

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Wahrnehmung der Qualität von Dolmetschleistungen durch Rezipient*innen im
Vergleich Mensch–Maschine

verfasst von | submitted by

Martin Weisz BA

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

Wien | Vienna, 2026

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

UA 070 331 342

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

Masterstudium Translation Deutsch Englisch

Betreut von | Supervisor:

Univ.-Prof. Mag. Dr. Franz Pöchhacker

Inhaltsverzeichnis

Abbildungsverzeichnis	4
Tabellenverzeichnis	4
Einleitung	5
1 Maschinelles Dolmetschen.....	9
1.1 Begriffsabgrenzung	12
1.2 Spracherkennung	15
1.3 Maschinelle Übersetzung.....	18
1.4 Sprachsynthese.....	22
1.5 Prosodie	26
1.6 Die Cascading- und End-to-end-Ansätze	29
1.7 Limitationen und Herausforderungen	33
2 Qualitätsbeurteilung von Dolmetschleistungen	36
2.1 Der Qualitätsbegriff in der Dolmetschwissenschaft	36
2.2 Qualitätsbewertung im Lichte aktueller technischer Entwicklungen	42
3 Methodik.....	46
3.1 Zielsetzung, Forschungsfragen und Hypothesen	46
3.2 Versuchsdesign und Teilnehmer*innen.....	48
3.3 Material.....	50
3.4 Fragebogen.....	52
3.5 Durchführung.....	57
3.6 Datenaufbereitung.....	59
3.7 Limitationen und Grenzen	61
4 Ergebnisse.....	63
4.1 Datengrundlage und Stichprobe.....	63
4.2 Wahrnehmung der maschinell erzeugten Dolmetschung	68
4.3 Wahrnehmung der von einem Menschen angefertigten Dolmetschung	75

4.4 Gruppenübergreifende Gegenüberstellung der Ergebnisse	80
4.5 Betrachtung von Korrelationen im Hinblick auf die Hypothesen	86
5 Diskussion und Schlussfolgerungen	90
Bibliographie.....	95
Anhänge.....	103
Anhang 1: Ausgangsrede, orthografische Transkription.....	103
Anhang 2: Von einem Menschen angefertigte Dolmetschung, erweiterte orthografische Transkription	103
Anhang 3: Maschinell angefertigte Dolmetschung, erweiterte orthografische Transkription	104
Anhang 4: Gegenüberstellung der Gruppen anhand M, SD und ΔM	105
Anhang 5: Antworten der MD-Gruppe auf den Fragebogen.....	106
Anhang 6: Antworten der AD-Gruppe auf den Fragebogen	108
Anhang 7: Antworten der MD-Gruppe auf die offenen Fragen	110
Anhang 8: Antworten der AD-Gruppe auf die offenen Fragen.....	112
Abstract Deutsch	114
Abstract Englisch	115

Abbildungsverzeichnis

Abbildung 1: Angabe der Erfahrung mit Simultandolmetschen in Prozent.....	66
Abbildung 2: Einschätzung, ob Dolmetschung maschinell generiert wurde, in Prozent	67
Abbildung 3: Darstellung der Mittelwertunterschiede	81

Tabellenverzeichnis

Tabelle 1: Demographische Angaben der Stichprobe.....	64
Tabelle 2: Zusammenfassung der Antworten der AD-Gruppe in Zahlen	70
Tabelle 3: Verteilung der Antworten der AD-Gruppe in Prozent.....	71
Tabelle 4: Zusammenfassung der Antworten der MD-Gruppe in Zahlen	76
Tabelle 5: Verteilung der Antworten der MD-Gruppe in Prozent	77
Tabelle 6: Gruppenvergleich betreffend Professionalität, Vertrauenswürdigkeit und Qualität	82
Tabelle 7: Gruppenvergleich im Hinblick auf die Präsentationsebene	82
Tabelle 8: Gruppenvergleich betreffend primär inhaltlich-strukturelle Aspekte	83
Tabelle 9: Reihung der Gruppenunterschiede nach Effektstärke d.....	85
Tabelle 10: Antwortverteilungen von Stimme bzw. Natürlichkeit in Relation zu Gesamtqualität.....	87
Tabelle 11: Antwortverteilung der Einschätzung zum maschinellen Ursprung relativ zur wahrgenommenen Qualität	88

Einleitung

Das Konzept des vollautomatischen Dolmetschens genießt seit geraumer Zeit große Faszination und wird als unvermeidbare Realität der nahen Zukunft gehandelt. Maschinen, die in jeder Situation und über schier jegliche sprachliche Barriere hinweg zu perfekten und nahezu ohne erhebliche Zeitverzögerung dargebotenen Dolmetschungen fähig sind, fanden ihren Weg en masse in literarische Werke und Filme. Der schnelle technologische Fortschritt der letzten Jahrzehnte scheint an diese der Fiktion entstammenden Maschinen den Anspruch und die Erwartungshaltung zu stellen, nicht erst in ferner Zukunft, sondern bereits in absehbarer Zeit in das moderne Leben einzuziehen. Tatsächlich finden individuelle dem automatischen Dolmetschen zugrunde liegende Technologien bereits in verschiedensten Teilen der Gesellschaft erfolgreich Anwendung - allerdings konnte noch kein vollautomatisches Dolmetschsystem diesen vielschichtigen Anforderungen und hohen Erwartungen gänzlich gerecht werden (vgl. Fantinuoli 2018: 5; 2024: 210; Härtel 2016: 68; Sperber et al. 2025: 23345; Waibel und Fügen 2008: 70).

Der Nutzen eines solchen Systems steht allerdings außer Frage. Die moderne vernetzte Globalgesellschaft brachte und bringt weiterhin mannigfaltige interlinguale Kommunikationsbedürfnisse hervor, die mitunter über *Linguae Francae* oder Sprachmittler*innen erfüllt werden können, jedoch aus Kostengründen oder aufgrund mangelnder Sprachbeherrschung zu großen Teilen gänzlich unbefriedigt bleiben (vgl. Fügen et al. 2007: 211).

Eine mögliche Lösung zur Überwindung dieser Barrieren sind bereits zum heutigen Tage verfügbare Systeme zur maschinellen Erzeugung von Dolmetschleistungen. Auch wenn derartige Systeme fortwährend mit neuen technologischen Entwicklungen weiter verfeinert, erweitert und überarbeitet werden, wird an sie längst der hohe Anspruch gestellt, reibungslose Kommunikation zu nahezu jedem Themengebiet in verschiedensten kommunikativen Situationen über eine Vielzahl an Sprachen hinweg zu ermöglichen (vgl. Fantinuoli 2025: 220).

Obwohl die beträchtliche Anzahl bereits entwickelter Dolmetschsysteme der beschriebenen Idealvorstellung noch nicht zur Gänze entsprechen können, heben sich diese Lösungen bereits in diverser Hinsicht von ihren menschlichen Gegenstücken ab. Maschinen arbeiten pausenlos und unermüdlich, zeichnen sich durch theoretisch unlimitierte Verarbeitungsgeschwindigkeit und potenziell ins Unermessliche skalierbare Speicherkapazitäten aus und können Inhalte bereits heute kontext- und situationsgebunden mitunter verständlich zwischen Sprachen vermitteln. Insbesondere mithilfe von Large Language Models (LLM) wie ChatGPT, Claude und Gemini, die in den letzten Jahren aufgrund ihrer eindrucksvollen Funktionalitäten die breite

Bevölkerung erreicht und in ihren Bann gezogen haben, kann durch die dieser Technologie zugrunde liegenden neuen technischen Übersetzungsansätze in einer insbesondere für Laien eindrucksvollen Weise erfolgreich zwischen Sprachen vermittelt werden. Bei heutzutage handelsüblichen Systemen zur Erzeugung maschineller Dolmetschungen stehen diesem durch LLM aufgezeigten Leistungspotenzial bislang jedoch trotz der bereits erzielten Fortschritte noch mannigfaltige Defizite und Herausforderungen gegenüber (vgl. Fantinuoli 2025: 210; Fügen et al. 2007: 211; Taylor 2009: 48).

Insbesondere die Darbietung von Dolmetschungen scheint sowohl für Rezipient*innen von großer Bedeutung als auch einer jener Aspekte maschineller Dolmetschsysteme zu sein, die noch ein großes, nicht annähernd ausgeschöpftes Entwicklungspotenzial aufweisen (vgl. Taylor 2009: 3, 47). Künstliche Stimmen muten zumeist befremdlich und unnatürlich an, werden rasch als irritierend und unangenehm wahrgenommen, sind der Informationsaufnahme gar abträglich und verleihen andernfalls soliden Dolmetschleistungen einen bitteren Beigeschmack (vgl. Taylor 2009: 3, 47). Selbst im Falle angemessen vermittelter Inhalte und ungeachtet etwaiger weiterer zugrunde liegender Vorteile und Stärken maschineller Dolmetschungen können Rezipient*innen nicht natürlich klingenden Dolmetschungen mangelnde Qualität beimessen (vgl. Taylor 2009: 3, 47).

Während die Entwicklungen und der technische Fortschritt der jüngsten Vergangenheit indes von einer stets voranschreitenden Verbesserung der Funktionalitäten und unentwegter Ausmerzungen existierender Mängel dieser Systeme zeugen sowie natürlich anmutende Stimmen und Sprache durch LLM und verwandte Systeme bereits künstlich generiert werden können, wird in der vorliegenden Arbeit ein zum heutigen Tage verfügbares System zur Erzeugung maschineller Dolmetschleistungen hinsichtlich der gegenwärtigen Beschaffenheit und aktuellen Funktionsweise getestet sowie analysiert.

Vor diesem Hintergrund ist es das Ziel der vorliegenden Arbeit, zu erfassen und zu betrachten, wie die Qualität einer mit einem bereits heute verfügbaren Dolmetschsystem generierten Dolmetschung von nicht als Dolmetscher*innen tätigen Personen wahrgenommen wird. Die auf diesem Wege erfassten Ergebnisse werden in der Folge mit den von einer zweiten Gruppe erhaltenen Angaben verglichen, der eine von einer professionellen Dolmetscherin angefertigte Dolmetschleistung derselben Ausgangsrede vorgelegt wurde. Durch idente systematische Erhebung der Wahrnehmungsurteile per Fragebogen sollen Muster und Unterschiede

aufgedeckt sowie schließlich entsprechende Schlüsse gezogen werden. Dabei wird davon ausgegangen, dass Aspekte, die primär die Darbietung und Präsentation der Dolmetschleistung betreffen, besonders bedeutend für die von den Rezipient*innen beigemessene Qualität sind, dass eine maschinell, nicht natürlich anmutende Stimme der wahrgenommenen Qualität abträglich ist und dass die Qualität der Dolmetschung umso negativer bewertet wird, je sicherer die Teilnehmenden sich darin sind, dass ihnen eine maschinell erzeugte Leistung vorgelegt wurde.

An diese Themeneinleitung anschließend werden im ersten Kapitel die für das Verständnis der vorliegenden Arbeit erforderlichen theoretischen Grundlagen maschinellen Dolmetschens behandelt. Dabei werden in dem für die Arbeit erforderlichen Ausmaß die grundlegenden Entwicklungen dieser Systeme betrachtet, um zugrunde liegende Technologien und anhaftende Gegebenheiten im heutigen technologischen Umfeld einordnen zu können und ein Verständnis der damit verbundenen Möglichkeiten zu erhalten. Es werden ferner die zahlreichen Technologien und Prozesse beleuchtet, die in maschinelle Dolmetschsysteme einfließen, um Funktionalitäten, Möglichkeiten, Problematiken und Potenziale aufzuzeigen. Da der Begriff des maschinellen Dolmetschens und seine Verwendung terminologisch nicht vollends unproblematisch sind, müssen darüber hinaus entsprechende begriffliche Überlegungen angestellt werden. Als ein zentraler Faktor für die Darbietung und Präsentation von Dolmetschleistungen wird auch die Prosodie und ihre Bedeutung im Komplex des maschinellen Dolmetschens eingeordnet. Da die Bereiche des maschinellen Dolmetschens und Übersetzens von rasanten und schnelllebigen technologischen Entwicklungen und Innovationen geprägt sind, werden ferner die aktuellsten Neuerungen im Zusammenhang mit der grundlegenden Architektur und den Gestaltungsmöglichkeiten von Systemen zur vollautomatischen Erzeugung von Dolmetschleistungen beleuchtet, damit der derzeitige technologische Kontext korrekt erfasst werden kann. Schließlich werden fortwährende Herausforderungen und Probleme beleuchtet, mit denen derartige Systeme und deren Entwickler gegenwärtig konfrontiert sind.

Das darauffolgende Kapitel wird der komplexen Thematik der Qualität und Qualitätsbeurteilung von Dolmetschungen gewidmet. Die historische und gegenwärtige Situation in der Dolmetschwissenschaft werden skizziert, Möglichkeiten und Gegebenheiten im Zusammenhang mit aktuellen technischen Entwicklungen im Hinblick auf die Qualitätsbewertung werden thematisiert und der in der vorliegenden Arbeit zur Anwendung gelangende Fokus auf die durch Rezipient*innen wahrgenommene Qualität maschinell generierter Dolmetschungen wird hergeleitet. Es wird betrachtet, wie Laien Dolmetschleistungen bewerten, welche Maßstäbe sie anwenden können und schließlich welche Kategorien in der vorliegenden Arbeit primär zur

Anwendung gelangen werden, um die zu erfassenden Qualitätsurteile der Teilnehmenden zu erhalten.

Danach werden das Forschungsdesign und die in der vorliegenden Arbeit angewendete Methodik dargelegt. Dabei werden das Versuchsdesign sowie alle für das Verständnis erforderlichen theoretischen Grundlagen und Informationen zur Durchführung der Erhebung beleuchtet. Die für die Befragung herangezogenen Dolmetschleistungen werden behandelt und es wird festgehalten, auf welches Erhebungsinstrument für die Befragung zurückgegriffen wurde, welche Daten dabei extrahiert werden und wie diese letztlich zueinander in Relation gesetzt werden. Schließlich wird festgehalten, mit welchen Limitationen die Analyse und Erhebung in der vorliegenden Arbeit behaftet sind – insbesondere im Hinblick auf die Generalisierbarkeit der erhaltenen Ergebnisse.

Die auf diesem Wege erhaltenen Informationen, Daten und Angaben werden schließlich in Anlehnung an die im Vorfeld formulierten Forschungsfragen und Hypothesen festgehalten und ausgewertet. Abschließend werden die Ergebnisse diskutiert und Schlussfolgerungen daraus gezogen.

1 Maschinelles Dolmetschen

Im vorliegenden Kapitel wird das Konzept von maschinellm Dolmetschen beleuchtet und so der theoretische Grundstein der vorliegenden Arbeit gelegt. Der Begriff wird für die Zwecke der dieser Arbeit zugrunde liegenden Analyse abgesteckt, gegenüber dem maschinellen Übersetzen und verwandten Konzepten abgegrenzt, und der Aufbau und die Funktionsweise von maschinellen Dolmetschsystemen, einschließlich der verschiedenen Modelle, werden in dem für das Verständnis der vorliegenden Arbeit erforderlichen Umfang behandelt. Darüber hinaus werden die Teilaspekte des Prozesses des maschinellen Dolmetschens unter gebührender Berücksichtigung der unterschiedlichen Ansätze zu deren Kombination aufgeschlüsselt und näher beleuchtet. Zum Abschluss dieses Kapitels wird ebenso näher auf die Limitationen und Herausforderungen im Zusammenhang mit der maschinellen Erzeugung von Dolmetschleistungen eingegangen.

