzungen produzieren können. Laut Krenz/Ramlow (2008: 52-53) findet vollautomatische maschinelle Übersetzung im engeren Sinne de facto keine Anwendung. Des Weiteren ist die Idee eines Übersetzungssystems, welches keiner Überarbeitung bedarf, laut Ramlow (2009: 116) insofern absurd, weil Humanübersetzungen bei welchen eine ausgezeichnete Qualität erwartet/gefordert wird, genauso wenig ohne die Unterstützung einer zweiten Übersetzerin/eines zweiten Übersetzers in Form einer Revision auskommen. Zusammenfassend lässt sich also feststellen, dass für von einem maschinellen Übersetzungssystem erzeugte Übersetzungen von *hoher* Qualität eine Überarbeitung/Revision immer ein fundamentales Element des Übersetzungsprozesses ist.

Nichtsdestotrotz gibt es heutzutage eine Vielzahl von Situationen, in denen MÜ-Systeme ohne menschliche Unterstützung eingesetzt werden. Diese Situationen können aufgrund von Zeitmangel, Kostengründen oder wegen einer Kombination dieser beiden Faktoren entstehen. Unter solchen Umständen gibt es schlichtweg keinen Platz für menschliches Eingreifen. Auch wenn vollautomatische Systeme Verwendung finden, bedeutet dies aber nicht, dass diese MÜ-Systeme „perfekt" sind. Zahlreiche Unternehmen und akademische WissenschaftlerInnen, welche MÜ-Systeme nutzen, sind bemüht, diese Systeme stetig weiterzuentwickeln und zu verbessern. Dies ist ein Indiz dafür, dass die Übersetzungsqualität von MÜ-Systemen noch weiter optimiert werden kann (vgl. Way 2018: 159-160).

4 Strategien und Ansätze der maschinellen Übersetzung

MÜ-Systeme bedienen sich einer Vielzahl von verschiedenen Strategien und Ansätzen, um eine Übersetzung zu generieren. Während in der Vergangenheit ausschließlich regelbasierte Übersetzungsstrategien für MÜ-Systeme verfügbar waren, sind seit dem ALPAC-Bericht neuere Ansätze entwickelt worden. Ältere Systeme, die heutzutage nicht mehr, oder nur noch kaum verwendet werden, werden in diesem Kapitel genauso beschrieben wie aktuellere Ansätze, wie zum Beispiel der neuronale Ansatz, der neueste Ansatz im Bereich der maschinellen Übersetzung. Ein Überblick über die verschiedenen Ansätze gibt folgende Grafik:

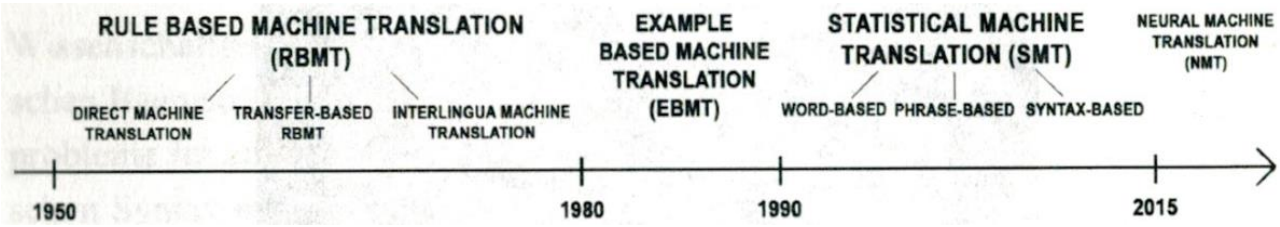

Abbildung 3: Ansätze der maschinellen Übersetzung (vgl. Bánik et al 2019: 32).

4. 1 Regelbasierte Übersetzungsstrategien

Bei regelbasierten Übersetzungsstrategien orientieren sich MÜ-Systeme vorwiegend an linguistischen Regeln, um einen Ausgangstext in die Zielsprache zu übersetzen. Hierbei wird zwischen direkten und indirekten Systemen differenziert.

4. 1. 1 Direkte Systeme

Alle MÜ-Systeme der ersten Generation wandten direkte Systeme an. Diese Herangehensweise war gekennzeichnet durch einen kompletten Mangel von Zwischenschritten beim Übersetzungsprozess; der Ausgangstext wurde „direkt“ in die Zielsprache übersetzt. In Anbetracht der Tatsache, dass MÜ-Systeme der ersten Generation vorrangig auf Computern der späten 50er und frühen 60er Jahre verwendet wurden, welche selbst im Vergleich mit den bescheidensten der heutigen elektronischen Taschenrechner primitiv wirken, überrascht es nicht, dass das direkte System zahlreiche Makel aufwies. Charakteristisch für diese frühen Systeme waren „Wort-für-Wort“-Übersetzungen und dementsprechende Fehltranslationen (vgl. Hutchins/Somers 1992: 71-72). Eine bestenfalls oberflächliche syntaktische Analyse und

ungeeignete Interpretationen der Ausganstexte führten zu einer äußerst geringen Qualität der Übersetzungen (vgl. Ramlow 2009: 74-75).

4. 1. 2 Indirekte Systeme

Das Scheitern der direkten Systeme führte zur Entwicklung von fortschrittlicheren linguistischen Modellen für den Übersetzungsprozess. Es wurde entschieden, mehr Fokus auf die Analyse des Ausgangstextes zu legen – anders als beim direkten System besteht der Übersetzungsprozess beim indirekten System aus mehreren Phasen. Es wird zwischen interlinguabasierten und transferbasierten Systemen unterschieden.

4. 1. 2. 1 Interlinguabasierte Methode

Beim interlinguabasierten System wird zunächst der Ausgangstext analysiert und in einer, wie der Name bereits andeutet, Interlingua, also Zwischensprache, dargestellt. In der Folge wird dann, ausgehend von dieser Interlingua, der Zieltext erstellt. Als eine solche Interlingua können natürliche Sprachen, künstliche wie Esperanto oder eine universelle sprachenunabhängige Repräsentation dienen. Als ein bedeutender Vorteil dieser Herangehensweise kann man das unproblematische Erstellen von mehrsprachigen Systemen sehen (vgl. Krenz/Ramlow 2008: 39).

Während das Hinzufügen neuer Sprachen in einem Interlinguasystem einfach ist, stellt das Finden einer „wahrlich universellen“ und sprachenunabhängigen Zwischensprache eine große Herausforderung dar, der bereits Linguisten und renommierte Philosophen des 17. Jahrhunderts nicht gewachsen waren (siehe Kapitel 2.1) (vgl. Hutchins/Somers 1992: 75). Dieses Problem wurde bis heute noch nicht gelöst. Tatsächlich handelt es sich bei allen bisherigen interlinguabasierten Systemen bloß um experimentelle Systeme – in der Praxis werden sie nicht verwendet.

4. 1. 2. 2 Transferbasierte Systeme

Übersetzungsprozesse bei transferbasierten Systemen finden in drei Phasen statt. Zuerst wird eine sprachenabhängige, abstrakte Repräsentation eines Satzes in einer Ausgangssprache erzeugt. Anschließend findet ein Transfer von der ausgangsprachlichen Repräsentation des Satzes in eine Satzrepräsentation der Zielsprache statt und zuletzt wird, aufbauend auf die

abstrakte Repräsentation, der Satz in der Zielsprache generiert. Um diesen Prozess zu ermöglichen, benötigt dieses System eine Grammatik, in der zweisprachige Transferregeln vorhanden sind (vgl. Ramlow 2009: 75).

Im Gegensatz zu interlinguabasierten Systemen kristallisiert sich hier ein klarer Nachteil bezüglich mehrsprachiger Systeme heraus. Das Hinzufügen einer weiteren Sprache bedeutet nämlich nicht nur zwei zusätzliche Analysen und Satzgenerierungen, sondern auch weitere Transfers – die genaue Anzahl dieser hängt von der Anzahl der Sprachen im bestehenden System ab (vgl. Hutchins/Somers 1992: 75-76).

Transferbasierte Systeme finden bei bidirektionalen Übersetzungen innerhalb eines Sprachenpaares oder bei unidirektionalen Übersetzungen für eine begrenzte Anzahl von Sprachen Verwendung (vgl. Krenz/Ramlow 2008: 42).

4. 2 Neuere Ansätze

Neben den regelbasierten Übersetzungsstrategien, bei denen die Analyse sprachlicher Regeln im Vordergrund stehen, stellen die korpusbasierte Methode und die wissensbasierte Methode neuere Ansätze für MÜ-Systeme dar. Der neuronale Ansatz ist zwar auch ein korpusbasierter, jedoch wird diesem Ansatz aufgrund der heutigen Dominanz von NMÜ-Systemen im Bereich der MÜ ein eigenes Unterkapitel gewidmet.

4. 2. 1 Korpusbasierter Ansatz

Beim korpusbasierten Ansatz wird beim Übersetzungsprozess auf große, zweisprachige Textkorpora zurückgegriffen. Hierbei unterscheidet man zwischen beispielbasierter und statistischer maschineller Übersetzung.

Bei ersterer wird bei der Übersetzung auf ein umfassendes zweisprachiges Korpus an zuvor gespeicherten Beispielübersetzungen zurückgegriffen. Den Strukturen aus der Ausgangssprache werden automatisch zielsprachliche Strukturen zugeteilt, die diesen am meisten ähneln oder sogar identisch sind (vgl. Krenz/Ramlow 2008: 42).

Handelt es sich um einen reinen beispielbasierten Ansatz, wird die Übersetzung ausschließlich, und ohne Abgleich von sprachlichen Regeln, durch den Vergleich von Sätzen vom zweisprachigen Korpus erzeugt. Falls eine vollständige Übersetzung jedoch nicht möglich ist – etwa, weil keiner der gespeicherten Beispielsätze der Korpora für eine Übersetzung

herangezogen werden kann, kann die beispielbasierte Methode mit einem regelbasierten Ansatz kombiniert werden. In einem solchen Fall werden ganze Sätze oder Satzteile, welche zuvor nicht übersetzt worden konnten, linguistisch analysiert und anschließend mithilfe der Analyseergebnisse ergänzt (vgl. Ramlow 2009: 80).

Beim statistischen Ansatz werden Textkorpora auf berechenbare Zusammenhänge untersucht. Von diesen ausgehend wird versucht, wiederkehrende Muster für die Übersetzung zu prognostizieren. Diese berechenbaren Zusammenhänge sind unter anderem die Länge von Sätzen, das gemeinsame Auftreten von Wörtern im Ausgangs- oder Zieltext und die Position von Wörtern im Satz (vgl. Ramlow 2009: 81). Bei diesem Ansatz ist nicht nur kein menschliches Eingreifen vorgesehen; auch von der Verwendung von linguistischen Informationen wird abgesehen (vgl. Krenz/Ramlow 2008: 43). Vor dem Aufstieg der neuronalen maschinellen Übersetzung in den 2010er Jahren dominierte der phrasenbasierte statistische Ansatz den Bereich der MÜ. Andere statistische Ansätze sind wort- und syntaxbasierte Ansätze (vgl. Bentivogli et al 2016: 1; Bánik et al. 2019: 32).

4. 2. 2 Wissensbasierter Ansatz

Auf der Suche nach moderneren, verbesserten Ansätzen in der MÜ nach der Zäsur des ALPAC-Berichts setzten viele ihre Hoffnung in den Einsatz von Künstlicher Intelligenz (KI). Bereits in den frühen 1970er Jahren waren WissenschaftlerInnen aus dem Bereich der künstlichen Intelligenz an der Stanford Universität an MÜ-Projekten beteiligt. Eines der wesentlichsten Probleme, mit welchem sich WissenschaftlerInnen beschäftigten, war das der sogenannten „semantischen Barriere“. Dies war auch das Hauptargument für den Einsatz von KI – MÜ-Systeme müssen die Bedeutung von Texten verstehen, um adäquat übersetzen zu können. Hierbei wurde angenommen, dass eine linguistische Analyse nicht ausreichte, um zufriedenstellende Ergebnisse zu erreichen; MÜ-Systeme sollen künftig auch mithilfe von „Weltwissen“ Texte übersetzen, daher der Begriff „wissensbasierter Ansatz“. Ähnlich einem interlinguabasierten System wird also eine semantische Analyse vorgenommen und eine sprachenunabhängige Repräsentation der Bedeutung des Textes ermittelt. Hinzu kommt jedoch in diesem Fall noch eine Interpretation unter Einbezug von Weltwissen (vgl. Hutchins/Somers 1992: 313-314).

Laut Thurmair (2017: 41) liegen die größten Probleme bei diesen Systemen „in der mangelnden Robustheit in Fällen von inkorrektem oder defizientem Input, in ihrer Schwäche im Umgang mit zu vielen alternativen Analysemöglichkeiten (Ermitteln der jeweils korrekten Alternative) und v. a. in ihrer Schwäche bei der Transfer-Selektion“.

4. 2. 3 Neuronale maschinelle Übersetzung

Die neuronale maschinelle Übersetzung (NMÜ) ist ein relativ neuer Ansatz. Wie bereits in Kapitel 2.5 erwähnt, gelangen NMÜ-Systemen 2016 bei der Conference on Machine Translation große Erfolge. Seitdem dominieren NMÜ-Systeme die MÜ-Industrie. Forcada (2017: 292) beschreibt NMÜ folgendermaßen:

> Neural machine translation is a new breed of corpus-based machine translation (also called data-driven or, less often, corpus-driven machine translation). It is trained on huge corpora of pairs of source-language segments (usually sentences) and their translations, that is, basically from huge translation memories containing hundreds of thousands or even millions of translation units. In this sense, it is similar to the statistical machine translation technology that was the state of the art until very recently, but uses a completely different computational approach: neural networks. (Forcada 2017: 292)

Tatsächlich haben NMÜ-Systeme, trotz ihres Namens, nur eine indirekte Verbindung mit jenen Neuronen, die sich in menschlichen Gehirnen befinden. NMÜ-Systeme basieren auf einem Netzwerk von (künstlichen) Neuronen. Dieses Netzwerk besteht aus tausenden von künstlichen Einheiten, die Neuronen insofern ähneln, indem ihre Ausgabe oder Aktivierung von Reizen abhängt, die sie von anderen Neuronen erhalten. Ferner spielt die Stärke der Verknüpfungen untereinander eine Rolle.

Ähnlich wie bei statistischen MÜ-Systemen arbeiten NMÜ-Systeme auch mit zwei- beziehungsweise mehrsprachigen Textkorpora. Übersetzte Abschnitte werden ebenfalls nach dem Wahrscheinlichkeitsprinzip berechnet. Ein wesentlicher Unterschied ist aber, neben dem neuronalen Netz, auf welchem die NMÜ aufbaut, der Berechnungsvorgang der Wahrscheinlichkeit. Anders als beim statistischen Ansatz wird zuerst der Kontext des Ausgangstextes analysiert; erst anhand dieser Analyse wird in der Folge die Wahrscheinlichkeit für die Wörter im Zieltext berechnet. Ferner weist die NMÜ jedem einzelnen Wort zusätzliche sprachliche Informationen wie Deklination, Konjugation und Kontext zu. Diese Informationen werden im NMÜ-System gespeichert und erlernt. Dieser Vorgang wird Training genannt. Anschließend wird das Gelernte mittels der neuronalen Netze kodiert, um die Wahrscheinlichkeit für die „beste Übersetzung“ zu maximieren (vgl. Castilho et al 2017: 111; Forcada 2017: 300-301).

In einer Vielzahl von verschiedenen Studien konnten NMÜ-Systeme bessere Ergebnisse erzielen als phrasenbasierte statistische MÜ-Systeme. Unter anderem enthielten Übersetzungen von NMÜ-Systemen weniger Wortstellungs-, lexikalische und morphologische Fehler, was den Posteditions-Aufwand verringerte (vgl. Popović 2017: 209-210). Grundsätzliche Vorteile sind außerdem laut Bentivogli et al (2016: 1) eine Vereinfachung gegenüber früheren Paradigmen und ein effizienterer Einsatz von Personal- und Datenressourcen im Vergleich zu regelbasierter MÜ. Im Zuge dieser positiven Studien wechselte auch das bekannte Übersetzungsprogramm

Google Übersetzer zu einem NMÜ-System (vgl. Moorkens 2018: 377). Zwar ist der Übersetzungsprozess weniger transparent als bei früheren Systemen, aber Bentivogli et al. (2016: 6) zufolge stellt die NMÜ dennoch einen deutlichen Fortschritt im Vergleich zu vorherigen MÜ-Systemen dar. Laut Wu et al. (2016: 1-2) weisen NMÜ-Systeme drei weitere wesentliche Schwächen auf: langsames Training, Ineffektivität im Umgang mit seltenen Wörtern und manchmal das Unvermögen, alle Wörter des Ausgangssatzes zu übersetzen. Erstens erfordert es in der Regel einen beträchtlichen Zeit- und Rechenaufwand, ein NMÜ-System auf einem großen Übersetzungsdatensatz zu trainieren, wodurch die experimentelle Durchlaufzeit und die Innovationsgeschwindigkeit verlangsamt werden. Zweitens weisen NMÜ-Systeme bei der Übersetzung seltener oder ungewöhnlicher Wörter Mängel auf. Der Wortschatz neuronaler Modelle ist in der Regel auf 30.000 bis 50.000 Wörter begrenzt, doch da Übersetzung ein „offenes" Vokabular-Problem ist, sind NMÜ-Übersetzungsmodelle oftmals überfordert, insbesondere bei Sprachen mit produktiven Wortbildungsprozessen wie Deutsch (vgl. Sennrich et al 2016: 1). Schließlich erzeugen NMÜ-Systeme manchmal Zielsätze, die nicht alle Bestandteile des Eingangstextes übersetzen, was zu verwirrenden Übersetzungen führen kann (vgl. Wu et al 2016: 2).

5 Probleme der maschinellen Übersetzung

> Was MÜ […] nicht oder nur sehr eingeschränkt leisten kann, ist: fehlertolerant sein, ahnen, interpretieren, vermuten, assozieren, antizipieren oder gar paraphrasieren, zwischen den Zeilen lesen, Unübersetztes (richtig) verarbeiten oder auch nur Sinnvolles von Sinnlosem trennen. Ohne menschliches Weltwissen und ohne (künstliche) Intelligenz kann das Stück Software weder selbstständig (dazu)lernen, noch Entscheidungen treffen. (vgl. Burchardt/Porsiel 2017: 16)

Dieses Kapitel befasst sich mit grundlegenden Schwierigkeiten und Hindernissen im Bereich der maschinellen Übersetzung. Menschliche Sprache weist eine Vielzahl an Ambiguitäten auf – sowohl auf morphologischer als auch auf lexikalischer und syntaktischer Ebene. Menschlichen ÜbersetzerInnen ist es aufgrund ihres sprachlichen, aber auch aufgrund ihres außersprachlichen Wissen ohne weiteres möglich, dieses komplexe und vielschichtige Gewirr an Mehrdeutigkeiten aufzulösen. Maschinellen Übersetzungssystemen steht dieses immense Wissen jedoch gar nicht, oder nur begrenzt zur Verfügung. Dieses Welt- und Kulturwissen wird jedoch benötigt, wenn Wörter mehr als eine Bedeutung haben. In so einem Fall wird von einer Ambiguität oder Mehrdeutigkeit gesprochen. Mehrdeutigkeiten sind ein häufiges Phänomen menschlicher Sprachen. Eine Vielzahl von Worten ist mehrdeutig – was konsequenterweise bedeutet, dass Sätze meist mehrere verschiedene Bedeutungen haben. Dies ist für den Übersetzungsprozess nicht nur insofern ungünstig, weil alternative Bedeutungen oftmals gar nicht intendiert sind, sondern weil sich Mehrdeutigkeiten „multiplizieren". Besteht ein Satz aus zwei doppeldeutigen Worten, könnte ein Satz vier Bedeutungen haben (vgl. Arnold et al. 1994:

112). Besteht ein Satz aus einer noch größeren Anzahl von mehrdeutigen Worten, wird leicht eine beachtliche Nummer an möglichen Bedeutungen erreicht. Alle Möglichkeiten zu analysieren und sich im Anschluss dann für eine einzige Variante zu entscheiden, kann sich als problematisch herausstellen – vor allem für Maschinen. Diese Mehrdeutigkeit der natürlichen Sprache und die Inkongruenz der Lexik und Syntax zwischen den Sprachen stellen laut Bánik et al. (2019: 32) die wesentlichsten linguistischen Probleme der maschinellen Übersetzung dar. In diesem Kapitel wird grob zwischen morphologischer, lexikalischer und syntaktischer Mehrdeutigkeit unterschieden.

5. 1 Morphologische Mehrdeutigkeit

Wird eine regelbasierte maschinelle Analyse des Ausgangstextes vorgenommen, werden zuerst die grammatischen Kategorien wie Substantiv, Verb, Adjektiv, etc., bestimmt. Hierbei gibt es zwei Möglichkeiten. Wird entschieden, jedes Wort mit all seinen möglichen Flexionsformen in das Wörterbuch aufzunehmen, so ist eine morphologische Analyse verzichtbar (vgl. Krenz/Ramlow 2008: 56). In diesem Fall werden meist zwei Gründe angeführt, die diese Entscheidung rechtfertigen sollen: Einerseits wird auf die stetig sinkende Bearbeitungszeit durch moderne Computern hingewiesen, welche in der Vergangenheit bei umfangreichen Wörterbücher noch problematisch lang war. Denn Wörterbücher sind, was die Menge an enthaltenen Informationen betrifft, die größten Komponenten eines regelbasierten MÜ-Systems (bei statistischen bzw. korpusbasierten Methoden spielen Wörterbücher eine weniger wichtige Rolle). Ferner schränken, mehr als jede andere Komponente, der Umfang und die Qualität des Wörterbuchs den Spielraum und die Reichweite eines Systems sowie die zu erwartende Qualität der Übersetzung ein (vgl. Eisold 2017: 109; Arnold et al. 1994: 87-88). Andererseits muss bei dieser Variante auch auf keine möglichen, komplexen Flexionsparadigmen Rücksicht genommen werden.

Eine morphologische Analyse hingegen hat den entscheidenden Mehrwert, unbekannte Wörter leichter erkennen zu können. Indem grammatikalische Flexionsendungen analysiert werden, ist es möglich, auf die syntaktische Funktion eines Wortes in einem Satz zu schließen, selbst wenn das fragwürdige Wort selbst nicht übersetzt werden kann.

Für maschinelles Übersetzen sind Flexion, Derivation und Komposition die entscheidenden Wortbildungsverfahren. Die Analyse von Flexionsendungen weist dann Probleme auf, wenn man diesen mehr als eine Interpretation/Bedeutung zuweisen kann (vgl. Hutchins/Somers 1992: 82-83). Beispielsweise wird dem flektierten deutschen Verb *trinkst* problemlos eine einzige Interpretation zugewiesen, da die Flexionsendung *-st* keine Zweifel darüber aufkommen lässt, dass es sich hierbei um die 2. Person Singular Präsens Indikativ-

Verbform handelt. Die Infinitiv-Form *suchen* lässt allerdings mehr Interpretationsspielraum zu – sowohl 1. Person Plural Präsens Indikativ als auch 3. Person Plural Präsens Konjunktiv sind hierbei neben der Infinitiv-Form möglich. Für Verben mit einer *-en*-Endung ist demzufolge also eine Analyse des Kontextes vonnöten, um eine Disambiguierung vornehmen zu können (vgl. Krenz/Ramlow 2008: 57).

Derivation, oder auch Ableitung, bezeichnet ein Verfahren der Wortbildung. Viele Sprachen weisen hierbei Regelmäßigkeiten auf, welche ausgenutzt werden können, um den Umfang von Wörterbüchern zu reduzieren. Beispielsweise wird aus der Kombination eines Adjektivs mit dem Suffix *-heit* ein Nomen – *frei* + *heit*: „Freiheit"; *besonnen* + *heit*: „Besonnenheit" (vgl. Hutchins/Somers 1992: 83) Allerdings gibt es im Deutschen auch mehrdeutige Suffixe und Affixe. Zusätzlich erschwerend ist die Tatsache, dass Wortbildungsverfahren in der Ausgangssprache nicht in gleichem Maße wie in der Zielsprache anwendbar sind (vgl. Krenz/Ramlow 2008: 58).

Komposition, oder Wortzusammensetzung, beschreibt die Bildung eines neuen Wortes durch zwei bereits bestehenden Wörter/Wortstämme. Die Bedeutung und die Übersetzung ergeben sich oftmals aus den Wortbestandteilen. Problematisch sind hingegen Komposita, die auf mehr als eine Art segmentiert werden können und demzufolge mehrdeutig sind (vgl. Krenz/Ramlow 2008: 58-59). Beispielsweise ist Tischdecke eine offensichtliche Zusammensetzung der beiden Wörter *Tisch* und *Decke* und bloß auf eine Art zu segmentieren – dementsprechend wird dem Kompositum nur eine Bedeutung zugewiesen. Folgende Beispiele zeigen aber, dass eine korrekte Interpretation nicht immer so einfach ist:

(a) „Beileid": *Bei* + *Leid* oder aber auch *Beil* + *Eid*
(b) „Kulturgeschichte": *Kultur* + *Geschichte* oder aber auch *Kult* + *Urgeschichte*
(c) „Wachtraum": *Wach* + *Traum* oder aber auch *Wacht* + *Raum*

Während bei den ersten beiden Komposita Fehldeutungen mithilfe einer semantischen Analyse vermieden werden können, sind bei Wachtraum tatsächlich zwei semantische Lesarten möglich – ein Beispiel für echte Mehrdeutigkeit vor (vgl. Krenz/Ramlow 2008: 59).

5. 2 Lexikalische Mehrdeutigkeit

Während bei morphologischer Ambiguität ein Wort auf mehr als eine Weise analysiert werden konnte, beschreibt lexikalische Ambiguität Wörter, welche auf mehrere Arten interpretiert werden können. Lexikalische Mehrdeutigkeit umfasst mehrdeutige Kategorienzugehörigkeit, Homonymie, Polysemie und Mehrwortlexeme.

Mehrdeutige Kategorienzugehörigkeit bedeutet, wie der Name bereits vermuten lässt, dass ein Wort, abhängig vom Kontext, mehr als einer grammatikalischen Kategorie zugehörig sein kann. Als Beispiel für eine solche mehrdeutige Kategorienzugehörigkeit benutzen Hutchins / Somers (1992: 85) das englische Wort *round* – in den folgenden Sätzen findet es als Nomen, Verb, Adjektiv, Präposition, Partikel und Adverb Verwendung.

(1a) *Liverpool were defeated in the first round.*

(1b) *The boy started to round up the cattle.*

(1c) *I want to purchase a round table.*

(1d) *We are going on a cruise round the globe.*

(1e) *A glass of cold water soon brought him round.*

(1f) *The house measured forty feet round.*

Die Schwierigkeiten, die für ein maschinelles Analyseprogramm bei mehrdeutiger Kategorienzugehörigkeit auftreten, werden bei diesem Beispiel evident.

Ein Homograph oder Homogramm beschreibt ein Wort, welches mehrere, gänzlich verschiedene Bedeutungen hat, obwohl es gleich geschrieben wird. Das Wort *Arm* kann zum Beispiel ein Körperteil, oder aber auch einen Mangel an Reichtum beschreiben. Polysemie beschreibt ein ähnliches Phänomen – Polyseme umfassen Wörter, welche zwar mehrere Bedeutungen haben, die allerdings untereinander ähnlich sind. Eine strikte Grenzziehung zwischen diesen beiden Begriffen ist jedoch schwierig und im Rahmen der MÜ auch größtenteils irrelevant (Hutchins/Somers 1992: 86-87).

Mehrwortlexeme sind als komplexe Einheiten im Wörterbuch verzeichnet und bestehen aus mehreren Wörtern. Unter diesem Begriff fallen auch Kollokationen und idiomatische Wendungen (vgl. Krenz/Ramlow 2008: 62). Die vollständige Bedeutung von idiomatischen Wendungen, oder auch Redensarten, kann nicht allein durch das Verstehen der einzelnen Bestandteile eines Satzes erfasst werden. Genau hierin liegt die Schwierigkeit für MÜ-Systeme (vgl. Austermühl 2001: 172). Beispielsweise wäre eine korrekte Übersetzung aller Wörter in der Redewendung „*Das ist mir Wurst*.“ ins Englische zwar „*This is me sausage*.“ doch die tatsächliche Bedeutung der idiomatischen Wendung geht in diesem Falle verloren. Kollokationen wie „*in Betracht ziehen*“ oder „*in der Tat*“ und konventionalisierte idiomatische

Wendungen werden deshalb als feste Wendungen im Wörterbuch verzeichnet, um Problemen im Übersetzungsprozess vorzubeugen (vgl. Krenz/Ramlow 2008: 62). Alternativ besteht die Möglichkeit, Mehrwortlexeme grundsätzlich zu vermeiden, oder diese bei der Präedition zu ersetzen, um die Fehlerwahrscheinlichkeit zu verringern (vgl. Austermühl 2001: 172-173). Lexikalische Fehler sind die häufigsten und auffallendsten in einer Übersetzung (vgl. Bánik et al. 2019: 88).

5. 3 Syntaktische Mehrdeutigkeit

Während lexikalische Ambiguität versucht, mittels Analyse die Bedeutung einzelner Wörter zu präzisieren, beschäftigt sich syntaktische Ambiguität mit der Interpretation von syntaktischen Strukturen und Sätzen (vgl. Hutchins/Somers 1992: 88). Die syntaktische Analyse ist einer der wesentlichsten Bestandteile von maschinellen Analyseverfahren, da die syntaktischen Regeln einer Sprache im Vergleich zu semantischen oder ähnlichen Phänomenen begrenzter sind und somit zumindest theoretisch am einfachsten künstlich simulierbar sind.

Oftmals können Sätze, welche syntaktisch mehrdeutig sind, nur disambiguiert werden, wenn kontextuelle Informationen und/oder Weltwissen miteinbezogen wird (vgl. Krenz/Ramlow 2008: 64). Folgendes Beispiel wird bei Hutchins / Somers (1992: 88) angeführt:

(2) *Time flies like an arrow.*

Für menschliche LeserInnen ist es ein Leichtes, hier den Satz auf mehrere Weisen zu deuten, wie folgende Paraphrasierungen zeigen:

(2a) *The passage of time is as quick as an arrow.*

(2b) *A species of flies called ‘time flies’ enjoy an arrow.*

(2c) *Measure the speed of flies like you would an arrow.*

Das Erkennen der Ambiguität in diesem Satz mag auf dem ersten Blick nicht offensichtlich sein – vor allem nicht, wenn der Satz in einem bestimmten Kontext eingebettet ist. Da MÜ-Systeme allerdings nur begrenzt über kontextuelle Informationen verfügen, ist es für diese somit nicht immer einfach, die korrekte Interpretation auszuwählen.

Neben kontextuellen Informationen muss für manche Sätze auch Weltwissen vorhanden sein, um diese korrekt interpretieren zu können, wie Krenz und Ramlow (2008: 65) bei folgendem Beispiel beweisen:

(3) *Jagdverbot für Jungrobben empfohlen.*

Hier bestehen zwei Interpretationsmöglichkeiten:

(3a) *Jemand hat empfohlen, das Jagen von Jungrobben zu verbieten.*

(3b) *Jemand hat Jungrobben ein Jagdverbot empfohlen.*

Anders als bei (2) ist bei diesem Satz offensichtlich, dass bloß (3a) die richtige Interpretation sein kann, da Jungrobben selbstverständlich nicht in der Lage sind, zu jagen. Allerdings verfügen MÜ-Systeme über signifikant geringeres Wissen über die Welt als über Kontext und Semantik. Extralinguistische Probleme zu beschreiben ist sehr kompliziert, weil ihre Kenntnis nur schwer kodifiziert werden kann (vgl. Bánik et al. 2019: 32).

6 Aktueller Forschungsstand der MÜ für den Bereich Literaturübersetzen

In diesem Kapitel wird der aktuelle Forschungsstand der maschinellen Übersetzung für den Bereich der Literaturübersetzung analysiert. Es wird auf eine Auswahl der relevantesten Studien ab 2010 eingegangen. Unter anderem werden Studien zur Lyrikübersetzung und Studien, bei denen literarische Texte maschinell übersetzt und teilweise auch posteditiert wurden, untersucht.

6. 1 Forschung ab 2010

In der Forschung zur maschinellen Übersetzung wird der Fokus meist auf sogenannte Fachtexte (beispielsweise fachsprachliche Texte aus den Gebieten der Medizin, Technik, Rechtswissenschaften, etc.) gelegt. Somit gab es bislang nur eine sehr begrenzte Anzahl von Arbeiten, welche sich mit der Anwendung von MÜ bei literarischen Texten beschäftigten. Dies lässt sich damit erklären, dass bei nicht literarischen Texten das vorrangige Ziel des Übersetzungsprozesses lediglich die Übertragung der Bedeutung des Ausgangssatzes in die Zielsprache ist, ohne notwendigerweise stilistische Feinheiten der Ausgangssprache widerzuspiegeln. Derartige Übersetzungen sind, laut Toral und Way (2018: 264-265), für literarische Texte überhaupt nicht geeignet. Diese Meinung begründen die beiden damit, dass bei LeserInnen von literarischen Texten eine gänzlich andere Erwartungshaltung existiert; es reicht demzufolge nicht aus, wenn die Übersetzung bloß den Sinn des Ausgangstextes überträgt; es wird zusätzlich erwartet, dass unter anderem auch das Leseerlebnis des

Originaltextes erhalten bleibt. Somit gilt die Literaturübersetzung als die größte Herausforderung für den Bereich der maschinellen Übersetzung. Hier ist allerdings anzumerken, dass es sehr wohl Situationen gibt, in denen „wortwörtliche“ Übersetzungen für literarische Textausschnitte geeignet sind. Ferner ist Fachübersetzen nicht mit wörtlichem Übersetzen gleichzusetzen.

Das Interesse in der Computerlinguistik an der Verarbeitung von literarischen Texten wuchs in den letzten Jahren stetig an. Beispielsweise wurde 2012 der Workshop Computational Linguistics for Literature eingeführt, welcher seitdem jährlich stattfindet. Ein populärer Forschungsschwerpunkt ist laut Toral und Way (2018: 265) die automatische Identifizierung von Textausschnitten, wie zum Beispiel figurative Mittel wie Metaphern, Humor und Ironie. Allerdings gibt es dafür eine eher begrenzte Anzahl von Arbeiten zur Anwendung von MÜ bei literarischen Texten, die in Folge untersucht werden.

6. 2 Forschung zur Lyrikübersetzung

Forschungen zur maschinellen Lyrikübersetzung gibt es laut Meyer-Sickendiek et al. (2018: 1-2) seit 2010, eine Auswahl dieser sei hier kurz beschrieben.