Maschinelles Dolmetschen bezeichnet den weitgespannten Komplex an Vorgängen und Prozessen, durch die gesprochene Sprache in eine Maschine eingespeist und folglich in einer vorgegebenen Zielsprache lautsprachlich – beziehungsweise auch als Text – wiedergegeben wird (vgl. Barfuß 2017: 12; Fantinuoli 2018: 5, 2024: 210; Lu und Fantinuoli 2025: 2; Härtel 2016: 68). Zahlreiche Technologien und Abläufe werden in maschinellen Dolmetschsystemen vereint, um Lautsprache automatisch zu registrieren, zu verarbeiten sowie letztlich zu generieren und somit zwischen Sprachen zu vermitteln (vgl. Fantinuoli 2018: 5, 2024: 210; Härtel 2016: 68; Taylor 2009: 48). Eine Vielzahl derartiger maschineller Dolmetschsysteme ist gegenwärtig bereits verfügbar und findet jeweils in gezielter Ausrichtung auf bestimmte Zwecke und angestrebte Funktionalitäten Anwendung. Die durchwegs zielgerichtete Ausgestaltung – etwa auf Alltagskonversationen gegenüber Konferenzsettings – sowie die individuelle Konzipierung dieser Dolmetschsysteme sind vonnöten, um die angestrebten Ziele, Aufgaben und Funktionsweisen vor dem Hintergrund der inhärenten Komplexität maschineller Dolmetschungen tatsächlich erreichen zu können. Wenngleich diesem Umstand geschuldet kein Dolmetschsystem einem anderen gleicht, können die zugrunde liegenden Prozesse allgemein und softwareübergreifend in die drei beziehungsweise vier Teilprozesse der Spracherkennung sowie -verarbeitung, der maschinellen Übersetzung und letztlich der lautsprachlichen Wiedergabe synthetischer Sprache unterschieden werden, sofern eine Textausgabe nicht ausreichend ist oder nicht angestrebt wird (vgl. Hamon et al. 2007: 1; Lu und Fantinuoli 2025: 2; Waibel und Fügen 2008: 71).

Auch wenn maschinelles Dolmetschen erst seit einigen Jahren dank rasanter technologischer Fortschritte große Aufmerksamkeit genießt, wird an Systemen des maschinellen Dolmetschens bereits seit mehreren Jahrzehnten gearbeitet (vgl. Fantinuoli 2018: 5f, 2024: 211; Lu und Fantinuoli 2025: 15). Erstmals wurden derartige Systeme im späten 20. Jahrhundert entwickelt, wo sie jedoch aufgrund deutlich eingeschränkter Funktionalitäten nur in begrenztem Maße nutzbar waren (vgl. Horváth 2021: 175). Die Entwicklung, Ausgestaltung und Einsatzmöglichkeiten maschineller Dolmetschsysteme waren von Anfang an eng an den Fortschritt der ihnen inhärenten Technologien gekoppelt, insbesondere an die automatische Spracherkennung und die maschinelle Übersetzung (vgl. Fantinuoli 2025: 211). In beiden Bereichen vollzog sich seither ein grundlegender Wandel von regelbasierten und statistischen Ansätzen hin zu komplexen, auf Deep Learning beruhenden neuronalen und zunehmend multimodalen Modellen (vgl. Fantinuoli 2025: 212; Lu und Fantinuoli 2025: 1). Dieser Übergang und die zugrunde liegenden technologischen Fortschritte waren entscheidend dafür, dass maschinelles Dolmetschen heute nicht mehr ausschließlich als experimentelles Forschungsfeld angesehen wird, sondern zunehmend auch in der Praxis einsetzbare Technologien und Systeme hervorbringt (vgl. Fantinuoli 2025: 210, 212).

Insbesondere in jüngerer Zeit hat sich die Entwicklung der Technologien, die das Rückgrat des maschinellen Dolmetschens bilden, deutlich beschleunigt (vgl. Fantinuoli 2025: 211). Fortschritte in der automatischen Spracherkennung, in der Verarbeitung natürlicher Sprache (Natural Language Processing – NLP), im Bereich des Deep Learning, bei dem aus großen Datenmengen komplexe Muster entdeckt werden können, in der neuronalen maschinellen Übersetzung (Neural Machine Translation – NMT) sowie zuletzt im Bereich der generativen künstlichen Intelligenz (KI) haben die Weiterentwicklung maschineller Dolmetschsysteme in hohem Maße vorangetrieben und das technologische Ökosystem des Dolmetschens maßgeblich verändert (vgl. Fantinuoli 2018: 5f; Lu und Fantinuoli 2025: 1f). Zuletzt hat speziell das Aufkommen der generativen KI, insbesondere der sogenannten „Large Language Models“ (LLM), tiefgreifende Entwicklungsschübe in diesem Feld ausgelöst (vgl. Lu und Fantinuoli 2025: 3). Im Zuge dessen wurde nicht nur die Leistungsfähigkeit dieser Systeme und Technologien kontinuierlich verbessert, sondern ihre Ausgestaltung und Architektur grundlegend neu ausgerichtet. An die Stelle der ursprünglich genutzten Systeme, die auf dem sogenannten *Cascading*-Ansatz beruhen und in denen einzelne Teilprozesse nacheinander ablaufen, treten zunehmend *End-to-End*-Ansätze, in denen mehrere Verarbeitungsschritte innerhalb eines einzigen inte-

grierten Systems zusammengeführt werden (vgl. Lu und Fantinuoli 2025: 1). Diese beiden Ansätze sowie die wichtigsten technologischen Innovationen werden nachstehend in Kapitel 1.6 näher beleuchtet.

Infolge der auf den vorangehenden Seiten skizzierten Entwicklungen und Fortschritte werden Systeme des maschinellen Dolmetschens mittlerweile in realen Anwendungskontexten eingesetzt und sind nun in der Lage, gesprochene Äußerungen nahezu in Echtzeit zwischen Sprachen zu übertragen (vgl. Fantinuoli 2025: 210, 212). Zugleich hat sich ihre Fähigkeit zum Umgang mit komplexeren ausgangssprachlichen Eigenschaften sowie mit unterschiedlichsten Kommunikationssituationen erheblich erhöht (vgl. Lu und Fantinuoli 2025: 2f). Verbesserte Akzeptanz, geringere Kosten, höhere Portabilität und gesteigerte Benutzerfreundlichkeit haben ferner dazu beigetragen, dass die Einsatzmöglichkeiten maschineller Dolmetschsysteme um eine Vielzahl unterschiedlicher Fachgebiete, Modi und Szenarien erweitert wurden (vgl. Lu und Fantinuoli 2025: 2f). Vor diesem Hintergrund kommt der Technologie des maschinellen Dolmetschens und den entsprechenden Systemen und Applikationen gegenwärtig dank der damit bedingten Neuausrichtung des globalen Dolmetschökosystems große Aufmerksamkeit zu (vgl. Fantinuoli 2018: 5f, 2024: 211; Lu und Fantinuoli 2025: 15).

Trotz dieser Fortschritte bleibt das Ziel, unabhängig von Sprachenpaar, Dolmetschmodus sowie Kommunikationssituation und -kontext eine dem menschlichen Dolmetschen vergleichbare Qualität zu erreichen, weiterhin unerfüllt (vgl. Fantinuoli 2018: 5; 2024: 210; Sperber et al. 2025: 23345). Nach wie vor bestehen entlang des gesamten Arbeitsablaufs des maschinellen Dolmetschens Hindernisse und Probleme, wodurch weder im konsekutiven noch im simultanen Modus Leistungen auf dem Niveau menschlicher Dolmetscher*innen erzielt werden können (vgl. Fantinuoli 2018: 5; 2024: 210; Horváth 2021: 177; Sperber et al. 2025: 23345). Darüber hinaus mutet die mit maschinelltem Dolmetschen verbundene Zielvorstellung eines uneingeschränkten sprachübergreifenden Zugangs zu gesprochener Kommunikation allgemein hin als erstrebenswert an, birgt jedoch zugleich erhebliche Risiken (vgl. Fantinuoli 2025: 220). Zahlreiche Herausforderungen gehen über rein technische oder sprachliche Limitationen hinaus und betreffen die funktionale und auch gesellschaftliche Einordnung dieser Systeme (vgl. Fantinuoli 2025: 210, 220). Auf diese Limitationen und Herausforderungen wird in Kapitel 1.7 näher eingegangen.

Im Anschluss an diese allgemeine Einführung in maschinelles Dolmetschen werden auf den folgenden Seiten dieses Kapitels die einzelnen Bestandteile maschineller Dolmetschsysteme in dem für die Zwecke und für das Verständnis dieser Arbeit notwendigen Umfang näher beleuchtet. Wie in Kapitel 1.6 näher ausgeführt, hat der unablässige technische Fortschritt Systeme hervorgebracht, die mehrere der inhärenten Teilprozesse reibungs- und nahtlos in einer Weise miteinander integrieren, die eine strenge Trennung beziehungsweise Aufteilung auf technischer Ebene hinfällig macht und somit umfassendere hybride Teilprozesse entstehen lässt. Dennoch werden in den Kapiteln 1.2, 1.3 und 1.4 die zentralen bei Systemen zur Erzeugung maschineller Dolmetschleistungen zum Tragen kommenden Teilprozesse individuell im Detail behandelt, auch wenn sie je nach Ausgestaltung gegebener Systeme mitunter nicht mehr ohne Weiteres voneinander getrennt werden können beziehungsweise müssen, da diese Teilprozesse nach wie vor – wenngleich in verschiedensten Formen – im Kern einer jeden maschinellen Dolmetschung stehen. Ebenso werden die Geschichte des maschinellen Dolmetschens sowie einschlägige technische Entwicklungen und Innovationen nur insoweit beleuchtet, als dies für das verfolgte Forschungsziel notwendig ist. Zuallererst gilt es jedoch, die im Kern der vorliegenden Arbeit stehenden Begrifflichkeiten rund um das maschinelle Dolmetschen voneinander abzugrenzen.

1.1 Begriffsabgrenzung

In der vorliegenden Arbeit finden sowohl der Begriff des maschinellen Dolmetschens als auch jener des maschinellen Übersetzens Verwendung. Die Abgrenzung dieser Begrifflichkeiten und die Entscheidung, in welchem Kontext sie jeweils anzuwenden sind, sind allerdings nicht ohne Tücken. Die Terminologie für technikbasierte Formen der Übertragung gesprochener Sprache kommt in der Fachliteratur nicht einheitlich zur Anwendung. Insbesondere im Deutschen fehlt bislang eine klar standardisierte Nomenklatur, was auch damit zusammenhängt, dass ein großer Teil der einschlägigen Forschung auf Englisch publiziert wird. Hinzu kommt, dass sprachliche Unterschiede zwischen dem Deutschen und dem Englischen zu terminologischen Verschiebungen beitragen können (vgl. Pöchlacker 2024: 11). Bevor maschinelles Dolmetschen nun eingehender thematisiert wird, werden zunächst die einschlägigen Benennungen eingeordnet und voneinander abgegrenzt. Dies dient auch der Einordnung dieser Arbeit im Verhältnis zu anderen wissenschaftlichen Beiträgen mit verwandtem Forschungsgegenstand, die in englischer Sprache publiziert werden.

Je nach disziplinärem Kontext kommen für den automatischen interlingualen Transfer gesprochener Sprache unterschiedliche Bezeichnungen zur Anwendung. In der englischsprachigen Literatur sind insbesondere die Benennungen *speech translation*, *speech-to-speech translation*, *spoken language translation*, *automatic interpreting* und *machine interpreting* verbreitet (vgl. Fantinuoli 2025: 210; Lu und Fantinuoli 2025: 2; Pöchhacker 2024: 6, 10). Während *speech-to-speech translation* vor allem in der Informatik Technologien bezeichnet, die gesprochene Äußerungen einer Sprache in gesprochene Äußerungen einer anderen Sprache überführen, wird englischsprachige Forschung insgesamt häufiger unter *speech translation*, *speech-to-speech translation* oder *spoken language translation* als unter *machine interpreting* geführt (vgl. Fantinuoli 2025: 210; Lu und Fantinuoli 2025: 6; Pöchhacker 2024: 6, 10). Die terminologische Vielfalt ist hierbei nicht nur Ausdruck disziplinärer Unterschiede, sondern auch des Vorhandenseins unterschiedlicher Auffassungen darüber, was im abstrakten Gesamtsystem der Translation als Übersetzung beziehungsweise Dolmetschung zu gelten hat. Wie Pöchhacker mit Blick auf die englische Benennung hervorhebt, führt die Frage, ob *maschinelle Übersetzung* als *Übersetzung* gelten kann, unmittelbar zu der allgemeineren Frage, was grundsätzlich unter „Translation“ zu verstehen ist (vgl. Pöchhacker 2024: 1). Gerade in der Translationswissenschaft sind die einschlägigen Definitionen der Translation weit gefasst und nicht eindeutig gegeneinander abgegrenzt (vgl. Pöchhacker 2024: 1). Entsprechend ist auch die Frage, ob automatisierte sprachübertragende Systeme unter den Begriff des Dolmetschens fallen, nicht allein terminologischer, sondern zugleich konzeptueller Natur.

In Teilen der in englischer Sprache verfassten Forschung wird daher vorgeschlagen, den Begriff *interpreting* terminologisch auf menschlich kontrollierte und in Echtzeit produzierte Leistungen zu beschränken und für vollautomatische Systeme stattdessen Bezeichnungen wie *automatic speech translation* oder *automatic spoken language translation* zu verwenden (vgl. Horváth 2021: 174f, 177; Pöchhacker 2024: 16ff). Dieser Auffassung zufolge ist Dolmetschen wesentlich durch Merkmale wie Sinnkonstitution, Situationsgebundenheit, Dialogizität und Multimodalität geprägt, die von automatischer Sprachübertragung bislang nicht in vergleichbarer Weise umgesetzt werden (vgl. Pöchhacker 2024: 16). Demgegenüber ist *machine interpreting* in technologie- und translationswissenschaftlich orientierten Arbeiten als Bezeichnung durchaus etabliert und wird teils als Sammelbegriff neben *speech-to-speech translation* beziehungsweise *speech translation* verwendet (vgl. Fantinuoli 2018: 5, 2024: 210f). Die Benennung umfasst insbesondere auf eine Echtzeitanwendung ausgerichtete Systeme und kann somit als

eine spezifischere, an der unmittelbaren Nutzung orientierte Bezeichnung innerhalb des weiteren Feldes der *speech-to-speech translation* verstanden werden (vgl. Fantinuoli 2025: 211). Zudem kann *speech-to-speech translation* sowohl die zugrunde liegende Technologie als auch deren Anwendung bezeichnen, während *machine interpreting* primär auf die Anwendung bezogen ist (vgl. Fantinuoli 2025: 211). Auch neuere Arbeiten fassen *machine interpreting* breit und bezeichnen darunter Formen der Übertragung gesprochener Eingaben in schriftliche und/oder gesprochene Ausgaben, die durch Unmittelbarkeit, Echtzeitanwendung und Überschneidungen mit den Tätigkeiten menschlicher Dolmetscher*innen und Re-Speaker*innen gekennzeichnet sind (vgl. Lu und Fantinuoli 2025: 2). Obwohl diesen Ausführungen entsprechend durchaus eine terminologische Logik erkannt werden kann, besteht insgesamt kein Konsens über eine einheitliche Benennung.

Da sich somit weder im Forschungsdiskurs noch in konzeptueller Hinsicht ein einheitlicher Gebrauch feststellen lässt, wird in der vorliegenden Arbeit keine abschließende Entscheidung über diese Grundsatzfrage gefällt, sondern eine funktionale Entscheidung über die zur Anwendung kommende Terminologie. Eine Begründung für die Verwendung des Oberbegriffs *maschinelles Dolmetschen* lässt sich aus klassischen Bestimmungen des Dolmetschens ableiten, die vor allem auf Unmittelbarkeit, Einmaligkeit der Darbietung und eingeschränkte Korrigierbarkeit abstellen. So definiert Kade (1968: 34f) Dolmetschen als den Transfer eines lediglich einmalig dargebotenen ausgangssprachlichen Textes in einen zielsprachlichen Text, der nur eingeschränkt kontrollierbar und aufgrund des Zeitdrucks kaum korrigierbar ist. Auch im Kontext maschineller Dolmetschsysteme werden Ausgangstexte zwar üblicherweise nur einmal vorgetragen, können jedoch in der registrierten Form für die nachfolgende maschinelle Verarbeitung zumindest zwischengespeichert werden und sind somit theoretisch wiederholbar (vgl. Härtel 2016: 68). Dies ist – wie aus den anschließenden Unterkapiteln hervorgehen wird – dem Umstand geschuldet, dass es sich beim maschinellen Dolmetschen um keine isolierte Technologie handelt, sondern um einen komplexen multimodularen Prozess, in dem spezialisierte Teilkomponenten verschiedene Aufgaben übernehmen. Inwieweit eine gegebene Softwarelösung tatsächlich die Speicherung und Wiederholbarkeit der erfassten Eingaben ermöglicht, hängt von der Beschaffenheit des Systems ab und kann nicht pauschal beantwortet werden (vgl. Härtel 2016: 68). Dies trägt dazu bei, dass im Hinblick auf maschinell erzeugte Dolmetschleistungen *maschinelles Dolmetschen* terminologisch nicht ohne Weiteres von *maschinellem Übersetzen* abgegrenzt werden kann. Werden allerdings maschinelle Dolmetschsysteme, ihre grundlegende

Funktionalität und die Gesamtheit möglicher Verwendungszwecke betrachtet, so stehen die Aspekte der Simultanität sowie der einmaligen, hinterher nicht mehr korrigierbaren und lautsprachlichen Darbietung im Vordergrund. Folglich erscheint eine Bezeichnung des Gesamtkomplexes zugrunde liegender Vorgänge als maschinelles Dolmetschen vertretbar und findet in der vorliegenden Arbeit aus pragmatischen Gründen zur Benennung dieses Gesamtprozesses Anwendung.

Zugleich ist diesen Überlegungen entsprechend auf der Ebene der einzelnen Verarbeitungsschritte eine differenziertere terminologische Betrachtung erforderlich. Um das maschinelle Dolmetschen gesprochener Äußerungen zu ermöglichen, greifen in automatischen Dolmetschsystemen zahlreiche Technologien und Prozesse ineinander (vgl. Taylor 2009: 48). Wie in Kapitel 1.3 näher erläutert wird, können bestimmte Subprozesse, die sich auf den Transfer erkannter und aufbereiteter Elemente der Ausgangssprache in Elemente der Zielsprache vor deren Ausgabe durch Sprachsynthese beschränken, unter dem Begriff des *maschinellen Übersetzens* subsumiert werden (vgl. Fügen et al. 2007: 212; Härtel 2016: 69). Die terminologische Differenzierung zwischen dem Gesamtprozess und den darin integrierten Teilprozessen spiegelt sich folglich auch in der technischen Architektur dieser Systeme wider. Die Verwendung des Oberbegriffs *maschinelles Dolmetschen* für den Gesamtprozess steht somit nicht im Widerspruch zu einer engeren Bezeichnung des unter *maschinelles Übersetzen* geführten Teilprozesses. Vor diesem Hintergrund wird in den nachstehenden Kapiteln die Funktionsweise dieser Systeme anhand ihrer grundlegenden Komponenten und Verarbeitungsschritte näher dargelegt.

1.2 Spracherkennung

Am Beginn eines jeden Dolmetschsystems kann die Spracherkennung einschließlich zugehöriger unterstützender Subprozesse angesiedelt werden (vgl. Barfuß 2017: 12). Lautsprache wird automatisch registriert, analysiert, kategorisiert, aufbereitet und gegebenenfalls korrigiert. Aus diesen Vorgängen resultierend werden ursprüngliche ausgangssprachliche Äußerungen typischerweise in schriftlicher Repräsentation festgehalten. Es handelt sich hierbei allerdings nicht um eine Technologie, die lediglich der automatischen Transkription gesprochener Sprache dient. Vielmehr können unter anderem je nach Verwendungszweck und -ziel des jeweiligen Systems auch Informationen über einzelne oder mehrere Sprecher*innen, etwaige Sprecher*innenwechsel, verwendete Sprachen und die Verwendung von Schlüsselwörtern von Relevanz sein und gegebenenfalls gesammelt werden (vgl. Euler 2006: 15).

Die vielfältigen Einsatzmöglichkeiten und Technologien heben den inhärenten Charakter maschineller Dolmetschsysteme hervor, bei Ausrichtung auf vorgegebene und klar formulierte Zielsetzungen optimale Funktionalität zu erreichen. So sind die Spracherkennungsprozesse mancher Dolmetschsysteme darauf ausgelegt und ausgerichtet, Sprecher*innen und deren Wechsel automatisch zu erkennen, während andere wiederum aufgrund der Ausrichtung und Abstimmung auf Einzelpersonen ohnehin nur von diesen Individuen sachgemäß bedient werden können (vgl. Fügen et al. 2007: 215, 228f; Tran et al. 2025: 630). Ebenso geht aus der angestrebten Funktionalität gegebener Systeme hervor, ob Ausgangs- und Zielsprache automatisch erkannt werden können, ob diese im Vorfeld vorzugeben sind und für welche Sprachen das System grundsätzlich geeignet ist (vgl. Härtel 2016:71f; Horváth 2021: 176).

Unabhängig von Zweck und Zielsetzung benötigen Spracherkennungssysteme Datenbanken, die je nach angestrebten Einsatzmöglichkeiten mit Wörtern in einer bestimmten Quantität und Qualität gefüllt werden müssen (vgl. Härtel 2016: 72). Registrierte Audioinformationen werden mit diesen Einträgen abgeglichen, um Ansatzpunkte für nachfolgende Analysen, Kategorisierungen und Segmentierungen zu schaffen (vgl. Härtel 2016: 72). Die konkrete Ausgestaltung dieser Datenbanken könnte von Rechen- und Speicherkapazitäten und deren Limitationen geleitet werden, jedoch auch mit einer geringen Zahl an Worteinträgen auf eine bessere Erkennungsgenauigkeit abzielen (vgl. Härtel 2016: 72). Fachsprachliche Monologe in konferenztypischen Situationen, die für die vorliegende Arbeit relevant sind, erfordern Datenbanken mit einer beträchtlichen Anzahl an Einträgen (vgl. Härtel 2016: 72).

Die Aufnahme möglichst vieler oder gar aller möglicherweise bedeutsamen Wörter in die eingespeisten Datenbanken ist dementsprechend nicht für alle maschinellen Dolmetschsysteme zwingendermaßen zielführend. Zudem sind neben der Grundform eingetragener Wörter in vielen Fällen auch Flexionen aufzunehmen, um der Maschine einen sachgemäßen Rückgriff gar erst zu ermöglichen (vgl. Härtel 2016: 73). Typischerweise sind überdies jene kommunikativen Situationen, die Dolmetschleistungen erfordern, von einer Vielzahl an Eigennamen und Termini aus diversen Fachbereichen und -gebieten gekennzeichnet (vgl. Kutz 2012: 183). Eine zweckgebundene Ausrichtung des maschinellen Dolmetschsystems ist auch in diesen Situationen für eine optimale Funktionalität nützlich, indem möglicherweise relevante Eigennamen, beispielsweise in der Form von nennenswerten Teilnehmer*innen und Sprecher*innen sowie Fachtermini, die sich aus der zugrunde liegenden Thematik ergeben, eingespeist werden (vgl. Fügen et al. 2007: 215). In Ermangelung von Entsprechungen in Datenbanken, zum Beispiel

für nicht erfasste Termini oder nicht berücksichtigte Personen, ist es technisch durchaus denkbar, maschinelle Dolmetschsysteme das Internet nach entsprechenden Informationen durchforschten zu lassen (vgl. Härtel 2016: 73).

Dementsprechend können maschinelle Dolmetschsysteme je nach angestrebter Funktion und Zweck in Ausrichtung und Funktionsweise stark variieren. Ungeachtet der Zielsetzung besteht das Grundproblem der automatischen Spracherkennungsprozesse darin, aus der Flut an Informationen den Kerngehalt herauszufiltern, der die zu vermittelnde Botschaft enthält (vgl. Taylor 2009: 22). Lautsprachliche Äußerungen können hierbei allerdings nicht unmittelbar als Einzelwörter oder -silben erfasst werden (vgl. Hübener 1997: 9f). Empfangene Signale müssen erst in rascher Abfolge gemessen, aufbereitet und verarbeitet werden, während jene Informationen ausgeblendet werden, denen keine kommunikationsrelevanten Inhalte beigemessen werden (vgl. Härtel 2016: 74; Pfister und Kaufmann 2008: 40). Diese folglich digitalisierte Datenmenge und anhaftende Zusatzinformationen werden mathematisch auf ihren Kerngehalt analysiert, reduziert und anschließend segmentiert (vgl. Härtel 2016: 74ff). Sofern diese Möglichkeit bei einem gegebenen System Berücksichtigung findet, können bei der Spracherkennung auch zusätzlich Prosodie-Daten erfasst werden, um Grenzen einzelner Wörter oder ganzer Sätze zu erfassen (vgl. Nöth 1991: 48; Sasu 2025: 85, 88). Die Bedeutung und Beschaffenheit von Prosodie wird in Kapitel 1.5 vor allem in Hinblick auf Einsatzmöglichkeiten und ihre Rolle in der Synthese künstlicher Sprache – und folglich im gesamten Dolmetschprozess – näher beleuchtet. Nach erfolgter Durchführung aller zugrunde liegenden Abläufe wird eine an Informationen möglichst dichte Grundlage, üblicherweise einschließlich ganzer Wörter mit passender Lexik und vollständiger Satzkonstruktionen, für die darauffolgende Weiterverarbeitung der maschinellen Übersetzung geschaffen.

Der letztlich erfolgreichen Erkennung von Lautsprache steht allerdings eine Fülle an Hindernissen im Wege. Einzelne Vorgänge können hierbei mit Schwierigkeiten behaftet sein, die auch fortwährende und weitreichende Folgen für die nachfolgenden Prozesse der maschinellen Übersetzung und der Sprachsynthese haben können (vgl. Horváth 2021: 176). Allgemein hin wird Sprache meist spontan und ohne tiefgreifende Planung geäußert, da Menschen üblicherweise keine ausgedehnten Pausen einlegen, bis ihnen in der Rolle als Sprecher*innen perfekte Formulierungen in den Sinn kommen (vgl. Taylor 2009: 30). Diesem Umstand entspringt die Vielzahl an Hesitationen, Füllwörtern, grammatisch unvollständigen oder inkorrekten Sätzen, Pausen und Neustarts, die in alltäglicher Kommunikation keine Seltenheit darstellen (vgl. Fantinuoli 2025: 215; Lu und Fantinuoli 2025: 8; Taylor 2009: 30). Allerdings werden

auch diese Elemente gemeinsam mit Räuspern, Niesen, Husten, Trinken, Umgebungsgeräuschen und ähnlichen Klängen von Maschinen registriert, wodurch zwangsläufig der Verlust von Informationen begünstigt wird (vgl. Petrick 2009: 2; Pfister und Kaufmann 2008: 362). Je geringer der Umfang der entsprechenden Hintergrundinformationen und Datensätze für die betreffenden Sprachen ist, desto problematischer gestaltet sich ferner der Umgang damit für das System (vgl. Fantinuoli 2025: 215).

Bei ausschließlicher Erkennung und Erfassung von Audioinformation fällt überdies eine Reihe weiterer möglicherweise relevanter Informationen unter den Tisch (vgl. Taylor 2009: 30). Nonverbale und paraverbale Signale unterstützen die Kommunikation, vermitteln mitunter essenzielle Inhalte, übersteigen jedoch die Aufnahmekapazitäten handelsüblicher maschineller Dolmetschsysteme (vgl. Härtel 2016: 83; Horváth 2021: 176). Die fehlende Beachtung von Gestik, Mimik, Distanzverhalten und dergleichen kann demgemäß ebenfalls den Verlust von Informationen bedeuten (vgl. Horváth 2021: 176; Kutz 2012: 253ff). Eine sachgemäße Beachtung auch dieser kommunikativen Aspekte durch maschinelle Dolmetschsysteme liegt allerdings durchaus im Rahmen des Denk- und Umsetzbaren. Bei Beiträgen im Rahmen von Sitzungen des Europäischen Parlaments, von denen einer als Ausgangsmaterial der für die vorliegende Arbeit ausgewählten Dolmetschleistungen diente, können derartige Merkmale von Dolmetscher*innen visuell wahrgenommen und berücksichtigt werden.

Aufgrund der enormen Komplexität maschineller Spracherkennung – zu deren Störfaktoren diverse Aspekte wie die Physiologie und der Gemütszustand von Sprecher*innen zählen – können jederzeit bei allen zugrunde liegenden Subprozessen Fehlinterpretationen und -analysen unterlaufen (vgl. Taylor 2009: 22). Die Reduktion dieser Faktoren auf ein Minimum bedingt den Einsatz vielschichtiger sowie problem- und zielgerichteter Technologien (vgl. Härtel 2016: 70, 74f), die nicht im Fokus der vorliegenden Arbeit liegen und daher nicht näher behandelt werden.

1.3 Maschinelle Übersetzung

Nach abgeschlossener Aufbereitung der registrierten – und je nach Ausgestaltung des betreffenden Systems in Textform umgewandelten – Daten können diese maschinell in ein anderes Sprachsystem übersetzt werden (vgl. Horváth 2021: 176; Waibel und Fügen 2007: 71). Auch dieser Teilprozess eines maschinellen Dolmetschsystems ist von einer Vielfalt an Technologien, unterschiedlichen Zielsetzungen und dementsprechend verschiedenen Ansätzen geprägt.