Greene et al. (2010) verwendeten einen phrasenbasierten maschinellen Übersetzungsansatz, um italienische Gedichte in englische „Übersetzungsverbände“ zu übersetzen, wobei sie diese Verbände nach der besten Übersetzung durchsuchten, die einem vorgegebenen metrischen Muster gehorcht. In ihrer Arbeit schreiben Greene et al., dass Menschen Vorteile gegenüber Maschinen besitzen, wenn es um „kreatives Schreiben“ (Gedichte, Geschichten, Witze, etc.) geht. Greene et al. halten weiters fest, dass Menschen einerseits grammatikalisch korrekte Äußerungen konstruieren können und andererseits „Freude und Herzschmerz“ erleben. Während Maschinen seit 2010 zwar definitiv die Fähigkeit erlernt haben, grammatikalisch vollständige Sätze zu produzieren, ist es Maschinen jedoch (noch) nicht gelungen, Emotionen zu empfinden. Allerdings haben Maschinen laut Greene et al. auch ihre Vorzüge – Maschinen gelingt es beispielsweise mit Leichtigkeit konkrete Aufgaben zu erledigen, wie etwa Wörter zu finden, die sich auf andere reimen, mit dem Buchstaben „s“ beginnen und dreisilbig sind. Außerdem sind Maschinen in der Lage, umfangreiche Textkorpora zu analysieren und zu speichern.

Ghazvininejad et al. (2018) haben das erste neuronale Gedichtübersetzungssystem entwickelt. In ihrer Arbeit wurde ein französisches Quellgedicht in ein englisches Zielgedicht übersetzt. Menschliche Evaluierungen wurden an den generierten Gedichten durchgeführt und

laut eigenen Angaben wurde die Übersetzungsqualität in 78,2% der Fälle als akzeptabel eingestuft.

Die Arbeit von Meyer-Sickendiek et al. (2018) beschäftigt sich mit der Übersetzung von Gedichten in freien Versen. Anstatt also die metrischen Versfüße (Jambus, Trochäus usw.) und die Reimpaare zu identifizieren, machte diese Studie die ersten Schritte zur automatischen Übersetzung von Lyrik mit einem Fokus auf die für moderne und postmoderne Lyrik typischen rhythmischen Merkmale. Essenziell für diese Studie war das Verwenden von *lyrikline*, dem weltweit größten Korpus von Lyrikübersetzungen (vgl. Meyer-Sickendiek et al. (2018: 2).

6. 3 Weitere Studien

Voigt und Jurafsky (2012) untersuchten die Rolle der Referenzkohäsion in Übersetzungen und fanden heraus, dass literarische Texte dichtere Referenzketten aufweisen. Sie schlossen daraus, dass die Einbeziehung von Diskursmerkmalen über die Satzebene hinaus ein wichtiger Forschungsschwerpunkt für die Anwendung von MÜ bei literarischen Texte sein müsse.

Ähnlich wie bei der vorliegenden Arbeit, verwendete Besacier (2014) MÜ, gefolgt von einer Postedition (durch nicht-professionelle Übersetzer) um eine Kurzgeschichte vom Englischen ins Französische zu übersetzen. Ein solcher Arbeitsablauf wurde als nützliche und kostengünstige Alternative für die Übersetzung literarischer Werke angesehen, wenn auch auf Kosten der Übersetzungsqualität.

Toral und Way (2015) veröffentlichten 2015 eine vergleichbare Studie zu jener von Besacier (2014). Allerdings wurde bei Besacier (2014) „bloß“ eine Kurzgeschichte übersetzt, während Toral und Way einen Roman übersetzten. Außerdem wurden Besaciers Posteditionen von nicht-professionellen Übersetzern erstellt, während Toral und Ways MÜ-Ergebnisse mit Übersetzungen von professionellen ÜbersetzerInnen verglichen wurden.

Toral und Ways aktuelle Studie von 2018 basiert auf deren Forschungsarbeit von 2015. Ein wesentlicher Unterschied ist jedoch, dass bei der aktuellen Studie vorrangig ein neuronales, maschinelles Übersetzungssystem verwendet wurde, statt einem phrasenbasierten, wie bei der Arbeit von 2015. In der aktuellen Studie sind die Autoren außerdem der Meinung, dass das Aufkommen zweier neuer Technologien die Situation für MÜ in der Literaturübersetzung ändern könnte – die wachsende Beliebtheit von E-Books und das Aufkommen der NMÜ.

Durch das kontinuierliche Erscheinen von Büchern (Originalromanen sowie Übersetzungen) in digitalem Format ergibt sich nun eine schier unerschöpfliche Quelle an zweisprachigen Paralleltexten, welche die wichtigste Ressource für das Training von korpusbasierten MÜ-Systemen ist. Aufgrunddessen könnten nun also MÜ-Systeme entwickelt werden, welche genau auf die speziellen Anforderungen von Romanen zugeschnitten sind. Dies sollte also zu einer verbesserten Leistung von neuen MÜ-Systemen führen, da in der Vergangenheit bereits öfters bestätigt worden war, dass korpusbasierte Systeme nur dann optimal arbeiten, wenn sie mit Daten trainiert werden, die jenen ähnlich sind, auf welche die Systeme später auch angewendet werden (vgl. Toral/Way 2018: 264).

NMÜ ist laut Toral und Way (2018: 264) für literarische Texte aufgrund folgender zwei Erkenntnisse von besonderem Interesse: Erstens scheint die Leistung von NMÜ-Systemen besonders vielversprechend für lexikalisch reichhaltige Texte zu sein, was bei literarischen Texten oftmals der Fall ist. Zweitens wird behauptet, dass NMÜ in der Lage ist, statt einer wörtlichen Übersetzung das kulturelle Äquivalent in einer anderen Sprache finden.

Im empirischen Teil ihrer Studie wurden jeweils ein phrasenbasiertes MÜ-System und ein NMÜ-System mit einer großen Menge an literarischen Texten mit insgesamt mehr als 100 Millionen Wörtern trainiert. Im Anschluss daran wurden Übersetzungen von zwölf Romanen vom Englischen ins Katalanische produziert. Laut der automatischen Bewertungsmetrik BLEU schnitt das NMÜ-System bei allen Romanen signifikant besser ab als das phrasenbasierte System. Ferner fand auch eine menschliche Evaluation statt, bei der Textausschnitte der Übersetzungen der beiden MÜ-Systeme und einer menschlichen Übersetzung aus drei ausgewählten Büchern bewertet wurden. Auch bei der Humanbewertung schnitt das NMÜ-System besser ab als das phrasenbasierte. Interessanterweise waren die Übersetzungen des NMÜ-Systems, laut der menschlichen Evaluation, in 17% - 34% der Fälle (abhängig vom Buch und von der ausgewählten Textpassage) von äquivalenter Qualität wie die Übersetzungen, welche von professionellen ÜbersetzerInnen gefertigt wurden (vgl. Toral/Way 2018).

Folglich wäre die Hypothese der vorliegenden Arbeit zumindest teilweise widerlegt, da bereits die Übersetzung eines NMÜ-Systems in ihrer Studie in 17-34% der Fälle so zufriedenstellend war wie die von professionellen ÜbersetzerInnen, ohne posteditiert worden zu sein und da in der vorliegenden Arbeit davon ausgegangen wird, dass die Zieltexte, welche von einem MÜ-System produziert werden, nicht dieselbe Qualität aufweisen, wie die von HumanübersetzerInnen. Weitere Studien zum Thema Postedition werden in Kapitel 8 untersucht.

Matusov (2019) adaptierte in seiner Studie ein NMÜ-System an literarische Texte und übersetzte Geschichten vom Deutschen ins Englische und vom Englischen ins Russische. Die Fehleranalyse und Bewertung der Übersetzungen übernahm ein/e bilinguale/r EvaluatorIn.

Zweck dieser Studie war es, zu beweisen, dass maschinelle Übersetzung sehr wohl auch für literarische Texte geeignet ist. Die Ergebnisse zeigten, dass bis zu 30 % der untersuchten Segmente, meist kurze Sätze, als akzeptabel eingestuft wurden und nur ein Korrekturlesen durch einen einsprachigen Korrekturleser der Zielsprache erfordern würden. Längere Sätze wurden von der Syntax her gut übersetzt, so dass lediglich punktuelle und minimale Posteditionen notwendig wären. Die Übersetzung vom Deutschen ins Englische erzielte bessere Ergebnisse als jene vom Englischen ins Russische; die Qualität war oft hoch genug, um die Geschichte zu verstehen und sogar „zu genießen". Matusov (2019) schließt daraus, dass MÜ von Literatur nicht nur für die Unterstützung professioneller LiteraturübersetzerInnen in einem Postediting-Szenario nützlich sein könnte, sondern auch, um weitgehend unentdeckte fremdsprachige Bücher sofort online für LeserInnen weltweit verfügbar zu machen, beispielsweise durch eine Übersetzung ins Englische.

Guerberof-Arenas und Torals Studie (2020) rückte die Messgröße Kreativität in den Mittelpunkt. Eine Geschichte wurde aus dem Englischen ins Katalanische in drei Modalitäten übersetzt. Die Übersetzung erfolgte in drei Varianten: Die Geschichte wurde nur maschinell übersetzt oder zunächst maschinell übersetzt und anschließend posteditiert oder der Text wurde ohne jegliche Hilfsmittel humanübersetzt. Eine Gruppe von 88 katalanischen TeilnehmerInnen las die Geschichten, ohne zu wissen, welche Übersetzungsvariante ihnen vorlag und füllte im Anschluss daran einen Fragebogen aus. Die Forschungsfrage der Studie lautete: „Wie wirkt sich die Kreativität in verschiedenen Übersetzungsmodalitäten auf das Leseerlebnis aus?" In der Studie erzielte die reine Humanübersetzung den „höchsten Kreativitätswert" und wurde auch am häufigsten als die beste Übersetzung eingeschätzt. Die Autoren der Studie nahmen an, dass ÜbersetzerInnen eine kreativere/originellere Übersetzung produzieren, wenn sie nicht durch eine MÜ-Ausgabe eingeschränkt sind. Allerdings geht aus den Ergebnissen der Studie auch hervor, dass die LeserInnen die Postedition marginal lieber oder zumindest genauso gern lasen wie die Humanübersetzungen. Weiters wurde ein „klarer Zusammenhang" zwischen der Übersetzungsrezeption und dem Interesse und Vergnügen der LeserInnen an einer Geschichte festgestellt. Guerberof-Arenas und Toral gehen aufgrund ihrer Forschungsergebnisse davon aus, dass die Kreativität am höchsten ist, wenn professionelle Übersetzer in den Übersetzungsprozess eingreifen, insbesondere wenn ohne Hilfsmittel gearbeitet wird und, dass Kreativität beim Übersetzen der Faktor sein könnte, der das Lesevergnügen und die Rezeption übersetzter literarischer Texte steigert.

Der Forschungsüberblick in diesem Unterkapitel besteht aus den wichtigsten Vorarbeiten für die vorliegende Arbeit – ausgewählt aufgrund Ähnlichkeiten beziehungsweise Parallelen zur Studie, die in dieser Arbeit durchgeführt wird. Einige dieser Gemeinsamkeiten werden in Kapitel 10. 4. 4 hervorgehoben und in Verbindung mit den Ergebnissen der vorliegenden Arbeit gebracht.

7 Translation Quality Assessment

Dieses Kapitel beschäftigt sich mit der Bewertung von MÜ-Systemen und ihrem Output, den Übersetzungen. Zunächst wird erklärt, was der Begriff Translation Quality Assessment umfasst und wieso objektive Evaluierungen von MÜ-Systemen so problematisch sind. Die verschiedenen Evaluierungsansätze – menschliche Evaluierung und automatisierte Evaluierung – und die wichtigsten Messgrößen werden diskutiert. Da im empirischen Teil der vorliegenden Arbeit ebenfalls eine menschliche Evaluierung stattfindet, werden in diesem Kapitel unter anderem jene Kriterien hervorgehoben, welche im späteren Verlauf essentiell für den Evaluierungsprozess sein werden. Zuletzt wird auch auf Fehlerklassifizierung eingegangen.

7. 1 Begriffserklärung TQA – Translation Quality Assessment

Das Hauptziel von Translation Quality Assessment (TQA) besteht darin, sicherzustellen, dass ein bestimmtes Qualitätsniveau von übersetzten Texten erreicht, aufrechterhalten und KundInnen, BenutzerInnen, LeserInnen etc. zur Verfügung gestellt wird. Die Beurteilung der Übersetzungsqualität ist auch entscheidend für die Überprüfung von ÜbersetzerInnen in Ausbildung, für die berufliche Zertifizierung und Akkreditierung sowie für die Rekrutierung von ÜbersetzerInnen. Außerdem kann TQA zur Messung der Leistung von Übersetzungstechnologien - insbesondere der maschinellen Übersetzung - verwendet werden, um die Produktivität von ÜbersetzerInnen in technologisierten Arbeitsabläufen zu bewerten und andere verwandte Aspekte zu untersuchen, wie zum Beispiel den Postediting-Aufwand und die Qualitätseinschätzung (vgl. Doherty 2017: 1; Castilho et al. 2018: 10).

Translation Quality Management ist ein viel diskutiertes Thema, was in Anbetracht der großen Bedeutung, die der Bewertung der Übersetzungsqualität in der Übersetzungswissenschaft und insbesondere in der Übersetzungstechnologie sowie in der gesamten Übersetzungs- und Lokalisierungsindustrie zukommt, wenig überraschend ist. Aufgrund von Ressourcenbeschränkungen fallen TQA-Prozesse sehr unterschiedlich aus und sind oftmals mit Einschränkungen verbunden. Gleichzeitig haben die Entwicklung und der weit verbreitete Einsatz von MÜ-Technologie zu einer Fülle unterschiedlich operationalisierter Definitionen von Übersetzungsqualität geführt (vgl. Castilho et al. 2018: 10).

Doch wie wird die Qualität eines Textes beurteilt? ÜbersetzerInnen wissen, mit großer Wahrscheinlichkeit besser als viele andere, dass ein Dokument von zwei verschiedenen Personen niemals gleich übersetzt werden wird. Nimmt man nun eine beliebige Anzahl an Übersetzungen des selben Textes, wird man nicht nur feststellen, dass alle Übersetzungen anders sind, sondern auch, dass man einige davon als gute, andere als schlechte und einige sogar

als sehr gute Texte erachten kann (vgl. White 2003: 213). Doch mittels welcher Kriterien wird entschieden, ob ein Text gut übersetzt worden ist oder nicht? Viel hängt davon ab, unter welchen Bedingungen eine Übersetzung angefertigt worden ist und für welche/n RezipientIn (vgl. Hutchins/Somers 1992: 2). Doch selbst wenn von diesen zahlreichen „guten" Übersetzungen genau eine ausgewählt wird und als „korrekte" beziehungsweise „richtige" Übersetzung klassifiziert wird, würde eine solche Feststellung konsequenterweise bedeuten, dass die vielen anderen guten Übersetzungen „falsch" sind, oder zumindest weniger richtig. Dementsprechend lässt sich also feststellen, dass mehr als eine Art und Weise existiert, um einen Text zu übersetzen und es Menschen schwerfällt, eine nicht von Subjektivität beeinträchtige Entscheidung zu treffen. Während dies zwar ein Zeugnis für die üppige Variabilität und bemerkenswerten Fantasie der menschlichen Sprache ist, welche Teil eines jeden Übersetzungsprozesses sind, macht es die Aufgabe, MÜ-Systeme zu evaluieren nicht einfacher (vgl. White 2003: 213).

7. 2 Evaluierungsansätze für die Bewertung von MÜ

Es gibt zwei Hauptansätze für die Bewertung von MÜ: Der erste stützt sich auf die menschliche Auswertung eines maschinell erzeugten Textes und der zweite versucht, diese Evaluation zu automatisieren. Zunächst werden im folgenden Kapitel die bedeutendsten Messgrößen diskutiert, die für die menschliche Evaluierung wichtig sind. Anschließend werden die wesentlichen Unterschiede zwischen der automatisierten und der menschlichen Bewertung erläutert.

7. 2. 1 Menschliche Evaluierung

Menschen analysieren eine Reihe von Kriterien, mithilfe welcher die Qualität des Outputs von MÜ-Systemen beurteilt wird. Der manuellen Bewertung wird des Öfteren vorgeworfen, subjektiv zu sein. Weitere Kritikpunkte sind die Kosten und die Dauer der Ausführung. Evaluierungen, welche von Menschen durchgeführt werden, haben aber den Vorteil, komplexe sprachliche Phänomene beurteilen zu können, welche über die bloße Angemessenheit und Sprachflüssigkeit einer Übersetzung hinaus gehen. Ferner kann der Fokus auf eine Fehlertypologie gelegt werden – dies setzt allerdings voraus, dass die EvaluatorInnen über entsprechende übersetzerische und sprachwissenschaftliche Kenntnisse verfügen. Manuelle Ansätze unterscheiden sich auch in Bezug auf die Fähigkeiten der EvaluatorInnen: Während zum Beispiel Grammatik und Sprachgewandtheit allein anhand des Zieltextes beurteilt werden können (weshalb zumindest theoretisch eine umfassende Vertrautheit mit der Zielsprache

erforderlich ist), setzt die Beurteilung der Genauigkeit und Angemessenheit zweisprachige Kompetenzen voraus. All diese Faktoren können sich auf die Kosten, die Länge und die Gesamtkomplexität der menschlichen MÜ-Evaluierung auswirken (vgl. Castilho et al. 2018: 25).

TQA wird am häufigsten unter den Gesichtspunkten der Angemessenheit und des Sprachflusses durchgeführt, obwohl auch manchmal andere Messgrößen wie Lesbarkeit, Verständlichkeit und Akzeptabilität untersucht werden (vgl. Castilho et al. 2018: 17). Die in diesem Unterkapitel erwähnten Messgrößen Grammatik, Genauigkeit, Angemessenheit und Flüssigkeit sind alles Kriterien, anhand derer die Zieltexte der TeilnehmerInnen des empirischen Teiles der vorliegenden Arbeit im späteren Verlauf evaluiert werden. Die Evaluatorin besitzt exzellente Kenntnisse in den Sprachen Deutsch und Englisch (im Zuge der Studie wird vom Englischen ins Deutsche übersetzt). Somit ist sie fähig, die Genauigkeit und Angemessenheit der Übersetzungen zu beurteilen.

7. 2. 1. 1 Angemessenheit und Flüssigkeit

Angemessenheit (beziehungsweise „Genauigkeit“ oder „Treue) ist ein bewährtes TQA-Maß für den MÜ-Output und wird typischerweise als das Ausmaß definiert, in dem der MÜ-Output die Bedeutung des quellsprachlichen Inputs erfasst (vgl. Doherty 2017: 133). Eine Übersetzung auf ihre Flüssigkeit (beziehungsweise „Sprachfluss“) zu analysieren ist eine der traditionellsten Methoden, um die Qualität zu beurteilen. Die Flüssigkeit eines übersetzten Satzes wird oftmals anhand der grammatikalischen Fehler, falsch übersetzten Wörter und nicht übersetzten Wörter ermittelt (vgl. Arnold et al. 1994: 169-170). In einem solchen Fall bewerten menschliche EvaluatorInnen den Output typischerweise Segment für Segment (d.h. ohne Kontext) und auf Satzebene, wobei das Segment des Ausgangstextes daneben präsentiert wird. Bei beiden Kriterien wird häufig eine Ordinalskala verwendet, wobei menschliche EvaluatorInnen beispielsweise angeben können, dass die maschinell generierte Übersetzung "die gesamte Bedeutung", "den größten Teil der Bedeutung" ... "keine der Bedeutungen" usw. des Ausgangssegments erfasst hat (vgl. Doherty 2017: 133).

Wie bereits erwähnt, werden die Kriterien Angemessenheit und Flüssigkeit im späteren Verlauf noch von größerer Bedeutung sein. Im Evaluierungsbogen wurden die Kriterien Genauigkeit und Angemessenheit in einer einzigen Frage zusammengefasst:

„*Wiedergabe der Bedeutung: Genauigkeit bzw. Angemessenheit*

Wie gut bzw. genau wurden die Inhalte, wurde die Bedeutung des Textes wiedergegeben? Wurden Bedeutungen verändert, verzerrt?“

Auch bei der vorliegenden Arbeit wurde Gebrauch von einer Ordinalskala gemacht – die Evaluatorin konnte zwischen den Antwortmöglichkeiten *Sehr zufriedenstellend; Zufriedenstellend; Durchschnittlich; Eher unzufriedenstellend* und *Unzufriedenstellend* entscheiden und außerdem zusätzliche Anmerkungen hinzufügen. Diese Ordinalskala war für alle Kriterien gleich – Anmerkungen waren ebenfalls bei allen Messgrößen möglich. Das Kriterium der Flüssigkeit wurde in einer separaten Frage evaluiert.

7. 2. 1. 2 Lesbarkeit und Verständlichkeit

Im Allgemeinen bezieht sich Lesbarkeit auf die Schwierigkeit, mit der ein bestimmter Text von unterschiedlichen RezipientInnen gelesen werden kann. Viele verschiedene Messgrößen der Lesbarkeit existieren, und sie basieren meist auf linguistischen Merkmalen wie Worthäufigkeit und Satzlänge, zusätzlich zu außersprachlichen Merkmalen wie Formatierung und Abstände. Die Verständlichkeit gibt an, wie verständlich der Zieltext für die/den LeserIn ist. Während die Lesbarkeit vom Text selbst abhängt, kann die Verständlichkeit je nach dem spezifischen Leser/der spezifischen Leserin des Textes (z.B. dem Grad seiner Ausbildung und Vertrautheit mit dem Lesen bestimmter Textarten) variieren (vgl. Castilho et al. 2018: 19).

Die Kriterien Lesbarkeit und Verständlichkeit werden im Zuge der Studie nicht von weiterer Bedeutung sein, da die produzierten Zieltexte nicht anhand dieser Messgrößen evaluiert wurden.

7. 2. 1. 3 Akzeptabilität

Der Begriff "Akzeptabilität" (oder „Akzeptanz“) wird in verschiedenen Bereichen wie Linguistik, Übersetzung und HCI (Human-Computer Interaction) verwendet, weswegen dem Begriff unterschiedliche Definitionen zugeordnet wurden. Im Kontext der TQA beschreibt Akzeptabilität, zu welchem Grad der Ziel- oder Ausgabetext den Bedürfnissen und Erwartungen der LeserInnen entspricht (vgl. Castilho et al. 2018: 20). Laut Roturier (2006: 4) bezieht sich Akzeptanz im Rahmen der MÜ außerdem nicht nur auf die Relevanz, die ein Text für seine EmpfängerInnen hat, sondern auch auf die Art und Weise, in der seine textlichen Eigenschaften von EmpfängerInnen akzeptiert, toleriert oder abgelehnt werden. Im Bereich der HCI hingegen empfinden BenutzerInnen eine Übersetzung als akzeptabel, wenn sie mit dieser in der Lage sind, Aufgaben zu erfüllen, ungeachtet der darin enthaltenen Mängel. Ein Text wird als weniger akzeptabel empfunden, wenn Mängel in der Übersetzung die Fähigkeit, den Text zu verwenden, beeinträchtigen (vgl. Castilho 2016: 58).

Auch auf eine Evaluierung des Aspektes der Akzeptabilität wird im empirischen Teil der vorliegenden Arbeit verzichtet. Genau wie bei den Messgrößen der Lesbarkeit und Verständlichkeit erklärt sich dieser Verzicht dadurch, dass dieses Kriterium für die Studie der vorliegenden Arbeit zu allgemein ist – da die Evaluation der Studie auf die besonderen Merkmale von literarischen Texten maßgeschneidert wurde, wurden differenziertere Kriterien entworfen, was die Verwendung der Messgrößen Akzeptabilität, Lesbarkeit und Verständlichkeit obsolet macht.

7. 2. 2 Automatisierte Evaluierung

Die Evaluierung von MÜ durch den Menschen ist ressourcenintensiv und kann beträchtlich länger dauern als die Evaluierung von Humanübersetzungen, da die Qualität des MÜ-Outputs in größerem Maße variieren kann. Der Aspekt der Subjektivität ist ein weiteres, beträchtliches Problem bei der manuellen Bewertung. Was zusätzlich zur steigenden Beliebtheit von automatisierter Bewertung beiträgt, ist die Notwendigkeit, häufige, ressourcensparende repetitive Evaluierungen im Bereich der MÜ-Forschung durchführen zu müssen (vgl. Doherty 2017: 134).

Die automatische Evaluierung beruht meist auf automatischen Evaluierungsmetriken (AEMs), bei denen es sich um Skripte oder Softwareprogramme handelt, die TQA unter Verwendung eines expliziten und formalisierten Katalogs linguistischer Kriterien zur Evaluierung des MÜ-Outputs durchführen (vgl. Doherty 2017: 134). Beliebte AEMs sind unter anderem “GTM” (General Text Matcher), “METEOR” (Metric for Evaluation of Translation with Explicit ORdering) und “TER” (Translation Edit Rate”); laut Castilho et al. (2018: 25-26) ist jedoch BLEU (Bilingual Evaluation Understudy) der De-Facto-Standard für die meisten Forschungszwecke. Automatische MÜ-Evaluierung ist derzeit ein florierendes Forschungsgebiet, mit zahlreichen WissenschaftlerInnen, die an der Bewertung und Verbesserung von automatischen Metriken arbeiten und neue Metriken vorschlagen. Dies hat zur Folge, dass neuere Metriken in einigen Aspekten sogar noch bessere Ergebnisse liefern als BLEU; ein Indiz dafür, dass die Suche nach den besten Ansätzen und automatischen Metriken zur Bewertung der MÜ-Qualität noch lange nicht beendet ist.

Eines der Hauptargumente für die Verwendung automatischer Metriken ist die minimale menschliche Arbeit, die für diese erforderlich ist, und die vermeintliche Objektivität, die mit der geringen menschlichen Intervention einhergeht. Im Gegensatz zu menschlichen Evaluierungen wird hier die Notwendigkeit von (zweisprachigen) EvaluatorInnen zur Bewertung der Übersetzung umgangen, was die Beurteilung in der Regel wesentlich kostengünstiger macht. Mit dieser Einsparung steigt allerdings auch das Risiko, dass die

Genauigkeit darunter leidet. Ferner darf nicht darauf vergessen werden, dass ÜbersetzerInnen benötigt werden, um die im Prozess verwendete(n) Referenzübersetzung(en) zu erstellen; dies führt an sich schon ein subtiles Element der Subjektivität und Variabilität ein (vgl. Castilho et al. 2018: 26). Ein weiteres Problem ist laut Way (2018: 166-167) außerdem: „MT developers are forced to (wrongly) assume human translations to be perfect when conducting automatic MT evaluation". Doch selbst wenn KorrekturleserInnen im vollständig verwalteten Übersetzungs-, Lektorats- und Korrekturlesezyklus vorhanden sind, treten gelegentlich Fehler auf – und manchmal sind diese fehlerhaften menschlichen Übersetzungen genau jene, mit denen der Output von MÜ-Systemen verglichen wird.

7. 3 Fehlerklassifizierung

Während all diese Evaluierungen und Metriken sehr wertvolle Informationen darstellen und bei der kontinuierlichen Verbesserung von MÜ-Systemen helfen, sind MÜ-ForscherInnen laut Popović (2018: 130-131) von weiteren Informationen abhängig, um bestimmte Fragen beantworten zu können wie etwa: Was sind die größten Probleme in einem Übersetzungssystem? Was sind die besonderen Stärken und Schwächen des Systems? Verbessert eine bestimmte Modifikation einen Aspekt des Systems, obwohl sie die Gesamtbewertung möglicherweise nicht verbessert? Übertrifft ein schlechter bewertetes MÜ-System ein besser bewertetes in einem bestimmten Gesichtspunkt? Eine Korrelation zwischen diesen Fragen und den Gesamtqualitätswerten einer Evaluierung zu finden ist kein leichtes Unterfangen. Aus diesem Grund sind Fehlerklassifizierungs- und Analysetechniken entstanden, die Fehler in einem übersetzten Text identifizieren und klassifizieren, um eine bessere Grundlage für Entscheidungen und Beurteilungen zu schaffen. Ziel der Fehleranalyse ist meist, ein Fehlerprofil für eine Übersetzungsausgabe zu erhalten, sprich die Verteilung der Fehler über die definierten Fehlerklassen wird ermittelt. Eine weitere Anwendung ist der Vergleich verschiedener Übersetzungsergebnisse, also die Ermittlung von Fehlerverteilungen über verschiedene Übersetzungen für jede Fehlerklasse. Darüber hinaus können auch spezifischere Analysen durchgeführt werden, wie beispielsweise Beziehungen zwischen bestimmten Fehlertypen und BenutzerInnen-/PosteditorInnen-Präferenzen und Auswirkungen verschiedener Fehlertypen auf verschiedene Aspekte des Nachbearbeitungsaufwands.

Ähnlich wie die Gesamtevaluation ist auch die Fehlerklassifizierung keine einfache Aufgabe. Sie kann manuell, automatisch oder mit einer kombinierten Methode (halbautomatisch) durchgeführt werden. Dabei können verschiedene Informationsquellen neben der analysierten Übersetzung verwendet werden, wie zum Beispiel quellsprachliche Texte, Referenzübersetzungen oder auch posteditierte Übersetzungen. Allein die Definition einer geeigneten Menge von Fehlerklassen (einer Fehlertypologie oder Taxonomie) ist bereits eine anspruchsvolle Aufgabe, wie im Anschluss gezeigt wird (vgl. Popović 2018: 131).

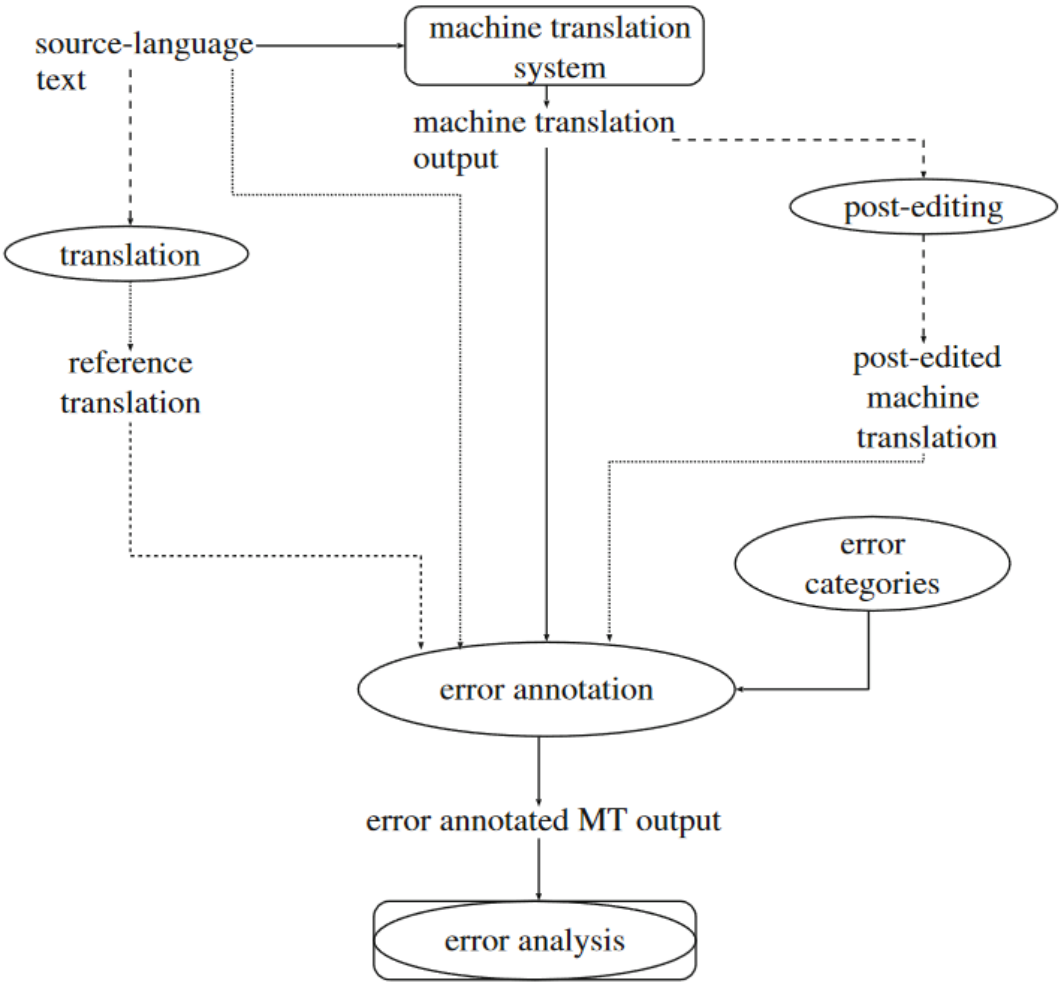

Abbildung 4: Allgemeiner Ablauf der manuellen Fehlerkennzeichnung; das Rechteck kennzeichnet den automatischen Prozess, die Ellipse den manuellen Prozess. Abbildung von Popović (2018: 132)

Der allgemeine Prozess der manuellen Fehlerklassifikation ist in Abbildung 4 dargestellt. Für jede Aufgabe und Herangehensweise sollte im Vorfeld ein Satz von Fehlerkategorien klar definiert werden. Dies ist aus mehreren Gründen eine anspruchsvolle Aufgabe: Die Fehler sollten alle Vor- und Nachteile des MÜ-Systems widerspiegeln, die sowohl für die jeweilige Aufgabe als auch für die beteiligten Sprachen wichtig sind und die Fehlertypen sollten sowohl linguistische als auch übersetzerische Aspekte abdecken. Trotz einer Reihe von Arbeiten in dieser Richtung, gibt es noch keine allgemeinen Regeln für die Definition von Fehlerkategorien (vgl. Popović 2018: 132) Im folgenden Unterkapitel wird ein

kurzer Überblick über Fehlertypologien für die manuelle Klassifizierung gegeben, die in den vergangenen Jahren für verschiedene Anwendungsbereiche verwendet wurden.

7. 3. 1 Fehlertypologien für die manuelle Klassifizierung

Federico et al. (2014) verwenden eine Typologie mit einer Reihe von grundlegenden Fehlerklassen für die Analyse von MÜ aus dem Englischen ins Arabische, Chinesische und Russische. Die Fehlerklassen waren: *morphologische Fehler; lexikalische Fehler, Hinzufügungen; Auslassungen; Fall- und Zeichensetzungsfehler; Satzstellungsfehler* und *zu viele Fehler*. Für jedes Segment markierten BearbeiterInnen fehlerhafte Wörter und vergaben eine Gesamtqualitätsbewertung. Diese Anmerkungen wurden verwendet, um den Einfluss einzelner Fehlertypen und deren Kombinationen auf die Gesamtqualitätsbewertung mit Hilfe von Mixed-Effects-Modellen zu untersuchen. Die größte Korrelation wurde für lexikalische Fehler und fehlende Wörter beobachtet. Ein sehr interessantes Ergebnis ist, dass die menschliche Qualitätswahrnehmung nicht unbedingt von der Häufigkeit eines bestimmten Fehlertyps abhängt; ein Satz mit einer niedrigen Gesamtbewertung kann durchaus weniger fehlende Wörter und/oder lexikalische Fehler enthalten als ein anderer Satz mit einer höheren Bewertung.