Der Erfolg maschineller Dolmetschsysteme hängt grundsätzlich zu einem großen Teil von Fortschritten und Innovationen im Bereich des maschinellen Übersetzens ab (vgl. Horváth 2021: 176). Dementsprechend haben sich die Systeme des maschinellen Übersetzens – analog zum maschinellen Dolmetschen – von ursprünglich regelbasierten Ansätzen, über korpusbasierte, statistische Modelle hin zu auf künstlichen neuronalen Netzwerken gestützten Systemen entwickelt und zuletzt LLM für sich erschlossen (vgl. Horváth 2021: 176; Lu und Fantinuoli 2025: 5). Da der Fokus der vorliegenden Arbeit auf den lautsprachlichen Aspekten maschinell erzeugter Dolmetschleistungen liegt, werden diese Ansätze im Folgenden lediglich in den für das Verständnis erforderlichen Grundzügen theoretisch beleuchtet.

Im Rahmen des sogenannten regelbasierten Ansatzes zur Automatisierung maschineller Übersetzungen dienen der Maschine beigebrachte Regeln der Analyse von Daten, aus denen folglich ein Text in der Zielsprache erzeugt wird (vgl. Barfuß 2017: 12; Fantinuoli 2025: 212; Ramlow 2009: 59). Dieser ursprünglich verfolgte Ansatz war jedoch aufgrund der schieren Komplexität der Sprachen, der Begrenzungen im Hinblick auf verfügbare Rechenleistung und Daten sowie des mit der Schaffung angemessener Regeln einhergehenden Arbeitsaufwandes in seiner Anpassungsfähigkeit und Erweiterbarkeit stark begrenzt (vgl. Son und Kim 2023: 1, 3). Der korpusbasierte Ansatz stützt sich hingegen in der Generierung zielsprachlicher Texte ausschließlich auf ein Korpus von verfügbaren und erschließbaren Paralleltexten (vgl. Barfuß 2017: 12; Fantinuoli 2025: 212; Härtel 2016: 87f). Dabei wird aus den zugrunde liegenden Informationen jene Aneinanderreihung zielsprachlicher Wörter statistisch errechnet, bei der es sich um die mit größter Wahrscheinlichkeit passende Übersetzung handeln soll (vgl. Son und Kim 2023: 3). Allerdings werden bei mit dieser Methodik angefertigten Übersetzungen inhaltliche Verbindungen in ausgangssprachlichen Texten und der jeweilige Kontext typischerweise nicht im erforderlichen Maße berücksichtigt (vgl. Fantinuoli 2025: 212; Son und Kim 2023: 3). Beide Ansätze und Systeme können in individueller Ausgestaltung kombiniert werden. Dementsprechend können Übersetzungen auf der Grundlage grammatischer Regeln erzeugt und anschließend statistisch überarbeitet oder für eine regelbasierte Nachbearbeitung statistisch vorverarbeitet werden (vgl. Härtel 2016: 89; Koehn 2010: 82ff).

Allen Bestrebungen zur vollautomatischen Erzeugung von Übersetzungen ist ein enormer Bedarf an Daten gemein, der für einen korrekten und sachgemäßen Umgang mit Sprache bei maschinellen Übersetzungen – wie bei jedem anderen Vorgang, Teilschritt und Subprozess eines maschinellen Dolmetschsystems – gedeckt werden muss (vgl. Glück 2010: 373; Lu und Fantinuoli 2025: 4). Tatsächlich ist die Menge weltweit und jederzeit verfügbarer Daten rapide

angewachsen (vgl. Klein et al. 2013: 319). Diese neu entstandene und unaufhörlich wachsende Datenmenge immensen Ausmaßes trägt die Bezeichnung „Big Data“ und kann für maschinelle Übersetzungen beziehungsweise Dolmetschungen herangezogen werden, um etwaige Muster, Trends und Beziehungen diverser Art aufzudecken (vgl. Joshi und Marthandan 2018: 1). Die Herausforderung im Umgang mit Big Data besteht darin, diesen Datenkomplex, der für traditionelle Datenbanken bei Weitem zu umfassend ist, sachgemäß zu erschließen und für maschinelle Übersetzungs- beziehungsweise Dolmetschsysteme nutzbar zu machen (vgl. Klein et al. 2013: 319; Marz et al. 2016: 15).

Eine Möglichkeit, diese Daten für diverse Zwecke verfügbar zu machen, bieten künstliche neuronale Netzwerke, die in der Form von mathematischen Nachbildungen komplexer Abläufe des menschlichen Gehirns zur Identifizierung von sonst schwer entdeck- und nachvollziehbaren Mustern und Gesetzmäßigkeiten Verwendung finden (vgl. Barfuß 2017: 14f; Chang 2011). Mit der Entwicklung entsprechender Technologien wurde damit begonnen, künstliche neuronale Netzwerke auch für maschinelle Übersetzungen zu erschließen (vgl. Son und Kim 2023: 2). Insbesondere im Hinblick auf inhaltliche Verflechtungen sowie den jeweiligen zugrunde liegenden Kontext des ausgangssprachlichen Materials lassen sich mit dieser Technologie bessere Übersetzungsleistungen erzielen als mit den davor bestehenden Ansätzen (vgl. Son und Kim 2023: 3).

Mithilfe des sogenannten *Context-Dependent Deep Neural Network Hidden Markov Model* wurde das Potenzial dieser Technologie genutzt, um Erkennungsraten und Übersetzungsgenauigkeiten kontinuierlich zu erhöhen (vgl. Barfuß 2017: 15; Klie 2015: 13). Diese durch autonomes Lernen angetriebene unentwegte Weiterentwicklung aller Teilprozesse und Technologien maschineller Dolmetschsysteme führte zusehends auch zur Ausmerzung von Fehlerquellen und zur Erschließung neuer Funktionalitäten (vgl. Barfuß 2017: 16). Rasch danach wurden die auf Hidden Markov Modellen basierenden Modelle allerdings zunehmend von Deep-Learning-basierten Ansätzen verdrängt (vgl. Peinl 2021: 4). Infolge der mithilfe dieser Technologien erzielten Fortschritte bei maschinellen Übersetzungen rückten schließlich Möglichkeiten zur Steigerung von Skalierbarkeit und Effizienz, die Fähigkeit der Systeme zur Anpassung an verschiedene Themengebiete und das Ziel eines möglichst menschenähnlichen Verständnisses zugrunde liegender Inhalte in das Zentrum der Forschung auf diesem Gebiet (vgl. Son und Kim 2023: 2).

Diesen Zielsetzungen entsprechend werden vormalig eingesetzte Ansätze infolge des technologischen Fortschritts und der Entwicklung neuer Systeme nicht einfach ersetzt, sondern können vielmehr in diese integriert und mit ihnen kombiniert werden (vgl. Son und Kim 2023: 3). Durch eine solche Kombination verschiedener Ansätze in Ausrichtung auf konkrete Zielsetzungen sollen ihre Limitationen überwunden und Übersetzungen höherer Qualität erzielt werden (vgl. Son und Kim 2023: 3). Diesen Entwicklungen entsprechend wird ein vielschichtiger, umfassender Ansatz für maschinelle Übersetzung verfolgt, der mitunter auf den Fähigkeiten der Generativen KI fußt und zuletzt neben NMT auch LLM – wie etwa ChatGPT – für sich erschlossen hat (vgl. Fantinuoli 2025: 214; Lu und Fantinuoli 2025: 5). Mit dem Erfolg dieser neuen Technologie und ihrer weitverbreiteten Übernahme durch die Allgemeinbevölkerung kamen auch den dadurch neu eröffneten Anwendungsmöglichkeiten im Bereich der maschinellen Übersetzung großes Interesse zu (vgl. Son und Kim 2023: 3). Dank der inhärenten Fähigkeit zum autonomen Lernen können LLM mit verschiedensten Themengebieten umgehen und mitunter bemerkenswerte Übersetzungsleistungen erzielen (vgl. Son und Kim 2023: 3). Entscheidend für die Qualität der mithilfe von LLM angefertigten Übersetzungen ist hierbei das zugrunde liegende Sprachenpaar, wobei in manchen bessere Ergebnisse erzielt werden als in anderen (vgl. Son und Kim 2023: 15). LLM können Übersetzungen hoher Qualität gänzlich ohne Rückgriff auf entsprechende Korpora während des Übersetzungsvorgangs anfertigen, müssen zu diesem Zweck jedoch während ihrer Entwicklungsphase entsprechend trainiert worden sein (vgl. Li et al. 2024: 576). Qualität und Umfang des Trainingsmaterials sind ausschlaggebend für die in weiterer Folge in den entsprechenden Sprachenpaaren erzielten Übersetzungsleistungen (vgl. Horváth 2021: 177; Li et al. 2024: 583; Lu und Fantinuoli 2025: 4; Son und Kim 2023: 15).

Trotz der Erfolge werden auf anderen Technologien basierende Systeme nicht einfach durch die auf LLM gestützten Übersetzungsleistungen und -möglichkeiten ersetzt. Einer Studie aus dem Jahr 2023 zufolge konnten mithilfe von Google Translate und Microsoft Translator bei Übersetzungen aus dem Englischen in eine andere Sprache bessere Ergebnisse erzielt werden als mit dem LLM GPT-4 (vgl. Son und Kim 2023: 14). Herkömmliche Systeme können demzufolge – insbesondere bei Übersetzungen aus dem Englischen in andere Sprachen – in bestimmten Kontexten und Situationen mit LLM mithalten (vgl. Son und Kim 2023: 14). An LLM wird jedoch die Erwartung gestellt, mit der weiteren technologischen Entwicklung über eine stets wachsende Anzahl an Themengebieten und Sprachenpaaren hinweg noch genauere Über-

setzungen erzielen zu können (vgl. Fantinuoli 2025: 216; Son und Kim 2023: 2). Bislang beschränkt sich die experimentelle Forschung in diesem Bereich jedoch auf bestimmte Sprachen, wodurch noch nicht abschließend geklärt werden konnte, welche Leistungen entsprechende Systeme bei jenen Sprachen theoretisch realisieren können, für die Daten in geringerem Umfang vorhanden sind (vgl. Lu und Fantinuoli 2025: 16). Zur zukünftigen Entwicklung von LLM kommt erschwerend hinzu, dass ihre exakte Funktionsweise nicht gänzlich transparent ist – auch nicht für deren Entwickler (vgl. Wang und Wang 2025: 2). Trotz der bestehenden Limitationen und Herausforderungen lassen sich mithilfe von maschineller Übersetzung – abhängig von Ausgestaltung und Zielsetzung des betreffenden Systems – als für die maschinelle Dolmetschung entscheidenden Subprozess bereits Übersetzungen hoher Qualität erzeugen.

1.4 Sprachsynthese

Im Rahmen der oben dargelegten Prozesse werden ausgangssprachliche Äußerungen erkannt, verarbeitet und letztlich wird in Form einer maschinellen Übersetzung eine zielsprachliche Basis geschaffen, die Nutzer*innen nun lautsprachlich verfügbar gemacht werden muss. Die Ausgabe einer maschinell angefertigten Übersetzung könnte zwar in rein schriftlicher Textform erfolgen, erfordert allerdings die Möglichkeit einer lautsprachlichen Wiedergabe, um dem zugrunde liegenden Soft- und Hardwarekomplex den Charakter eines maschinellen Dolmetschsystems zu verleihen. Erreicht wird dies durch die synthetische Generierung von Sprache, auch Sprachsynthese genannt (vgl. Black und Lenzo 2004: 761; Dutoit 1997: 13; Härtel 2016: 94; Horváth 2021: 177; Taylor 2009: 1). Die Sprachsynthese ist ein Gebiet, das in technischer Hinsicht von raschen Innovationen, Entwicklungen und diversen Techniken geprägt ist. Daher soll im vorliegenden Kapitel nicht vollumfassend geschildert werden, wie Sprache auf technischer Ebene synthetisiert werden kann, sondern vielmehr sollen Aspekte, Probleme, grundlegende Funktionsweisen und maßgebliche Entwicklungen im Zusammenhang mit maschinellem Dolmetschen aufgezeigt werden, um eine angemessene Grundlage für das der Arbeit zugrunde liegende Forschungsziel zu schaffen.

Für die künstliche Nachbildung akustischer Sprache muss die durch vorgelagerte Prozesse registrierte Grundlage erst in kleinere Bestandteile aufgeteilt, klassifiziert, disambiguiert und bereinigt werden, indem unter anderem der zielsprachliche Text in ein lesbareres Format – bei manchen Systemen etwa reine Kleinbuchstaben – konvertiert wird oder enthaltene Homographe eindeutig gemacht werden (vgl. Härtel 2016: 93; Taylor 2009: 64). Für die Sprachausgabe der somit erhaltenen Daten kann die Anatomie des Menschen ohne große Speicherplatzerfordernisse nachgeahmt werden (vgl. Härtel 2016: 94f). Natürlich klingende Sprache bedarf im

Rahmen dieser Erzeugungsvariante allerdings einer Reihe überaus komplexer Regeln, die in Nachahmung menschlicher Anatomie nur schwer zu realisieren ist (vgl. Härtel 2016: 94f). Alternativ kann auf bereits aufgezeichnete lautsprachliche Äußerungen von Menschen zurückgegriffen werden, die erst in ihre Bestandteile aufgegliedert und für auszugebende Äußerungen wieder entsprechend zusammengefügt werden (vgl. Taylor 2009: 412). Da die zugrunde liegenden Aufnahmen von Menschen erstellt werden, kann der erzeugten Sprache auf diesem Wege schon seit Langem ein vergleichsweise natürlich anmutender Charakter verliehen werden (vgl. Olive et al. 1998: 194).

Grundsätzlich sind Sprachsynthesysteme schon seit über 40 Jahren im Stande, zielsprachliche Äußerungen auf eine in gewissem Maße verständliche Art und Weise zu vermitteln (vgl. Taylor 2009: 48). Zudem lassen sich heutzutage mithilfe von LLM wie ChatGPT – wenn gleich in erster Linie noch abseits des maschinellen Dolmetschens – bereits überaus natürlich klingende Sprachausgaben erzeugen (vgl. Xie et al. 2025: 764f). Die Erzeugung einer tatsächlich menschlich klingenden und nicht als künstliche Nachempfindung identifizierbaren Stimme bleibt allerdings im Kontext der maschinellen Dolmetschung weiterhin eine unerreichte Idealvorstellung (vgl. Lu und Fantinuoli 2025: 8; Taylor 2009: 48). Trotz der grundsätzlich enormen technischen Fortschritte – sowohl bei maschinellen Dolmetschsystemen insgesamt als auch bei der Sprachsynthese im Einzelnen – kann eine natürlich und menschlich anmutende Qualität bei der zielsprachlichen Wiedergabe im Zuge einer maschinellen Dolmetschung noch nicht durchgängig erzielt werden (vgl. Lu und Fantinuoli 2025: 8). Eine der dafür maßgeblichen Faktoren ist die ordnungsgemäße Berücksichtigung von Prosodie (vgl. Lu und Fantinuoli 2025: 8).