Lommel et al (2014a) haben mit MQM (Multidimensional Quality Framework) ein flexibles System entwickelt, bei welchem der Schwerpunkt auf der analytischen Qualität liegt. Somit konzentriert sich die Qualitätsbewertung auf die Identifizierung bestimmter Probleme/Fehler im übersetzten Text und auf deren Kategorisierung. MQM definiert über 80 Problem-/Fehlertypen welche streng hierarchisch in einer baumähnlichen Struktur (siehe Abbildung 5) angeordnet sind. Einige der Oberkategorien sind *Angemessenheit, Flüssigkeit, Stil* und *Terminologie*. Jeder dieser Oberkategorien sind Unterkategorien untergeordnet – beispielsweise gehören Auslassungen oder Hinzufügungen zur Kategorie *Angemessenheit* und Satzstellung und Kohärenz zu *Flüssigkeit*. Die bewertenden Personen (im Originaltext heißen sie „Annotators“) wurden angehalten, die Kategorie für Fehler so spezifisch wie möglich zu wählen und Kategorien auf höherer Ebene nur dann zu verwenden, wenn es nicht möglich war, eine Kategorie tiefer in der Hierarchie auszuwählen. Wenn ein Problem beispielsweise als "Wortfolge" kategorisiert werden konnte, konnte es auch als "Grammatik" kategorisiert werden, aber die bewertenden Personen wurden angewiesen, "Wortfolge" zu verwenden, da diese Kategorie spezifischer und somit aussagekräftiger war. Der Hauptzweck dieses Systems ist laut Popović (2018: 136) eine große Menge an hierarchischen Fehlerkategorien zu haben, welche die Auswahl jener beliebigen Teilmenge ermöglicht, die für die jeweilige Aufgabe geeignet ist. Diese Metrik wird bereits für die Evaluierung und den Vergleich von MÜ-Systemen verwendet. Lommel et al. (2014b) verwendeten MQM zum Vergleich von

regelbasierten und phrasenbasierten Systemen für mehrere Sprachpaare und Domänen; Klubicka et al. (2017) verwendeten MQM zum Vergleich von phrasenbasierten und neuronalen Systemen für die Sprachkombination Englisch-Kroatisch. Lommel et al (2014b) behaupten, die mittels MQM durchgeführte Qualitätsanalyse könne eine wertvolle Ergänzung zu automatischen Methoden wie BLEU und METEOR sein. Eine solche Analyse sei zwar arbeitsintensiv in der Durchführung, biete aber Einblick in die Ursachen von Fehlern und schlage mögliche Lösungen vor.

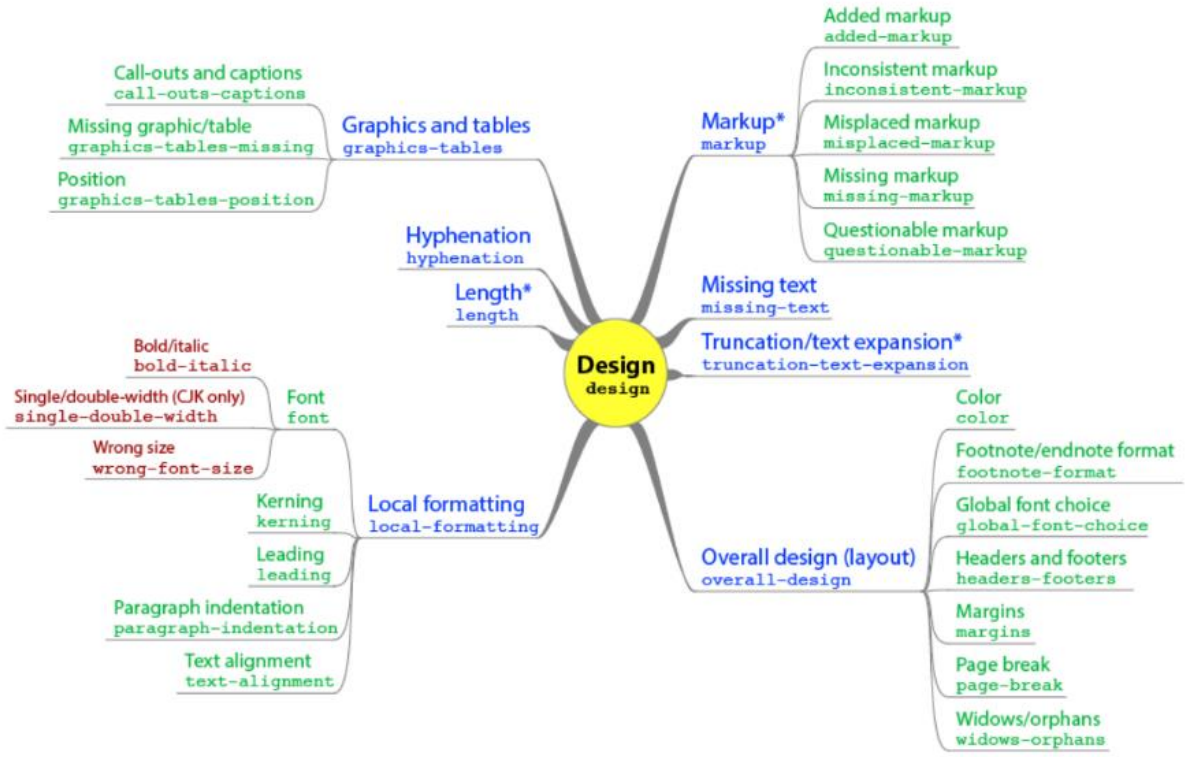

Abbildung 5: Hierarchische Baumstruktur des MQM-System. Abbildung von Lommel et al. (2015).

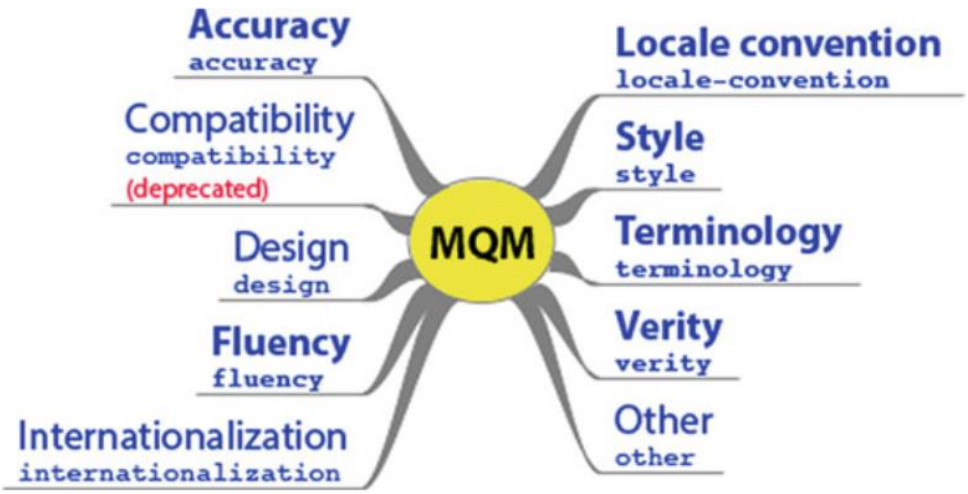

Abbildung 6: Darstellung der Oberkategorien im MQM-System. Abbildung von Lommel et al. (2015).

7. 3. 2 Herausforderungen der manuellen Fehlerklassifizierung

Der vorangegangene Abschnitt hat gezeigt, dass die Fehlerklassifikation einen großen Anwendungsbereich hat und eine Reihe von Fragen beantworten kann, die für die Verbesserung und Entwicklung von MÜ-Systemen sowie für ein besseres Verständnis der menschlichen Bewertungskriterien und des Postediting wichtig sind. Allerdings wurde auch offenkundig, dass diese Prozesse zeit- und ressourcenintensiv und voller herausfordernder Teilaufgaben sind. Einige dieser besonderen Herausforderungen sind laut Popović (2018: 139) unter anderem das Profil der BewerterInnen, Konsistenz/Folgerichtigkeit, Anzahl der Fehlerklassen und die genaue Definition der Fehlerklassen.

Müssen/sollen die BewerterInnen bloß die Zielsprache beherrschen oder sollen sie fließend in beiden Sprachen sein? Unabhängig vom Ausbildungsstand der BewerterInnen sind genaue Richtlinien und ein intensives Training notwendig, um eine ausreichende Übereinstimmung zwischen den einzelnen BewerterInnen zu erreichen und zuverlässige Ergebnisse zu erhalten. Bereits geringe Unterschiede in den Qualifikationen, Fertigkeiten und Präferenzen zwischen den einzelnen BewerterInnen können zu Unstimmigkeiten in der Konsistenz führen. Allerdings lassen sich auch in optimalen Szenarien bestimmte Inkonsistenzen nicht vollständig vermeiden (vgl. Popović 2018: 139).

Detailliertere Fehlertypologien liefern in der Regel bessere Informationen über die Fehler und Probleme, wie dies zum Beispiel beim vertiefenden Hierarchiesystem von MQM der Fall ist. Mehr Fehlerkategorien bedeuten jedoch einen höheren kognitiven Aufwand für die BewerterInnen und führen außerdem zu einer geringeren Konsistenz. Allerdings ist nicht nur die Anzahl, sondern auch die genaue Definition der Fehlerkategorien sehr wichtig. Zum Beispiel kann selbst eine einfache Unterscheidung zwischen Adäquatheit und Flüssigkeit aufgrund bestimmter Arten von grammatikalischen Fehlern schwierig sein (vgl. Popović 2018: 139-140).

7. 3. 3 Automatische Fehlerklassifizierung

Eine naheliegende Methode, den Aufwand manueller Fehlerklassifikationen zu reduzieren, ist die Automatisierung der Prozesse. Die Vor- und Nachteile sind ähnlich wie bei automatischen Bewertungsmetriken – die automatische Variante ist im Normalfall schneller, kostengünstiger und konsistenter, aber auch ungenauer, anfälliger für die Vergabe falscher „errortags“ und stark abhängig von Referenzübersetzungen.

Bereits 2006 schlugen Popović et al. ein System vor, in dem Satzstellungs- und Flexionsfehler automatisch geschätzt wurden. Dieses System arbeitete mit sogenannten WER (Word Error Rate) - und PER (Position-independent Word Error Rate)-Raten. WER setzt genau die gleiche Reihenfolge der Wörter in Hypothesen- und Referenzsegmenten voraus. PER hingegen vernachlässigt die Wortreihenfolge vollständig und misst nur den Unterschied in der Anzahl der Wörter, die in Hypothesen- und Referenzsegmenten vorkommen. Für beide Metriken wird die resultierende Anzahl der Fehler durch die Anzahl der Wörter in der Referenz geteilt. Dadurch sollen sich die Satzstellungsfehler in der Differenz zwischen WER und PER widerspiegeln und die Flexionsfehler mit der Differenz zwischen PER der Originalwörter und PER der Lemmata korrelieren (vgl. Popović 2018: 142).

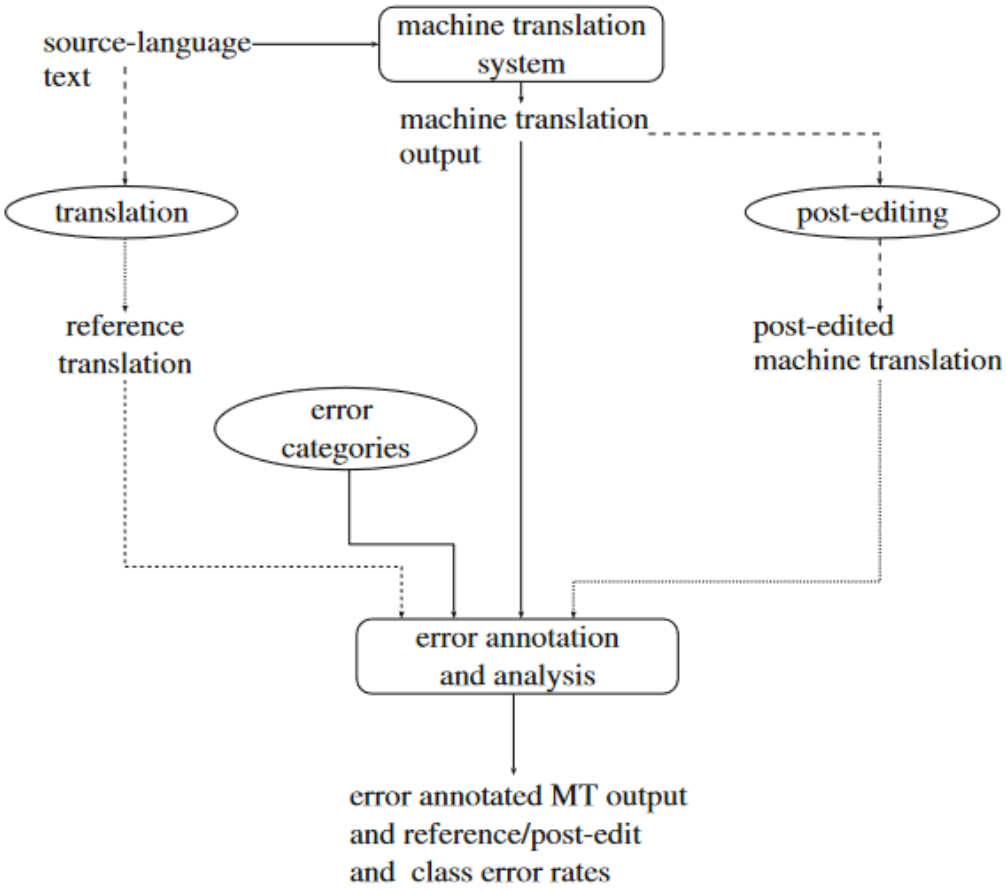

Abbildung 7: Ablauf der automatischen Fehlerklassifizierung; Rechtecke kennzeichnen den automatischen Prozess, die Ellipse den manuellen Prozess. Abbildung von Popović (2018: 143).

Ein 2010 entwickeltes, automatisches Fehlerklassifikationstool ist *Addicter* (Automatic Detection and Display of Common Translation Errors). Dieses System arbeitet mit einer Reihe von verschiedenen Fehlerklassen: Flexionsfehler, Satzstellungsfehler, Auslassungen, Hinzufügungen und Fehlübersetzungen. Zusätzlich erkennt Addicter unübersetzbare Wörter. Lemmata werden benötigt, um Flexionsfehler von lexikalischen Fehlern zu unterscheiden. Das Tool wurde durch einen Vergleich der automatisch generierten Ergebnisse mit denen einer

manuellen Fehlerklassifikation getestet und wies ausreichend hohe Korrelationen auf, um für eine Reihe von Aufgaben menschliche BewerterInnen ersetzen zu können (vgl. Popović 2018: 143).

Die Evaluationskriterien, welche in der Studie der vorliegenden Arbeit verwendet werden, stützen sich teilweise auf Kriterien, welche in früheren Forschungsarbeiten angewendet wurden.

8 Postedition

In diesem Kapitel wird zunächst erklärt, was der Begriff Postedition (PE) bedeutet. Außerdem wird auf die Fähigkeiten und Kenntnisse eingegangen, welche PosteditorInnen besitzen sollten. Abschließend wird auf Forschungsarbeiten eingegangen, welche, ähnlich wie die Studie in der vorliegenden Arbeit, sich mit der Postedition von MÜ beschäftigten.

8. 1 Was ist Postedition?

Laut Wallberg (2017: 160) müssen Übersetzungen von MÜ-Systemen kontrolliert und bearbeitet werden. Dies fällt in das Aufgabengebiet von PosteditorInnen. Doch was genau umfasst der Begriff Postedition? Postedition wird am häufigsten mit MÜ in Verbindung gebracht und ist jener Begriff, der für die Korrektur von maschinellen Übersetzungsausgaben durch menschliche EditorInnen verwendet wird. Die Aufgabe von PosteditorInnen ist es, Texte, die von einem MÜ-System aus einer Ausgangssprache in eine Zielsprache übersetzt wurden, zu bearbeiten, zu modifizieren und/oder zu korrigieren (vgl. Allen 2003: 297). Während Posteditionen in der Regel zwar von Menschen durchgeführt werden, ist automatisches Posteditieren auch möglich (vgl. Hansen-Schirra et al. 2017: 176).

Auf welche Art, Weise und Intensität die Postedition eines Textes durchgeführt wird, hängt laut Allen (2003: 301) unter anderem von folgenden Aspekten ab:

- Art der AuftraggeberInnen/Zielgruppe
- Umfang der voraussichtlich zu bearbeitenden Texte
- Posteditions-Zeit
- Verwendungszweck des Textes
- erwünschte/erwartete Qualität des Textes

Im Allgemeinen wird allerdings nur zwischen *Light* und *Full Postediting* unterschieden.

Das Ziel von *Light Postediting* (auch *Minimal Postediting* oder *Rapid Postediting* genannt) ist es, einen maschinell generierten Text verständlich zu machen (Garcia 2011: 218). Stilistische und syntaktische Fehler, sowie Formatierungsfehler sind zu vernachlässigen, solange der Inhalt des Textes korrekt übermittelt wird. Diese Art des Postediting dient meist als eine Art Orientierungs- und Entscheidungshilfe; für eine Publikation des Textes ist die Qualität meist nicht ausreichend. Vorrangig ist hierbei die Informationsgewinnung (vgl. Garcia 2011: 218; Hansen-Schirra et al. 2017: 177-178).

Full Postediting (oder vollständige Postedition) hingegen zielt darauf ab, einen publizierbaren Text zu produzieren; dies impliziert ein hohes Qualitätsniveau des produzierten Textes. Auf Aspekte der Grammatik, Syntax, Terminologie und Formatierung wird viel Wert gelegt. Stilistische Perfektion jedoch wird auch bei dieser Posteditionsart nicht erwartet (vgl. Allen 2003: 306; Hansen-Schirra et al. 2017: 178; Seewald-Heeg 2017: 170).

8. 2 Erforderliche Kompetenzen von PosteditorInnen

Das Postediting von maschinell erzeugten Texten soll, ähnlich wie bei menschlichen Übersetzungen, von ausgebildeten ÜbersetzerInnen durchgeführt werden und nicht von LaiInnen, die beider involvierter Sprachen mächtig sind, da diese leicht inhaltliche Fehler übersehen könnten. Selbstverständlich gilt das ebenso für LaiInnen, die „bloß“ die Zielsprache beherrschen. Studien, bei denen PosteditorInnen nur Zugriff auf den maschinell übersetzten Output hatten, aber nicht auf den Ausgangstext, haben gezeigt, dass zwar tatsächlich gut lesbare Texte produziert wurden, aber inhaltliche Fehler selbst von ausgebildeten ÜbersetzerInnen übersehen wurden (vgl. Hansen-Schirra et al. 2017: 176) Verständnis des Ausgangstextes ist also eine Voraussetzung für eine vollständige Postedition, doch über welche Kompetenzen müssen PosteditorInnen neben der Sprachkompetenz noch verfügen? Die sogenannte „DIN ISO 18587: 2017 Übersetzungsdienstleistungen – Posteditieren maschinell erstellter Übersetzungen – Anforderungen“-Norm empfiehlt, dass PosteditorInnen folgende Fertigkeiten und Kompetenzen besitzen sollten:

- Übersetzerische Kompetenz
- Sprachliche und textliche Kompetenz in der Ausgangs- und Zielsprache
- Kompetenz beim Recherchieren, bei der Informationsgewinnung und -verarbeitung
- Kulturelle Kompetenz
- Technische Kompetenz
- Sachgebietskompetenz

Diese Kompetenzen können laut der Norm durch Vorlage von Dokumenten nachgewiesen werden, die belegen, dass die/der PosteditorIn

- ein Hochschulstudium im Übersetzen, Linguistik oder ein Sprachstudium abgeschlossen hat.
- ein Hochschulstudium auf einem anderen Gebiet abgeschlossen hat und darüber hinaus über zwei Jahre Berufserfahrung als ÜbersetzerIn verfügt.
- über fünf Jahre Berufserfahrung als ÜbersetzerIn verfügt.

Ferner müssen PosteditorInnen laut der Norm über folgende Fertigkeiten verfügen:

- Grundlegende Kenntnisse über MÜ-Technologie und Systeme für computerunterstütztes Übersetzen
- Grundlegende Kenntnisse darüber, wie Terminologie-Management-Systeme mit MÜ-Systemen interagieren
- Betrachtungen zur Wirtschaftlichkeit (Zeit und Kosten)
- Fertigkeit, strukturiertes Feedback über wiederkehrende Fehler im MÜ-Text geben zu können, um das MÜ-System zu verbessern

Laut Wallberg (2017: 166) sind die Anforderungen der Norm gut umsetzbar und viele Dienstleister orientieren sich an dieser. Seewald-Heeg (2017: 169) zufolge sind eine positive Einstellung zur MÜ, genauso wie eine realistische Einschätzung der Möglichkeiten der MÜ weitere Voraussetzungen für eine effiziente PE. Es sei angemerkt, dass sich all die angeführten Kompetenzen in diesem Unterkapitel auf den Bereich des Fachübersetzten beziehen und nicht auf den Literaturbereich. Für diesen Bereich sind unter Umständen andere Fertigkeiten vonnöten.

8. 3 Bisherige Forschung zu Postedition

Forschungsarbeiten in der Vergangenheit zum Thema Postedition konzentrierten sich vor allem auf den Vergleich von posteditierten MÜ-Produkten mit der maschinellen Rohübersetzung und einer Humanübersetzung im Hinblick auf den Ablauf, die Geschwindigkeit, die Qualität und die Benutzerakzeptanz (vgl. Temizöz 2016). Guerra (2003) untersuchte bereits 2003, ob die Postedition eines maschinell übersetzten Textes schneller sei als die Übersetzung des Quellentextes von einer/m menschlicher/n ÜbersetzerIn. Guerra kam zum Schluss, dass eine beträchtliche Zeitersparnis erzielt werden konnte, wenn die Ausgangstexte zuerst maschinell übersetzt und dann posteditiert wurden, anstatt ausschließlich von Menschen übersetzt zu werden. Bei ihren Versuchstexten handelte es sich um „Marketingbroschüren“. Ausgehend von

einem Arbeitsaufwand von 2000-3000 übersetzten Wörtern pro Tag könnte ein/e durchschnittliche/r ÜbersetzerIn laut Guerras Forschungsergebnissen zwischen 1 Stunde 50 Minuten und 2 Stunden 45 Minuten täglich einsparen, wenn diese/r konsequent posteditiert, statt „nur“ zu übersetzen.

Bowker und Ehgoetz (2007) führten eine Studie durch, in welcher die Akzeptanz von Posteditionen untersucht wurde. Die Studie ergab, dass etwas mehr als zwei Drittel der ProbandInnen posteditierte MÜ-Ausgaben gegenüber menschlichen Übersetzungen bevorzugte. Bei den zu übersetzenden Texten handelte es sich um Verwaltungsakten, also um rein informative Texte.

Fiederer und O'Brien (2009) untersuchten, ob posteditierte Texte von einer geringeren Qualität seien als menschliche Übersetzungen. Übersetzt wurde ein englischsprachiges Benutzerhandbuch ins Deutsche. In der Studie schnitten Posteditionen in den Kategorien Klarheit und Genauigkeit zumindest genauso gut ab wie Humanübersetzungen, teilweise sogar besser. Beim Qualitätsparameter Stil wurden hingegen die menschlichen Übersetzungen den posteditierten Produkten vorgezogen.

Ähnlich wie die in Kapitel 6. 3 besprochenen Studien, widerlegen die gerade angeführten Arbeiten die Hypothese, dass Posteditionen im Vergleich zu Humanübersetzungen zwangsläufig eine niedrigere Qualität aufweisen. Allerdings muss an dieser Stelle festgehalten werden, dass es sich bei den Versuchstexten dieser Studien nicht um literarische Texte handelte, sondern um technische und rechtliche Texte sowie Texte aus Werbebroschüren. Immerhin bestätigten Erkenntnisse dieser Studien zumindest teilweise die Hypothese der vorliegenden Arbeit; Postedition von maschinell generierten Texten ist signifikant schneller als der Prozess einer Humanübersetzung.

Im Bereich der maschinellen Übersetzung, in welchem es laufend neue Entwicklungen gibt, darf das Erscheinungsjahr älterer Studien nicht außer Acht gelassen werden. Vor weniger als 15 Jahren beispielsweise galten phrasenbasierte Systeme noch als das Nonplusultra bei MÜ-Systemen. Es sei also darauf hingewiesen, dass wissenschaftliche Studien in diesem Bereich (ähnlich wie in vielen anderen computertechnischen Bereichen) stets bloß als eine Momentaufnahme zu verstehen sind. Dementsprechend sind die Erkenntnisse, welche in der Studie der vorliegenden Arbeit gemacht wurden, ebenfalls lediglich eine Momentaufzeichnung aus dem Jahre 2021, welche in einer womöglich bereits sehr nahen Zukunft als überholt und gegenstandslos betrachtet werden könnte.

9 Empirische Untersuchung

In diesem Kapitel wird die Untersuchungsmethode dargestellt. Hierunter fällt der Aufbau der Untersuchung, die Auswahl der ProbandInnen und des Versuchstextes sowie die Vorstellung des Programms DeepL.

9. 1 Vorgehensweise und Vorstellung der ProbandInnen

An der Untersuchung haben fünf Studierende des Masterstudiums der Translation an der Universität Wien teilgenommen. Voraussetzungen waren ein abgeschlossenes oder aktives Studium des Schwerpunktes „Übersetzen in Literatur - Medien – Kunst“ und die Arbeitssprachen Deutsch und Englisch. Als Studierende dieses Schwerpunktes besitzen die ausgewählten Personen das Wissen und die praktische Fähigkeit, die Spezifik literarischer und künstlerischer Texte zu erkennen und dementsprechend eine Reihe von Stilmitteln und Sprachregistern in Übersetzungsprozessen anzuwenden (vgl. transvienna 2021). Drei der ProbandInnen befanden sich zum Zeitpunkt der Studie (Dezember 2020/Jänner 2021) im fünften Semester des Masterstudiums, die vierte im dritten Semester.

Der Versuchstext wurde in zwei ungefähr gleich großen Hälften aufgeteilt, in Text A und Text B. Vier der Testpersonen übersetzten jeweils einen Teil und posteditierten den anderen Teil, welcher zuvor von DeepL „rohübersetzt“ wurde. Übersetzt wurde stets von Englisch nach Deutsch. Hierbei gab es zwei Gruppen mit jeweils zwei ProbandInnen. Gruppe A hatte den Auftrag, Text A zu übersetzen und Text B zu posteditieren und Gruppe B posteditierte Text A und übersetzte Text B. Alle TeilnehmerInnen der beiden Gruppen erhielten genaue Anweisungen, den englischen Originaltext, einen Fragebogen und entweder Teil A oder B als Rohübersetzung per Mail. Allen TeilnehmerInnen stand es frei, Hilfsmittel wie Wörterbücher oder dergleichen zu verwenden. Es gab kein Zeitlimit und die TeilnehmerInnen führten die Aufgabe an ihren üblichen Arbeitsplätzen (zuhause oder im Büro) und auf ihren eigenen Computern durch. Alle TeilnehmerInnen mussten genau dokumentieren, wieviel Zeit sie für die einzelnen Texte benötigten. Die Beantwortung des Fragebogens wurde nicht in der Zeitaufzeichnung berücksichtigt.

Die fünfte Person, welche in keine der beiden Gruppen war, war die Evaluatorin. Die Evaluatorin studierte an der Freien Universität Berlin im Bachelor Anglistik und Komparatistik mit einem Auslandsaufenthalt an der University of Bath. Darauf folgte ein Masterstudium in Translationswissenschaft an der Universität Wien mit dem Schwerpunkt Literatur-Medien-Kunst. Schon während ihres Studiums begann sie, freiberuflich als Übersetzerin zu arbeiten.

Sie hat bereits sechs Werke übersetzt und plant, 2021 ihr erstes eigenes Werk zu veröffentlichen. Sie lebt und arbeitet in Wien.

Der Evaluatorin wurden genaue Anweisungen, der englische Originaltext und alle Übersetzungen und Posteditionen der beiden Gruppen per Mail gesendet. Genau wie bei den anderen ProbandInnen stand es der Evaluatorin frei zu entscheiden, wo, wie und wann sie arbeitet. Alle Übersetzungen und Posteditionen wurden von ihr evaluiert. Bewertet wurden unter anderem die Orthographie, Lexik, Syntax und Stilistik der Zieltexte auf einer Skala von 1-5 (1 = Sehr zufriedenstellend, 5 = Unzufriedenstellend). Der Evaluatorin wurde nicht mitgeteilt, bei welchen Texten es sich um Posteditionen und bei welchen um Übersetzungen handelte. Dies sollte unter anderem dazu beitragen, dass die Beurteilung so objektiv wie möglich vonstattengeht. Nach der Bewertung eines jeden Textes gab sie an, ob sie hinter dem Text eine Übersetzung oder eine Postedition vermutete.

Dreh- und Angelpunkt eines jeden TQA welches nicht automatisch durchgeführt wird, sind selbstverständlich die EvaluatorInnen. An TQA können je nach Szenario sowohl professionelle als auch nicht professionelle EvaluatorInnen beteiligt sein: Während intuitiv davon ausgegangen wird, dass „Profis“ zuverlässigere Ergebnisse liefern, können „Amateure“ bei einigen TQA-Aufgaben genauso hilfreich sein. Laut Castilho et al. (2018: 23) sind vor allem im MÜ-Forschungskontext professionelle EvaluatorInnen eher die Ausnahme als die Regel, da Ressourcenbeschränkungen es oft ForscherInnen erschweren, auf ausgebildete professionelle EvaluatorInnen zuzugreifen. Unabhängig davon, ob es sich um eine/n professionelle/n EvaluatorIn handelt oder nicht, haben die Fähigkeiten, das Wissen, die Expertise und sogar persönliche Vorlieben der evaluierenden Person einen nicht zu unterschätzenden Einfluss auf die Evaluierung und damit konsequenterweise auf das gesamte Forschungsergebnis. Somit wird an dieser Stelle darauf hingewiesen, dass jede menschliche Evaluation stets ein subjektives Element enthält, welches sich schwer oder gar nicht vermeiden lässt. Vorzüge und Nachteile einer menschlichen Evaluierung wurden in Kapitel 7. 2 besprochen. Im späteren Verlauf der Arbeit werden oft Inhalte aus den Posteditionen und Übersetzungen hervorgehoben, bei denen es sich zwar nicht per se um Fehler handelt, sondern um Wörter/Satzteile/Sätze, die die Evaluatorin unter Umständen anders verfasst hätte. Eine genaue Untersuchung dieser Fehler und „Unebenheiten“ findet sich in Kapitel 10. 4. 3.

Alle ProbandInnen haben vor und nach der Übersetzung, Postedition oder Evaluierung einen Fragebogen ausgefüllt. Mittels dieser Fragebögen wurden Daten über die Profile der Probandinnen und ihrer Wahrnehmung des Testprozesses erhoben. Die Frage- und Evaluierungsbögen werden später genauer behandelt.

9. 2 Der Versuchstext

Beim Versuchstext handelt es sich um eine englischsprachige Kurzgeschichte von der Autorin Abigail Allen. Der Titel lautet „Lace Curtain“ und die Geschichte umfasst 1106 Wörter. Die deutsche Rohübersetzung von DeepL ist 1115 Wörter lang. Der Text wurde für die empirische Studie in zwei ungefähr gleich große Hälften geteilt. Teil A umfasst 577 Wörter, während Teil B 538 Wörter lang ist. Sowohl der vollständige englischsprachige Text als auch die deutsche Übersetzung von DeepL finden sich im Anhang.[1]

9. 3 DeepL – Neuronale maschinelle Übersetzung online

DeepL ist seit August 2017 auf dem Übersetzermarkt und ist von der DeepL GmbH mit Sitz in Köln entwickelt worden. Zuvor hieß das Unternehmen Linguee GmbH – dieses wurde bereits 2008 von Gereon Frahling und Leonard Fink gegründet. Die Nutzung des Online-Übersetzungsdienst ist bis zu einer Textlänge von 5000 Zeichen kostenfrei. DeepL verwendet „Convolutional Neural Networks“, ein künstliches neuronales Netz, welches mit über einer Milliarde übersetzter Sätze trainiert wurde, die von der Übersetzungssuchmaschine Linguee bereitgestellt wurde. Somit liefert DeepL laut eigenen Angaben weltweit die besten Übersetzungen und übertrifft alle anderen Übersetzungssysteme (vgl. [a]Deepl 2021; Meinhold 2021).

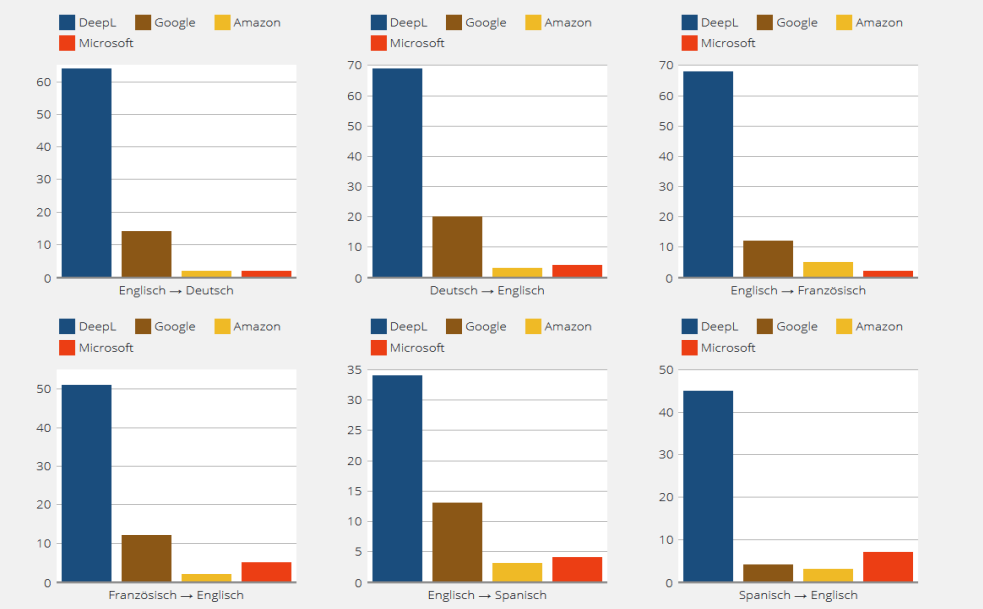

Abbildung 8: DeepL im Vergleich (vgl. [b]DeepL 2021)

[1] An dieser Stelle möchte ich mich herzlichst bei der Autorin Abigail Allen für die Erlaubnis bedanken, ihre Kurzgeschichte „Lace Curtain“ in dieser Arbeit abdrucken zu dürfen.