Ungeachtet dessen, inwiefern prosodische Daten und Informationen in der Sprachsynthese berücksichtigt werden, bedarf es ausgeklügelter Techniken und Prozesse, um aus Text ein durchgehendes akustisches Signal zu erzeugen (vgl. Taylor 2009: 192). Korrekte und legitime Möglichkeiten zur Aussprache einzelner Wörter können ohne weitere technologische Unterstützung oder Aufbereitung weder von der ihr anhaftenden Bedeutung noch von der schriftsprachlichen Repräsentation erschlossen werden (vgl. Horváth 2021: 177; Taylor 2009: 207). Um Text maschinell auf die erwünschte Art und Weise akustisch wiederzugeben, müssen diese Beziehungen zwischen Schrift- und Lautsprache festgehalten werden (vgl. Burkhardt 2001: 36, 38; Taylor 2009: 207f). Jene Wörter einschließlich aller Schriftformen und Aussprachemöglichkeiten, die dem Dolmetschsystem geläufig sein sollen, müssen im Vorfeld in Datenbanken eingespeist und festgehalten werden (vgl. Taylor 2009: 63). Dementsprechend wird in der Sprachsynthese auf die in den Wörterbüchern eingebetteten Paare von geschriebener Form und Aussprache

zurückgegriffen sowie unbekannte – sprich nicht in den Datenbanken enthaltene – Wörter über Algorithmen erschlossen (vgl. Taylor 2009: 192, 207f). Insbesondere zur Aufgliederung sowie annähernden Aussprache von nicht erfassten Eigennamen, Komposita oder auch Lehnwörtern sind diese überaus komplexen Algorithmen von großer Bedeutung (vgl. Härtel 2016: 93, 97f; Taylor 2009: 217f).

Für die anschließende lautsprachliche Wiedergabe ziehen maschinelle Dolmetschsysteme Einträge aus Datenbanken heran, anhand derer sie das gewünschte akustische Signal ausgeben können. Die Einträge dieser Datenbanken können die Form ganzer Wörter, einzelner Phoneme, Morpheme, Silben, Phrasen und mitunter auch vollständiger Sätze annehmen, die anschließend aneinandergereiht werden (vgl. Burkhardt 2001: 39f; Härtel 2016: 69, 93; Taylor 2009: 208). Je näher solch ein System an der Speicherung ganzer Wörter orientiert ist, desto mehr Einträge sind vonnöten, um alle gewünschten Benennungen abdecken zu können (vgl. Taylor 2009: 193ff). Finden allerdings kleinere Bestandteile unterhalb der Wortebene Verwendung, sind wesentlich weniger Einträge erforderlich, da mit einer geringen Anzahl an Einträgen und Elementen bereits eine Vielzahl an Wörtern konstruiert werden kann (vgl. Taylor 2009: 193ff). Mit einer solchen Vorgehensweise geht allerdings eine erhöhte Anfälligkeit für Ungenauigkeiten einher (vgl. Taylor 2009: 193ff).

Da in den zugrunde gelegten Wörterbüchern und Datenbanken alle Informationen eingebettet sein müssen, die zur Aussprache der maschinell generierten Übersetzung notwendig sind, müssen die zugehörigen Einträge von hoher Qualität sein, um eine einheitliche Aussprache zu gewährleisten und die Fehleranfälligkeit zu minimieren (vgl. Taylor 2009: 215, 217). Da Wörter auf verschiedenste Art und Weise ausgesprochen werden können, ist auch ein durchwegs einheitlicher Akzent nicht problemlos realisierbar und muss durch entsprechende Berücksichtigung in den Datenbanken gewährleistet werden (vgl. Taylor 2009: 215, 217). Selbst innerhalb einer gegebenen Akzentgruppe existieren zumeist mehrere korrekte, valide oder zumindest akzeptable Aussprachemöglichkeiten nebeneinander (vgl. Taylor 2009: 215, 217). Obwohl diese Umstände den Sprachsyntheseprozessen diesbezüglich einen relativ weit gefassten Spielraum gewähren, muss dennoch mit höchster Präzision gearbeitet werden, um den zur Verfügung stehenden Ermessensrahmen nicht zu übersteigen (vgl. Taylor 2009: 215, 217).

Auch hinsichtlich Homographie werden die Entwickler*innen maschineller Sprachsyntheseprozesse vor Herausforderungen gestellt. Idealerweise wird die zugrunde liegende Ambiguität bereits im Zuge der maschinellen Übersetzung weitestgehend und automatisch beseitigt

(vgl. Härtel 2016: 97). Auf Basis der während dieses Prozesses vorgenommenen Analysen können zusätzliche Informationen in die Synthesekomponenten eingespeist und somit die Genauigkeit und Qualität der durch das System ausgegebenen lautsprachlichen Dolmetschung verbessert werden (vgl. Härtel 2016: 97).

Zuletzt konnte insbesondere durch den Einsatz tiefer neuronaler Netze, Deep Learning und LLM die Qualität der Sprachsynthese im Allgemeinen stark verbessert werden, die abseits des Bereichs, der Bedürfnisse und der Erfordernisse des maschinellen Dolmetschens bereits sehr gute Leistungen erzielt (vgl. Peinl 2021: 1, 4; Xie et al. 2025: 764f). Damit die mithilfe dieser Technologien realisierte Audioausgabe tatsächlich hohe Qualitätsanforderungen erfüllen kann, ist der Zweck des jeweiligen Systems zu berücksichtigen (vgl. Ren et al. 2019: 5410). Darüber hinaus können bei manchen Systemen für die Erstellung von wenigen Minuten an Audiomaterial mehrere Stunden erforderlich sein (vgl. Ren et al. 2019: 5410, 5417). Um leistungsstarke Ergebnisse erzielen zu können, müssen die Systeme erst einem Training unterlaufen, was ebenso eine lange Zeit in Anspruch nehmen kann – mitunter sind bis zu 50 Stunden vonnöten, um eine hohe Qualität der Audioausgabe zu ermöglichen (vgl. Peinl 2021: 1). Im Mittelpunkt der Forschung stand in jüngster Zeit nicht mehr die Verbesserung der Qualität, sondern die Weiterentwicklung der Systeme im Hinblick auf Schnelligkeit bei möglichst gleichbleibender Qualität (vgl. Peinl 2021: 4). Da auf die Darbietung eines Ausgangstextes eine möglichst als angemessen angesehene *Décalage* – die Verzögerung zwischen dem Moment, in dem die*der Redner*in eine Aussage tätigt, und dem Zeitpunkt, zu dem die*der Rezipient*in die Dolmetschung empfängt – bis zur ausgangssprachlichen Wiedergabe folgen sollte, müssen im Rahmen von maschinellen Dolmetschungen Ergebnisse mit möglichst geringem Zeitverzug ausgegeben werden. Folglich sind jene Sprachsynthesysteme, die auf langen Verarbeitungsprozessen beruhen, für die maschinelle Erzeugung von Dolmetschleistungen nicht geeignet. Eine beständige Herausforderung im Feld der Sprachsynthese ist ferner dem Umstand geschuldet, dass sie im Vergleich zu den Prozessen der Spracherkennung und der maschinellen Übersetzung im Nachhinein nicht mehr korrigiert oder umgekehrt werden kann (vgl. Carstensen et al. 2001: 462). Um Fehlern bestmöglich vorzubeugen, bietet es sich dementsprechend an, den Fokus der Präventionsbemühungen auf Fehlerquellen aller vorgelagerten Prozesse zu legen, indem beispielsweise zusätzliche prosodische Informationen bereits in der Spracherkennung gesammelt werden (vgl. Carstensen et al. 2001: 465).

1.5 Prosodie

Nicht nur in den vorgelagerten Prozessen, sondern auch für die Sprachsynthese selbst ist Prosodie von immenser Bedeutung und Relevanz. Sie bezeichnet jenen Teil der menschlichen Kommunikation, der Emotionen Ausdruck verleiht, Wortbetonungen vermittelt, die Haltung der sprechenden Person widerspiegelt, Sätze in Phrasen unterteilt, den Rhythmus der Sätze steuert sowie Intonation, Höhe und Klang der Sprache kontrolliert (vgl. Batliner et al. 2000: 106; Härtel 2016: 80f; Taylor 2009: 111, 121). Prosodie ist hierbei nicht an Phoneme gekoppelt, trägt mitunter essenzielle Informationen und gibt mittels Pausenverhalten und angewendeter Sprachmelodie Aufschluss über Satzgrenzen (vgl. Batliner et al. 2000: 106; Härtel 2016: 81). Wenngleich bestimmte Regeln und Regelmäßigkeiten für die Anwendung von Prosodie in menschlicher Kommunikation zweifelsohne existieren, wie beispielsweise in der Gestalt von Betonungen der neu in ein Gespräch eingebrachten Informationen, gibt es in der Forschung dennoch keinen Konsens über die exakte Funktionsweise und Regelmäßigkeit von Prosodie (vgl. Batliner et al. 2000: 106; Taylor 2009: 118, 121). Dementsprechend existiert bislang kein einheitliches sowie universell anerkanntes Modell, und Studien zu dieser Thematik ziehen in individueller Zielsetzung unterschiedliche prosodiebezogene Parameter heran (vgl. Galdino et al. 2025: 465).

Wie Menschen unterschiedliche Geräusche zur lautsprachlichen Kommunikation produzieren, kann in gewissem Maße nachvollzogen und generalisiert werden. Noch gestaltet es sich allerdings in technischer Hinsicht schwierig, exakt festzustellen und zu erklären, wie sich die Signalenergie und Wellenformen von Sprache zwischen sprechenden Personen unterscheiden und welche Rolle Prosodie hierbei einnimmt (vgl. Neururer et al. 2024: 64f; Taylor 2009: 146). Prosodie wird äußerst individuell zur Anwendung gebracht und variiert in Ausprägung, Nutzung und Intensität je nach sprechender Person, ihren spontanen und persönlichen Entscheidungen sowie der jeweiligen kommunikativen Situation (vgl. Taylor 2009: 126). Der große in der Kommunikation zur Verfügung stehende Spielraum in Verbindung mit Dialekten und Varietäten erschwert die Ergründung, Erforschung und das Verstehen von Prosodie zusätzlich (vgl. Härtel 2016: 81). Dementsprechend existiert eine Vielzahl an Modellen, Theorien und Erklärungsversuchen für Phänomene der Prosodie (vgl. Galdino et al. 2025: 465; Taylor 2009: 259).

Aufgrund der Unsicherheit darüber, wie Prosodie zu analysieren, zu modellieren und zu synthetisieren ist, sowie der tragenden Rolle, die Prosodie in menschlicher Kommunikation

einnimmt, wird in Sprachsyntheseprozessen maschineller Dolmetschsysteme versucht, sich einer natürlich anmutenden Verwendung in Ausrichtung auf bestimmte vordefinierte Zielsetzungen gezielt anzunähern (vgl. Härtel 2016: 92; Taylor 2009: 111). Zum genauen Verständnis solcher realisierbaren Einsatzmöglichkeiten sowie der inhärenten Limitationen von Prosodie im Kontext von maschinellen Dolmetschsystemen wird im Folgenden nun anhand einer möglichen Unterteilung näher auf ihre Beschaffenheit und Charakteristika eingegangen.

Der Großteil der bereits beschriebenen und allgemeinsprachlich der Prosodie zugeschriebenen Funktionen kann unter dem Begriff der affektiven Prosodie subsumiert werden. In menschlicher Kommunikation werden Emotionen, Empfindungen und der eigene Gemütszustand primär auf diesem Wege vermittelt (vgl. Taylor 2009: 123). Beispielsweise werden Gesprächspartner*innen eher mithilfe des Tonfalls und nur selten anhand der Wortwahl über empfundene Emotionen wie Wut, Freude oder Trauer in Kenntnis gesetzt (vgl. Taylor 2009: 123). Wenngleich mögliche Ausdrucksformen von Ärger und Zorn vielen Sprach- und Kulturräumen gemein sind, bestehen in anderen Bereichen gleichfalls erhebliche Unterschiede in der Verwendung und Wahrnehmung von Prosodie (vgl. Taylor 2009: 123). Da dieser Aspekt menschlicher Kommunikation zumal wesentliche Informationen vermittelt, sind Generalisierungen dementsprechend nur schwer in maschinellen Dolmetschsystemen realisierbar (vgl. Crumpton und Bethel 2016: 271f; Taylor 2009: 123).

Auch die übrigen der Prosodie zurechenbaren Aspekte können in der Kommunikation von großer Bedeutung sein und im Gegensatz zur affektiven Prosodie ohne direkten Bezug auf Emotionen Einfluss auf Tonhöhe, Stimmqualität und Sprechrhythmus nehmen (vgl. Härtel 2016: 80; Taylor 2009: 124, 128). Bei der Aussprache von Wörtern und Sätzen zur Anwendung gebrachte Melodien und Akzente, die Lautstärke, die Sprechgeschwindigkeit und das Pausenverhalten sind ein wesentlicher Bestandteil menschlicher Kommunikation, der mehrere Sprachsegmente umfasst (vgl. Black und Lenzo 2004: 761; Härtel 2016: 94; Taylor 2009: 128f). Diese sogenannten Suprasegmentalia sind in jeder lautsprachlichen Äußerung anzutreffen und können anhand klar generalisierbarer Regeln, wie beispielsweise dem Senken der Stimme gegen Satzende, mittels Syntax schon seit Längerem synthetisch erfolgreich und verlässlich nachempfunden werden (vgl. Black und Lenzo 2004: 761; Härtel 2016: 94; Taylor 2009: 128f).

In manchen Situationen kann die gezielte Verwendung von Prosodie erforderlich sein, obwohl damit keine der bereits dargelegten Funktionen erfüllt wird. Beispielsweise könnte ein erhöhter Lärmpegel eine ebenso erhobene Stimme erfordern, um Zuhörer*innen zu erreichen

(vgl. Taylor 2009: 125f). Da mit diesem Aspekt von Prosodie keine zusätzlichen oder essenziellen Informationen vermittelt werden und seine Anwendung sowie Form gänzlich optionaler Natur sind, ist es sowohl problematisch als auch vernachlässigbar, entsprechende Daten zu erheben und in maschinellen Dolmetschsystemen zu berücksichtigen (vgl. Taylor 2009: 125f, 142, 259).

Diesen Darlegungen entsprechend können verschiedene Aspekte der Prosodie im Kontext des maschinellen Dolmetschens sowohl für das Verständnis als auch für die Vollständigkeit der zu vermittelnden Informationen von immenser Bedeutung sein. Der gezielte Einsatz von Prosodie trägt ferner dazu bei, synthetische Sprache natürlicher klingen zu lassen, und spielt zugleich eine zentrale Rolle bei der Verbesserung der Sprachsynthese insgesamt (vgl. Crumpton und Bethel 2016: 271; Pöchhacker 2024: 8). Demgemäß ist eine entsprechende Berücksichtigung von Prosodie in maschinellen Dolmetschsystemen für eine optimale Informationsvermittlung durchaus zweckdienlich, gestaltet sich allerdings aufgrund von vielschichtigen Überlegungen und technischen Problematiken schwierig. Trotz bemerkenswerter technologischer Fortschritte steht diese Technologie weiterhin vor komplexen Herausforderungen, insbesondere im Hinblick darauf, die Nuancen gesprochener Sprache sowie kontextuelle Feinheiten präzise zu erfassen (vgl. Fantinuoli 2025: 221).