In von DeepL durchgeführten Blindtests wurden 119 Absätze aus unterschiedlichen Themengebieten von den Übersetzungssystemen von DeepL, Google, Amazon und Microsoft übersetzt. Professionelle ÜbersetzerInnen hatten die Aufgabe, die Qualität der Übersetzungen zu evaluieren ohne zu wissen, welche Übersetzungen von welchem System stammten. Abbildung 8 zeigt die Ergebnisse dieser Tests. Die Studien wurden im Jänner 2020 durchgeführt (vgl [b]DeepL 2021).

Seit März 2020 sind auch Übersetzungen aus den Sprachen Chinesisch und Japanisch möglich. März 2021 wurde der DeepL Übersetzer um 13 europäische Sprachen erweitert. Somit lassen sich aktuell (April 2021) Texte aus und in 24 Sprachen übersetzen: Bulgarisch Chinesisch, Dänisch, Deutsch, Englisch, Estnisch, Finnisch, Französisch, Griechisch, Italienisch, Lettisch, Litauisch, Japanisch, Niederländisch, Polnisch, Portugiesisch, Rumänisch, Russisch und Schwedisch, Slowakisch, Slowenisch, Spanisch, Tschechisch und Ungarisch.

Ähnlich wie bei Google Übersetzer gibt es links ein Eingabefeld für die zu übersetzenden Texte und rechts das Feld für die Übersetzungen in der ausgewählten Sprache (vgl. Berger 2020). Wird ein Wort markiert oder angeklickt, schlägt das Programm alternative Formulierungen und Synonyme vor – ein wesentlicher Unterschied zum System von Google.

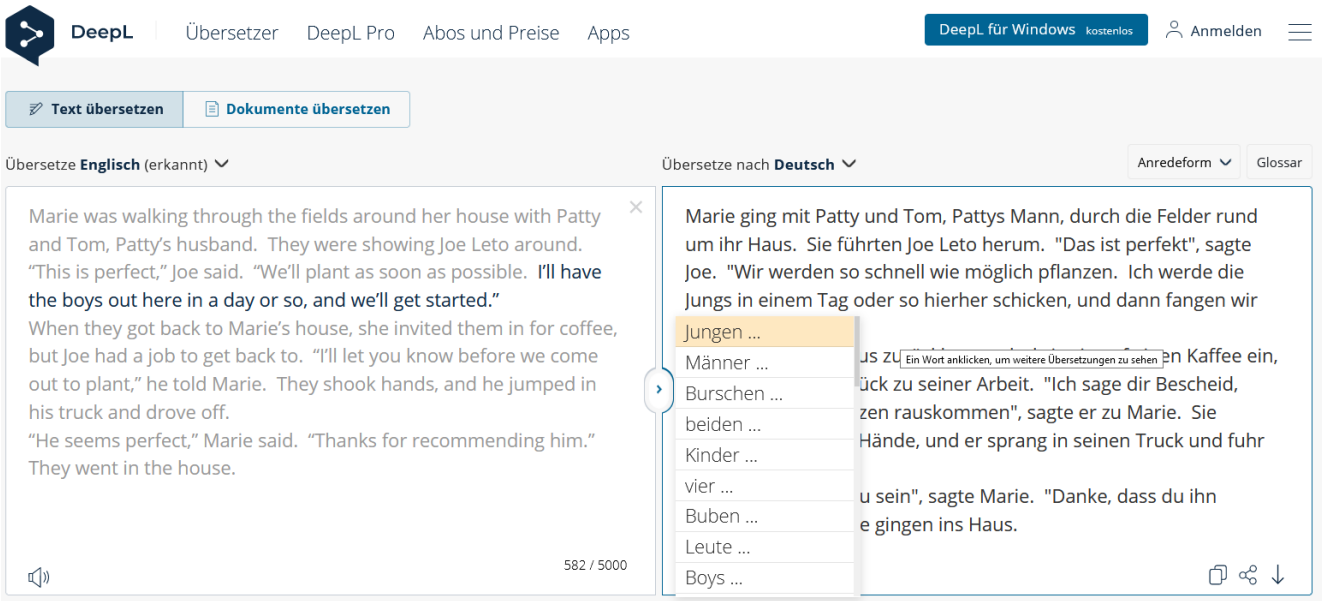

Abbildung 9: Das Eingabefeld von DeepL. Der Mauszeiger verweilt auf dem Wort „Jungs". Das Programm schlägt eine Reihe von alternativen Begriffen vor (vgl. DeepL 2021)

Seit März 2018 gibt es neben der kostenlosen Nutzung auch die Möglichkeit des kostenpflichtigen Aboservice DeepL Pro. Hierbei gibt es verschiedene Pakete für EinzelnutzerInnen, Teams mit mindestens 3 NutzerInnen und EntwicklerInnen. DeepL Pro unterscheidet sich insbesondere hinsichtlich Datensicherheit und Begrenzungen der Zeichenanzahl von der kostenfreien Variante (vgl. [b]DeepL 2021). Für die empirische Studie der vorliegenden Arbeit wurde die kostenfreie Version von DeepL verwendet.

10 Durchführung der Untersuchung

Dieses Kapitel geht auf die empirische Studie der vorliegenden Arbeit ein. Zunächst werden die Fragebögen der Gruppen A und B erläutert. Im Anschluss daran werden der Fragebogen und die Evaluierungsbögen der Evaluatorin genauer untersucht. Der Zeitaufwand, die Qualität der Zieltexte, die Bewertung jeder einzelner Messgröße und Beziehungen und Ähnlichkeiten zu früheren Studien werden genau untersucht. Dieses Kapitel schließt mit einem Fazit und Ausblick in die Zukunft.

10. 1 Fragebögen der Gruppen A und B

Im Rahmen der Studie der vorliegenden Arbeit erhielten alle TeilnehmerInnen des Experiments einen Fragebogen. Die Fragebögen bezogen sich auf sozialstatistische Merkmale der TeilnehmerInnen und erfragten Fakten, Verhalten und Meinungen der ProbandInnen. Die Fragebögen enthielten offene Fragen, Ja-/Nein-Fragen und Likert-Skala-Fragen. Die Likert-Skala ist eine psychometrische Skala und findet überwiegend in den Sozialwissenschaften Anwendung. Die Antwortmöglichkeiten zu den gestellten Fragen waren größtenteils mindestens fünfstufig und jede Antwort maß eine unterschiedliche Ausprägung des gemessenen Merkmals (vgl. Döring/Bortz 2014: 268-269). Einige der Fragen aus diesen Fragebögen wurden bereits in ähnlicher Form von Nietzke (2019) besprochen.

Es gab keine zeitliche Beschränkungen für die Beantwortung des Fragebogens. Die Fragebögen der Gruppen A und B umfassten 18 Fragen. Der Fragebogen der Evaluatorin umfasste 7 Fragen. Alle TeilnehmerInnen der Studie mussten die Fragen 1-5 vor der eigentlichen Arbeit (Postedition/Übersetzung/Evaluierung) beantworten. Die restlichen Fragen wurden nach dem Beenden der Arbeit beantwortet. Die ersten Fragen betrafen die maschinelle Übersetzung allgemein. Die restlichen Fragen behandeln konkret den Versuchstext, die Übersetzungen und/oder die Posteditionen. Bei jeder Frage gab es mehrere Antwortmöglichkeiten. Mehrfachnennungen waren möglich. Außerdem bestand bei jeder Frage die Möglichkeit, eine alternative Antwort zu geben und/oder die Antwort zu begründen. Alle TeilnehmerInnen erhielten die Fragebögen als ein Word-Dokument per E-Mail. Die Fragebögen finden sich im Anhang.

In Folge werden die Antworten der Fragebögen von Gruppe A und B besprochen. Auf den Fragebogen der Evaluatorin wird im nächsten Unterkapitel eingegangen. Da die ersten fünf Fragen jedoch für alle gleich waren und um Redundanz zu vermeiden, werden auch die Antworten der Evaluatorin für diese ersten Fragen im folgenden Teil inkludiert.

Im Folgenden werden die Inhalte der Fragebögen (Fragen sowie Antwortmöglichkeiten) kursiv geschrieben. Außerdem beschreiben die Begriffe ProbandInnen und Testpersonen in der Folge stets TeilnehmerInnen der Gruppen A und/oder B, aber nicht die Evaluatorin. Alternative Antwortmöglichkeiten werden unter Anführungszeichen erwähnt.

Frage 1: *Wie oft benutzen Sie maschinelle Übersetzung? (Dazu zählen Google Übersetzer, DeepL und andere)*

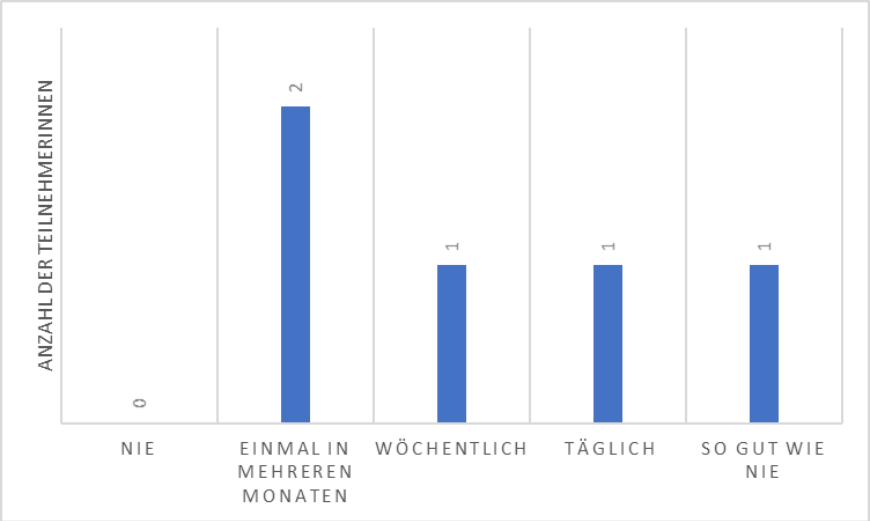

Abbildung 10: Visualisierung der Frage 1 (eigene Grafik, erstellt am 16.04.2021)

Die ProbandInnen konnten zwischen den Antworten *Nie; Einmal in mehreren Monaten; Wöchentlich* und *Täglich* wählen. Die ProbandInnen antworteten sehr unterschiedlich. Jeweils eine/r antwortete einmal *Einmal in mehreren Monat, Wöchentlich, Täglich* und eine Person benutze MÜ sogar „so gut wie nie“. Die Evaluatorin benutzt maschinelle Übersetzung *Einmal in mehreren Monaten*. Ob jene Personen, die MÜ häufiger benutzen, aufgrund dessen bessere Posteditionen produzieren können, wird sich zeigen.

Frage 2: *Wie zufrieden sind Sie grundsätzlich mit dem Output von maschineller Übersetzung?*

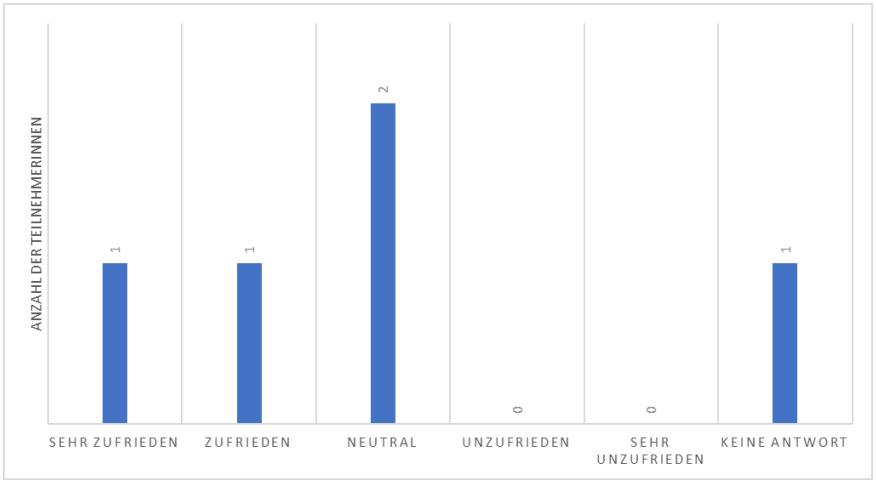

Abbildung 11: Visualisierung der Frage 2 (eigene Grafik, erstellt am 16.04.2021)

Sehr zufrieden; Zufrieden; Neutral; Unzufrieden; Sehr unzufrieden und *Keine Antwort* waren die Antwortmöglichkeiten. Insgesamt sind die TeilnehmerInnen dieser Studie zufrieden mit dem Output von MÜ. Zwei Personen antworteten *Neutral*, eine *Zufrieden* und eine Person ist sogar *Sehr zufrieden*. Interessanterweise hängt die Zufriedenheit der ProbandInnen laut den vorliegenden Antworten nicht direkt mit der Häufigkeit der Nutzung zusammen – jene Testperson, die *Sehr zufrieden* ist, hat zuvor geantwortet, MÜ „so gut wie nie" zu benutzen.

Die Evaluatorin kreuzte keine der angeführten Antwortmöglichkeiten an, führte ihre Antwort aber dafür ein wenig aus – sie verwende MÜ nur im privaten Bereich und hierbei reiche ihr meist das Übersetzungsprogramm von Google, „Google Translate", da es ihr nur um das Verständnis geht. Allerdings wusste sie von „Experimenten, dass DeepL etwa für wissenschaftliche Journalbeiträge absolut unzulänglich ist."

Frage 3: *Wie sinnvoll ist Ihrer Meinung nach die Anwendung von maschineller Übersetzung für professionelle Übersetzungsdienste (im außerliterarischen Bereich)?*

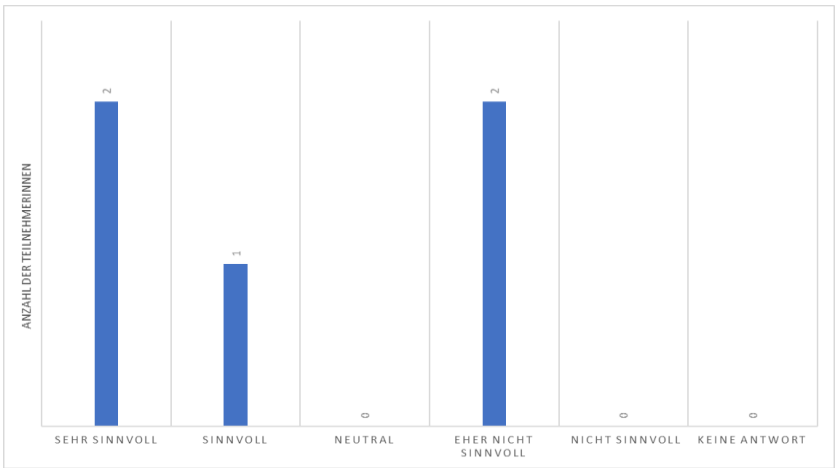

Abbildung 12: Visualisierung der Frage 3 (eigene Grafik, erstellt am 16.04.2021)

Sehr sinnvoll, Sinnvoll, Neutral, Eher nicht sinnvoll, Nicht sinnvoll und *Keine Antwort* konnte geantwortet werden. Während zwei Mal *Sehr sinnvoll* und einmal *Sinnvoll* angekreuzt wurde, waren eine Testperson und die Evaluatorin weniger überzeugt und antworteten *Eher nicht sinnvoll*. Letztere fügte eine Bemerkung hinzu: Ihrer Meinung nach kann/könnte MÜ potenziell als Unterstützung einen Wert haben, jedoch solle MÜ zumindest bei dem heutigen Stand der maschinellen Übersetzung nicht alleinstehen.

Frage 4: *Wie sinnvoll ist Ihrer Meinung nach die Anwendung von maschineller Übersetzung im Bereich des literarischen Übersetzens?*

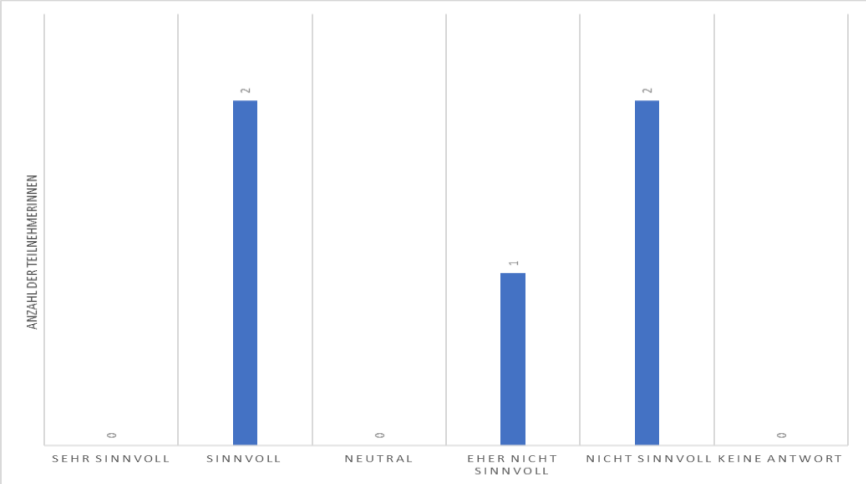

Abbildung 13: Visualisierung der Frage 4 (eigene Grafik, erstellt am 16.04.2021)

Sehr sinnvoll, Sinnvoll, Neutral, Eher nicht sinnvoll, Nicht sinnvoll und *Keine Antwort* standen zur Auswahl. Auch bei dieser Frage gingen die Meinungen auseinander. Zwei Testpersonen antworteten *Sinnvoll*, einmal davon jedoch mit der Anmerkung, dass „gelernte PosteditorInnen notwendig seien“. Ein/e ProbandIn antwortete *Eher nicht sinnvoll* und die letzte ProbandIn und die Evaluatorin waren der Anwendung von MÜ gegenüber noch negativer eingestellt und antworteten beide *Nicht sinnvoll*. Die Evaluatorin begründete ihre Antwort so: „Literatur zeichnet sich unter anderem durch ihre ‚Leerstellen‘ und/oder Mehrdeutigkeiten aus, die der/die Leser:in füllen bzw. interpretieren muss/kann/soll/darf. Die maschinelle Übersetzung wird (wahrscheinlich nie) diese Leerstellen und/oder Mehrdeutigkeiten ‚nachbilden‘ können.“ Im weiteren Verlauf der Studie kamen allerdings keine Leerstellen und/oder Mehrdeutigkeiten vor – dies lag wohl vor allem am Schreibstil des Versuchstextes.

Frage 5: *Wie sind Ihre bisherigen Erfahrungen mit dem Programm DeepL?*

Es konnte zwischen den Antwortmöglichkeiten *Sehr zufriedenstellend, Zufriedenstellend, Neutral, Unzufriedenstellend, Sehr unzufriedenstellend* und *Keine bisherige Erfahrungen* gewählt werden. Scheinbar ist keine/r der StudienteilnehmerInnen explizit unzufrieden mit DeepL – zwei Mal wurde *sehr zufriedenstellend*, einmal *zufriedenstellend* und einmal *neutral* geantwortet. *Neutral* war ebenfalls die Antwort der Evaluatorin – allerdings hatte sie anscheinend noch nicht allzu viel Erfahrung mit dem Programm, denn sie fügte die Anmerkung hinzu, dass sie bloß „einmal damit herumexperimentiert" habe.

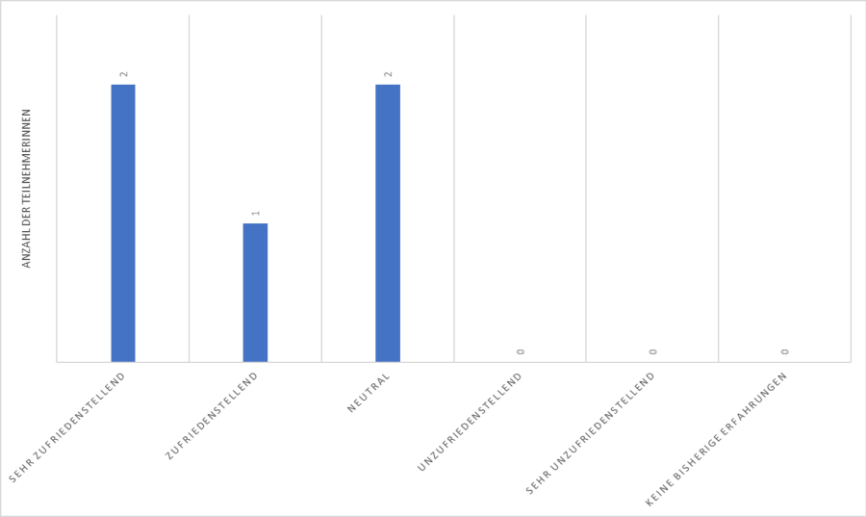

Abbildung 14: Visualisierung der Frage 5 (eigene Grafik, erstellt am 16.04.2021)

Frage 6a: *Wieviel Zeit haben Sie für die Postedition benötigt?*

Frage 6b: *Wieviel Zeit haben Sie für die Übersetzung benötigt?*

Die Zeitangabe sollte so genau wie möglich und in Minuten angegeben werden. Die Beantwortung des Fragebogens wurde hierbei nicht mitgezählt. Wie bereits in der Einleitung erwartet, war die Postedition für fast alle ProbandInnen mit weniger Zeitaufwand verbunden als die Übersetzung. Bloß TeilnehmerIn 3 sticht hier heraus, da bei ihr/ihm das Gegenteil der Fall ist. Bei allen anderen hingegen ist es sogar ein recht signifikanter Unterschied – sowohl TeilnehmerIn 1 als auch 2 benötigten für ihre Übersetzungen mehr als doppelt so viel Zeit als für ihre Posteditionen. Während TeilnehmerIn 1 um 150% mehr Zeit benötigte, waren es bei TeilnehmerIn 2 sogar 180%. TeilnehmerIn 4 hingegen brauchte „nur" 15 Minuten mehr für ihre/seine Übersetzung als für die Postedition – das ist immerhin noch eine Erhöhung des Zeitaufwandes um 33,33%. Sonderfall T3 benötigte für die Postedition 16,67% mehr Zeit (also 10 Minuten) als für ihre/seine Übersetzung. T3 war nach eigenen Angaben selbst überrascht und hätte eigentlich erwartet, dass sie/er bei der Postedition schneller sein würde.

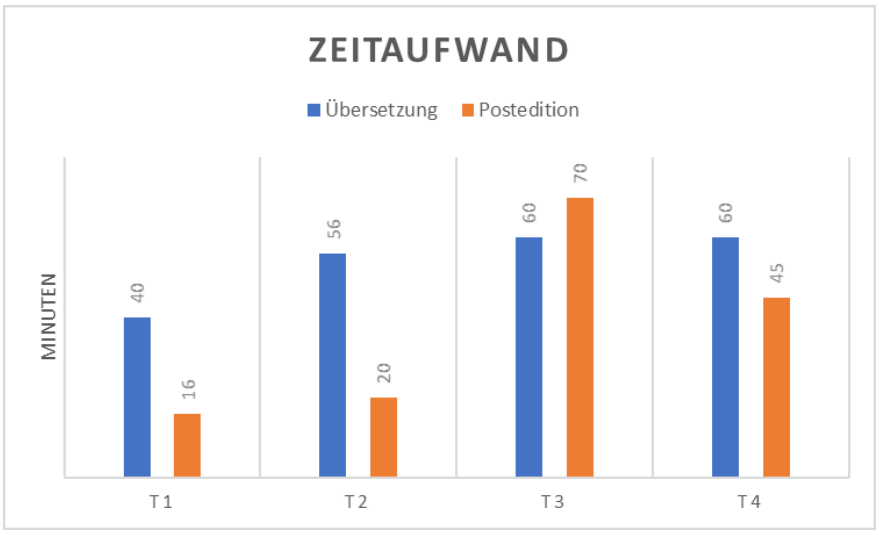

Abbildung 15: Zeitaufwand der ProbandInnen in Minuten (eigene Grafik, erstellt am 16.04.2021)

Frage 7: *Als wie „schwierig" zu übersetzen würden Sie die Kurzgeschichte beschreiben?*

Hier waren folgende Antwortmöglichkeiten gegeben: *Sehr schwierig, Schwierig, Neutral, Leicht, Sehr leicht* und *Keine Antwort*. Die ProbandInnen fanden insgesamt, dass der Versuchstext nicht schwierig zu übersetzen war. Jeweils zwei Mal wurden die Antworten *Leicht* und *Neutral* angekreuzt. In der Tat wurde bei der Auswahl des Versuchstextes auf eine nicht zu komplexe Syntax, Lexik, und Stilistik, sowie auf eine begrenzte Länge geachtet.

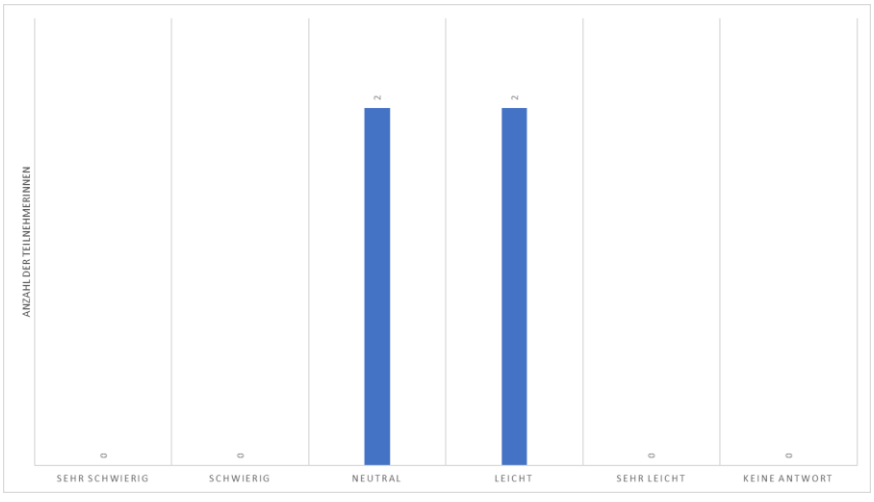

Abbildung 16: Visualisierung der Frage 7 (eigene Grafik, erstellt am 16.04.2021)

Frage 8: *Wie zufrieden waren Sie mit dem Output, also der „Rohübersetzung" des Programms DeepL?*

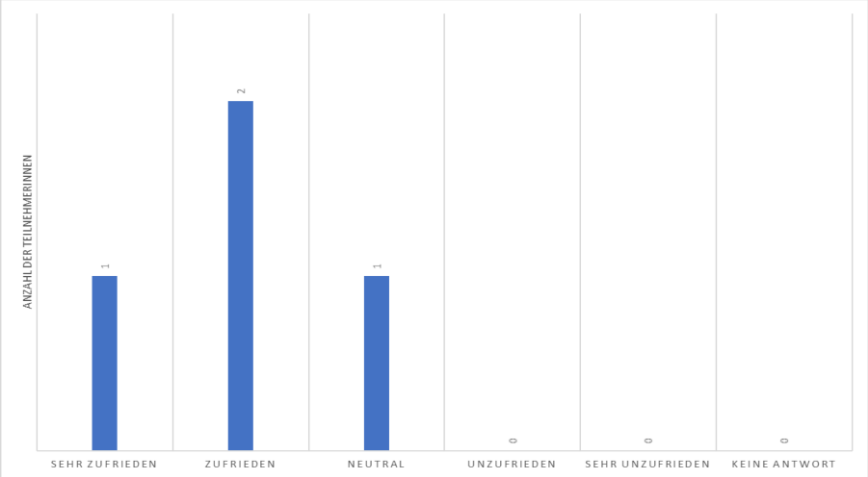

Abbildung 17: Visualisierung der Frage 8 (eigene Grafik, erstellt am 16.04.2021)

Bei dieser Frage konnten die ProbandInnen zwischen *Sehr zufrieden*, *Zufrieden*, *Neutral*, *Unzufrieden*, *Sehr unzufrieden*, und *Keine Antwort* entscheiden. Bloß eine Person antwortete *Neutral*; zwei waren *Zufrieden* und eine Testperson war sogar *Sehr zufrieden*. Eine der zufriedenen Personen war jedoch „gar nicht so glücklich darüber, wie schön der Text nach außen wirkt." Laut ihr/ihm wird es nämlich schwieriger, Fehler zu erkennen, je flüssiger sich der Text liest.

Frage 9: *Wie zufrieden sind Sie mit dem von Ihnen posteditieren Zieltext?*

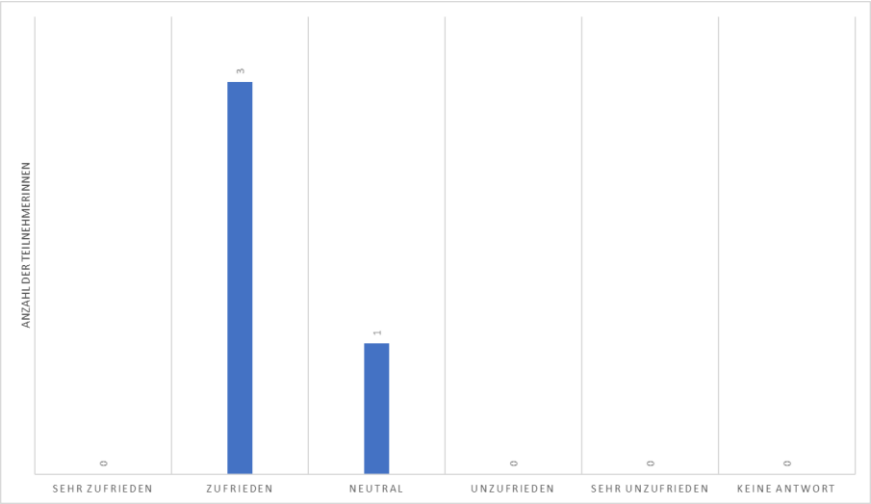

Abbildung 18: Visualisierung der Frage 9 (eigene Grafik, erstellt am 16.04.2021)

Genau wie bei der vorherigen Frage sind auch hier die Optionen *Sehr zufrieden, Zufrieden, Neutral, Unzufrieden, Sehr unzufrieden,* und *Keine Antwort*. Drei der vier Probandinnen gaben an, *Zufrieden* mit ihren posteditierten Zieltext zu sein. Die vierte Person antwortete *Neutral*.

Frage 10: *Wie zufrieden sind Sie mit dem von Ihnen direkt übersetzten Zieltext?*

Bei Frage 10 gab es die selben Antwortmöglichkeiten wie bei den vorherigen Fragen – *Sehr zufrieden, Zufrieden, Neutral, Unzufrieden, Sehr unzufrieden* und *Keine Antwort*. Hier haben alle ProbandInnen dieselbe Antwort gewählt – alle TeilnehmerInnen gaben an, *Zufrieden* mit den von ihren direkt übersetzten Texten zu sein. Im späteren Verlauf dieser Arbeit wird sich zeigen, wie zufriedenstellend die Zieltexte laut der Evaluatorin waren.

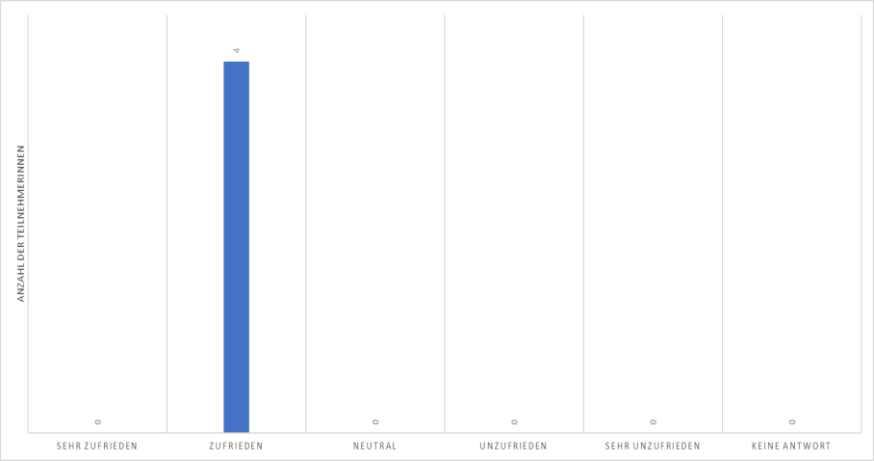

Abbildung 19: Visualisierung der Frage 10 (eigene Grafik, erstellt am 16.04.2021)

Frage 11: *Hätten Sie es bevorzugt, die Kurzgeschichte „nur" zu übersetzen, also nicht zu posteditieren?*

Frage 11 konnte mit *Ja, Nein,* oder *Nicht sicher* beantwortet werden. Bloß eine der vier ProbandInnen hätte es nicht bevorzugt, die Kurzgeschichte nur zu übersetzten und antwortete *Nein*. Zwei Personen antworteten *Ja* und eine dieser Personen ist der Meinung, dass „durch den Einsatz maschineller Übersetzungstools der Brainstorming-Schritt im Übersetzungsprozess entfällt und die Übersetzungsmöglichkeiten eingeschränkt werden, weil man durch das Produkt der Maschinenübersetzung beeinflusst wird." Die verbleibende Person war sich *Nicht sicher* und meinte hierzu: „Reines Übersetzen macht mehr Spaß, aber wenn es um Zeit geht, hat das Posteditieren klare Vorteile."

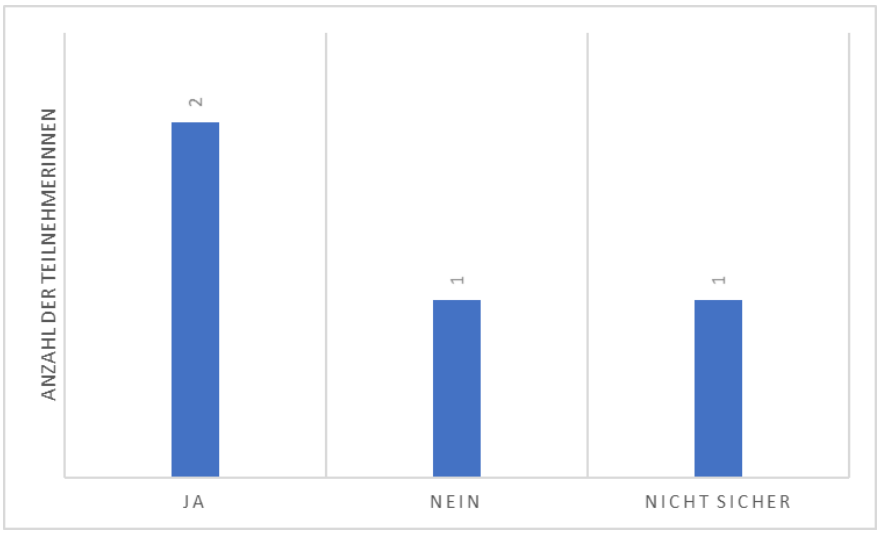

Abbildung 20: Visualisierung der Frage 11 (eigene Grafik, erstellt am 16.04.2021)

Frage 12: *Hätten Sie es bevorzugt, die gesamte Kurzgeschichte zu posteditieren?*

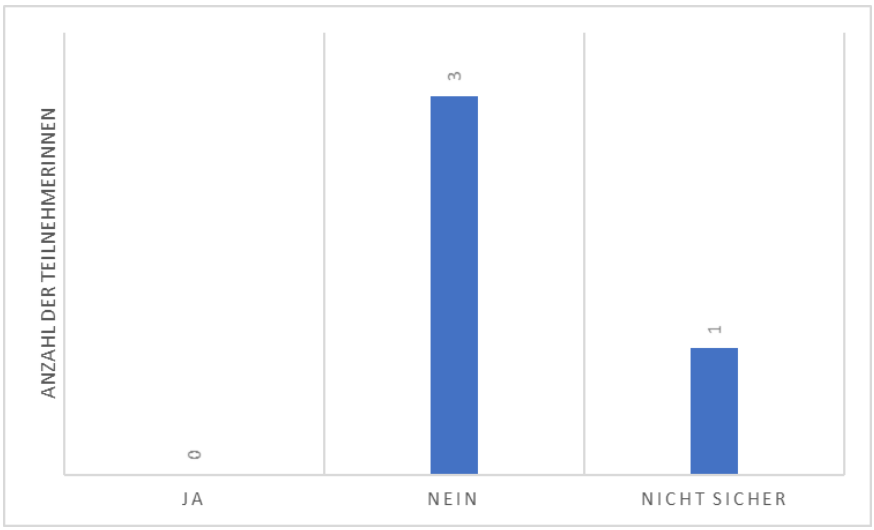

Abbildung 21: Visualisierung der Frage 12 (eigene Grafik, erstellt am 16.04.2021)

Auch bei Frage 12 wurden die Antwortmöglichkeiten *Ja* und *Nein* von einem *Nicht sicher* ergänzt. Jene zwei, die Frage 11 mit *Ja* beantwortet haben, haben Frage 12 mit *Nein* beantwortet. Eine dieser Personen fügte eine Bemerkung hinzu: „Bei einem vorübersetzten Text orientiert man sich stark an dem Ergebnis, das das Übersetzungstool ‚ausspuckt', man wird im Vorhinein in eine Richtung gelenkt und zieht meiner Meinung nach demnach nicht so viele andere Übersetzungsoptionen in Erwägung, sondern versucht, die Rohübersetzung, die man vor sich hat, zu optimieren." Dieselbe Person, die bei der vorherigen Frage *Nicht sicher* geantwortet hat, antwortete hier ebenfalls mich *Nicht sicher*. Kurioserweise hat jene Person, welche bei der vorherigen Frage die Antwortmöglichkeit *Nein* ankreuzte, bei Frage 12 auch nein angekreuzt.

Ein erdenklicher Grund für diesen „Widerspruch“ könnte sein, dass es diese TeilnehmerIn tatsächlich bevorzugt, eine Texthälfte zu posteditieren und die andere Hälfte zu übersetzen. Andere Gründe sind für mich nicht ersichtlich.

Frage 13: *Sind Sie der Meinung, dass Sie mehr Zeit benötigt hätten, wenn Sie die Kurzgeschichte „nur“ übersetzt hätten?*

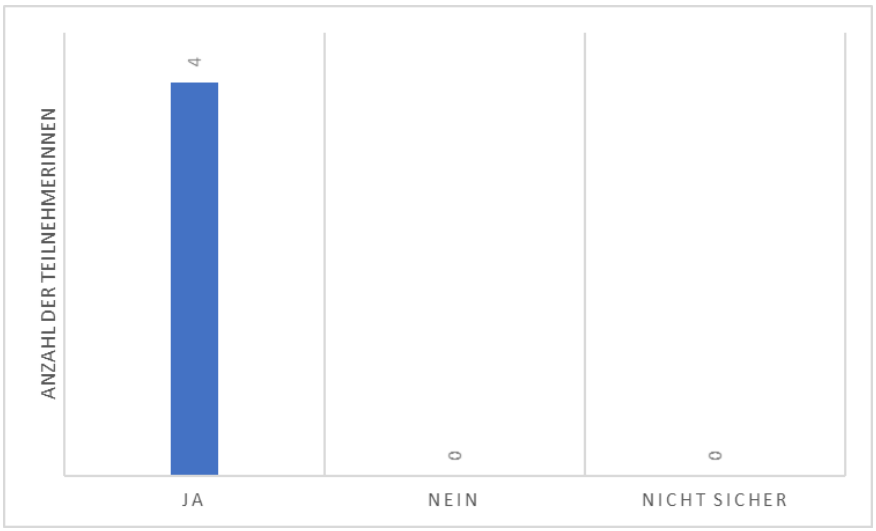

Abbildung 22: Visualisierung der Frage 13 (eigene Grafik, erstellt am 16.04.2021)

Es standen wieder die Antwortmöglichkeiten *Ja*, *Nein*, und *Nicht sicher* zur Auswahl. Bei dieser Frage waren sich alle ProbandInnen einig – jede/r gab an, dass sie/er länger gebraucht hätte, wenn sie/er die Kurzgeschichte nur übersetzt hätte. Interessanterweise beantwortete auch TeilnehmerIn 3, also die einzige Person, welche für die Postedition mehr Zeit benötigte, als für die Übersetzung, diese Frage mit *Ja*. Eine mögliche Erklärung für diesen scheinbaren Widerspruch ist, dass TeilnehmerIn 3 selbst nicht realisiert hat, dass sie/er für die Postedition mehr Zeit benötigt hat als für ihre/seine Übersetzung.

Frage 14: *Sind Sie der Meinung, dass Sie mehr Zeit benötigt hätten, wenn Sie die gesamte maschinell vorübersetzte Kurzgeschichte posteditiert hätten?*

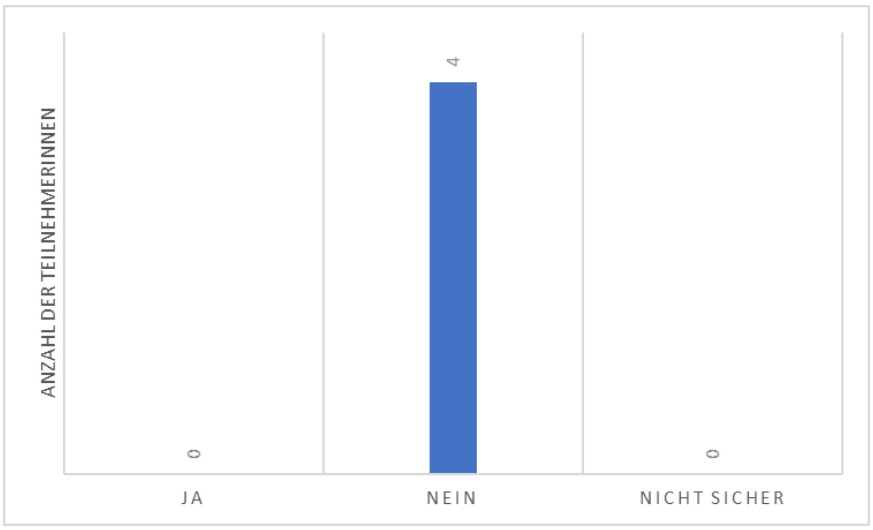

Abbildung 23: Visualisierung der Frage 14 (eigene Grafik, erstellt am 16.04.2021)

Erneut gab es die Auswahl zwischen *Ja*, *Nein*, und *Nicht sicher*. In Anbetracht der Antworten der Frage 13 sind die Antworten der ProbandInnen hier wenig überraschend – alle antworteten mit *Nein*.

Frage 15*: Sind Sie der Meinung, dass Sie einen qualitativ hochwertigeren Zieltext hätten produzieren können, wenn Sie „nur" übersetzt hätten?*

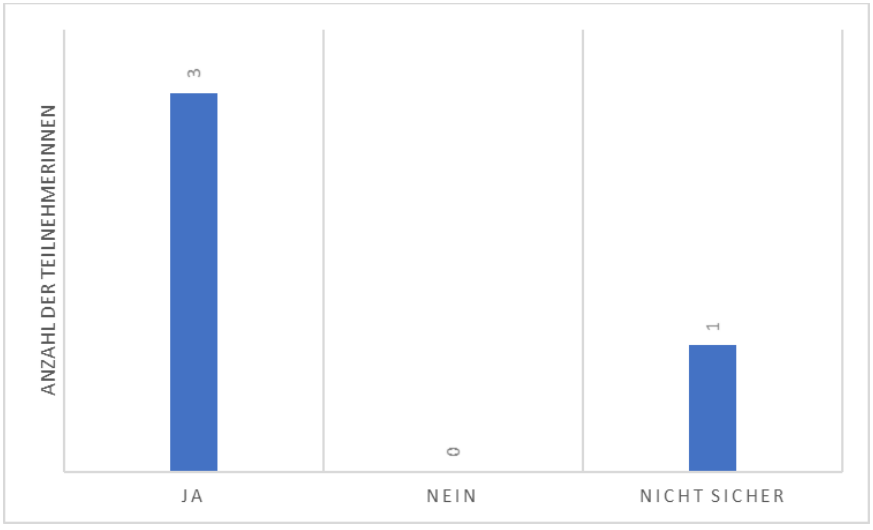

Abbildung 24: Visualisierung der Frage 15 (eigene Grafik, erstellt am 16.04.2021)

Abermals gab es die Antwortmöglichkeiten *Ja*, *Nein*, und *Nicht sicher*. Drei der Testpersonen waren der Meinung, dass sie in der Lage gewesen wären, einen besser Zieltext zu verfassen, wenn sie nur übersetzt hätten. Nur eine Person war sich *Nicht sicher* und meinte:

„Bei manchen Dingen fand ich die Übersetzung durch die Maschine fast treffender, und ich konnte beim Editing besser auf einzelne Details eingehen."

Frage 16: *Sind Sie der Meinung, dass Sie einen besseren Zieltext hätten produzieren können, wenn Sie die gesamte Kurzgeschichte posteditiert hätten?*

Ein weiteres Mal konnten die Testpersonen zwischen *Ja*, *Nein*, und *Nicht sicher* entscheiden. Dieselben drei Personen, welche Frage 15 mit *Ja* beantwortetet hatten, kreuzten bei Frage 16 *Nein* an. Naturgemäß wurde die Antwort *Nicht sicher* bei dieser Frage ebenfalls von derselben Person ausgewählt wie bei der vorherigen Frage. Sie/er fügte ihrer/seiner Antwort folgendes hinzu: „Schwierig, ich bin mir nicht sicher, ob manche Teile nicht doch etwas „übersetzt" klingen würden; ich bin beim Postediting doppelt so kritisch, was das angeht."

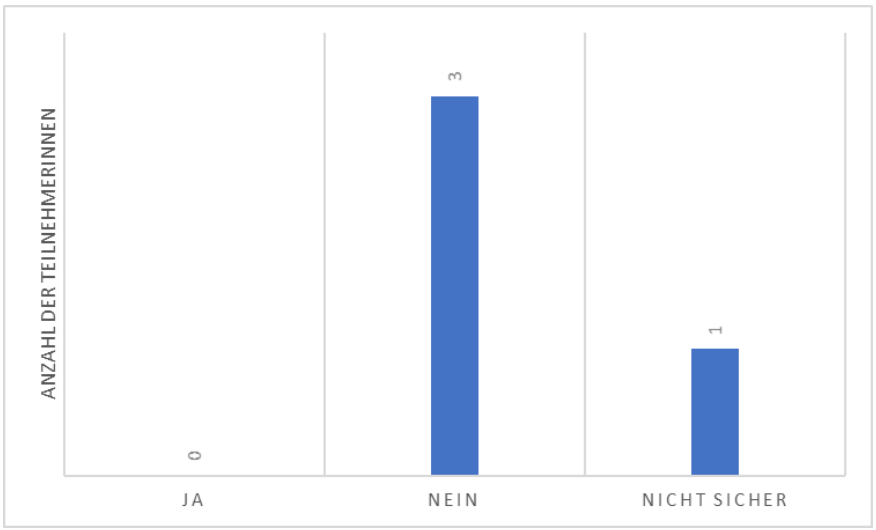

Abbildung 25: Visualisierung der Frage 16 (eigene Grafik, erstellt am 16.04.2021)

Frage 17: *Hatten Sie das Gefühl, dass Sie beim Posteditieren Ihre Kreativität nicht so entfalten konnten, wie wenn Sie nur übersetzt hätten?*

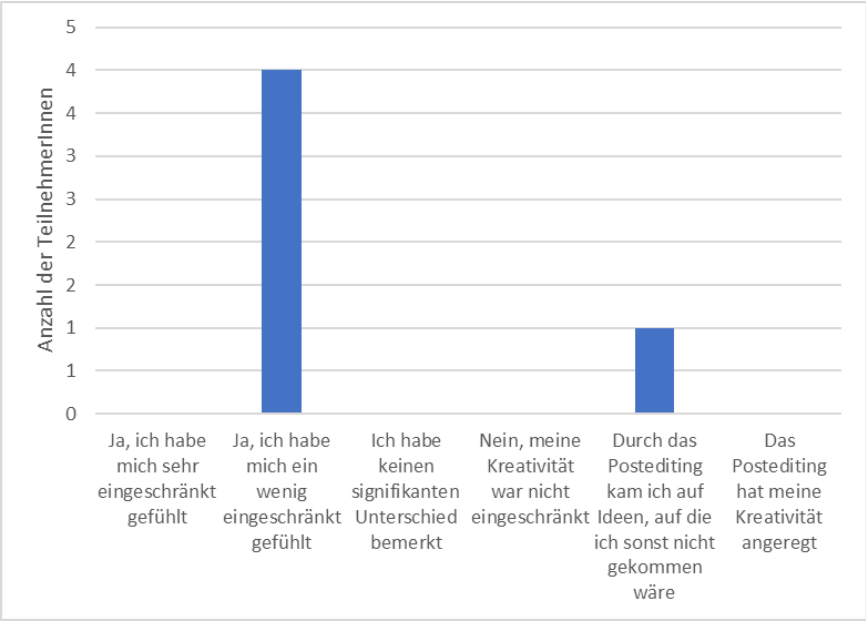

Abbildung 26: Visualisierung der Frage 17 (Abbildung von mir am 16.04.2021 erstellt)

Bei Frage 17 standen ausführlichere Antwortmöglichkeiten zur Auswahl:

- *Ja, ich habe mich sehr eingeschränkt gefühlt.*
- *Ja, ich habe mich ein wenig eingeschränkt gefühlt.*
- *Ich habe keinen signifikanten Unterschied bemerkt.*
- *Nein, meine Kreativität war nicht eingeschränkt.*
- *Durch das Postediting kam ich auf Ideen, auf die ich sonst nicht gekommen wäre.*
- *Das Postediting hat meine Kreativität angeregt.*

Bei dieser Frage waren sich alle ProbandInnen einig – jede/r der Testpersonen gab an, sich beim Posteditieren *ein wenig eingeschränkt gefühlt* zu haben. Interessanterweise gab es bei dieser Frage jedoch auch eine Mehrfachnennung. Eine Person kreuzte zusätzlich auch die Antwort „*Durch das Postediting kam ich auf Ideen, auf die ich sonst nicht gekommen wäre.*“ an. Bei dieser Person wurde anscheinend ihre/seine Kreativität beim Posteditieren gleichzeitig angeregt und eingeschränkt.

Frage 18: *Hat sich Ihre Meinung zu maschineller Übersetzung im Rahmen dieser Studie geändert?*

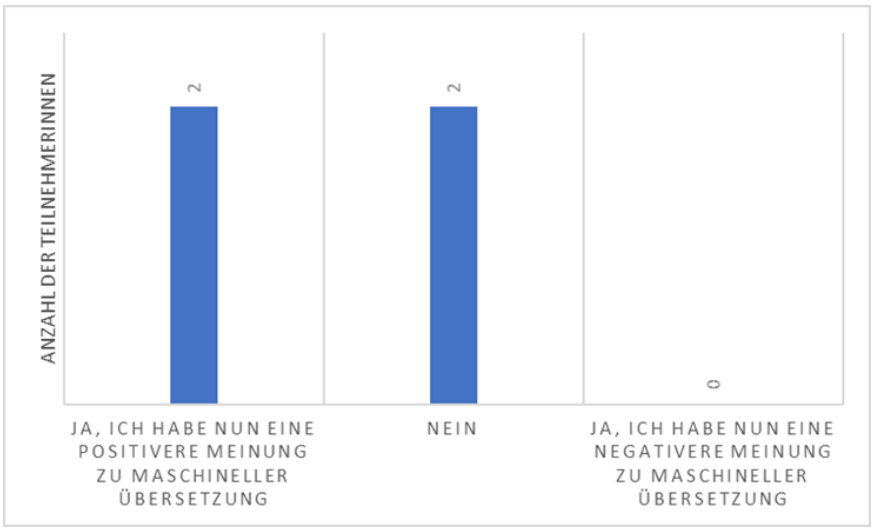

Abbildung 27: Visualisierung der Frage 18 (eigene Grafik, erstellt am 16.04.2021)

Bei der letzten Frage des Fragebogens gab es folgende drei Antwortmöglichkeiten:

- *Ja, ich habe nun eine positivere Meinung zu maschineller Übersetzung.*
- *Nein*
- *Ja, ich habe nun eine negativere Meinung zu maschineller Übersetzung.*

Zwei der ProbandInnen hatten ihre Meinung zu MÜ nicht geändert – sie kreuzten die Antwort *Nein* an. Eine/r dieser beiden gab an, dass sie/er sich „schon länger mit PE im literarischen Bereich beschäftigt." Sie/er hatte das Gefühl, dass ihre/seine Einstellung zu MÜ sehr text- und situationsabhängig sei. Die anderen zwei TeilnehmerInnen der Studie kamen durch das Experiment zu *einer positiveren Meinung zu maschineller Übersetzung*. Eine/r der beiden meinte, dass sie/er noch nie DeepL für literarische Texte benutzt habe und von der Leistung des Programms positiv überrascht sei. Laut dieser Person hat das Programm eine akzeptable Übersetzung erstellt, bei der kein großer Zeitaufwand notwendig war, um ihres/seines Erachtens nach eine zufriedenstellende Endversion anzufertigen.

Im Anschluss an die 18 Fragen hatten alle ProbandInnen noch die Gelegenheit, unter folgendem Punkt abschließende Bemerkungen sowie ihre Meinung zu teilen:

Bitte vergleichen Sie abschließend in eigenen Worten die beiden Textbearbeitungsmodi und Ihre Erfahrungen damit im Laufe dieser Studie:

T1: „Wie schon oben beschrieben haben beide Methoden ihre Vor- und Nachteile. Würde ich diesen Text für die Uni oder in der Freizeit übersetzen (und mich damit besonders darauf fokussieren, meine Kreativität zu fördern), würde ich ihn ‚nur' übersetzen. Bei einem Auftrag würde ich mich evtl. eher für Postediting entscheiden und bei Passagen, bei denen ich unsicher bin, Rat von KollegInnen oder anderen MuttersprachlerInnen einholen. Im Grunde

genommen sind reines Übersetzen und Postediting für mich zwar verbundene, aber doch unterschiedliche Skills."

T2: „Ich glaube, dass maschinelle Übersetzung für einige Textpassagen, bei denen man sich nicht sicher ist, wie man sie übersetzen soll, hilfreich sein kann. Jedoch wird dadurch die Kreativität schon sehr stark eingeschränkt und mir persönlich fällt es dann schwerer mir andere Übersetzungslösungen einfallen zu lassen. In dieser Studie fand ich die Rohübersetzung mittels DeepL relativ gut, jedoch weiß ich aus Erfahrung, dass dies besonders bei literarischen Texten nicht immer der Fall ist. Ich persönlich merke, dass es mir bei Texten, die ich ‚nur übersetze' leichter fällt, auf den Satzbau und mögliche Fehler zu achten, die mir bei der Postedition einer Rohübersetzung eventuell entgehen könnten. Außerdem finde ich, dass die Kreativität eine große Rolle bei guten Übersetzungen spielt und ich persönlich fühle mich da beim Posteditieren von Rohübersetzungen zu sehr eingeschränkt. Außerdem glaube ich, dass es mir beim Übersetzen leichter fällt genauer zu arbeiten, da ich beim Posteditieren oftmals das Gefühl habe, ich korrigiere nur eine Übersetzung, die nicht von mir stammt."

T3: „Insgesamt war ich positiv überrascht über das Ergebnis der Deep L-Übersetzung. Mit Google Translate habe ich weitaus schlechtere Erfahrungen gemacht. Was mir jedoch stark auffiel, war die Tatsache, dass das Programm die länderspezifische Zeichensetzung nicht berücksichtigt. Auch wurde die Satzstellung zum Teil eins zu eins aus dem Originaltext übernommen, worauf bei der Posteditierung geachtet werden muss. Beide Bearbeitungsmodi haben Vor- und Nachteile. Ich denke, bei einer literarischen Übersetzung sollte man sich nicht zu sehr auf maschinelle Übersetzungsprogramme verlassen, sondern am besten selbst übersetzen und eventuell bei möglichen Schwierigkeiten bestimmte Satzteile davon übersetzen lassen, um auf neue Ideen zu kommen."

T4: „Postediting: Durch die DeepL-Rohübersetzung wurde schon eine gewisse Richtung vorgegeben, wodurch man meiner Meinung nach einen „Tunnelblick" entwickelt und sich denkt, das ist eh in Ordnung, ich muss nur noch ein paar Kleinigkeiten überarbeiten. Satzbau und Grammatik habe ich eigentlich kaum geändert, die Wortwahl und generell die Stilistik mussten angepasst werden. Wenn man unter Zeitdruck steht, ist MT (damit meine ich explizit DeepL, und nicht andere furchtbare Tools wie Google Translator etc.) auf jeden Fall ein nützliches Werkzeug, dass als Fundament für die Endübersetzung dienen kann.

Eigenübersetzung: Bei dem traditionellen Übersetzungsverfahren ohne Heranziehung maschineller Übersetzungswerkzeuge werden erstmal verschiedenen Optionen erwägt, demnach dauert der Prozess auch länger. Die Vorlage war in diesem Fall kein allzu literarisch herausfordernder Text, daher kam es zu keinem wesentlichen zeitlichen Unterschied. Ich muss aber gestehen, dass ich in einem ‚professionellen' Szenario wahrscheinlich doppelt so viel Zeit in die Übersetzung investiert hätte…"

10. 2 Fragebogen der Evaluatorin

Bei den ersten fünf Fragen des Fragebogens der Evaluatorin handelt es sich, wie bereits erwähnt, um dieselben fünf Fragen, die auch im anderen Fragebogen enthalten sind. Aufgrund dessen wurden die Antworten der Evaluatorin zu diesen Fragen bereits im Unterkapitel 10. 1. besprochen. Demgemäß wird hier gleich auf Frage 6 eingegangen. Frage 6 wurde vor der Evaluierung beantwortet; Frage 7 wurde danach beantwortet.

Frage 6: *Sind Sie der Meinung, dass Sie einen Text, welcher posteditiert worden ist, von einem unterscheiden können, der „vollständig" übersetzt worden ist?*

Der Evaluatorin standen folgende Antwortmöglichkeiten zur Auswahl: *Nein, Eher nein, Ja, Eher ja* und *Nicht sicher*. Die Evaluatorin traf keine endgültige Entscheidung und kreuzte sowohl *Eher ja* und *Nicht sicher* an. In einer Anmerkung vermutet sie, „dass möglicherweise bei einem posteditierten Text gerade in der Lexik und Syntax Spuren einer etwas unglücklichen maschinellen Übersetzung zurückbleiben könnten."

Während des Evaluierungsprozesses, welcher im nächsten Unterkapitel genauer beschrieben wird, wurde der Evaluatorin nicht mitgeteilt, bei welchen Texten es sich um Übersetzungen und bei welchen es sich um Posteditionen handelt. Im Anschluss an jede Evaluation musste sie ihre Einschätzung bekanntgeben.

Frage 7: *Hat sich Ihre Meinung zu maschineller Übersetzung im Rahmen dieser Studie geändert?*

Die Evaluatorin hatte die Wahl zwischen

- *Ja, ich habe nun eine positivere Meinung zu maschineller Übersetzung.*
- *Nein*
- *Ja, ich habe nun eine negativere Meinung zu maschineller Übersetzung.*

Die Frage wurde schlicht mit einem *Nein* beantwortet – anscheinend hat sich die Meinung der Evaluatorin zu MÜ nicht (signifikant) geändert. Die Antwort wurde nicht weiter von ihr erläutert. Nach der siebten und letzten Frage des Fragebogens war es der Evaluatorin noch möglich, abschließende Bemerkungen hinzuzufügen:

„Ich denke, das gerade was grammatikalische Bezüge in einem Text so wie passende Auswahl von Verben betrifft bei der maschinellen Übersetzung gewisse Schwächen unvermeidlich sind, die dann auch bei der Postedition schwer „auszumerzen" sind, da sie, wenn man den Text schon ‚zu gut' kennt, vielleicht gar nicht mehr auffallen."

10. 3 Die Evaluierungsbögen

Die Evaluatorin erhielt pro ProbandIn jeweils einen Evaluierungsbogen und eine Übersetzung und Postedition. Die 4 Evaluierungsbögen waren ident. Nach genauem Lesen der Zieltexte wurden die Kriterien Grammatik, Lexik, Syntax, Stilistik, Flüssigkeit, Genauigkeit und Vollständigkeit mit Hilfe einer 5-Punkte-Skala bewertet. *Sehr zufriedenstellend* stellte den Wert 1 dar, während *Unzufriedenstellend* eine Wertigkeit von 5 hatte. Somit bedeutet eine **niedrige** Punktezahl eine **bessere** Bewertung und eine **hohe** Punktezahl dementsprechend eine **schlechtere** Bewertung. Während beispielsweise die Kategorie Grammatik eine Messgröße darstellt, die wichtig zur Beurteilung eines Textes und der übersetzerischen Kompetenz allgemein ist, wurden die Kategorien Syntax, Stilistik und Flüssigkeit ausgewählt, weil es sich bei diesen Bereichen um besonders relevante für die literarische Übersetzung handelt. Anmerkungen zu den jeweiligen Beurteilungen der Messwerte waren möglich. Es gab 7 Fragen pro Textteil. Somit bestand jeder Evaluierungsbogen aus insgesamt 14 Fragen. Nummeriert wurden die Fragen von 1a bis 7a und 1b bis 7b. Der Zeitaufwand für jeden Textteil wurde individuell aufgezeichnet. Nach Bewertung aller Kriterien musste die Evaluatorin einschätzen, ob es sich bei den entsprechenden Texten um Posteditionen oder Übersetzungen handelte. Die Evaluatorin hatte Gelegenheit, ihre Entscheidungen zu begründen und/oder zusätzliche Bemerkungen zu schreiben. Der vollständige Evaluierungsbogen findet sich im Anhang. Anmerkung: Der Evaluierungsbogen wurde zwar selbstständig und ohne fremde Hilfe erstellt, aber die von mir verwendeten Kriterien/Messgrößen wurden bereits zuvor in diversen Fehlertypologien verwendet, welche in Kapitel 7. 3 besprochen wurden.

10. 3. 1 Evaluierung von Text 1

Text 1 wurde von TeilnehmerIn 1 erstellt. Für die Evaluierung der Übersetzung hat die Evaluatorin 30 Minuten benötigt, für die der Postedition 20. Die Evaluatorin hat bei Text 1 nicht richtig erkannt, bei welchem Text es sich um die Übersetzung und bei welchem Text es sich um die Postedition handelt.

Die Übersetzung erreichte insgesamt eine Punktezahl von 25. Während die Übersetzung eher unzufriedenstellende bis durchschnittliche Werte in den Kategorien Grammatik, Lexik, Syntax, Stilistik und Flüssigkeit erzielte, wurde die Genauigkeit der Übersetzung sogar als *Unzufriedenstellend* gewertet. Begründet wurde diese Entscheidung mit einer Sinnesveränderung im Zieltext. Lediglich in der Kategorie Vollständigkeit wurde ein positiver Wert erreicht.

Die Postedition erreichte 16 Punkte, also fast 10 Punkte weniger als die Übersetzung. Positiv bewertet wurden die Grammatik, Syntax, Flüssigkeit und Vollständigkeit – die Genauigkeit wurde sogar als sehr zufriedenstellend beurteilt. Nur die Stilistik war *Durchschnittlich* und die Lexik *Eher unzufriedenstellend*.

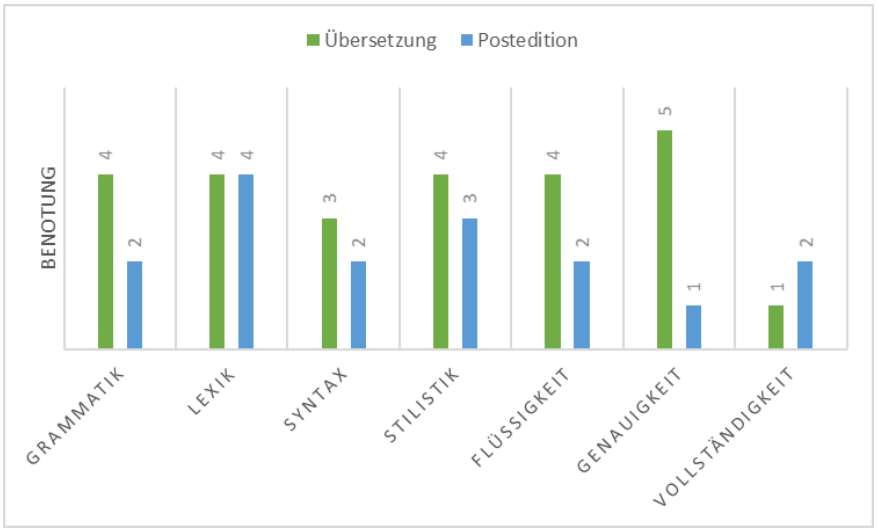

Abbildung 28: Evaluation von Text 1 (eigene Grafik, erstellt am 16.04.2021)

10. 3. 2 Evaluierung von Text 2

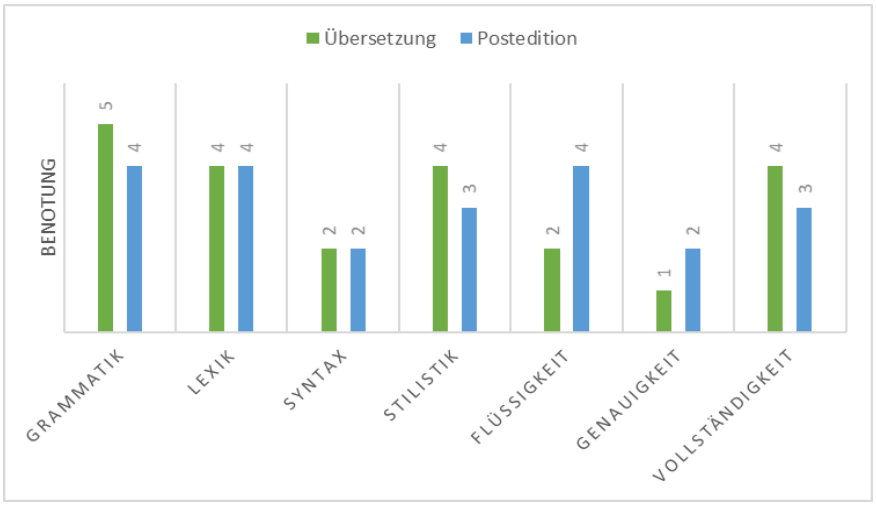

Abbildung 29: Evaluation von Text 2 (eigene Grafik, erstellt am 16.04.2021)

Text 2 wurde von TeilnehmerIn 2 übersetzt und posteditiert. Für die Evaluierung der Übersetzung hat die Evaluatorin 45 Minuten benötigt, für die Postedition 40. Die Evaluatorin hat bei diesem Text die Übersetzung und die Postedition richtig erkannt.

Die Gesamtpunktzahl der Übersetzung liegt bei 22. Die Lexik, Stilistik und Vollständigkeit waren *Eher unzufriedenstellend* und die Grammatik erhielt die schlechteste mögliche Bewertung; Zeiten- und Kommafehler wurden als Begründung für diese negative Beurteilung angeführt. Syntax und Flüssigkeit waren dafür *Zufriedenstellend* und die Genauigkeit erhielt die Bestnote.

Die Postedition von Text 2 erreichte ebenfalls 22 Punkte. Grammatik, Lexik und Flüssigkeit wurden von der Evaluatorin als *Eher unzufriedenstellend* bewertet. Stilistik und Vollständigkeit waren laut ihr *Durchschnittlich* und die Syntax und Genauigkeit waren *Zufriedenstellend*.

10. 3. 3 Evaluierung von Text 3

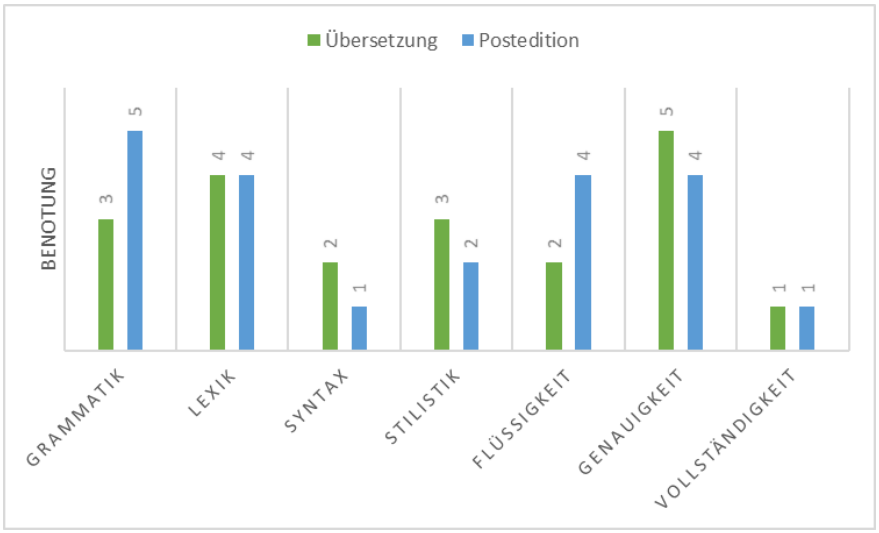

Abbildung 30: Evaluation von Text 3 (eigene Grafik, erstellt am 16.04.2021)

Text 3 wurde von TeilnehmerIn 3 verfasst. Der Zeitaufwand für die Evaluierung umfasste bei der Übersetzung 30 Minuten und bei der Postedition 40 Minuten. Dies ist somit der einzige Text, bei welchem die Evaluation der Postedition mehr Zeit in Anspruch nahm als für die Übersetzung. Auch bei diesem Text erkannte die Evaluatorin die Übersetzung und die Postedition richtig.