Eine zentrale Herausforderung im Bereich der Sprachsynthese besteht darin, Prosodie gezielt zur Vermittlung von Emotionen einzusetzen (vgl. Crumpton und Bethel 2016: 276). Damit verbunden ist die grundlegende Schwierigkeit, prosodische Informationen angemessen aus sprachlichen Daten zu erschließen. Insbesondere bei den auf dem *Cascading*-Ansatz beruhenden Systemen (siehe Kapitel 1.6) können entsprechende prosodische Elemente durch die genutzten Modelle zur Spracherkennung häufig nur unzureichend erfasst werden und wesentliche Informationen gehen folglich verloren (vgl. Fantinuoli 2025: 221). Aus reinen Textdaten – etwa nach einer maschinellen Übersetzung – lassen sich derartige Informationen ohne weitere Intervention nur eingeschränkt ableiten beziehungsweise generieren (vgl. Crumpton und Bethel 2016: 276; Horváth 2021: 177; Taylor 2009: 49, 127, 140). Dementsprechend müssen zunächst die sprachlichen Merkmale, die Menschen zur emotionalen Kommunikation nutzen, identifiziert und quantifiziert werden, damit synthetische Sprache emotionale Intentionen angemessen wiedergeben kann (vgl. Crumpton und Bethel 2016: 272). Zwar wurden verschiedene Markup-Sprachen entwickelt, um Texte mit für die Prosodie relevanten Metadaten anzureichern, doch bleibt deren praktische Anwendung begrenzt, da sowohl die konkrete Ausgestaltung synthetischer Prosodie als auch die systematische Nutzung des bestehenden Gestaltungsspielraums mit

Unsicherheiten behaftet sind (vgl. Crumpton und Bethel 2016: 272, 276; Horváth 2021: 177; Taylor 2009: 140).

Vor diesem Hintergrund erscheint es nachvollziehbar, dass in der Praxis lange Zeit mitunter auf affektive Prosodie verzichtet und stattdessen eine neutrale Sprechweise bevorzugt wurde, um Fehlinterpretationen zu vermeiden – selbst, wenn dadurch potenziell relevante Informationen verloren gingen (vgl. Taylor 2009: 142). Neuere Entwicklungen relativieren diese Einschränkungen jedoch insofern, als moderne Sprachsynthesysteme zunehmend auf Ansätzen beruhen, die prosodische Merkmale direkt aus umfangreichen Datensätzen lernen (vgl. Chan und Kuang 2024: 27; Liu et al. 2023: 7). Zudem liegt es mittlerweile im Rahmen des technisch Möglichen, einem Text zugrunde liegende semantische und kontextuelle Merkmale auch ohne Rückgriff auf Metadaten zur Verbesserung der Vorhersage von Prosodie zu berücksichtigen, und mithilfe neuronaler Systeme zur synthetischen Erzeugung von Sprache lassen sich nun deutlich natürlicher anmutende Sprachausgaben erzeugen (vgl. Lam et al. 2024: 776; Zeng et al. 2020: 4422f).

Potenzial zur weiteren Verbesserung besteht insbesondere darin, kontextuelle Daten, die über den reinen Text hinausgehen, künftig stärker in die Sprachsynthese einzubinden (vgl. Fantinuoli 2025: 221). Vor allem unter Verwendung von LLM könnten visuelle Daten – etwa in Bezug auf Sprecher*innen, Umgebung oder Kommunikationssituation – parallel zur ausgangssprachlichen Wiedergabe erfasst, analysiert und in den maschinellen Dolmetschprozess eingespeist werden (vgl. Fantinuoli 2025: 221). Durch die sachgemäße Einbeziehung derartiger Informationen in maschinelle Dolmetschsysteme könnten weitere Feinheiten und Nuancen berücksichtigt und schließlich der Übersetzungsprozess insgesamt sowie die Sprachsynthese erheblich verbessert werden (vgl. Fantinuoli 2025: 221).

1.6 Die Cascading- und End-to-end-Ansätze

Auf den vorangehenden Seiten wurde beleuchtet, wie Systeme zur Erzeugung maschineller Dolmetschungen basierend auf den ihnen zugrunde liegenden Teilprozessen funktionieren und mit welchen Besonderheiten, Funktionalitäten und Ausgestaltungsmöglichkeiten diese behaftet sind. Mit dem rasch voranschreitenden technischen Fortschritt geht allerdings ein ebenso rascher und unentwegt weiter voranschreitender Wandel einher, im Zuge dessen die beleuchteten Teilprozesse zusehends ineinander und miteinander integriert werden. Auch wenn die individuellen Subprozesse nach wie vor essenzielle Bestandteile von Dolmetschsystemen darstellen, können die Grenzen zwischen den einzelnen Komponenten mitunter nicht mehr ohne Weiteres

hart gezogen werden. Dementsprechend existieren für die konkrete Ausgestaltung maschineller Dolmetschsysteme und die Umsetzung beziehungsweise Einbindung der oben erläuterten Teilprozesse derzeit zwei grundlegende technische Ansätze, die in der Literatur als *Cascading* und als *End-to-End* bezeichnet werden (vgl. Fantinuoli 2025: 212; Horváth 2021: 175). Da im Deutschen auch im Hinblick auf diese Terminologie keine einheitlich genutzte Entsprechung existiert, werden in der vorliegenden Arbeit aus praktischen Gründen die in englischer Sprache geläufigen Benennungen verwendet. In diesem Unterkapitel wird näher auf diese beiden technischen Ansätze eingegangen, indem sie vorgestellt, die Unterschiede zwischen ihnen dargelegt und ihre Bedeutung für maschinelle Dolmetschsysteme beleuchtet werden.

Aufgrund der vielfältigen Zielsetzungen und unterschiedlichen Umsetzungsmöglichkeiten gleicht kein System zur Erzeugung maschineller Dolmetschungen dem anderen. Dementsprechend können verschiedene Dolmetschsysteme auf verschiedensten Punkten des Spektrums zwischen den beiden durch diese zwei Ansätze geprägten Ausgestaltungsmöglichkeiten angesiedelt werden (vgl. Fantinuoli 2025: 212). Analog zu den vorstehenden Ausführungen beinhalten die eher auf dem *Cascading*-Ansatz beruhenden Systeme in ihrer auf das Einfachste heruntergebrochenen Form zumindest die Teilprozesse der automatischen Spracherkennung zur Konvertierung gesprochener Sprache in Text, der maschinellen Übersetzung des auf diesem Wege erfassten Textes in die Zielsprache und der synthetischen Wiedergabe der ausgegebenen Übersetzung (vgl. Fantinuoli 2025: 213; Horváth 2021: 175; Pöchhacker 2024: 7). Dementsprechend werden bei dem *Cascading*-Ansatz die verschiedenen Teilprozesse in individueller Ausgestaltung aneinandergereiht und -gekoppelt (vgl. Fantinuoli 2025: 212). Das übergreifende Ziel dieser Systeme, einen gegebenen gesprochenen Text in der Ausgangssprache per maschineller Dolmetschung lautsprachlich in der Zielsprache wiederzugeben, wird bei diesem Ansatz auf Unterprozesse aufgeteilt, die allesamt durch spezialisierte sowie eigens dafür konzipierte Untersysteme durchgeführt werden und dabei aufeinander aufbauen (vgl. Fantinuoli 2025: 212f).

Die Ausgestaltung von auf dem *Cascading*-Ansatz beruhenden Systemen kann stark variieren und ist abhängig von den individuellen Anforderungen im Hinblick auf Zweck, Zielsetzung und Design sowie der erforderlichen Komplexität (vgl. Fantinuoli 2025: 213). Unter dieser Vielzahl von Modellen existieren auch überaus komplexe und ausgeklügelte Systeme, die beispielsweise eigens darauf ausgerichtete Komponenten umfassen können, die Décalage zu minimieren oder für durchgehende, ununterbrochene Übersetzungsdarbietungen zu sorgen, oder Komponenten zur Vereinfachung, Korrektur oder Beschleunigung der auszugebenen

Sprachsynthese (vgl. Fantinuoli 2025: 213). Da die den entsprechenden Prozessen zugrunde liegenden Technologien über einen langen Zeitraum entwickelt und entsprechend verfeinert wurden, können die auf dem *Cascading*-Ansatz beruhenden Systeme von hochentwickelten Technologien profitieren (vgl. Fantinuoli 2025: 213).

Derartige Systeme sind jedoch auch mit inhärenten Beschränkungen behaftet (vgl. Fantinuoli 2025: 213). Für auf dem *Cascading*-Ansatz beruhende maschinelle Dolmetschleistungen sind – entsprechend der Erläuterungen in Kapitel 1.3 – umfangreiche Datensätze erforderlich, um die verschiedenen für die Umsetzung der Teilprozesse erforderlichen Module zu trainieren (vgl. Lu und Fantinuoli 2025: 4). Dies ist mit hohen Trainingsanforderungen behaftet und die funktionale Aneinanderreihung der Subprozesse und -module ist ebenso mit hohem Aufwand verbunden (vgl. Fantinuoli 2025: 213). Außerdem werden etwaige Fehler, die an beliebiger Stelle im Komplex der zur maschinellen Erzeugung von Dolmetschleistungen bei auf dem *Cascading*-Ansatz beruhenden Systemen anfallen, in nachfolgenden beziehungsweise nachgereihten Subprozessen und -modulen ohne entsprechenden Filter oder automatische Korrektur aufgegriffen, fortgeführt und womöglich verschlimmert (vgl. Fantinuoli 2025: 213; Horváth 2021: 182).

Im Kern der auf dem *Cascading*-Ansatz beruhenden Systemen und der entsprechenden Technologien steht die maschinelle Übersetzung (vgl. Pöchhacker 2024: 7). Wie in der Einleitung zu Kapitel 1 erläutert, sind Systeme des maschinellen Dolmetschens den technologischen Entwicklungen der maschinellen Übersetzung gefolgt und im Zeitverlauf stets komplexer geworden (vgl. Fantinuoli 2025: 211f). Mit der steigenden Komplexität gingen ferner neue Zielsetzungen und Ausrichtungs- und Ausgestaltungsentscheidungen für die zu entwickelnden Systeme einher. Der Fokus rückte auf multimodalen Input und eine bessere Berücksichtigung des weiteren Kontexts der Kommunikationssituation, die eine maschinelle Dolmetschung erfordert (vgl. Pöchhacker 2024: 8f).

Zur Erfüllung dieser neuen Zielsetzung und Gestaltungsrichtung haben sich die Entwicklungsbemühungen zusehends von der schlichten Aneinanderreihung verschiedener Subkomponenten abgewendet und eher einem gesamtheitlicheren Ansatz zugewendet, bei dem die Grenzen zwischen den einzelnen Teilprozessen nicht mehr ohne Weiteres klar gezogen werden können. Dieser Ansatz wird in der englischsprachigen Literatur *End-to-End*-Ansatz genannt (vgl. Fantinuoli 2025: 214; Horváth 2021: 175f; Lu und Fantinuoli 2025: 1; Pöchhacker 2024: 8f). Da in der deutschen Literatur keine autoritative Definition in einheitlicher Verwendung

besteht, wird auch für dieses Konzept aus praktischen Gründen auf die englische Benennung zurückgegriffen.

Der *End-to-End*-Ansatz ist gekennzeichnet von geringer Latenz und Echtzeitanwendung zur kontextgebundenen und kommunikationssituationsabhängigen Vermittlung zwischen Sprachen und weist in dieser Hinsicht größere Ähnlichkeit zur Tätigkeit menschlicher Dolmetscher*innen auf als zuvor verfügbare Systeme (vgl. Lu und Fantinuoli 2025: 2). Bei auf dem *End-to-End*-Ansatz beruhenden Dolmetschsystemen wird die maschinelle Übersetzung ohne Zwischenschritte – wie etwa eine zwischenzeitliche Verschriftlichung des erkannten Textes – unmittelbar in die Spracherkennung eingebettet und integriert (vgl. Horváth 2021: 175f; Pöchhacker 2024: 9). Derzeit scheint sich der Fokus von *Cascading*-Systemen auf *End-to-End*-Systeme zu verlagern (vgl. Fantinuoli 2025: 221).

Bei der Erzeugung maschineller Dolmetschleistungen mithilfe von *End-to-End*-Systemen können diverse Sprachmerkmale des Ausgangsmaterials – wie etwa die Prosodie, Betonungen, Pausenverhalten, Sprechtempo – berücksichtigt werden und entsprechend in die Ausgabe des erzeugten Zieltextes einfließen (vgl. Fantinuoli 2025: 215, 221). Hierbei wird die Gesamtheit aller erkannten Merkmale berücksichtigt, zueinander in Relation gesetzt und in der Folge miteinander vernetzt (vgl. Fantinuoli 2025: 215). Durch die zentrale Einbindung von neuronaler maschineller Übersetzung soll ferner die Fortführung von Fehlern in nachgereichten Subprozessen beziehungsweise -komponenten verhindert werden (vgl. Pöchhacker 2024: 9).

Auch für *End-to-End*-Systeme sind umfangreiche Datensätze vonnöten, die mitunter schwierig zu beschaffen sind – insbesondere für weniger verbreitete Sprachen (vgl. Fantinuoli 2025: 215). Ohne entsprechende Daten, die in beiden Sprachen des jeweiligen Sprachenpaars auch mit nicht-sprachlichen beziehungsweise paraverbalen Merkmalen angereichert wurden, können die auf diese Daten zurückgreifenden Modelle und neuronale Netzwerke nicht sachgemäß trainiert und entwickelt werden (vgl. Fantinuoli 2025: 215; Lu und Fantinuoli 2025: 4). Folglich können solche Daten und Informationen weder berücksichtigt werden noch in die ausgegebene Dolmetschung einfließen (vgl. Fantinuoli 2025: 215).

Durch die Zuwendung zu neuronalen Netzwerken sowie Large Language Models und den auf *End-to-End*-Systemen gerückten Fokus der Forschungs- und Entwicklungsbemühungen konnten erhebliche Fortschritte im Bereich des maschinellen Dolmetschens und der Leistungsfähigkeit entsprechender Systeme erzielt werden (vgl. Lu und Fantinuoli 2025: 7f). Gestützt auf größere Kosteneffizienz, verbesserte Verlässlichkeit und die in den letzten Jahren

erreichte weite Verbreitung werden Systeme zur maschinellen Dolmetschung dank dieser technologischen Entwicklungen zusehends wettbewerbsfähig (vgl. Lu und Fantinuoli 2025: 8). Die so angestoßenen, tiefgreifenden Veränderungen auf dem Gebiet des maschinellen Dolmetschens stellen die gesamte Forschung, Lehre und Praxis des Dolmetschens vor einen Paradigmenwechsel, der zu einer grundlegenden Neuausrichtung des allgemeinen Dolmetschökosystems führen könnte (vgl. Lu und Fantinuoli 2025: 8, 15).

1.7 Limitationen und Herausforderungen

Den obenstehenden Ausführungen entsprechend werden Systeme zur automatischen Erzeugung von Dolmetschungen zusehends leistungstärker. Folglich wird an derartige Systeme der hohe Anspruch gestellt, über Sprachen hinweg uneingeschränkten Zugang zu gesprochenen Inhalten zu geben (vgl. Fantinuoli 2025: 220). Dieses Ziel erscheint erstrebenswert, ist jedoch auch mit Risiken behaftet, und maschinelles Dolmetschen geht trotz der großen technologischen Fortschritte weiterhin mit vielschichtigen Herausforderungen einher (vgl. Fantinuoli 2025: 221).