Die Übersetzung von Text 3 erzielte 20 Punkte. Syntax und Flüssigkeit wurden beide jeweils mit einer 2 benotet und die Vollständigkeit war laut Evaluatorin sogar *Sehr zufriedenstellend*. Stilistik und Grammatik erreichten beide einen durchschnittlichen Wert von 3 Punkten. Die Lexik war *Eher unzufriedenstellend* und die Genauigkeit war gar

Unzufriedenstellend – diese Beurteilung begründete die Evaluatorin mit einer Sinnesveränderung in einem Satz, weswegen laut ihr eine ganze Szene keinen Sinn ergab.

Die Postedition wurde minimal schlechter bewertet als die Übersetzung und erreichte eine Punkteanzahl von 21. Die Grammatik wurde mit einer 5 bewertet – Komma-, Zeitfehler und Bezugsprobleme führten zu dieser Bewertung. Lexik, Flüssigkeit und Genauigkeit waren ebenfalls *Eher unzufriedenstellend*. Dafür waren sowohl die Syntax als auch die Vollständigkeit *Sehr zufriedenstellend*. Die Stilistik der Postedition war ebenfalls *Zufriedenstellend*.

10. 3. 4 Evaluierung von Text 4

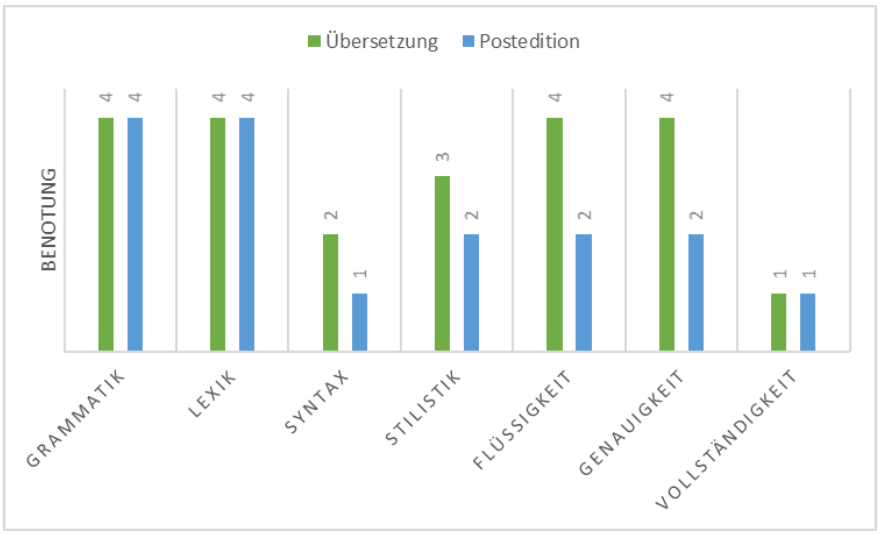

Abbildung 31: Evaluation von Text 4 (eigene Grafik, erstellt am 16.04.2021)

TeilnehmerIn 4 übersetzte und posteditierte Text 4. Mit 25 Minuten wurde knapp unter einer halben Stunde für die Evaluierung von der Übersetzung gebraucht. Für die Postedition wurde genau soviel Zeit benötigt. Übersetzung und Postedition wurden auch bei Text 4 korrekt identifiziert. Somit gelang es der Evaluatorin, alle Texte bis auf Text 1 richtig zu erkennen.

Die Übersetzung dieses Textes erreichte 22 Punkte. Zwar wurde keine einzige Kategorie als *Unzufriedenstellend* eingestuft, aber dafür wurden die Grammatik, Lexik, Flüssigkeit und Genauigkeit allesamt alle als *Eher unzufriedenstellend* bewertet. Bloß die Syntax war *Zufriedenstellend* und die Vollständigkeit *Sehr zufriedenstellend*. Laut der Evaluatorin war die Stilistik *Durchschnittlich.*

Mit 16 Punkten wurde die Postedition von Text 4 signifikant besser bewertet als die Übersetzung. Genauso wie bei der Übersetzung wurden die Grammatik und Lexik bloß mit einer 4 bewertet; allerdings waren dies die einzigen Messgrößen, die bei der Postedition eine schlechte Beurteilung erhielten. Stilistik, Flüssigkeit und Genauigkeit waren der Evaluatorin

zufolge *Zufriedenstellend* und die Syntax und Vollständigkeit des Textes erhielten sogar die beste mögliche Bewertung.

10. 4 Diskussion der Daten

In diesem Unterkapitel wird zunächst der Zeitaufwand der Übersetzungen, Posteditionen und Evaluationen erläutert. Aufgrund des verschwindend geringen Unterschiedes in der Textlänge wird keine Rücksicht darauf genommen, dass Gruppe A Textteil A und Gruppe B Textteil B übersetzt hat – alle Übersetzungen werden also gleich bewertet. Im Anschluss wird auf die Bewertung der Zieltexte eingegangen. Die sieben Kriterien, anhand derer alle Zieltexte evaluiert wurden, werden alle gleich gewertet. Dies bedeutet zum Beispiel, dass einer 1 in der Kategorie Stilistik genau derselbe Wert zugemessen wird wie einer 1 in der Kategorie Grammatik. Im Anschluss werden alle Daten, welche im Zuge dieser Studie erhoben wurden, diskutiert.

10. 4. 1 Der Zeitaufwand

Die ProbandInnen benötigten insgesamt **216 Minuten** für ihre **Übersetzungen**. Somit ergibt sich ein durchschnittlicher Zeitaufwand von genau 54 Minuten pro Übersetzung. Für die Posteditionen war über eine Stunde weniger vonnöten – insgesamt dauerten die **Posteditionen 151 Minuten**. Daraus ergibt sich ein zeitlicher Aufwand von durchschnittlich 37,75 Minuten pro Postedition. Somit benötigten die ProbandInnen für die Übersetzungen im Durchschnitt knapp eine Viertelstunde (um genau zu sein 16,25 Minuten) mehr als für die Postedition. In anderen Worten waren die TestteilnehmerInnen um **30% schneller** (abgerundet von 30,09%) wenn sie posteditierten. Dies ist zwar ein signifikanter Unterschied, doch dieses Ergebnis dürfte niemanden überraschen. Die ProbandInnen beantworteten Frage 14 immerhin einstimmig – alle waren davon überzeugt, dass sie/er schneller gewesen wäre, wenn sie/er den gesamte Text posteditiert hätte. Somit wurde auch die **Hypothese der Einleitung bestätigt** – der Posteditierungsprozess eines maschinell generierten Textes ist weniger zeitaufwändig als eine Humanübersetzung desselben Textes. Interessant sind die individuellen Zeitunterschiede: Bei den Übersetzungen benötigten zwei ProbandInnen genau 60 Minuten, eine/r 54 Minuten und eine Person fertigte ihre/seine in nur 40 Minuten an, also 20 Minuten schneller als die langsamsten ÜbersetzerInnen. Deutlich größer waren die Unterschiede bei den Posteditionen. Hier wurden jeweils einmal 16, 20, 45 und 70 Minuten gebraucht. Die beiden Personen, die die Übersetzungen in 60 Minuten produzierten, benötigten auch bei den Posteditionen mehr Zeit als die anderen TeilnehmerInnen: sie waren in 70 und 45 Minuten fertig. Die/Der schnellste ÜbersetzerIn war auch die/der schnellste PosteditorIn: sie/er war in 40 Minuten mit der

Übersetzung fertig und in 16 Minuten mit der Postedition. Fast genauso schnell mit der Postedition war jene Person, die 56 Minuten für die Übersetzung brauchte – sie/er war in 20 Minuten fertig. Somit ist der individuelle Zeitunterschied bei den Posteditionen beträchtlicher als bei den Übersetzungen.

Dieses wohl wenig überraschende Ergebnis deckt sich auch mit Forschungsergebnissen anderer Studien. Laut Guerras (2003) Studie könnten ÜbersetzerInnen pro Tag bis zu 2 Stunden und 45 Minuten einsparen, wenn sie (ausgehend von einer durchschnittlichen Übersetzungsleistung von 2000-3000 Wörtern pro Arbeitstag) posteditieren, statt nur zu übersetzen. Besaciers (2015) Einschätzung zufolge wurde mittels einer Postedition (im Vergleich zu einer Humanübersetzung) eine Zeitersparnis von circa 50% erreicht. Auch bei Nietzke (2018) benötigten ProbandInnen mehr Zeit für das „reine" Übersetzen als für das Posteditieren. Nur bei Besaciers (2015) Studie wurden literarische Texte behandelt, bei den anderen beiden Studien wurden nicht literarische Texte übersetzt/posteditiert.

Beim Zeitaufwand für die Evaluationen zeigt sich hingegen ein anderes Bild. Für die Bewertung aller Übersetzungen benötigte die Evaluatorin 130 Minuten und für die Posteditionen 125. Also wurden im Schnitt bloß 5 Minuten mehr für die Übersetzungen benötigt, das ist ein zeitlicher Mehraufwand von 4%. Für die Evaluierung der einzelnen Texte benötigte die Evaluatorin jeweils zwischen mindestens 20 und maximal 40 Minuten. Eine mögliche Erklärung für diesen geringen Unterschied könnte sein, dass der Aufwand, welcher für eine Evaluierung notwendig ist, unabhängig von der Qualität des Textes ist. Somit wäre die Evaluierung von einem qualitativ hochwertigen Text genauso zeitaufwändig wie die eines „schlechten" Textes. Demnach wären wahrscheinlich vor allem der Umfang, Inhalt und Stil eines Textes ausschlaggebend für die Evaluationszeit. Da sich die Texte in diesen Aspekten kaum unterschieden, würde dies erklären, warum der zeitliche Aufwand für die Posteditionen und Übersetzungen fast ident sind. Andere mögliche Ursachen sind allerdings nicht auszuschließen. Beispielsweise könnte die Kürze der Texte Auswirkungen darauf haben, dass sich Qualitätsunterschiede nicht so niedergeschlagen haben, wie das in umfangreicheren Texten der Fall wäre. Aufgrund der kleinen Studiengröße kann es außerdem sein, dass der aufgezeichnete Zeitaufwand der Evaluationen von der Norm abweicht – verallgemeinerbare Ergebnisse wären bei einer umfassenderen Studie mit mehr ProbandInnen und EvaluatorInnen möglich.

10. 4. 2 Die Qualität der Zieltexte

Werden alle „Noten“ zusammengezählt, welche die Übersetzungen erhielten, ergibt sich eine Gesamtzahl von 89. Dies bedeutet eine durchschnittliche Punktezahl von 22,25 pro Übersetzung. Bricht man diese Zahl auf alle 7 Kriterien herunter, ergibt sich eine Gesamtnote von 3,18. Dies bedeute auf der Skala, welche für diese Arbeit verwendet wurde, dass die Übersetzungen im Schnitt *Durchschnittlich* waren.

Bei den Posteditionen ergibt sich eine Punktezahl von insgesamt 75. Pro Postedition ergibt dies 18,75 Punkte. Dies bedeutet, dass jede Postedition im Schnitt eine Note von 2,67 erhielt. Somit ergibt sich ein Wert, welcher sich relativ genau zwischen *Zufriedenstellend* und *Durchschnittlich* einpendelt, mit einer leichten Tendenz zu *Durchschnittlich*. Im direkten Vergleich lässt sich darauf schließen, dass die Posteditionen um über einen halben Notengrad, beziehungsweise um **16%** (abgerundet von 16,04%) **besser** von der Evaluatorin **bewertet** worden sind als die Übersetzungen. Dies ist ein beträchtlicher und zugleich überraschender Unterschied. 3 der 4 ProbandInnen gaben zuvor im Fragebogen an, dass sie einen besseren Text hätten produzieren können, wenn sie die gesamte Geschichte übersetzt hätten. Außerdem meinten alle ProbandInnen *Zufrieden* mit ihren Übersetzungen zu sein, während bloß 3 der 4 dieselbe Meinung zu ihren Posteditionen hatten. Die ProbandInnen schätzen ihre Übersetzungen also fälschlicherweise besser ein als ihre Posteditionen.

Werden diese Daten mit den Erkenntnissen aus dem vorherigen Unterkapitel kombiniert, wird das Ergebnis noch deutlicher. Im Zuge dieser Studie wurden die Posteditionen signifikant besser von der Evaluatorin beurteilt als die Übersetzungen, obwohl diese im Schnitt um 30% schneller produziert worden sind. Dieses Ergebnis widerlegt die Hypothese der vorliegenden Arbeit – eingangs wurde nämlich davon ausgegangen, dass Posteditionen zwar weniger Zeitaufwand benötigen, dafür aber qualitativ schlechter sind als Humanübersetzungen. Somit wurden auch Erkenntnisse jener Studien bestätigt, welche zuvor in Kapitel 6.3 und 8. 3 behandelt wurden. Obwohl in Ways und Torals Studie (2018) das Posteditieren nicht Teil ihres Experimente waren, widerlegen ihre Studienergebnisse die Hypothese insofern, da in ihrer Forschungsarbeit Texte, welche „nur“ von einem MÜ-System produziert wurden, in bis zu 34% der Fälle bereits so zufriedenstellend waren wie die von HumanübersetzerInnen. Dieses Ergebnis widerlegt nämlich direkt eine zentrale Annahme der vorliegenden Arbeit – dass maschinelle Übersetzungssysteme nicht so qualitativ hochwertige Übersetzungen produzieren können wie HumanübersetzerInnen. Denn wenn das Produkt von MÜ-Systemen bereits ein so hohes Niveau erreicht, würde das einen anschließenden Posteditierungsprozess überflüssig machen. Allerdings wurde ein solches Qualitätsniveau nicht immer erreicht – also ist davon auszugehen, dass das MÜ-Produkt in einem Großteil der Fälle von einer/m PosteditorIn hätte aufgewertet werden können.

Bei den Studien von Bowker und Ehgoetz (2007) und Fiederer und O'Brien (2009) schnitten die Posteditionen ebenfalls besser, oder zumindest genauso gut ab wie Humanübersetzungen. Allerdings sei darauf hingewiesen, dass bei ersterer Studie Verwaltungsakten und bei letzterer ein Benutzerhandbuch übersetzt wurde; also wurde bei keiner der beiden Studien literarische Texte übersetzt.

Zusammenfassend lässt sich also sagen, dass das Ergebnis, dass Posteditionen besser abschneiden als Humanübersetzungen, kein Novum in der MÜ-Forschung darstellt. Doch wie lässt sich dieses Ergebnis erklären? Laut diesem Ergebnis würde nämlich wenig bis gar nichts dagegen sprechen in Zukunft alle Textsorten zuerst maschinell zu übersetzen und anschließend zu posteditieren, da eine solche Arbeitsweise anscheinend nicht nur zeiteffizienter ist, sondern auch qualitativ hochwertigere Texte produziert. Eine potentielle Ursache ist die mangelnde Erfahrung der ProbandInnen der vorliegenden Studie – da es sich „bloß" um Studierende handelt, kann es sein, dass andere Ergebnisse zustande gekommen wären, wenn die empirische Studie mit professionellen ÜbersetzerInnen wiederholt werden würde, welche auf eine jahrelange Erfahrung als LiteraturübersetzerInnen zurückblicken können. Bei der Untersuchung der Messgrößen im folgenden Unterkapitel wird nochmal verdeutlicht, welchen Einfluss die Auswahl der ProbandInnen hatte. Eine gründlichere Untersuchung und Erklärung der Forschungsergebnisse finden sich in Kapitel 10. 5.

10. 4. 3 Untersuchung der Messgrößen

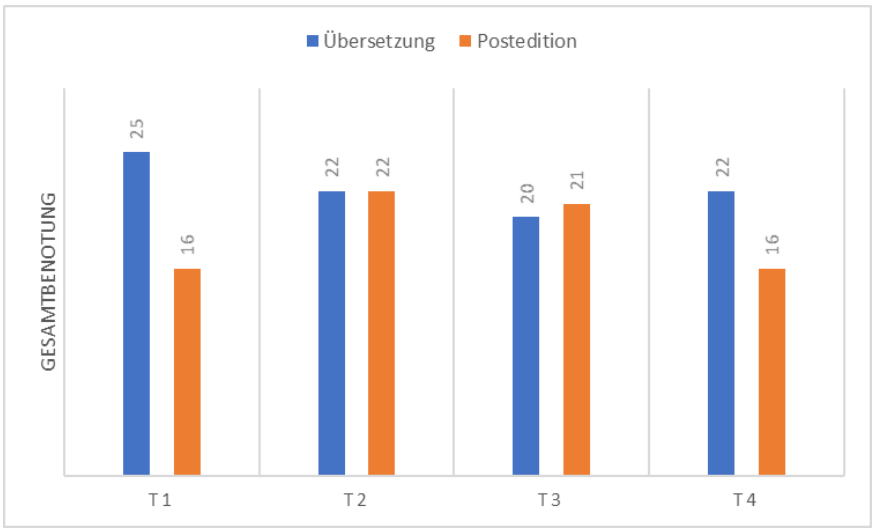

Abbildung 32: Evaluierung der Zieltexte im Vergleich. Die Zahlen stellen die Gesamtbenotung der Texte dar – je niedriger die Benotung, desto besser wurde der Zieltext bewertet (eigene Grafik, erstellt am 16.04.2021)

In diesem Unterkapitel wird auf die individuellen Fehler eines jeden Textes eingegangen und versucht, Beziehungen zu der Gesamtbewertung des jeweiligen Textes zu finden. Außerdem werden die häufigsten Fehler diskutiert, welche in den Texten vorkamen. Es wird nach den Kategorien, welche im Evaluierungsbogen verwendet wurden, vorgegangen. Der zeitliche

Aufwand wird hier nicht berücksichtigt. Zur Erinnerung: Alle Kriterien wurden mittels einer Ordinalskala bewertet. Die Evaluatorin hatte die Möglichkeit, Anmerkungen zu jedem bewerteten Aspekt anzugeben. Da in der vorliegenden Arbeit insbesondere das Übersetzen von literarischen Texten im Vordergrund steht, wird ein besonderes Augenmerk auf typische Problemfelder dieses Übersetzungsbereiches gelegt.

Grammatik/Orthographie

Die Frage im Evaluierungsbogen lautete:

„Wie beurteilen Sie die grammatikalische und orthographische Qualität des Textes?“

Die Übersetzungen erzielten in der Kategorie insgesamt 16 Punkte, was einen durchschnittliche Benotung von exakt 4 ergibt. Bei den Posteditionen sind es 15 und daraus ergibt sich die Note 3,75 – somit schnitten die Posteditionen in dieser Kategorie um 6,25% besser ab. Am häufigsten waren Zeitenfehler in den Übersetzungen – in 3 von 4 Fällen wurde mindestens ein Zeitfehler von der Evaluatorin angemerkt. Beispielsweise fehlt laut der Evaluatorin in folgender Textpassage (Auszug aus Text 2a; Übersetzung; Benotung der Grammatik: 5) mehrmals die Vorvergangenheit:

*„Das muss der Grund für ihren Anruf **sein**, als sie Marie **fragte**, ob sie vielleicht ein paar Eier übrig hätte, da ihre Hennen aufgehört hatten zu legen. Verständlicherweise wollte sie wissen, warum die Polizei auf Maries Grundstück ausgeschwärmt war und warum der Helikopter das Feld hinter ihrem Haus **umkreiste**.“*

Während es bei den Übersetzungen Zeitfehler waren, wurden bei den Posteditionen vor allem Bezugsfehler markiert. 3 von 4 Posteditionen wiesen Bezugsfehler auf (ein häufiger Fehler bei maschinell generierten Übersetzungen), ansonsten wurden Präpositionsfehler am öftesten angemerkt. Markierte Fehler waren unter anderem:

„*Die Tür hatte ein Fenster, das von einem Spitzenvorhang verdeckt war, und als sie **darauf** schaute, sah sie den Schatten eines Kopfes, der sich langsam dahinter erhob, als hätte die Person dort gehockt und auf den richtigen Moment gewartet, um ins Zimmer zu schauen*.“

„*Sie rannte zur Tür und schrie: "Verschwinde!", und die Person verschwand **aus dem Fenster***.“

*„Sie lag noch eine Weile still, **wachsam für** jegliche ungewöhnlichen Geräusche.“*

(Auszüge aus Text 2b, Postedition: Benotung der Grammatik: 4)

Wie sind diese Ergebnisse zu interpretieren? Bei den Übersetzungen fanden sich einige Zeitfehler – unter Umständen ist diese Anfälligkeit für Zeitfehler auf das Profil der

ProbandInnen zurückzuführen. Es ist anzunehmen, dass erfahrene, professionelle ÜbersetzerInnen nicht dazu tendieren, diese Art von Fehler zu begehen. Dies könnte ein Hinweis darauf sein, dass zur Vermeidung von solchen Fehlern dieser Thematik mehr Aufmerksamkeit im universitären Unterricht gewidmet werden sollte.

Lexik

Die Frage im Evaluierungsbogen lautete:

„Wie beurteilen Sie die lexikalische Qualität des Textes?“

Sowohl die Übersetzungen als auch die Posteditionen erzielten insgesamt 16 Punkte. Daraus ergibt sich eine durchschnittliche Note von 4 für die Lexik aller Texte. Bei den Übersetzungen führten „holprige und immer wieder unpassende Ausdrücke“ zu dieser Benotung. Bei folgendem Beispielsatz ist es der Evaluatorin etwa unverständlich, wieso ein Anglizismus verwendet wird.

*„Er will Heu pflanzen und hat all das **Equipment**, das er dafür benötigt.“*

(Auszug aus Text 1a, Übersetzung, Benotung der Lexik: 4)

*„Ich lass‘ Sie wissen, bevor wir zum **Pflanzen** kommen“, sagte er Marie.*

(Auszug aus Text 4b, Übersetzung, Benotung der Lexik: 4)

Unpassende Wortwahl laut der Evaluatorin - bei diesem Satz sollte statt dem Wort *„pflanzen“* das Verb *„säen“* stehen.

Ähnlich wie bei den Übersetzungen, wiesen auch die Posteditionen „ein paar Schwächen und Fehler auf“.

*„Sie gaben sich die Hand, und er sprang in seinen **Traktor** und fuhr los.“*

(Auszug aus Text 1b, Postedition, Benotung der Lexik: 4)

Dieser Satz wurde angemerkt, weil im Quellentext von einem Truck die Rede ist und dieser in der Postedition fälschlicherweise zu einem Traktor wird.

*„Ich habe jemanden gefunden, der dein Grundstück **mieten** will“, sagte Patty.*

(Auszug aus Text 3a, Postedition, Note 4)

Es sollte es laut der Evaluatorin *„pachten“* anstelle von *„mieten“* sein. Dies ergibt sich aus dem Kontext, da anzunehmen ist, dass das Grundstück für eine Betriebsfortführung genutzt wird.

Auch bei dieser Messgröße hätten andere TestteilnehmerInnen womöglich bessere Ergebnisse liefern können. Interessant ist, dass die Posteditionen in dieser Hinsicht genauso gut bewertet wurden wie die Übersetzungen. Vielleicht ist dies ein Hinweis darauf, dass die lexikalische Qualität bei Posteditionen abhängig vom Wortschatz der/s PosteditorIn ist.

Syntax

Die Frage im Evaluierungsbogen lautete:

„Wie beurteilen Sie die syntaktische Qualität des Textes?“

Mit einer Gesamtbenotung von 9 wurde die Syntax der Übersetzungen besser bewertet als die vorherigen Kriterien. Dies ergibt einen Durchschnitt von 2,25. Allerdings wurden die Posteditionen noch besser benotet – im Durchschnitt wurde die syntaktische Qualität der posteditierten Texte mit einer überragenden 1,5 benotet. Somit schnitten die Posteditionen in dieser Kategorie um 33,33% besser ab. Mit der Syntax der Übersetzungen war die Evaluatorin grundsätzlich zufrieden – Fehler per se fand sie laut eigenem Angaben zwar nicht, aber sie hat einige „ungewöhnliche“ Stellen markiert, die durch eine Umstellung idiomatischer klingen würden. An dieser Stelle sei gesagt, dass, wie die Frage bereits andeutet, es nicht ausschließlich um Syntax-Fehler im eigentlichen Sinne geht. In der Folge werden auch Sätze hervorgehoben, die zwar nicht falsch sind, sondern von der Evaluatorin anders formuliert worden wären. Selbstverständlich sei hier wieder auf die Subjektivität der Evaluatorin verwiesen.

„Patty gab die Eier in eine Plastiktüte und ***schob sie rüber zu Marie****.“*

Bei diesem Satz hätte die Evaluatorin *„…schob sie zu Marie rüber.“* bevorzugt.

*„****Nichts kann die*** *aus der Ruhe bringen.“*

(Auszüge aus Text 2a, Postedition, Benotung der Syntax: 2)

Ähnlich wie beim vorherigen Beispiel hat die Evaluatorin bloß etwas an der Satzstellung auszusetzen. Sie hätte „*Die kann nichts aus der Ruhe bringen*.“ geschrieben.

Wie die hervorragende Benotung vermuten lässt, fanden sich bei den Posteditionen ebenfalls kaum Fehler in der Syntax und nur wenige verbesserungswürdige Stellen.

*„Als sie zu Maries Haus zurückkamen, lud sie alle **ins Haus** auf einen Kaffee ein, aber Joe musste zurück zu seiner Arbeit.“*

(Auszug aus Text 2b, Postedition, Benotung der Syntax: 2)

Hier schlug die Evaluatorin zwei Alternativen vor: Entweder umdrehen und *„auf einen Kaffee ins Haus“* schreiben oder auf das *„ins Haus“* verzichten. Bei diesen „Korrekturen“ handelt es sich, wie bei der Stilistik (die nächste Messgröße, die untersucht wird), um typische Problemfelder der literarischen Übersetzung. Wie zuvor handelt es sich hierbei nicht um einen Fehler, sondern um einen Verbesserungsvorschlag, der bei fachsprachlichen Texten wie zum Beispiel medizinischen, technischen, juristischen etc. nicht notwendig wäre, da bei diesen das sogenannte Leseerlebnis nicht im Vordergrund steht.

Stilistik

Die Frage im Evaluierungsbogen lautete:

„Wie beurteilen Sie die stilistische Qualität des Textes? Vermittelt der Text den Eindruck von stilistischer Kohärenz? Wie gut ist das sprachliche Register getroffen?“

Mit einer Gesamtbewertung von 14 und einer durchschnittlichen Benotung von 3,5 wurde die Stilistik der Übersetzungen knapp besser als die Messgrößen Grammatik und Lexik beurteilt. Die Stilistik der Posteditionen erhielt insgesamt 10 Punkte, was einen Notendurchschnitt von 2,5 ergibt – somit wurde dieser Aspekt der Posteditionen um genau einen ganzen Notengrad, beziehungsweise um 28,5% (abgerundet von 28,57143%), besser bewertet. Laut der Evaluatorin wies jede der Übersetzungen zumindest stellenweise ein, für diesen Text, zu hohes Sprachregister auf. Es ist anzunehmen, dass ÜbersetzerInnen mit mehr translatorischer Erfahrung als Studierende diesen „Fehler“ eher nicht beziehungsweise gar nicht begangen hätten.

*„Marie war eine zierliche Frau, heute **jedoch** fühlte sie sich groß und ungeschickt, vielleicht **aufgrund all der Dinge**, die sie in letzter Zeit durchmachen musste.“*

*„Mein Nachbar Larry hat Rinder und **nun** auch einen Wasserbüffel.“*

(Auszüge aus Text 2a, Übersetzung, Benotung der Stilistik: 4)

Auch bei den Posteditionen wurde teilweise auf ein zu hohes Register zurückgegriffen, doch ansonsten wurde das Register grundsätzlich gut getroffen. Einige Schwächen wurden dennoch in Text 1b (Postedition; Benotung der Stilistik: 3) geortet.

*„Nachdem sie sich ein Nachthemd angezogen und die Zähne geputzt hatte, ging sie zurück ins Wohnzimmer und **zappte** sich durch die Kanäle des Fernsehers.“*

„Zappen“ stört laut der Evaluatorin den Lesefluss, sie empfiehlt *„durchschalten“.*

*„Sie rannte zur Tür und schrie: "**Verschwinde**!", und die Person **verschwand** aus dem Fenster.“*

Wortwiederholung von *„verschwinden“.* Die Evaluatorin würde *„Verschwinde!“* mit *„Hau ab“* oder ähnlichem ersetzen. Auch bei diesen Sätzen, wie in der zuvor besprochenen Kategorie der Syntax, handelt es sich nicht um falsche Sätze oder Fehler im eigentlichen Sinne. In der Tat vermied die Evaluatorin bei ihren Anmerkungen das Wort „falsch“ und empfahl bloß alternative Ausdrücke. Somit handelt es sich hierbei vor allem um persönliche Präferenzen der Evaluatorin, die die stilistische Qualität der Texte verbessern sollten.

Flüssigkeit

Die Frage im Evaluierungsbogen lautete:

„Wie flüssig bzw. gut lesbar ist der Text Ihrer Meinung nach?“

Die Flüssigkeit der Übersetzungen wurde insgesamt als durchschnittlich bewertet – mit einer Gesamtpunkteanzahl von 12 ergibt sich eine Note von 3 für alle Übersetzungen. Dasselbe gilt für die Posteditionen; genau, wie dies bereits bei der Kategorie Lexik der Fall war, schnitten die Posteditionen ebenso gut ab wie die Übersetzungen. Vor allem Syntaxschwierigkeiten und fehlende, beziehungsweise überflüssige Konjunktionen behinderten den Lesefluss bei den Übersetzungen.

*„Sie lag **für** eine Weile ruhig da, lauschte **nach** ungewöhnlichen Geräuschen.“*

(Auszug aus Text 3b, Übersetzung, Benotung der Flüssigkeit: 4)

Bei den Posteditionen empfand die Evaluatorin die teilweise „schräge“ Wortwahl als störend für den Lesefluss. Ferner „stolpert“ man laut ihr in einigen Fällen auch über „unpassende" Verben. Leider gab die Evaluatorin weder Beispiele für die schräge Wortwahl noch für die unpassenden Verben.

Wiedergabe der Bedeutung: Genauigkeit bzw. Angemessenheit

Die Frage im Evaluierungsbogen lautete:

„Wie gut bzw. genau wurden die Inhalte, wurde die Bedeutung des Textes wiedergegeben? Wurden Bedeutungen verändert, verzerrt?“

Die Angemessenheit von gleich 2 Übersetzungen wurde mit einer 5 benotet – insgesamt erzielten die Übersetzungen eine Punktezahl von 15, was pro Übersetzung einen Schnitt von 3,75 ergibt. Mit einer Gesamtpunktezahl von bloß 9 schnitten die Posteditionen in diesem Aspekt deutlich besser ab. Mit einem Durschnitt von 2,25 wurde die Genauigkeit der Posteditionen um genau eineinhalb Notengrade, beziehungsweise um beträchtliche 40% besser benotet. Vor allem Sinnesveränderungen im Zieltext waren ausschlaggebend für die beiden negativen Bewertungen der Angemessenheit bei den Übersetzungen.

*„Ich habe jemanden gefunden, um deinen Grund zu **verpachten**“, erklärte sie.*

(Auszug aus Text 1a, Übersetzung, Benotung der Genauigkeit: 5)

Kommentar der Evaluatorin dazu: „Im ZT steht jemand würde die Aufgabe des Verpachtens übernehmen, im AT aber ist die Rede davon, dass jemand den Grund pachten will. Letztendlich wird schon klar was gemeint ist, aber eigentlich ist der Rest dieser Szene damit sinnentleert.“

*„Als die **Jungen** zurück zu Maries Haus kamen, lud sie sie auf eine Tasse Kaffee ein, Joe hatte jedoch eine berufliche Verpflichtung, der er nachkommen musste.“*

(Auszug aus Text 3b, Übersetzung, Benotung der Genauigkeit: 5)

Das *„they“* des Ausgangstextes wird zu *„Jungen“*, was aber falsch ist, da sich das *„they“* auf die Akteure Marie, Joe, Patty und Tom bezieht und nicht auf die Söhne von Joe. Somit liegt erneut eine Sinnesveränderung vor, wodurch die restliche Szene im Zieltext keinen Sinn ergibt.

Die Bedeutung und Wirkung des Textes wurden in 3 der 4 Posteditionen adäquat wiedergegeben, was sich auch in den Noten widerspiegelt. Nur in einer einzigen Postedition gab es einige Passagen, welche nicht genau die Bedeutung des Ausgangstextes wiedergaben.

*„Sie fühlte sich bleiern und unbehaglich, als sie in Nancys Wohnzimmer saß, dem **grauenhaften, lauten** Ticken der Uhr auf dem Kaminsims lauschte und darauf wartete, dass Nancy die Eier, die Marie ihr mitgebracht hatte, in den Kühlschrank stellte.“*

(Auszug aus Text 3a, Postedition, Benotung der Genauigkeit: 4)

Im Zieltext ist es ein *„grauenhaftes, lautes Ticken“*, während im Ausgangstext *„grauenhaft lautes Ticken“* steht. Dies könnte auch als grammatikalischer Fehler gewertet werden (Adjektiv statt Adverb). Außerdem sei an dieser Stelle darauf hingewiesen, dass grammatikalische Fehler auch oft Sinnveränderungen nach sich ziehen können.

Vollständigkeit

Die Frage im Evaluierungsbogen lautete:

„Wurden Dinge ausgelassen/hinzugefügt?“

Keine andere Messgröße der Übersetzungen wurde so gut benotet wie die der Vollständigkeit. Mit einer Gesamtpunktzahl von 7 ergibt sich eine Durchschnittsnote von 1,75. Selbiges gilt für die Posteditionen – mit einer Punktezahl von ebenfalls 7 wurde derselbe Schnitt wie bei den Übersetzungen erreicht. Anders als bei den Übersetzungen ist dies aber nicht das am besten bewertete Bewertungskriterium – die Syntax wurde besser benotet mit einem Schnitt von 1,50. Drei der Übersetzungen erhielten in der Kategorie Vollständigkeit die Bestnote. Eine Übersetzung wurde mit einer 4 benotet – für diese schlechte Bewertung waren Probleme mit Inquit-Formeln ausschlaggebend.

„Marie war eine zierliche Frau, heute jedoch fühlte sie sich groß und ungeschickt, vielleicht aufgrund all der Dinge, die sie in letzter Zeit durchmachen musste.“

(Auszug aus Text 2a; Übersetzung, Benotung der Vollständigkeit: 4)

Bei diesem Satz wurde die Inquit-Formel *„she thought“* aus dem Ausgangstext gänzlich ausgelassen – der Satz lautet im Original: „*Marie was a petite woman, but she felt large and clumsy today, perhaps, **she thought**, because of everything she'd been through lately*".