Eine dieser komplexen Herausforderungen bei der maschinellen Dolmetschung gesprochener Sprache ist der Umgang mit Redewendungen, Umgangssprache und anderen kulturell behafteten Ausdrucksweisen. In Echtzeit soll gesprochene Sprache verarbeitet werden, die zu tiefst im situativen Kontext eingebettet ist und häufig Spontanausdrücke, Umgangssprache, Ungenauigkeiten und Auslassungen aufweisen (vgl. Lu und Fantinuoli 2025: 8). Wie oben dargelegt ist gesprochene Sprache darüber hinaus reich an para- und nonverbalen Inhalten wie Tonfall, Pausen, Intonation, Mimik, Gestik und Körperbewegung (vgl. Lu und Fantinuoli 2025: 8). Auch Ironie und Sarkasmus stellen Systeme zur Erzeugung maschineller Dolmetschungen vor Probleme, da beide Phänomene stark vom kulturellen Kontext geprägt und abhängig sind (vgl. Fantinuoli 2025: 216).

Kulturell behaftete Informationen können mitunter nicht ohne Weiteres zwischen Sprachen übertragen werden, da inhärente Bedeutungen oft tief und individuell in der Kultur und Geschichte der sie nutzenden Menschen verankert sind (vgl. Fantinuoli 2025: 216). Um derartige Inhalte sachgemäß dolmetschen zu können, müssen entsprechende Systeme diese Inhalte erst erkennen, die kommunikative Situation möglichst allumfänglich erfassen und schließlich relevante kulturelle und kontextuelle Hintergrundinformationen abrufen und korrekt anwenden können (vgl. Fantinuoli 2025: 216). Dementsprechend hängt die von maschinellen Dolmetschsystemen erzielbare Leistung in hohem Maße von deren Möglichkeiten und Fähigkeiten zur

Erfassung, Verarbeitung und Berücksichtigung kontextueller Hintergrundinformationen ab (vgl. Fantinuoli 2025: 216). Wird der größere Kontext unzureichend eingebunden, können Bedeutung und Absicht des Ausgangstextes inkorrekt erfasst und dementsprechend missverständlich wiedergegeben werden (vgl. Fantinuoli 2025: 216).

Eine weitere Herausforderung für maschinelle Dolmetschsysteme besteht in der Minimierung der Latenz beziehungsweise der *Décalage*, um tatsächlich in Echtzeit-Situationen sachgemäß eingesetzt werden zu können (vgl. Fantinuoli 2025: 217). Eine zu lange Verzögerung zwischen ausgangssprachlichen Äußerungen und deren Wiedergabe in der Zielsprache kann die Kommunikationssituation stören, zu Missverständnissen führen und sich nachteilig auf das Erlebnis der Rezipient*innen auswirken (vgl. Fantinuoli 2025: 217). Die zielsprachliche Wiedergabe sollte demnach möglichst rasch und im Hinblick auf Merkmale wie Redefluss und Pausenverhalten möglichst nah am Ausgangstext folgen (vgl. Fantinuoli 2025: 217).

Die *Décalage* wird im Zusammenhang mit maschinellen Dolmetschsystemen sowohl durch technische als auch sprachliche Faktoren bestimmt (vgl. Fantinuoli 2025: 217). Aus technischer Sicht beeinflussen die Rechenzeit und die Geschwindigkeit der Datenübertragung und -verarbeitung die *Décalage* bei maschinell angefertigten Dolmetschungen, die folglich etwa durch bessere Hardware oder eine effizientere Verarbeitung reduziert werden kann (vgl. Fantinuoli 2025: 217). Aus sprachlicher Sicht hängt die Latenz von der Segmentierung der Ausgangsrede ab (vgl. Fantinuoli 2025: 217). Dabei können beispielsweise erkannte Pausen herangezogen oder Semantik und Syntax in Echtzeit analysiert werden (vgl. Fantinuoli 2025: 217). Umfangreichere Segmente erhöhen die Latenz, jedoch – dank der so zusätzlich verfügbaren Informationen – auch die Genauigkeit der maschinellen Dolmetschung (vgl. Fantinuoli 2025: 217). Demgegenüber ermöglichen kürzere Segmente – zulasten der erzielbaren Genauigkeit – eine raschere Wiedergabe in der Zielsprache (vgl. Fantinuoli 2025: 217).

Ein weiterer nicht zu unterschätzender Faktor ist der verlässliche Einsatz von ausreichend leistungsstarken Systemen (vgl. Fantinuoli 2025: 217). Den obigen Ausführungen entsprechend werden die Fähigkeiten maschineller Dolmetschsysteme zur umfassenden Berücksichtigung verschiedenster Kommunikationssituationen und -kontexte stetig erweitert (vgl. Fantinuoli 2025: 217). Um in diesem Zusammenhang und bei nach wie vor möglichst geringer *Décalage* gute Leistungen zu erzielen, sind eine starke Rechenleistung, effiziente Softwarelösungen und eine entsprechend robuste Hardware vonnöten (vgl. Fantinuoli 2025: 217). Dies

gestaltet sich bei den auf dem *Cascading*-Ansatz beruhenden Systemen aufgrund der komplexen Verflechtung ihrer Teilprozesse besonders schwierig (vgl. Fantinuoli 2025: 217).

In diesem Kapitel wurde beleuchtet, wie maschinelles Dolmetschen – aufgeschlüsselt nach den fundamentalen Teilprozessen – grundsätzlich funktioniert, welche Ausgestaltungs- und Einsatzmöglichkeiten bestehen sowie mit welchen gegenwärtigen Limitationen und Potenzialen diese Systeme behaftet sind. Dabei wurden die Systeme zur Erzeugung maschineller Dolmetschleistungen in den aktuellen Kontext eingebettet und dementsprechend dargelegt, welche Leistungen derzeit realisiert werden können, welchen aktuellen Entwicklungen sie unterworfen sind, und welche Herausforderungen und Probleme es noch zu bewältigen gilt. Dank technologischer Fortschritte in vielen technologisch verwandten Bereichen – wie etwa der Sprachsynthese – kommt dem maschinellen Dolmetschen große Aufmerksamkeit zu und über diverse, individuell ausgestaltete Systeme werden zusehends neue Einsatzmöglichkeiten erschlossen. Wenngleich im Hinblick auf die Sprachsynthese, die einen der zentralen Bestandteile der Analyse der vorliegenden Arbeit darstellt, mit für das maschinelle Dolmetschen relevanten Technologien bereits großartige Ergebnisse erzielt werden können, sind mit derzeit erhältlichen Dolmetschsystemen im Allgemeinen noch keine dem Menschen gleichzusetzenden Leistungen erzielbar. Die Systeme werden kontinuierlich weiterentwickelt, um weitere Einsatzmöglichkeiten zu erschließen und sie zusehends wettbewerbsfähiger zu machen. Vor diesem Hintergrund soll im folgenden Kapitel betrachtet werden, wie die Qualität von Dolmetschleistungen – menschlichen und maschinellen Ursprungs – in der Dolmetschwissenschaft zu erfassen ist und welche Qualitätskriterien bei der Analyse der vorliegenden Arbeit zum Einsatz kommen.

2 Qualitätsbeurteilung von Dolmetschleistungen

Auf den vorstehenden Seiten wurden die genaue Funktionsweise, die Möglichkeiten, die Grenzen sowie anhaftende Problematiken und Eigenheiten des maschinellen Dolmetschens beleuchtet. Komplexe Technologien finden in der Gestalt von maschinellen Dolmetschsystemen bereits heute für Dolmetschleistungen in verschiedenen kommunikativen Situationen erfolgreich Anwendung. Inhalte können in gewissem Maße bereits verlässlich und korrekt maschinell vermittelt werden, jedoch besteht in diversen Angelegenheiten und Teilprozessen noch erhebliches Verbesserungs- und Weiterentwicklungspotenzial. Vor allem hinsichtlich der Berücksichtigung und Verwendung von Prosodie bleiben gegenwärtig verfügbare Ausgestaltungen maschineller Dolmetschsysteme unbefriedigend hinter den von der Gesellschaft hoch angesetzten Anforderungen und Erwartungen zurück, wenngleich menschlich anmutende Sprache abseits des Feldes der maschinellen Dolmetschung bereits synthetisch realisiert werden kann. Die Bewertung des Komplexes aus Vorzügen, Funktionalitäten und Defiziten maschineller Dolmetschsysteme ist sowohl für die Weiterentwicklung dieser Technologien als auch für die vorliegende Arbeit von zentraler Bedeutung (vgl. Härtel 2016: 84ff).

Doch bevor Aussagen über die Qualität von Dolmetschleistungen getroffen werden können, muss erst grundsätzlich erörtert und definiert werden, was unter dem Begriff „Qualität“ subsumiert und welche Bedeutung dem Qualitätsbegriff zugeschrieben wird. Wie aus diesem Kapitel hervorgehen wird, kann Qualität auf verschiedene Art und Weise definiert und zur Anwendung gebracht werden (vgl. Härtel 2016: 92). Um sowohl maschinell angefertigte als auch von menschlichen Dolmetscher*innen erzeugte Dolmetschleistungen beurteilen zu können, muss erst anhand von eigens definierten und zweckgebundenen Maßstäben eine Basis für die nachfolgende Bewertung etabliert werden (vgl. Grbić 2008: 237).

2.1 Der Qualitätsbegriff in der Dolmetschwissenschaft

Trotz der bedeutungsvollen Rolle, Funktion und Tragweite, die dem Konzept der Qualität in der und für die Disziplin des Dolmetschens zukommt, konnte dieser Begriff in der Dolmetschwissenschaft bislang in noch kein allgemein anerkanntes, einheitliches und universelles Regelwerk an Beurteilungs- und Evaluierungsschemata eingekleidet werden (vgl. Fantinuoli 2025: 218; Zwischenberger 2010: 128). Nach wie vor ist sich die einschlägige Fachliteratur uneins, wie Qualität zu erheben ist sowie welche Faktoren und Kriterien diesbezüglich auf welche Art und Weise kategorisch beachtenswert sind (vgl. Fantinuoli 2025: 218; Zwischenberger

2010: 128). Dementsprechend konnten von der Dolmetscher*innengemeinschaft klar und eindeutig quantifizierbare sowie allgemein anwendbare Kriterien noch nicht abschließend hervorgebracht werden (vgl. Wang und Wang 2025: 2; Zwischenberger 2013: 17). Trotz der großen Beachtung und Aufmerksamkeit, die diesem äußerst bedeutungsvollen Konzept zukommt, wurde es bislang noch nicht abschließend in einer einzigen, universell akzeptierten und anwendbaren Definition verankert (vgl. Fantinuoli 2025: 218; Zwischenberger 2010: 128).

Ungeachtet der fehlenden Standardisierung von Qualitätserhebungs- und Qualitätsbewertungskriterien ist diese Begrifflichkeit der Dolmetschwissenschaft keineswegs fremd. Nachdem Simultandolmetschen infolge der Nürnberger Prozesse international verstärkt beachtet und eingesetzt wurde, kam auch der Qualität von Dolmetschleistungen in Forschung, Lehre und Praxis entsprechend große Aufmerksamkeit zu (vgl. Zwischenberger 2010: 128). Ihre Beurteilung beziehungsweise Bewertung ist mittlerweile sowohl im akademischen Kontext als auch im Bereich des professionellen Dolmetschens fest verankert und hat eine der produktivsten Forschungsbereiche der Dolmetschwissenschaft hervorgebracht (vgl. Wang und Fantinuoli 2024: 327; Zwischenberger 2010: 128). Die Forschung zum Konzept der Qualität von Dolmetschleistungen umfasst mannigfaltige Forschungsansätze sowie eine Vielzahl an beleuchteten Themen (vgl. Pöhhacker 2005: 143).

Die Beurteilung beziehungsweise Bewertung der Qualität von Dolmetschleistungen ist ein zutiefst komplexes Unterfangen (vgl. Fantinuoli 2025: 218; Wang und Fantinuoli 2024: 327). Diverse Vorstellungen sowie Konzipierungsversuche wurden bereits hervorgebracht und diskutiert. Lange wurde in den zugrunde liegenden Qualitätsdiskussionen von einer abstrakten Idealvorstellung einer Dolmetschung ausgegangen, die jedoch nicht mit zufriedenstellenden qualitativen Indikatoren mess- und nachweisbar gemacht werden konnte (vgl. Grbić 2008: 232).

Dank der Mühen professioneller Dolmetscher*innen, der Existenz von Dolmetschverbänden und vielfältigen Forschungsbeiträgen konnten jedoch bestimmte Standards und Kriterien zur Identifizierung von Qualitätsmerkmalen in Forschung, Lehre und Praxis allmählich Fuß fassen sowie auch Beachtung und Anerkennung finden (vgl. Amini et al. 2013: 94; Zwischenberger 2010: 128f, 2013: 23). So wurde in verschiedensten Studien etwa ein Vorzug für Qualitätskriterien auf Ebene des Inhalts – wie logische Kohärenz und Übereinstimmung mit dem Ausgangstext – vor formellen und die Darbietung betreffenden Merkmalen – wie grammatische Korrektheit und Stimme – erkannt (vgl. Gile 1999: 51; Zwischenberger 2010: 129).

Einzelne Forschungsbeiträge können jedoch nur einen kleinen Ausschnitt des Konzepts der Qualität abdecken und es ist bei deren Lektüre und Auslegung stets zu prüfen, unter welchen Umständen und aus der Sicht welcher Personen das Konzept darin untersucht und behandelt wird (vgl. Zwischenberger 2010: 128). Bei der Beurteilung von Qualität ist demnach einer Vielzahl von Variablen und Blickpunkten Rechnung zu tragen, wodurch sich eine allumfassende und einheitliche Definition von Qualität, die alle Kommunikationssituationen und Perspektiven abdeckt, nur schwierig etablieren lässt (vgl. Zwischenberger 2010: 128). Folglich hängt das Konzept der Qualität zutiefst vom jeweiligen Kontext ab und bei ihrer Bewertung sind die jeweilige Situation und die vorliegenden Perspektiven zu berücksichtigen (vgl. Fantinuoli 2025: 218; Wu 2010: 21). Diesen Ausführungen entsprechend hat sich in der Dolmetschwissenschaft ein Konsens darüber etabliert, dass die Qualität von Dolmetschleistungen ein äußerst komplexes Konzept darstellt, bei dem verschiedenste Perspektiven zu berücksichtigen sind und das in keine universelle, situations- und kontextübergreifende Definition eingefasst werden kann (vgl. Pöchhacker 2012: 20). Qualität wird in der Dolmetschwissenschaft folglich nicht als einzelne und isoliert erschließbare Beschaffenheit gesehen, sondern als Komplex mehrerer Dimensionen und Blickpunkte, bei dem den verschiedenen Faktoren je nach Perspektive in unterschiedlichem Umfang kontextabhängig und subjektiv Bedeutung beigemessen wird. Der Komplex an Faktoren und Merkmalen, dem einschlägige Nutzergruppen besondere Aufmerksamkeit schenken, ändert sich je nachdem, wo und unter welchen Umständen die Dolmetschleistung erbracht wird (vgl. Wang und Fantinuoli 2024: 327). Dolmetschleistungen in einem Konferenzsetting, vor Gericht, im medizinischen Bereich, in der Politik oder im Bildungsbereich vereinen in sich und bedingen unterschiedliche Perspektiven, Erwartungshaltungen und Vorstellungen im Hinblick auf Qualität (vgl. Wang und Fantinuoli 2024: 327). Beispielsweise stehen im Zusammenhang mit Konferenzdolmetschleistungen typischerweise Qualitätskriterien wie Inhalt, Sprachgebrauch und Darbietung im Zentrum der Betrachtung (vgl. Wang und Fantinuoli 2024: 327).