*„Den anderen Rindern war das völlig egal“, **meinte** sie.*

*„Joe Leto“, **antwortete** Patty.*

*„Ich kenne ihn nicht“, **entgegnete** Marie.*

Bei diesen drei Beispiel wird deutlich, dass die Inquit-Formeln verändert wurden. Im Ausgangstext wurde an diesen Stellen stets *„said“* verwendet. Eine übermäßige Verwendung von *„said“* ist ein häufig zu beobachtendes Phänomen bei englischsprachiger Literatur. Im deutschsprachigen Raum würde eine dauernde Wiederholung von sagte/gesagt als störend empfunden werden. Während auch in der englischen Sprachen Wortwiederholungen im Normalfall vermieden werden sollten, wird für *„said“* anscheinend eine Ausnahme gemacht. Eine mögliche Erklärung dafür ist, dass davon ausgegangen wird, dass LeserInnen, die

beispielsweise in einen Roman „eingetaucht“ sind, die Inquit-Formel mit *„said“* gar nicht registrieren sollen, da die Verwendung von immer wechselnden Formeln ebendieses Lesevergnügen beeinträchtigen könnte (vgl. Trupkiewicz 2014). Ob also der Vollständigkeit halber stets *„said“* mit *„sagte“* übersetzt oder mit alternativen Ausdrücken ersetzt werden sollte, ist stets der/m jeweiligen ÜbersetzerIn überlassen; die Entscheidung variiert jedoch gewiss auch abhängig von genauer Textsorte, dem Kontext und persönlichen Präferenzen.

Laut der Evaluatorin gab es bei den Posteditionen im Gegensatz zu den Übersetzungen keine „Probleme“ mit den Inquit-Formeln, dafür aber einige Auslassungen und Hinzufügungen. Die folgenden Textpassagen sind allesamt Auszüge aus Text 2b (Postedition, Benotung der Vollständigkeit: 3).

*"Sie hatte **noch** nicht alle Details gehört, also wusste sie nicht, wo er verhaftet worden war.“*

Bei diesem Satz wurde das *„noch“* im Zieltext hinzugefügt, somit kommt es zu einer leichten Verzerrung des Sinnes.

*„Sie machte sich ein Käsesandwich und aß es vor dem Fernseher, **in der Hoffnung**, dass es am Ende der Nachrichten eine Wiederholung geben würde und sie mehr über den Autodieb hören würde, aber es wurde nichts mehr über ihn gesagt.“*

„In der Hoffnung“ steht nicht im Ausgangstext, hat aber keine wirkliche Sinnesveränderung zur Folge. Im Ausgangstext hieß es: „*She made a cheese sandwich and ate it in front of the TV, thinking there might be a recap at the end of the news and she would hear more about the car thief, but nothing more was said about him*.“

„Dann ging sie zu einem Fenster, das einen besseren Blick auf die Veranda und den Vorgarten bot.“

Der Satz im Ausgangstext lautet: *„Then she went to a window, not the one in the door, but one that afforded a better view of the porch and the front yard”*, somit wurde *“not the one in the door”* ausgelassen.

10. 4. 4 Beziehungen und Ähnlichkeiten mit früheren Studien

In Folge werden einige Studien, welche zuvor besprochen wurden, aufgrund von Ähnlichkeiten, Parallelen oder anderen Beziehungen zur Studie der vorliegenden Arbeit hervorgehoben.

Bei Besaciers (2015) Studie aus dem Jahre 2015 wurde ebenfalls eine noch nicht übersetzte Kurzgeschichte zuerst von einem MÜ-System „rohübersetzt“ und anschließend von einer/m StudentIn der Translation, posteditiert. Die Originalsprache der Kurzgeschichte war Englisch und übersetzt wurde ins Französische. Bei der Auswahl der Kurzgeschichte wurde bewusst eine ausgewählt, welche „Gemeinsamkeiten mit wissenschaftlichen und technischen Texten“ aufweist, da Besacier davon ausging, dass dadurch die „Kluft“ zwischen wissenschaftlichen und literarischen Texten beim Übersetzungsprozess geringer werden würde. Benutzt wurde ein phrasenbasiertes MÜ-System. Nach der Postedition wurde der Text von neun französischen MuttersprachlerInnen gelesen, welche in der Folge einen Fragebogen ausfüllten. Der Text wurde insgesamt als gut lesbar, verständlich und fehlerarm beurteilt. Die Meinung eines zehnten Lesers, dem eigentlichen Übersetzer der Werke des Autors, stimmte mit den Meinungen der anderen LeserInnen größtenteils überein. Der Übersetzer merkte an, dass: "es natürlich Unvollkommenheiten, ungeschickte Ausdrücke und spezifische Fehler gab, die korrigiert werden müssen". Benötigt wurden für den Posteditionsprozess 25 Stunden –der Übersetzer des Autors meinte, dass er zumindest doppelt so lange gebraucht hätte um „from scratch“ zu übersetzen. Somit geht Besacier davon aus, dass eine Zeitersparnis von 50% mittels Postediting erreicht werden könne. In der Conclusio meint Besacier außerdem, dass diese Methode kostengünstige Übersetzungen ermöglichen könnte, welche den Mehrwert für AutorInnen bietet, ihre Werke rasch in andere Sprachen übersetzen lassen zu können – mit dem Nachteil, dass geringe Qualitätsverluste, im Vergleich zu einer Humanübersetzung, mit dieser Entscheidung einhergehen würden.

Die Ergebnisse von Besaciers Studie hätten die Hypothese der vorliegenden Arbeit bestätigt – der Posteditionsprozess spart zwar eine beträchtliche Menge an Zeit, aber eine Humanübersetzung wäre qualitativ besser gewesen. Hervorzuheben ist hier allerdings, dass in der Studie von Besacier kein vollständiger menschlicher Übersetzungsprozess stattgefunden hat, welcher genaueren Aufschluss darüber gegeben hätte, wieviel Zeit genau durch das Postediting eingespart wurde und wo, beziehungsweise um wieviel die Humanübersetzung besser als die Postedition abschnitt.

Zu anderen Ergebnissen kamen die Studien von Bowker und Ehgoetz (2007) und Fiederer und O’Brien (2009). Bei beiden Studien wurden Posteditionen mit Humanübersetzungen verglichen. Die Posteditionen beider Studien erzielten, genauso wie die Posteditionen in der vorliegenden Arbeit, bessere Ergebnisse als die Humanübersetzungen. Zwar wurden bei diesen Arbeiten Verwaltungstexte und ein Benutzerhandbuch, also keine literarischen Texte, übersetzt, jedoch zeigen diese Ergebnisse einmal mehr, dass eine bessere Bewertung von Posteditionen gegenüber menschlichen Übersetzungen kein Novum in der Forschung von Postedition von MÜ darstellt.

In den Studien von Toral und Way (2018) und Matusov (2019) wurden keine Posteditionen durchgeführt. Bei beiden Studien wurden literarische Texte „nur“ maschinell übersetzt und im Anschluss, wie in der vorliegenden Arbeit, menschlich evaluiert. Das besondere bei Toral und Ways Studie (2018) ist, dass sogar nur maschinell generierte Übersetzungen (ohne danach posteditiert worden zu sein) in bis zu 34% der Fälle besser bewertet wurden als menschliche Übersetzungen. Matusov hingegen gab zwar an, dass nur 30% der untersuchten Übersetzungen als „akzeptabel“ eingestuft worden waren, aber anscheinend war die Qualität oftmals „hoch genug, um die Geschichte zu genießen“. Matusov (2019) empfiehlt auf Basis dieser Erkenntnisse genauso wie Besacier die Verwendung von MÜ für LiteraturübersetzerInnen und verweist ebenfalls auf den Vorteil, Neuerscheinungen beziehungsweise noch nicht übersetzte Werke rasch für viele LeserInnen (online) in zusätzlichen Sprachen verfügbar machen zu können.

Die Studie von Guerberof-Arenas und Toral (2020) hatte eine gänzlich andere Herangehensweise. Im Gegensatz zu den eben besprochenen Studien, stand die Messgröße der Kreativität im Mittelpunkt in der Arbeit von Guerberof-Arenas und Toral. Bei deren Arbeit wurde versucht, die Kreativität einer maschinell generierten Übersetzung, einer Postedition dieser maschinellen Übersetzung und einer reinen Humanübersetzung zu messen und miteinander zu vergleichen. Dieser „Kreativitätswert“ war bei den Humanübersetzungen am höchsten. Die AutorInnen dieser Studie gehen daher davon aus, dass Übersetzungen kreativer sind, wenn sie ohne (maschinelle) Hilfsmittel produziert werden. Außerdem sehen Guerberof-Arenas und Toral eine Korrelation zwischen Kreativität und Lesevergnügen. Obwohl die Humanübersetzung in den Kategorien „narratives Engagement“ und „Übersetzungsrezeption“ (im Original: *narrative engagement* und *translation reception*) am besten abschnitt, war der Lesegenuss der ProbandInnen der Studie bei der Lektüre der Posteditionen ein wenig höher, beziehungsweise mindestens genauso hoch wie bei der Lektüre der Humanübersetzungen. Daraus könnte geschlussfolgert werden, dass auch bei dieser Studie die Posteditionen zumindest genauso gut wie die reinen menschlichen Übersetzungen abschnitten.

Im anschließenden Kapitel werden all die zuvor gewonnenen Erkenntnisse für ein abschließendes Fazit zusammengefasst.

10. 5 Fazit und Ausblick in die Zukunft

Die Kriterien Grammatik, Lexik, Syntax, Stilistik, Flüssigkeit, Genauigkeit und Vollständigkeit wurden in der vorliegenden Studie anhand einer 5-stufigen-Ordinalskala bewertet. Der beste Wert, *Sehr zufriedenstellend* stellte den Wert 1 dar, während der schlechteste Wert *Unzufriedenstellend* eine Wertigkeit von 5 hatte. Also bedeutete eine niedrige Punktezahl eine

bessere Bewertung und eine hohe Punktezahl eine schlechtere Bewertung. Die Posteditionen wurden insgesamt um **16% besser bewertet** als die Übersetzungen. Bei den Posteditionen war der Zeitaufwand um **30% geringer** als bei den Übersetzungen. In allen Kategorien wurden die Posteditionen besser oder zumindest gleich gut wie die Übersetzungen bewertet. Betreffend die Kriterien Lexik, Flüssigkeit und Vollständigkeit erreichten die Posteditionen und Übersetzungen dieselbe Gesamtpunktezahl. Die Diskrepanz in der Bewertung war bei der Genauigkeit am größten – hier wurden die Posteditionen um 40% besser benotet. Am besten bewertet wurde bei den Übersetzungen die Vollständigkeit mit einer durchschnittlichen Note von 1,75 und bei den Posteditionen die Syntax mit einer durchschnittlichen Bewertung von 1,5. Am schlechtesten bewertet wurde bei beiden die Lexik mit einer durchschnittlichen Benotung von 4. Denselben Wert erzielten die Übersetzungen auch in der Kategorie Grammatik. Eine mögliche Erklärung für das schlechte Abschneiden könnte der Fokus der Studie sein. Unter Umständen haben die ProbandInnen zu viel Wert daraufgelegt, eine gute Postedition zu produzieren, wodurch sie achtloser beim Übersetzen waren. Dies würde bedeuten, dass die Testpersonen grundsätzlich in der Lage gewesen wären, Übersetzungen einer besseren Qualität zu produzieren. Ob diese Übersetzungen dann zumindest gleich gut oder sogar besser als die Posteditionen gewesen wären, bleibt offen. Eine weitere potentielle Ursache ist die mangelnde Erfahrung der ProbandInnen – da es sich „bloß“ um Studierende handelt, kann es sein, dass andere Ergebnisse zustande gekommen wären, wenn die empirische Studie mit professionellen ÜbersetzerInnen wiederholt werden würde, welche auf eine jahrelange Erfahrung als LiteraturübersetzerInnen zurückblicken können. Dies scheint auch die Untersuchung der Messgrößen in Kapitel 10. 4. 3 zu bestätigen, da beispielsweise grammatikalische Fehler eher nicht bei professionellen ÜbersetzerInnen zu finden sind. Auch eine schlechte Bewertung der Kategorie Lexik bei den Übersetzungen weist auf einen mangelnden Wortschatz der ProbandInnen hin. Die Syntax und Stilistik der Posteditionen wurden zwar auch besser benotet als die der Übersetzungen, doch dies könnte auf die persönlichen Präferenzen der Evaluatorin zurückzuführen sein. Vor allem bei diesen Kategorien wurden nicht nur Fehler hervorgehoben, sondern auch Ausdrücke und Satz- und Wortstellungen, die die Evaluatorin anders formuliert hätte. Somit ist es offensichtlich, dass bei einer so beschränkten Studie, mit nur einer einzigen Evaluatorin, die Fähigkeiten, das Wissen, die Expertise und sogar die persönlichen Vorlieben der evaluierenden Person einen nicht zu unterschätzenden Einfluss auf das gesamte Forschungsergebnis ausüben. Dies wäre auch eine weitere mögliche Begründung für die Forschungsergebnisse – ein/e andere/r EvaluatorIn käme mit denselben ProbandInnen möglicherweise zu anderen Ergebnissen. Auch die Größe der Studie war wohl ausschlaggebend für das Forschungsergebnis. Immerhin handelte es sich nur um eine kleine Studie mit vier ProbandInnen, einer Evaluatorin, einem einzigen MÜ-System, einem Versuchstext und einer einzigen Sprachenkombination. Aufgrund der kleinen Größe der Testgruppe haben individuelle Stärken, Schwächen und Präferenzen naturgemäß einen größeren Einfluss auf die Testergebnisse. Eine Wiederholung der Studie mit mehr EvaluatorInnen und mehreren,

heterogeneren Teilnehmergruppen (beispielsweise eine Gruppe bestehend aus Studierenden, einer Gruppe aus professionellen ÜbersetzerInnen und einer gemischten Gruppe) könnte Aufschluss darüber geben, wie maßgeblich das Profil der ProbandInnen und der Evaluatorin tatsächlich war. Alternativ könnte auch einfach der Versuchstext geändert werden – wie wären die Ergebnisse ausgefallen, wenn es sich um eine umfangreichere und kompliziertere Geschichte gehandelt hätte? Weitere interessante Fragen wären: Wie fallen die Ergebnisse mit einer anderen Sprachenkombination aus? Was passiert, wenn ein anderes MÜ-System beziehungsweise mehrere MÜ-Systeme verwendet werden? Ferner besteht auch die Möglichkeit, Kriterien unterschiedlich zu gewichten, beziehungsweise andere oder mehr Messgrößen in die Studie miteinzubeziehen. Beispielsweise könnte die Messgröße „Kreativität" (siehe Studie von Guerberof-Arenas und Toral Kapitel 6. 3) hinzugefügt werden oder Messgrößen, welche im literarischen Übersetzen von größerer Wichtigkeit sind als in anderen Bereichen, wie zum Beispiel Stil und Lexik, schwerer gewichtet werden als andere „unwichtigere" Messgrößen.

Die Tatsache, dass eine Vielzahl anderer Studien (besprochen in den Kapiteln 6. 3, 8. 3 und 10. 4. 4) zu ähnlichen Ergebnissen kamen könnte allerdings ein Hinweis darauf sein, dass das Profil der TestteilnehmerInnen nicht (allein) verantwortlich für die Ergebnisse dieser Studie ist. Immerhin zeigte die Studie von Toral und Way aus dem Jahre 2018 (siehe Kapitel 6. 3), dass NMÜ-Systeme, welche mit einem großen Korpus von literarischen Texten trainiert wurden, imstande sind, Übersetzungen (in über 30% der Fälle) von der Qualität einer Humanübersetzung zu produzieren, ohne posteditiert worden zu sein. Somit sollte die „Schuld" für das schlechte Abscheiden der Übersetzungen womöglich nicht beim Verhalten oder im Profil der ProbandInnen gesucht werden. Vielleicht sollte das Augenmerk eher auf der (überraschend) guten Leistung von DeepL liegen, einem Übersetzungssystem, welches online und frei verfügbar für jede/n ist. Schließlich ist anzunehmen, dass eine gute Rohübersetzung die Basis für jede gute Postedition ist. Einerseits wäre wohl mehr Zeit für die Posteditionen notwendig gewesen (was sich in der Zeitaufzeichnung niedergeschlagen hätte) und andererseits wären die Posteditionen nicht um so viel (immerhin 16%) besser als die Übersetzungen bewertet worden, wenn die maschinell generierte Übersetzung „schlecht" gewesen wäre. In der Tat zeigte sich bereits in den Studien von Toral und Way (2018) und Matusov (2019), wie bereits in den Kapiteln 6. 3 und 10. 4. 4 beschrieben, dass NMÜ-Systeme, welche gezielt auf literarische Texte trainiert wurden, selbstständig in der Lage sind, qualitativ hochwertige Übersetzungen von literarischen Texten zu produzieren.

Doch was sind neben einer guten Rohübersetzung weitere Voraussetzungen für eine gute Postedition? Spielen die Verwendungshäufigkeit von MÜ oder übersetzerische Fähigkeiten eine Rolle? In der Studie dieser Arbeit zeigt sich folgendes Bild: Die ProbandInnen erbrachten sehr unterschiedliche Leistungen; bei zwei TeilnehmerInnen waren die Posteditionen besser, bei einer/m war die Übersetzung besser und bei einer Person erzielten die

Postedition und Übersetzung genau dieselbe Punkteanzahl. TeilnehmerIn 4 und 1 produzierten Texte, welche jeweils 16 Punkte erhielten – die besten Texte. Beide Texte waren Posteditionen. Während T1 angab, nur 16 Minuten für die Postedition benötigt zu haben (der niedrigste Wert in der gesamten Zeitaufzeichnung), benötigte T4 45 Minuten. Offenbar besteht also kein direkter Zusammenhang zwischen Zeitaufwand und Qualität der Postedition. Im Fragebogen gab T1 an, MÜ ca. einmal in Monat zu verwenden. T4 kreuzte die Antwort *„Wöchentlich"* an. Auch die Verwendungshäufigkeit von MÜ spielte demnach wohl keine allzu große Rolle, denn jene/r TeilnehmerIn, die angab, MÜ *„Täglich"* zu verwenden, produzierte die Postedition, die 22 Punkte erhielt, also die am schlechtesten bewertete Postedition. T3, die einzige Person, die eine bessere Übersetzung als Postedition produzierte, produzierte auch die beste Übersetzung mit einer Punktezahl von 20. Ihre/Seine Postedition war minimal schlechter mit einer Gesamtwertung von 21, was beträchtlich schlechter war als die beiden besten Posteditionen mit einer Wertung von 16. Demnach lässt sich auch kein kausaler Zusammenhang zwischen Posteditions- und Übersetzungsfähigkeit finden. Mit den in dieser Studie erhobenen Daten lässt sich nicht erklären, wieso ausgerechnet T1 und T4 die besten Texte/Posteditionen produziert haben – womöglich haben diese beiden ProbandInnen in der Vergangenheit bereits (des Öfteren) posteditiert.

Den Ergebnissen der vorliegenden Studie zufolge wäre es also sinnvoll, maschinell übersetzte literarische Texte zu posteditieren, anstatt nur zu übersetzen. Eine solche Vorgehensweise würde offenbar nicht nur Zeit sparen, sondern auch noch bessere Texte produzieren. An dieser Stelle sei ein weiteres Mal gesagt, dass es sich zwar um eine ausgesprochen beschränkte Studie handelte, aber da sich die hier gewonnen Forschungsergebnisse größtenteils mit jenen aus anderen, umfangreicheren Studien (siehe Kapitel 6. 3, 8. 3 und 10. 4. 4) decken, wird im weiteren Verlauf angenommen, dass es sich um einigermaßen verlässliche Ergebnisse handelt. Falls also davon ausgegangen werden kann, dass das Posteditieren von maschinell generierten Übersetzungen nicht nur weniger zeitaufwändig (Zeitersparnis von 30%) ist, sondern auch noch zu qualitativ besseren Zieltexten (um 16% besser benotet) führt, stellt sich die Frage, welchen Mehrwert reines Humanübersetzen im Vergleich noch bietet? Denn selbst wenn durch das Posteditieren bloß gleich gute, oder sogar marginal schlechtere Texte entstehen würden, könnte der signifikante Zeitvorteil in bestimmten Szenarien trotzdem für einen Posteditierungsprozess sprechen. Eine mögliche Antwort auf diese Frage könnte die Studie von Guerberof-Arenas und Toral (2020) bieten. Laut dieser Studie sind reine Humanübersetzungen „kreativer" als reine maschinelle Übersetzungen und anschließende Posteditionen. Außerdem gibt es laut den Autoren dieser Arbeit einen Zusammenhang zwischen dem Lesevergnügen und der Kreativität eines literarischen Textes. Ein potentieller Mehrwert von Humanübersetzungen könnte also die Kreativität sein, ein Vorteil, den die MÜ der Menschheit bislang noch nicht abringen konnte. Dies könnte auch einer der Ursachen sein, wieso die Evaluatorin bei fast allen Texten, die sie evaluierte (mit einer

einzigen Ausnahme), richtigerweise erkannt hat, ob es sich um eine Übersetzung oder Postedition handelte. Das vorrangige Problem dieses „Vorteils“ ist, dass Kreativität außerordentlich schwer messbar ist. Wie andere für das literarische Übersetzen relevante Messgrößen ist Kreativität ein äußerst subjektives Kriterium. Trotz dieser Subjektivität kann wohl präsumiert werden, dass Kreativität ein Aspekt ist, in welchem Menschen Maschinen überlegen sind. Um die Diskussion weiter anzuregen, kann ferner davon ausgegangen werden, dass diese, genauso wie die anderen in dieser Arbeit erwähnten Studien, mit anderen ProbandInnen und/oder Versuchstexten und EvaluatorInnen zu anderen Ergebnissen gekommen wären. Immerhin waren die meisten StudienteilnehmerInnen Studierende und nicht professionelle ÜbersetzerInnen. Unter Umständen besteht also die Möglichkeit, dass Humanübersetzungen qualitativ sehr wohl Posteditionen überlegen sind. Außer Zweifel steht jedoch mit überragender Gewissheit, dass mittels eines Postediting-Prozesses Zeit eingespart wird. Diese Schlussfolgerung gilt als gewiss, da sogar DeepL, ein Übersetzungssystem welches online und gratis für jede/n verfügbar ist, so gute Rohübersetzungen erstellt, dass das Posteditieren solcher Übersetzungen zeiteffizienter ist als reines Übersetzen. Bei fortschrittlicheren NMÜ-Systemen, welche eigens auf literarische Texte trainiert worden sind, ist anzunehmen, dass durch das Posteditieren von Übersetzungen von diesen Systemen ein noch größerer Zeitgewinn entstehen könnte (siehe Toral und Ways Studie aus dem Jahr 2018). Alle diese Überlegungen miteinbegriffen kann also von folgendem Standpunkt ausgegangen werden: Posteditionen sind zwar zeiteffizienter, aber dafür sind Humanübersetzungen qualitativ besser. Großartig wären konkrete, quantifizierbare Werte, wie zum Beispiel: Posteditionen sparen zwar im Durchschnitt x Minuten Zeit, aber dafür sind menschliche Übersetzungen um y % besser. Derart verallgemeinerbare Ergebnisse sind allerdings praktisch unmöglich, da jede/r ÜbersetzerIn, EvaluatorIn, jedes MÜ-System und jeder Text verschieden sind. Da wir in einer sich ständig verändernden Welt und Gesellschaft leben, in welchen immer mehr Arbeitsprozesse automatisiert und beschleunigt werden, wäre es spannend herauszufinden, wieviel *Zeit translatorische Qualität* eigentlich *wert ist*. Denn im Berufsalltag von ÜbersetzerInnen ist es oft so, dass Arbeitsaufträge zunehmend schneller erledigt werden müssen und gleichzeitig schlechter bezahlt werden. In solchen und anderen Situationen könnten Posteditionen eine kosten- und zeiteffiziente Alternative zum reinen Übersetzen sein. In der Tat ist die Entscheidung für oder gegen Postedition im Grunde genommen eine Geldfrage – ist mir (beziehungsweise der/m AuftraggeberIn) x Zeit eine Qualitätssteigerung von y% wert? Alternativ könnte die Frage lauten: Ist mir eine Zeitersparnis von x Stunden/Minuten ein Qualitätsverlust von y wert? Aus diesen Gründen empfiehlt sich eine Posteditions-Ausbildung nicht nur für angehende ÜbersetzerInnen, sondern auch für professionelle TranslatorInnen. Eine solche Ausbildung würde das Fertigkeits-Repertoire von TranslatorInnen vergrößern, sodass sie in bestimmten Szenarien, in welchen eine zeitnahe Durchführung des Arbeitsauftrages von höchster Priorität ist, auf ein verlässliches „Werkzeug“ zurückgreifen können. Immerhin ist eine hohe Anpassungsfähigkeit von TranslatorInnen immer schon

überlebensnotwendig gewesen. Scheitern könnte diese Anpassung an den persönlichen Vorurteilen, die ÜbersetzerInnen oft gegen MÜ hegen. Die ProbandInnen der vorliegenden Arbeit waren MÜ gegenüber zwar nicht grundsätzlich abgeneigt (siehe Beantwortung der Fragebögen in Kapitel 10. 1), dennoch dachten alle ProbandInnen irrtümlicherweise, dass ihre Übersetzungen zumindest genauso gut, wenn nicht sogar besser als die Posteditionen gelungen seien. Während nur drei der vier ProbandInnen angaben, zufrieden mit ihren posteditierten Zieltext zu sein, gaben alle ProbandInnen an, zufrieden mit den von ihnen direkt übersetzten Texten zu sein. Dies deckt sich auch mit der Beantwortung von Frage 15*: „Sind Sie der Meinung, dass Sie einen qualitativ hochwertigeren Zieltext hätten produzieren können, wenn Sie „nur" übersetzt hätten?"*. Drei der vier ProbandInnen waren der Meinung, dass sie in der Lage wären, einen besseren Zieltext zu verfassen, wenn sie nur übersetzt hätten. Diese Antworten zeugen von einer ungerechtfertigten Voreingenommenheit gegenüber MÜ.

Auch die letzte Frage (Frage 18) beschäftigte sich mit der Einstellung gegenüber MÜ: *„Hat sich Ihre Meinung zu maschineller Übersetzung im Rahmen dieser Studie geändert?"* Die Frage wurde folgendermaßen beantwortet: „Zwei der ProbandInnen haben ihre Meinung zu MÜ nicht geändert – sie haben die Antwort Nein angekreuzt. Eine/r dieser beiden gab an, dass sie/er sich „schon länger mit PE im literarischen Bereich beschäftigt. Sie/er hat das Gefühl, dass ihre/seine Einstellung sehr text- und situationsabhängig ist. Die anderen zwei TeilnehmerInnen der Studie haben durch das Experiment nun eine positivere Meinung zu maschineller Übersetzung. Eine/r der beiden meint, dass sie/er noch nie DeepL für literarische Texte benutzt hat und von dem Ergebnis positiv überrascht war. Laut dieser Person hat das Programm eine akzeptable Übersetzung erstellt, bei der kein großer Zeitaufwand notwendig war, um ihres/seines Erachtens nach eine zufriedenstellende Endversion anzufertigen." Vor allem der letzte Kommentar beweist, dass noch viel Aufholbedarf bei der Inklusion von Themen wie MÜ und Postedition in den Ausbildungsprogrammen und Lehrveranstaltungen für angehende LiteraturübersetzerInnen existiert.

Nicht nur die Ergebnisse der in dieser Arbeit durchgeführten Studie, sondern auch die Geschichte der maschinellen Übersetzung, welche in Kapitel 2 beschrieben wurde, lassen folgende Schlussfolgerung vermuten: Es ist anzunehmen, dass MÜ-Systeme stetig besser werden – sie werden kontinuierlich schneller, präziser, günstiger und arbeiten mit wachsenden Textkorpora. Gleichzeitig ist davon auszugehen, dass sich die Fähigkeit von menschlichen ÜbersetzerInnen, Übersetzungen von hoher Qualität zu erzeugen, über die letzten Jahrzehnte wenig oder kaum verändert hat. Hier ist vom „reinen" Übersetzen die Rede, also einem Übersetzungsprozess ohne der Unterstützung etwaiger Computerprogramme. Selbstverständlich verhalf die Einführung von CAT-Tools ÜbersetzerInnen zu einer Steigerung der Effizienz und Produktivität, doch ist anzunehmen, dass die Entwicklung von CAT-Tools und MÜ-Systemen Hand in Hand geht. Somit wird die Diskrepanz in der Qualität zwischen MÜ und Humanübersetzungen fortlaufend schwinden. Bis zu welchem Punkt diese

Entwicklung führen wird und ob MÜ-Systeme HumanübersetzerInnen jemals in allen Parametern übertreffen werden, bleibt abzuwarten. Womöglich ist dieser Zeitpunkt in bestimmten Bereichen bald erreicht. Ob die „letzte Bastion“, das literarische Übersetzen, mit ihrem mächtigen Mauerwerk aus Humor, Lyrik, Rhythmen, Verbildlichungen und Kreativität jemals bezwungen werden wird, wird die Zeit zeigen. Unabhängig davon dürfen Vorurteile und vielleicht sogar Stolz und Eitelkeit nicht dafür sorgen, dass sich ÜbersetzerInnen von technologischen Fortschritten im Bereich der MÜ abwenden.

11 Conclusio

Ziel dieser Arbeit war es, die Möglichkeiten und Grenzen der maschinellen Übersetzung zu erforschen. Die Forschungsfrage lautete:

Ist es mittels Postedition einer vom Online-Tool DeepL erzeugten Übersetzung möglich, gleichwertige oder sogar bessere deutsche Übersetzungen einer englischen Kurzgeschichte zu erzeugen im Vergleich zu Übersetzungen von HumanübersetzerInnen?
In welchen Aspekten schneiden HumanübersetzerInnen besser und in welchen schlechter ab?

Um diese Frage zu beantworten, haben vier Studierende des Masterstudiums der Translation mit dem Schwerpunkt „Übersetzen in Literatur - Medien – Kunst“ der Universität Wien eine englische Kurzgeschichte zur Hälfte ins Deutsche übersetzt. Die andere Hälfte wurde mittels des Online-Tools DeepL rohübersetzt und anschließend von den Studierenden posteditiert. Alle Zieltexte wurden von einer Evaluatorin, einer professionellen Übersetzerin, bewertet. Alle fünf TeilnehmerInnen der Studie füllten einen Fragebogen aus. Die Evaluatorin füllte zusätzlich pro Zieltext einen Evaluierungsbogen aus, also insgesamt vier Evaluierungsbögen. Nach genauem Lesen der Zieltexte wurden die Kriterien Grammatik, Lexik, Syntax, Stilistik, Flüssigkeit, Genauigkeit und Vollständigkeit mit Hilfe einer 5-stufigen-Ordinalskala bewertet. Neben diesen Kriterien wurde auch der Zeitaufwand für die Übersetzungen und Posteditionen jedes Textes gemessen. Der Evaluatorin war nicht bekannt, bei welchen Texten es sich um Übersetzungen und bei welchen es sich um Posteditionen handelte, um eine möglichst unparteiische Beurteilung zu erlauben. Die Testergebnisse ergaben, dass die Posteditionen im Durchschnitt nicht nur um 30% schneller produziert wurden, sondern auch um 16% besser bewertet wurden als die Übersetzungen. Somit wurde die Hälfte der eingangs beschriebenen Hypothese bestätigt: Mittels einer Postedition gelingt eine signifikante Zeitersparnis. Allerdings erzielten die Posteditionen in der Studie der vorliegenden Arbeit bessere Ergebnisse als die reinen Humanübersetzungen, also anders als eingangs angenommen. In allen Kategorien schnitten die Posteditionen besser oder zumindest genauso gut ab wie die Übersetzungen. In

den Kategorien Lexik, Flüssigkeit und Vollständigkeit erreichten die Posteditionen und Übersetzungen dieselbe Gesamtpunktezahl. Die Grammatik, Syntax, Stilistik und Genauigkeit der Posteditionen wurden besser benotet als die der Übersetzungen. Aufgrund der geringen Größe der Studie sind diese Ergebnisse jedoch mit Vorsicht zu genießen. Verallgemeinerbare beziehungsweise gänzlich andere Ergebnisse wären womöglich mit mehr/anderen ProbandInnen, EvaluatorInnen, Übersetzungssystemen, Kriterien oder Versuchstexten zustande gekommen.

Zu guter Letzt können die Ergebnisse der im Rahmen dieser Arbeit durchgeführten Studie zu einer Empfehlung für das Curriculum des Masterstudiums Translation an der Universität Wien führen. Meines Wissens existiert momentan (April 2021) keine Lehrveranstaltung im derzeitigen Curriculum, welche sich explizit mit dem Thema Postedition auseinandersetzt und die im Kapitel 8 besprochenen Inhalte Studierenden näherbringt. In einem Zeitalter, in welchem maschinelle Übersetzungssysteme stetig besser werden, müssen sich (angehende) ÜbersetzerInnen den fortlaufend ändernden Umständen anpassen, um am hart umkämpften Translationsmarkt wettbewerbsfähig zu sein. Mehr als je zuvor ist die Auseinandersetzung mit CAT-Tools zu einer Notwendigkeit geworden. Somit empfehle ich die Einführung zumindest einer Lehrveranstaltung welche sich eingehend mit Postedition beschäftigt. Nur so können AbsolventInnen jenen hohen Ansprüchen gerecht werden, die auf der Webseite des Zentrums für Translationswissenschaften wie folgt formuliert sind:

„Unter sich ständig wandelnden gesellschaftlichen und technologischen Bedingungen sind die Absolventinnen und Absolventen des Masterstudiums befähigt, (selbst-)verantwortlich in einer globalisierten Gesellschaft translatorisch zu handeln, erworbenes Wissen zu verarbeiten, ihre Fähigkeiten anzuwenden und zu vermitteln, sich selbständig weiterzuentwickeln, sich flexibel an neue Tätigkeitsfelder anzupassen und sie kompetent mitzugestalten.“

(vgl. transvienna 2021)

12 Literaturverzeichnis

ALPAC. 1966. *Language and machines. Computers in translation and linguistics. A Report by the Automatic Language Processing Advisory Committee*. Washington, D.C.:National Research Council.

Arnold, Douglas/Balkan, Lorna/Lee Humphreys/Meijer, Siety/Sadler, Louisa. 1994. *Machine Translation An Introductory Guide*. Oxford: Blackwell Publishers.

Austermühl, Frank. 2001. *Electronic Tools for Translators. Manchester*: St. Jerome Publishing.

Bánik, Tomáš/Benko, Ľubomír/Machová, Renáta/Munk, Michal/Munková, Daša. 2019. *Wie irrt die Maschine? Probleme der maschinellen Übersetzung*. Hamburg: Verlag Dr. Kovač.

Bentivogli, Luisa/Bisazza, Arianna/Cettolo, Mauro/Federico, Marcello. 2016. Neural versus Phrase-Based Machine Translation Quality: a Case Study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* 16: 1, 257–267.

Berger, Daniel. 2020. *KI-Übersetzer DeepL unterstützt Japanisch und Chinesisch*. In: https://www.heise.de/newsticker/meldung/KI-Uebersetzer-DeepL-unterstuetzt-Japanisch-und-Chinesisch-4686000.html, Stand: 15.02.2021.

Besacier Laurent. 2014. *Automated translation of a literary work: a pilot study*. In: https://www.researchgate.net/publication/278379215_Automated_translation_of_a_literary_work_a_pilot_study#pf6. Stand: 28.03.2021.

Blatt, Achim/Freigang, Karl-Heinz/Schmitz, Klaus-Dirk/Thome, Gisela. 1985. *Computer und Übersetzen Eine Einführung*. Hildesheim: Georg Olms Verlag.

Bowker, Lynne/Ehgoetz, Melissa. 2007. Exploring User Acceptance of Machine Translation Output: A Recipient Evaluation. In: Dorothy Kenny & Kyongjoo Ryou (Hg.) *Across Boundaries: International Perspectives on Translation*. Newcastle: Cambridge Scholars Publishing, 209–224.

Burchardt, Aljoscha/Porsiel, Jörg. 2017. Vorwort: Was kann maschinelle Übersetzung und was nicht? In: Porsiel, Jörg (Hg.), 11-18.

Castilho, Sheila. 2016. *Measuring acceptability of machine translated enterprise content.* Dublin City University, PhD thesis.

Castilho, Sheila/Moorkens, Joss/Gaspari, Federico/Calixto, Iacer/Tinsley, John/Way, Andy. 2017. Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics* 108: 2, 109–120.

Castilho, Sheila/Doherty, Stephen/Gaspari, Federico/Moorkens, Joss. 2018. Approaches to Human and Machine Translation Quality Assessment. In: Moorkens, Joss/Castilho, Sheila/Gaspari, Federico/Doherty, Stephen (Hg.), 9-38.

[a]DeepL. 2021. In: https://www.deepl.com/press.html, Stand: 15.02.2021.

[b]DeepL. 2021. In: https://www.deepl.com/pro#single, Stand: 15.02.2021.

Doherty, Stephen. 2017. Issues in human and automatic translation quality assessment. In: Kenny, Dorothy (Hg.) *Human issues in translation technology*. London: Routledge, 131-148.

Doug, Arnold. 2003. Why Translation is difficult for computers. In: Somers, Harold (Hg.) *Computers and Translations A translator's guide*. Amsterdam/Philadelphia: John Benjamins.

Eisold, Christian. 2017. Zur Rolle der Terminologie in der maschinellen Übersetzung. In: Porsiel, Jörg (Hg.), 109-125.

Federico, Marcello/Negri, Matteo/Bentivogli, Luisa/Turchi, Marco. 2014. *Assessing the impact of translation errors on machine translation quality with mixed-effects models*. In: https://emnlp2014.org/papers/pdf/EMNLP2014172.pdf. Stand: 10.03.2021.

Fiederer, Rebecca/O'Brien, Sharon. 2009. Quality and machine translation: A realistic objective? In: https://www.jostrans.org/issue11/art_fiederer_obrien.pdf. Stand: 10.03.2021.

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

Garcia, Ignacio. 2011. Translating by Postedition: is it the way forward? In: *Machine Translation* 25: 3, 217-237.

Genzel, Dmitriy/Uszkoreit, Jakob/ Och, Franz . 2010. *"Poetic" Statistical Machine Translation: Rhyme and Meter*. In: https://www.aclweb.org/anthology/D10-1000.pdf. Stand: 27.03.2021.

Ghazvininejad, Marjan/Choi, Yejin/Knight, Kevin. 2018. Neural Poetry Translation. In: https://www.aclweb.org/anthology/N18-2011.pdf. Stand: 27.03.2021.

Goshawke, Walter/Kelly, Ian/Wigg, David. 1987. *Computer translation of natural language*. Wilmslow: Sigma Press.

Greene, Erica/Bodrumlu, Tugba/Knight, Kevin. 2010. *Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation*. In: https://www.aclweb.org/anthology/D10-1051/. Stand: 10.03.2021.

Guerberof-Arenas Ana/Toral Antonio. 2020. *The impact of post-editing and machine translation on creativity and reading experience*. In: https://www.jbe-platform.com/content/journals/10.1075/ts.20035.gue. Stand: 05.04.2021.

Guerra, Martínez, Lorena. 2003. *Human translation versus machine translation and full postediting of raw machine translation output*. In: http://sceuromix.com/enlaces/MASTER%20IN%20TRANSLATION%20STUDIES%20BY%20LORENA%20GUERRA-2003.pdf. Stand: 10.03.2021.

Hutchins, W. John/Somers, Harold L. 1992. *An Introduction to Machine Translation*. London: Academic Press Limited.

Hutchins, W. John. 2000. *Early Years in Machine Translation Memoirs and biographies of pioneers*. Amsterdam/Philadelphia: John Benjamins.

Kirchhoff, Katrin/Capurro, Daniel/Turner Anne. 2012. *Evaluating user preferences in machine translation using conjoint analysis*. In: https://www.aclweb.org/anthology/2012.eamt-1.pdf. Stand: 10.03.2021.

Koehn, Philipp. 2017. *Statistical Machine Translation Draft of Chapter 13: Neural Machine Translation*. In: https://arxiv.org/abs/1709.07809, Stand: 11.08.2020.

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

Lehrberger, John/Bourbeau, Laurent. 1988. *Machine Translation: Linguistic characteristics of MT systems and general methodology of evaluation*. Amsterdam: John Benjamins.

Lommel, Arle/Burchardt, Aljoscha/ Popović, Maja/Harris, Kim/Avramidis, Eleftherios/ Uszkoreit, Hans. 2014a. Using a new analytic measure for the annotation and analysis of MT errors on real data. *Proceedings of the 17th annual conference of the European Association for Machine Translation* 17: 165–172.

Lommel, Arle/Burchardt, Aljoscha/ Popović, Maja. 2014b. *Assessing Inter-Annotator Agreement for Translation Error Annotation*. In: https://www.dfki.de/fileadmin/user_upload/import/7445_LREC-Lommel-Burchardt-Popovic.pdf, Stand: 10.03.2021.

Lommel, Arle/Burchardt, Aljoscha/Görög, Attila/ Uszkoreit, Hans/Melby, Alan. 2015. *Multidimensional Quality Metrics (MQM) Issue types*. In: http://www.qt21.eu/mqm-definition/issues-list-2015-05-27.html. Stand: 10.03.2021.

Matusov, Evgeny. 2019. *The Challenges of Using Neural Machine Translation for Literature*. In: https://www.aclweb.org/anthology/W19-73.pdf. Stand: 05.04.2021.

Meinhold, Annabelle. 2021. *Google Translate vs. DeepL*. In: https://www.cio.de/a/google-translate-vs-deepl,3575364, Stand: 15.02.2021.

Meyer-Sickendiek, Burkhard/Hussein, Hussein/Baumann, Baumann. 2018. Towards the Creation of a Poetry Translation Mapping System. *Archives of Data Science, Series A* 5: 21-36.

Moorkens, Joss/Castilho, Sheila/Gaspari, Federico/Doherty, Stephen (Hg.) 2018. *Translation Quality Assessment From Principles to Practice*. Cham: Springer.

Popović, Maja. 2017. Comparing Language Related Issues for NMT and PBMT between German and English. *The Prague Bulletin of Mathematical Linguistics* 108: 2, 209–220.

Popović, Maja. 2018. Error Classification and Analysis for Machine Translation Quality Assessment. In: Moorkens, Joss/Castilho, Sheila/Gaspari, Federico/Doherty, Stephen (Hg.), 129-159.

Porsiel, Jörg (Hg.) 2017. *Maschinelle Übersetzung Grundlagen für den professionellen Einsatz*. Berlin: BDÜ Fachverlag.

Ramlow, Markus. 2009. *Die maschinelle Simulierbareit des Humanübersetzens Evaluation von Mensch-Maschine-Interaktion und der Translatqualität der Technik*. Berlin: Frank & Timme.

Roturier, Johann. 2006. *An investigation into the impact of controlled English rules on the comprehensibility, usefulness and acceptability of machine-translated technical documentation for French and German users.* Dublin City University, PhD thesis.

Temizöz, Özlem. 2016. Postediting machine translation output: subject-matter experts versus professional translators. *Perspectives* 24:4, 646-665.

Thurmair, Gregor. 2017. Wissensbetriebene Maschinelle Übersetzung. In: Porsiel, Jörg (Hg.), 29-43.

Toral, Antonio/Way, Andy. 2015. Machine-assisted translation of literary text: A case study. In: https://www.researchgate.net/publication/290209944_Machine-assisted_translation_of_literary_text_A_case_study. Stand: 31.03.2021.

transvienna. 2021. In: https://transvienna.univie.ac.at/studium/masterstudium-translation/, Stand: 16.03.2021.

Trujillo, Arturo. 1999. *Translation Engines: Techniques for Machine Translation*. London: Springer.

Trupkiewicz, Eleanore. 2014. Keep it Simple: Keys to Realistic Dialogue (Part II). In: https://www.writersdigest.com/write-better-fiction/keep-it-simple-keys-to-realistic-dialogue-part-ii. Stand: 08.04.2021.

Voigt, Rob/Jurafsky, Dan. 2012. Towards a literary machine translation: the role of referential cohesion. In: https://www.aclweb.org/anthology/W12-25.pdf. Stand: 10.03.2021.

Way, Andy/Toral, Antonio. 2018. Quality Expectations of Machine Translation. In: Moorkens, Joss/Castilho, Sheila/Gaspari, Federico/Doherty, Stephen (Hg.), 159-178.

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

Wu, Yonghui/Schuster, Mike/Chen, Zhifeng/Le, Quoc/Norouzi, Mohammad. 2016. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. In: https://arxiv.org/abs/1609.08144, Stand: 11.08.2020.

13 Abbildungsverzeichnis

Abbildung 1: „**Die Geschichte der maschinellen Übersetzung**“ 4

Quelle: Eigene Grafik, erstellt am 13.08.2020

Abbildung 2: „**Beziehung zwischen der Intervention des Menschen und der Automatisierung der Übersetzung**“ 11

Quelle: Bánik et al. (2019: 21)

Abbildung 3: „**Ansätze der maschinellen Übersetzung**“ 14

Quelle: Bánik et al (2019: 32)

Abbildung 4: „**Allgemeiner Ablauf der manuellen Fehlerkennzeichnung**“ 35

Quelle: Popović (2018: 132)

Abbildung 5: „**Hierarchische Baumstruktur des MQM-System**“..... 37

Quelle: Lommel et al. (2015)

Abbildung 6: „**Darstellung der Oberkategorien im MQM-System**“..... 37

Quelle: Quelle: Lommel et al. (2015)

Abbildung 7: „**Ablauf der automatischen Fehlerklassifizierung**“..... 39

Quelle: Popović (2018: 143)

Abbildung 8: „**DeepL im Vergleich**“ 46

Quelle: https://www.deepl.com/quality.html, Stand: 15.02.2021

Abbildung 9: „**DeepL Eingabefenster und Synonyme**“..... 47

Quelle: https://www.deepl.com/translator, Stand: 15.02.2021

Abbildung 10: „**Visualisierung der Frage 1**“..... 49

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 11: „**Visualisierung der Frage 2**“..... 50

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 12: „**Visualisierung der Frage 3**“.....50

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 13: „**Visualisierung der Frage 4**“.....51

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 14: „**Visualisierung der Frage 5**“.....52

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 15: „**Zeitaufwand der ProbandInnen**“.....53

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 16: „**Visualisierung der Frage 7**“.....53

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 17: „**Visualisierung der Frage 8**“.....54

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 18: „**Visualisierung der Frage 9**“.....54

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 19: „**Visualisierung der Frage 10**“.....55

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 20: „**Visualisierung der Frage 11**“.....56

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 21: „**Visualisierung der Frage 12**“.....56

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 22: „**Visualisierung der Frage 13**“.....57

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 23: „**Visualisierung der Frage 14**“.....58

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 24: „**Visualisierung der Frage 15**“.....58

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 25: „**Visualisierung der Frage 16**“.....59

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 26: „**Visualisierung der Frage 17**“.....60

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 27: „**Visualisierung der Frage 18**“.....61

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 28: „**Evaluation von Text 1**“ 65

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 29: „**Evaluation von Text 2**“ 65

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 30: „**Evaluation von Text 3**“ 66

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 31: „**Evaluation von Text 4**“ 67

Quelle: Eigene Grafik, erstellt am 16.04.2021

Abbildung 32: „**Evaluierung der Zieltexte im Vergleich**“ 71

Quelle: Eigene Grafik, erstellt am 16.04.2021

Anhang

Der Versuchstext (Originalfassung)

Lace Curtain

By Abigail Allen

Marie was a petite woman, but she felt large and clumsy today, perhaps, she thought, because of everything she'd been through lately—first her sudden preoccupation with Ray-Ray and then the business with the police showing up on her land, holding her at gunpoint, and then apologizing and saying it had all been a mistake before they ran around her house, got in their car, and left.

She felt leaden and awkward as she sat in Nancy's living room, listening to the atrociously loud ticking of the clock on the mantelpiece and waiting for Nancy to put the eggs Marie had brought her in the refrigerator. Nancy wanted to know what was going on, she thought. That must be why she'd phoned Marie saying her hens had quit laying, asking Marie if she had any eggs to spare. She, understandably, wanted to know why the police had swarmed onto Marie's place, why the helicopter had been hovering over the field behind her house.

Nancy returned to the living room with two cups of coffee. She gave one to Marie and sat down in the chair opposite her. "Thank you so much for the eggs," she said. "I'll pay you back when my hens start laying again. That helicopter flying around here the other day scared them, I think."

"I don't doubt it," Marie said. "We'll probably never know what it was all about." She could see the disappointment on Nancy's face. "I wish I had an explanation for it," she said.

"Larry said it scared his water buffalo too," Nancy said. She was a frail, elderly woman. Larry was her son. He raised cattle on their farm and had recently bought the water buffalo. "The other cattle couldn't have cared less," she said. "Nothing bothers them."

Marie smiled. "I've seen Larry riding around on the water buffalo," she said. "It seems like a patient animal."

"Larry loves the thing," Nancy said. "He named her Daisy."

"Daisy," Marie said. "I think it suits her."

"She's smart," Nancy said, sipping her coffee. "Larry thinks she understands everything he says."

Marie stopped by the convenience store to replace the eggs she'd given Nancy. Her cousin Patty, who worked there, was at the cash register. "I found somebody to lease your place," Patty said. "He wants to plant hay, and he's got all the equipment he needs."

Marie paid for the eggs. "Who is he?" she said.

"Joe Leto," Patty said. "He used to help my dad sometimes."

"I don't know him," Marie said. "Is he trustworthy?"

Patty put the eggs in a plastic bag and slid it toward Marie. "Absolutely," she said. "I wouldn't recommend him otherwise. He'll take good care of your land, and the best part is he works with his sons. It's a family operation."

"That's good," Marie said. "It'll be nice to do something with the land again. I'll enjoy watching them harvest the hay and bale it at the end of the season."

"The cattle farmers will be happy too," Patty said. "There's never enough hay around here."

"My neighbor Larry has cattle, and now he has a water buffalo. I guess they eat hay too." She picked up the bag with the eggs in it as another customer entered the store. "His mother said their hens had quit laying because of the business with the helicopter the other day."

Marie was walking through the fields around her house with Patty and Tom, Patty's husband. They were showing Joe Leto around. "This is perfect," Joe said. "We'll plant as soon as possible. I'll have the boys out here in a day or so, and we'll get started."

When they got back to Marie's house, she invited them in for coffee, but Joe had a job to get back to. "I'll let you know before we come out to plant," he told Marie. They shook hands, and he jumped in his truck and drove off.

"He seems perfect," Marie said. "Thanks for recommending him." They went in the house.

"I'd rather have a beer," Tom said as he took a seat on the sofa. "It's too hot to drink coffee."

Marie went to the kitchen and looked in the refrigerator. "You're in luck," she called to Tom. "I've got three bottles of India pale ale, so we can each have one."

"I really would like one," Patty said.

Tom held the bottle against his cheek before he drank from it. "Nice and cold," he said.

After Tom and Patty left, she turned on the news and walked into the kitchen as she listened to it. She was taking some Swiss cheese out of the refrigerator when she heard the newscaster saying someone named Troy Leto had been arrested for car theft.

"Leto," she said. She hadn't heard all the details, so she didn't know where he'd been arrested. She hoped he wasn't one of Joe's sons. She made a cheese sandwich and ate it in front of the TV, thinking there might be a recap at the end of the news and she would hear more about the

car thief, but nothing more was said about him. After she put on a nightgown and brushed her teeth, she went back to the living room and flipped through the channels on the TV.

She wondered whether she had locked the front door after Tom and Patty left. There was a window in the door, covered by a lace curtain, and when she looked at it, she saw the shadow of someone's head rising slowly behind it, as though the person had been crouching there, waiting for the right moment to look into the room.

She ran to the door, screaming, "Get out of here!" and the person disappeared from the window. She made sure the door was locked. Then she went to a window, not the one in the door, but one that afforded a better view of the porch and the front yard. She saw a tall, thin man running down the driveway to the road. Then she couldn't see him anymore. First her husband had left, then the police had converged on her property, and now this. She went to the bedroom. The doors and windows were locked. There was no way anybody could break in, she thought as she was climbing under the covers. If somebody did, she would sneak into the kitchen and get a butcher knife to defend herself with. She lay still for a while, alert for any unusual sounds. Maybe she would use a fillet knife, she thought as she was falling asleep.

[1106 Wörter]

Quelle: https://bigbridge.org/BB19/fiction/Abigail_Allen.html (Stand: 02.01.2021)

Übersetzung von DeepL

Spitzenvorhang

Von Abigail Allen

Marie war eine zierliche Frau, aber heute fühlte sie sich groß und unbeholfen, vielleicht, so dachte sie, wegen allem, was sie in letzter Zeit durchgemacht hatte - zuerst ihre plötzliche Beschäftigung mit Ray-Ray und dann die Sache mit der Polizei, die auf ihrem Grundstück auftauchte, sie mit vorgehaltener Waffe festhielt und sich dann entschuldigte und sagte, es sei alles ein Fehler gewesen, bevor sie um ihr Haus herumliefen, in ihr Auto stiegen und wegfuhren. Sie fühlte sich bleiern und unbeholfen, als sie in Nancys Wohnzimmer saß, dem grauenhaft lauten Ticken der Uhr auf dem Kaminsims lauschte und darauf wartete, dass Nancy die Eier, die Marie ihr mitgebracht hatte, in den Kühlschrank stellte. Nancy wollte wissen, was los war, dachte sie. Deshalb hatte sie Marie angerufen und gesagt, dass ihre Hühner aufgehört hatten zu legen, und Marie gefragt, ob sie Eier übrig hätte. Sie wollte verständlicherweise wissen,

warum die Polizei auf Maries Haus ausgeschwärmt war, warum der Hubschrauber über dem Feld hinter ihrem Haus geschwebt hatte.

Nancy kehrte mit zwei Tassen Kaffee ins Wohnzimmer zurück. Sie gab Marie eine und setzte sich auf den Stuhl ihr gegenüber. "Vielen Dank für die Eier", sagte sie. "Ich zahle es dir zurück, wenn meine Hühner wieder anfangen zu legen. Der Hubschrauber, der neulich hier herumflog, hat sie erschreckt, glaube ich."

"Daran zweifle ich nicht", sagte Marie. "Wir werden wahrscheinlich nie erfahren, was es damit auf sich hatte." Sie konnte die Enttäuschung auf Nancys Gesicht sehen. "Ich wünschte, ich hätte eine Erklärung dafür", sagte sie.

"Larry sagte, dass es seinen Wasserbüffel auch erschreckt hat", sagte Nancy. Sie war eine gebrechliche, ältere Frau. Larry war ihr Sohn. Er züchtete Rinder auf ihrer Farm und hatte kürzlich den Wasserbüffel gekauft. "Den anderen Rindern war das völlig egal", sagte sie. "Nichts stört sie."

Marie lächelte. "Ich habe Larry auf dem Wasserbüffel herumreiten sehen", sagte sie. "Es scheint ein geduldiges Tier zu sein."

"Larry liebt das Ding", sagte Nancy. "Er hat sie Daisy genannt."

"Daisy", sagte Marie. "Ich glaube, das passt zu ihr."

"Sie ist klug", sagte Nancy und nippte an ihrem Kaffee. "Larry denkt, sie versteht alles, was er sagt."

Marie ging in den Supermarkt, um die Eier zu ersetzen, die sie Nancy geschenkt hatte. Ihre Cousine Patty, die dort arbeitete, stand an der Kasse. "Ich habe jemanden gefunden, der dein Haus pachtet", sagte Patty. "Er will Heu anbauen und hat alle Geräte, die er braucht."

Marie bezahlte die Eier. "Wer ist er?", fragte sie.

"Joe Leto", sagte Patty. "Er hat meinem Vater manchmal geholfen."

"Ich kenne ihn nicht", sagte Marie. "Ist er vertrauenswürdig?"

Patty steckte die Eier in eine Plastiktüte und schob sie Marie zu. "Auf jeden Fall", sagte sie. "Ich würde ihn nicht anders empfehlen. Er wird sich gut um Ihr Land kümmern, und das Beste daran ist, dass er mit seinen Söhnen zusammenarbeitet. Es ist ein Familienbetrieb."

"Das ist gut", sagte Marie. "Es wird schön sein, wieder etwas mit dem Land zu tun zu haben. Ich werde es genießen, ihnen zuzusehen, wie sie das Heu ernten und es am Ende der Saison zu Ballen pressen."

"Die Viehzüchter werden auch glücklich sein", sagte Patty. "Hier gibt es nie genug Heu."

"Mein Nachbar Larry hat Rinder, und jetzt hat er einen Wasserbüffel. Ich schätze, die fressen auch Heu." Sie hob die Tüte mit den Eiern darin auf, als ein anderer Kunde den Laden betrat.

"Seine Mutter sagte, ihre Hühner hätten aufgehört zu legen, wegen der Sache mit dem Hubschrauber neulich."

Marie ging mit Patty und Tom, Pattys Mann, durch die Felder rund um ihr Haus. Sie führten Joe Leto herum. "Das ist perfekt", sagte Joe. "Wir werden so schnell wie möglich pflanzen. Ich werde die Jungs in einem Tag oder so hierher schicken, und dann fangen wir an."

Als sie zu Maries Haus zurückkamen, lud sie sie auf einen Kaffee ein, aber Joe musste zurück zu seiner Arbeit. "Ich sage dir Bescheid, bevor wir zum Pflanzen rauskommen", sagte er zu Marie. Sie schüttelten sich die Hände, und er sprang in seinen Truck und fuhr los.

"Er scheint perfekt zu sein", sagte Marie. "Danke, dass du ihn empfohlen hast." Sie gingen ins Haus.

"Ich würde lieber ein Bier trinken", sagte Tom, als er auf dem Sofa Platz nahm. "Es ist zu heiß, um Kaffee zu trinken."

Marie ging in die Küche und schaute in den Kühlschrank. "Du hast Glück", rief sie Tom zu. "Ich habe drei Flaschen India Pale Ale, wir können also jeder eine haben."

"Ich würde wirklich gern eins haben", sagte Patty.

Tom hielt die Flasche an seine Wange, bevor er daraus trank. "Schön kalt", sagte er.

Nachdem Tom und Patty gegangen waren, schaltete sie die Nachrichten ein und ging in die Küche, während sie den Nachrichten zuhörte. Sie nahm gerade etwas Schweizer Käse aus dem Kühlschrank, als sie den Nachrichtensprecher sagen hörte, dass jemand namens Troy Leto wegen Autodiebstahls verhaftet worden war.

"Leto", sagte sie. Sie hatte noch nicht alle Details gehört, also wusste sie nicht, wo er verhaftet worden war. Sie hoffte, dass er nicht einer von Joes Söhnen war. Sie machte sich ein Käsesandwich und aß es vor dem Fernseher, in der Hoffnung, dass es am Ende der Nachrichten eine Wiederholung geben würde und sie mehr über den Autodieb hören würde, aber es wurde nichts mehr über ihn gesagt. Nachdem sie sich ein Nachthemd angezogen und die Zähne geputzt hatte, ging sie zurück ins Wohnzimmer und blätterte durch die Kanäle des Fernsehers.

Sie fragte sich, ob sie die Haustür abgeschlossen hatte, nachdem Tom und Patty gegangen waren. In der Tür befand sich ein Fenster, das von einem Spitzenvorhang verdeckt war, und als sie darauf schaute, sah sie den Schatten eines Kopfes, der sich langsam dahinter erhob, als hätte die Person dort gehockt und auf den richtigen Moment gewartet, um ins Zimmer zu schauen.

Sie rannte zur Tür und schrie: "Raus hier!", und die Person verschwand aus dem Fenster. Sie vergewisserte sich, dass die Tür verschlossen war. Dann ging sie zu einem Fenster, nicht zu dem in der Tür, sondern zu einem, das einen besseren Blick auf die Veranda und den Vorgarten bot. Sie sah einen großen, dünnen Mann, der die Auffahrt zur Straße hinunterlief. Dann konnte

sie ihn nicht mehr sehen. Erst war ihr Mann abgehauen, dann war die Polizei auf ihrem Grundstück aufgetaucht, und jetzt das. Sie ging ins Schlafzimmer. Die Türen und Fenster waren verschlossen. Es gab keine Möglichkeit, dass jemand einbrechen konnte, dachte sie, während sie unter die Decke kroch. Und wenn doch, würde sie sich in die Küche schleichen und ein Fleischermesser holen, um sich damit zu verteidigen. Sie lag noch eine Weile still, wachsam für irgendwelche ungewöhnlichen Geräusche. Vielleicht würde sie ein Filetiermesser benutzen, dachte sie, als sie einschlief.

[1115 Wörter]

Quelle: https://www.deepl.com/translator (Stand: 15.02.2021)

Fragebogen der ProbandInnen

Fragebogen-Nr:

Gruppe:

(wird vom Forschenden ausgefüllt)

Fragebogen zum Thema Postedition

Titel der Masterarbeit:

„Cui bono?"

Post-Editing von maschinellen Übersetzungen in der Literaturübersetzung.

Im Rahmen meiner Masterarbeit beschäftigte ich mich mit folgender Forschungsfrage:

„Ist das Post-Editing von literarischen Übersetzungen, die von DeepL vorübersetzt wurden, produktiver als die reine Humanübersetzung?"

Name:

Datum:

Teil 1 - bitte beantworten Sie die Fragen 1-5, BEVOR Sie mit der Postedition/Übersetzung beginnen:

Hinweis: Setzen Sie ein Kreuz, wenn die Antwort zutrifft; Mehrfachnennungen sind möglich.

Frage 1

Wie oft benutzen Sie maschinelle Übersetzung? (Dazu zählen Google Übersetzer, DeepL und andere)

- ☐ Nie
- ☐ Einmal in mehreren Monaten
- ☐ Wöchentlich
- ☐ Täglich
- ☐ Andere Antwortmöglichkeiten: (bitte ausführen)

Frage 2

Wie zufrieden sind Sie grundsätzlich mit dem Output von maschineller Übersetzung?

- ☐ Sehr zufrieden
- ☐ Zufrieden
- ☐ Neutral
- ☐ Unzufrieden
- ☐ Sehr unzufrieden
- ☐ Keine Antwort
- ☐ Andere Antwortmöglichkeiten: (bitte ausführen)

Frage 3

Wie sinnvoll ist Ihrer Meinung nach die Anwendung von maschineller Übersetzung für professionelle Übersetzungsdienste (im außerliterarischen Bereich)?

- ☐ Sehr sinnvoll
- ☐ Sinnvoll
- ☐ Neutral
- ☐ Eher nicht sinnvoll
- ☐ Nicht sinnvoll
- ☐ Keine Antwort
- ☐ Andere Antwortmöglichkeiten: (bitte ausführen)

Frage 4

Wie sinnvoll ist Ihrer Meinung nach die Anwendung von maschineller Übersetzung im Bereich des literarischen Übersetzens?

- ☐ Sehr sinnvoll
- ☐ Sinnvoll
- ☐ Neutral
- ☐ Eher nicht sinnvoll
- ☐ Nicht sinnvoll
- ☐ Keine Antwort
- ☐ Andere Antwortmöglichkeiten: (bitte ausführen)

Frage 5

Wie sind Ihre bisherigen Erfahrungen mit dem Programm DeepL?

- ☐ Sehr zufriedenstellend
- ☐ Zufriedenstellend
- ☐ Neutral
- ☐ Unzufriedenstellend
- ☐ Sehr unzufriedenstellend
- ☐ Keine bisherige Erfahrungen
- ☐ Andere Antwortmöglichkeiten: (bitte ausführen)

Teil 2 – bitte beantworten Sie die folgenden Fragen erst, NACHDEM Sie die Postedition UND die Übersetzung vorgenommen haben:

Frage 6a

Wieviel Zeit haben Sie für die Postedition benötigt?

Dauer: (Bitte in Stunden – Minuten angeben)

Frage 6b

Wieviel Zeit haben Sie für die Übersetzung benötigt?

Dauer: (Bitte in Stunden – Minuten angeben)

Frage 7

Als wie „schwierig“ zu übersetzen würden Sie die Kurzgeschichte beschreiben?

- ☐ Sehr schwierig
- ☐ Schwierig
- ☐ Neutral
- ☐ Leicht
- ☐ Sehr leicht
- ☐ Keine Antwort
- ☐ Andere Antwortmöglichkeiten:

(bitte ausführen)

Frage 8

Wie zufrieden waren Sie mit dem Output, also der „Rohübersetzung“ des Programms DeepL?

- ☐ Sehr zufrieden
- ☐ Zufrieden
- ☐ Neutral
- ☐ Unzufrieden
- ☐ Sehr unzufrieden
- ☐ Keine Antwort
- ☐ Andere Antwortmöglichkeiten:

(bitte ausführen)

Frage 9

Wie zufrieden sind Sie mit dem von Ihnen posteditieren Zieltext?

- ☐ Sehr zufrieden
- ☐ Zufrieden
- ☐ Neutral
- ☐ Unzufrieden
- ☐ Sehr unzufrieden
- ☐ Keine Antwort
- ☐ Andere Antwortmöglichkeiten:

(bitte ausführen)

Frage 10

Wie zufrieden sind Sie mit dem von Ihnen direkt übersetzten Zieltext?

- ☐ Sehr zufrieden
- ☐ Zufrieden
- ☐ Neutral
- ☐ Unzufrieden
- ☐ Sehr unzufrieden
- ☐ Keine Antwort
- ☐ Andere Antwortmöglichkeiten:

(bitte ausführen)

Frage 11

Hätten Sie es bevorzugt, die Kurzgeschichte „nur“ zu übersetzen, also nicht zu posteditieren?

☐ Ja
☐ Nein
☐ Nicht sicher

Anmerkungen:

Frage 12

Hätten Sie es bevorzugt, die gesamte Kurzgeschichte zu posteditieren?

☐ Ja
☐ Nein
☐ Nicht sicher

Begründung (optional):

Frage 13

Sind Sie der Meinung, dass Sie mehr Zeit benötigt hätten, wenn Sie die Kurzgeschichte „nur“ übersetzt hätten?

☐ Ja
☐ Nein
☐ Nicht sicher

Anmerkungen:

Frage 14

Sind Sie der Meinung, dass Sie mehr Zeit benötigt hätten, wenn Sie die gesamte maschinell vorübersetzte Kurzgeschichte posteditiert hätten?

☐ Ja
☐ Nein
☐ Nicht sicher

Begründung (optional):

Frage 15

Sind Sie der Meinung, dass Sie einen qualitativ hochwertigeren Zieltext hätten produzieren können, wenn Sie „nur“ übersetzt hätten?

☐ Ja
☐ Nein
☐ Nicht sicher

Anmerkungen:

Frage 16

Sind Sie der Meinung, dass Sie einen besseren Zieltext hätten produzieren können, wenn Sie die gesamte Kurzgeschichte posteditiert hätten?

☐ Ja
☐ Nein
☐ Nicht sicher

Begründung (optional):

Frage 17

Hatten Sie das Gefühl, dass Sie beim Posteditieren Ihre Kreativität nicht so entfalten konnten, wie wenn Sie nur übersetzt hätten?

☐ Ja, ich habe mich sehr eingeschränkt gefühlt.
☐ Ja, ich habe mich ein wenig eingeschränkt gefühlt.
☐ Ich habe keinen signifikanten Unterschied bemerkt.
☐ Nein, meine Kreativität war nicht eingeschränkt.
☐ Durch das Postediting kam ich auf Ideen, auf die ich sonst nicht gekommen wäre.
☐ Das Postediting hat meine Kreativität angeregt.
☐ Andere Antwortmöglichkeiten:
(bitte ausführen)

Anmerkungen:

Frage 18

Hat sich Ihre Meinung zu maschineller Übersetzung im Rahmen dieser Studie geändert?

☐ Ja, ich habe nun eine positivere Meinung zu maschineller Übersetzung.
☐ Nein
☐ Ja, ich habe nun eine negativere Meinung zu maschineller Übersetzung.
☐ Andere Antwortmöglichkeiten:
(bitte ausführen)

Begründung (optional):

Bitte vergleichen Sie abschließend in eigenen Worten die beiden Textbearbeitungsmodi und Ihre Erfahrungen damit im Laufe dieser Studie:

Abstract

Die vorliegende Arbeit beschäftigt sich mit den Themen maschinelle Übersetzung, Postedition und literarisches Übersetzen. Eine englische Kurzgeschichte wurde mittels dem Online-Übersetzungstool „DeepL“ ins Deutsche übersetzt und in der Folge posteditiert. Die Posteditionen wurden mit Humanübersetzungen verglichen. Alle Zieltexte wurden von vier Studierenden der Translation der Universität Wien produziert. Die Evaluierung wurde von einer professionellen Übersetzerin durchgeführt. Im Durchschnitt wurden die Posteditionen um 30% schneller produziert und um 16% besser bewertet als die Humanübersetzungen. In allen Kategorien schnitten die posteditierten Texte besser, oder zumindest genauso gut ab wie die Humanübersetzungen. Die verwendeten Messgrößen waren Grammatik, Lexik, Syntax, Stilistik, Flüssigkeit, Genauigkeit und Vollständigkeit.

This thesis explores the topics of machine translation, postediting and literary translation. Using the online translation tool "DeepL" an English short story was translated into German and then postedited. The postedited texts were compared with human translations. All of the translations and postedited texts were produced by four translation students of the University of Vienna. The evaluation was carried out by a professional translator. On average, the postedited texts were produced 30% faster and rated 16% better than the human translations. In all measured variables the postedited texts performed better, or at least as well as the human translations. The metrics used were grammar, lexis, syntax, stylistics, fluency, accuracy and completeness.