Wie die Qualität von Dolmetschungen wahrgenommen wird, lässt sich demnach nicht anhand eines festen, allumfassenden Maßstabs verstehen, sondern ergibt sich vielmehr aus der Betrachtung des Zusammenspiels zwischen der konkreten Leistung und den individuellen Erwartungen der Rezipient*innen unter Berücksichtigung der Kommunikationssituation (vgl. Fantinuoli 2025: 218; Horváth 2021: 180). Aufgrund der abweichenden Perspektiven, Bedürfnisse und Vorstellungen hat jede Person unterschiedliche Maßstäbe dafür, welche Dolmetschleistung als angemessen oder gelungen gilt (vgl. Fantinuoli 2025: 218; Horváth 2021: 180; Wang und Fantinuoli 2024: 327). Dementsprechend ist Qualität und deren Beurteilung stets

relativ und subjektiv – was als hochwertige Dolmetschleistung wahrgenommen wird, hängt zutiefst vom Kontext, von den Eigenarten und Anforderungen der Rezipient*innen sowie von ihren individuellen Vorstellungen einer gelungenen Dolmetschung ab (vgl. Fantinuoli 2025: 218; Wang und Fantinuoli 2024: 327).

Diesen Überlegungen entsprechend und in Ermangelung eines einheitlichen übergreifenden Regelwerks beziehungsweise entsprechender Richtlinien zur Qualitätsbewertung überlässt die Dolmetscher*innengemeinschaft die Beurteilung von Dolmetschleistungen der subjektiven Wahrnehmung und den individuellen Qualitätsvorstellungen der Rezipient*innen (vgl. Amini et al. 2013: 90; Grbić 2008: 234; Kurz: 2001: 405; Wu 2010: 25). In freiem Ermessen evaluieren Rezipient*innen von Dolmetschleistungen deren Qualität nach persönlichen Maßstäben – häufig in Bezug auf die Diskrepanz der erwarteten und erreichten Wirkung des dargebotenen Zieltextes (vgl. Amini et al. 2013: 94; Grbić 2008: 233; Kurz 2001: 405; Zwischenberger 2013: 38). Üblicherweise haben Rezipient*innen ferner – abhängig von Sprachkompetenz und situationsabhängigen Rezeptionsmöglichkeiten – keine Kenntnis vom Ausgangsmaterial und können daher in dieser Hinsicht und auf dessen Grundlage keine Aussagen treffen, Vergleiche anstellen oder Schlüsse ziehen (vgl. Kurz 2001: 403). Aus diesem Grund können Rezipient*innen nur sehr begrenzt nachvollziehen, inwiefern die Inhalte der Ausgangsrede in der Dolmetschung Erwähnung fanden und mit welchen anhaftenden Problematiken die Dolmetscher*innen zu kämpfen hatten (vgl. Kalina 2002: 123; Mack 2002: 111). Abgesehen von groben Einschätzungen hinsichtlich Kohärenz und Plausibilität kann sich das Zielpublikum kein umfassendes Bild von der dargebotenen Dolmetschleistung machen und dementsprechend in dieser Hinsicht nur beschränkt gehaltvolle Aussagen über die Qualität treffen (vgl. Kalina 2002: 123). Darüber hinaus sind die subjektiven Einschätzungen der Rezipient*innen aufgrund von abweichenden und individuellen Qualitätsvorstellungen nur äußerst schwer zu vereinheitlichen und zu quantifizieren (vgl. Grbić 2008: 234; Mack 2002: 111).

Da Rezipient*innen von Dolmetschleistungen demzufolge nur eingeschränkt zugrunde liegende Qualitätskriterien und die Angemessenheit vermittelter Inhalte – zumindest in Relation zum Ausgangstext – beurteilen können, müssen sie sich auf jene Aspekte konzentrieren, die sich ihnen durch Rezeption der Dolmetschleistung unmittelbar erschließen (vgl. Christodoulides und Lenglet 2014: 1002). Zu diesen Aspekten gehören beispielsweise die Flüssigkeit der Wiedergabe, die Tonlage und -qualität der Stimme, markante Lücken beziehungsweise Pausen in der Rede oder die Aussprache. Dabei bewerten Rezipient*innen Dolmetschqualität nicht neutral, sondern – den obigen Ausführungen entsprechend – in Relation zu ihren individuellen

Erwartungen an Aspekte wie etwa Verständlichkeit, Natürlichkeit, Genauigkeit und Professionalität (vgl. Amini et al. 2013: 94; Fantinuoli 2025: 218; Kurz 2001: 405). Vor allem hinsichtlich automatischer Dolmetschsysteme legen Rezipient*innen großen Wert auf die Darbietung, den Klang und die Präsentation der maschinell erbrachten Leistung (vgl. Taylor 2009: 3, 47). Rezipient*innen scheinen dabei die hohe Anforderung an maschinelle Dolmetschsysteme zu stellen, natürliche Sprache produzieren zu können (vgl. Taylor 2009: 3). Dahingegen werden mechanisch und unnatürlich anmutende synthetische Stimmen als äußerst unangenehm, störend und ablenkend empfunden, zu Ungunsten der wahrgenommenen Qualität (vgl. Taylor 2009: 3, 47f). Demgemäß ist die Bereitstellung einer möglichst menschlich klingenden Stimme und natürlichen Präsentation eine der zentralen Bestrebungen von Entwicklern maschineller Dolmetschsysteme (vgl. Taylor 2009: 47).

Als Konsequenz liegt der Fokus der vorliegenden Arbeit nicht auf der Qualität der dargebotenen Dolmetschleistung per se, sondern auf der subjektiven Wahrnehmung dieser Dolmetschleistung seitens der Rezipient*innen. Aspekte wie inhaltliche Übereinstimmung und Äquivalenz zwischen Ausgangs- und Zieltext, Vollständigkeit, fachliche und terminologische Korrektheit oder grammatische Fehlerfreiheit – wie sie etwa bei der Ausbildung, Lehre und Praxis der Dolmetschwissenschaft typischerweise zum Tragen kommen – setzen einen Vergleich zwischen Ausgangs- und Zieltext voraus, sind für das Forschungsziel der vorliegenden Arbeit zur Gänze unerheblich und werden dementsprechend bei der Analyse außer Acht gelassen (vgl. Zwischenberger 2010: 131, 140). Darüber hinaus wird in der vorliegenden Arbeit ausschließlich die zielsprachige Qualität betrachtet. Da im Mittelpunkt der vorliegenden Arbeit die Qualitätswahrnehmung durch Rezipient*innen steht, wird Qualität dabei nicht als rein text- oder normbezogene Beschaffenheit der Dolmetschleistung angesehen, sondern als subjektive, anhand individueller Erwartungen und der jeweils wahrnehmbaren Präsentationsmerkmale vermittelte zieltext- beziehungsweise rezeptionsbezogene Bewertung ausgelegt. Über allgemeine Präferenzen oder Erwartungshaltungen werden weder Rückschlüsse noch Aussagen getroffen. Vor diesem Hintergrund wird im Folgenden auf die bei der vorliegenden Arbeit und der Untersuchung der Qualitätswahrnehmung zum Tragen kommenden Qualitätskriterien eingegangen.

Im Mittelpunkt der Analyse stehen zieltext- und rezeptionsbezogene Qualitätskriterien, die Rezipient*innen rein auf Grundlage der dargebotenen Dolmetschleistung beurteilen können. Diese für die vorliegende Arbeit ausgewählten Kriterien betreffen unter anderem die Verständlichkeit, Flüssigkeit und Nachvollziehbarkeit der Dolmetschung sowie den Gesamteindruck der Dolmetschqualität (vgl. Kurz 2001: 399f; Zwischenberger 2010: 137, 139f).

Die für die vorliegende Arbeit herangezogenen Kriterien der Kohärenz und sprachlichen Korrektheit beschreiben, in welchem Maße die Rezipient*innen dem Inhalt der Dolmetschung folgen können. Kohärenz betrifft hierbei den inneren logischen Zusammenhang der Wiedergabe und die Frage, ob sie den Rezipient*innen stimmig, schlüssig und nachvollziehbar erscheint. Sprachliche Korrektheit bezieht sich wiederum darauf, in welchem Ausmaß die dargebotene zielsprachliche Wiedergabe – insbesondere im Hinblick auf Faktoren wie Grammatik, Wortwahl, Satzbau – als angemessen wahrgenommen wurde (vgl. Kurz 2001: 398-401; Zwischenberger 2010: 136, 139f).

Darüber hinaus werden Merkmale wie das Pausenverhalten, Verzögerungen und das Sprechtempo berücksichtigt. Diese Faktoren beeinflussen, wie leicht der Dolmetschung gefolgt werden kann. Sie können maßgeblich dazu beitragen, in welchem Umfang die Dolmetschung als angenehm oder störungsfrei wahrgenommen wird oder ob ihre Rezeption mit erheblichem kognitivem Aufwand verbunden war (vgl. Christodoulides und Lenglet 2014: 1002; Rennert 2010: 103ff). Insbesondere für Rezipient*innen, die die Qualität ohne unmittelbaren Bezug zum Ausgangstext beurteilen, sind derartige unmittelbar hörbaren Merkmale besonders wichtig (vgl. Christodoulides und Lenglet 2014: 1002).

Besondere Bedeutung kommt in der vorliegenden Arbeit auch den Merkmalen der Stimmqualität und der der rezipierten Leistung beigemessenen Natürlichkeit zu. Diese Kriterien sind für die der vorliegenden Arbeit zugrunde liegenden Analyse erheblich, da der Vergleich zwischen den menschlich und maschinell angefertigten Dolmetschleistungen neben den sprachlichen, auch auditive und paralinguistische Unterschiede umfassen soll. Eine Wiedergabe, die etwa monoton, synthetisch beziehungsweise wenig natürlich anmutet, könnte die wahrgenommene Qualität verringern, selbst wenn sie inhaltlich nachvollziehbar bleibt (vgl. Taylor 2009: 3, 47f). Der Aspekt der Natürlichkeit wird in der vorliegenden Arbeit daher als Eindruck einer möglichst menschlich anmutenden, prosodisch angemessenen und als angenehm rezipierten Dolmetschleistung betrachtet (vgl. Cooper et al. 2024: 1; Kurz 2001: 399f).

Ergänzend dazu werden Professionalität und Vertrauenswürdigkeit als globale Qualitätsmerkmale berücksichtigt. Damit sollen verschiedene Aspekte der Wahrnehmung gebündelt und Aufschluss darüber gegeben werden, ob die Dolmetschleistung als zuverlässig, angemessen beziehungsweise akzeptabel wahrgenommen wurde. Diesen Ausführungen und den vorstehenden theoretischen Überlegungen entsprechend fließt die wahrgenommene Qualität in die der vorliegenden Arbeit zugrunde liegende Analyse nicht als eindimensionales Gesamturteil

ein, sondern als Komplex von verständnisbezogenen, sprachlichen, stimmlichen, prosodischen und die Darbietung und den Gesamteindruck betreffenden Einzelkriterien und -merkmalen, die das Zielpublikum bei der Rezeption der Dolmetschleistungen unmittelbar wahrnehmen können.

2.2 Qualitätsbewertung im Lichte aktueller technischer Entwicklungen

Neben der durch Menschen vorgenommenen Beurteilung von Qualität wird in der Forschung zunehmend auf automatische Bewertungsmodelle gesetzt. Da diese Systeme einen integralen Bestandteil des technischen und technologischen Komplexes rund um maschinelles Dolmetschen darstellen und in einem Spannungsverhältnis mit der rezipient*innenbasierten Bewertung stehen, werden einige in diesem Zusammenhang erhebliche Aspekte im Folgenden knapp beleuchtet, um der für die vorliegende Arbeit getroffenen Entscheidung, auf Qualitätsurteile durch Rezipient*innen zu setzen, einen umfangreicheren und weiter gespannten Rahmen zu bieten.

Für die Zwecke der Konzipierung und kontinuierlichen Weiterentwicklung maschineller Dolmetschsysteme ist die automatische Erkennung, Registrierung und Klassifizierung etwaiger Fehler, Unstimmigkeiten oder Verbesserungspotenziale unerlässlich (vgl. Härtel 2016: 84ff). Eine Vielzahl entsprechender automatischer Evaluierungssysteme wurde bereits entwickelt und findet zur Beurteilung maschineller Dolmetschsysteme einschließlich aller Teilprozesse Anwendung (vgl. Fügen et al. 2007: 218f). Anhand der durch diese Testmöglichkeiten geschaffenen systemübergreifenden und eindeutig quantifizierbaren Grundlage können nicht nur einzelne Dolmetschsysteme isoliert betrachtet, sondern auch Vergleiche zwischen unterschiedlichen Lösungen angestellt werden (vgl. Härtel 2016: 84ff). Wenngleich sich die Erstellung solcher objektiver Bewertungsinstrumente und -maßstäbe mitunter schwierig gestaltet, sind sie in ihrer Anwendung kosteneffizient, schnell und erfordern nur bedingt menschliche Unterstützung (vgl. Fünfer 2013: 49; Härtel 2016: 91). Im Rahmen dieser automatischen Optimierung von Systemen zur Erzeugung maschineller Dolmetschungen kann ebenso die Qualität von maschinell angefertigten Dolmetschungen unmittelbar automatisch maschinell bewertet werden (vgl. Lu und Fantinuoli 2025: 5).

Die Qualität von maschinell generierten Dolmetschleistungen – wie auch die der von Menschen erstellten – kann grundsätzlich von Menschen oder automatisch durch Maschinen bewertet werden (vgl. Han und Lu 2021: 75; Lu und Fantinuoli 2025: 5). Automatisch gemessen, getestet und evaluiert werden können neben Aspekten wie Fehlerquoten – sprich der Menge an bestätigten inkorrekten Verarbeitungen von Daten in einem beliebigen Teilprozess

des Systems – auch Zeiterfordernisse für die maschinelle Verarbeitung aufgenommener Informationen, die Natürlichkeit und Verständlichkeit sowie ob etwaige Durchsuchungsprozesse der Datenbanken von Erfolg gekrönt waren sowie mit angemessenen Ergebnissen belohnt wurden (vgl. Fügen et al. 2007: 218f; Taylor 2009: 523). Konkret zur automatischen Beurteilung von Qualität maschinell generierter Dolmetschleistungen werden in der Forschung und Praxis unterschiedliche Metriken und Modelle wie BLEU, chrF, COMET, METEOR, TER und WER herangezogen, auf die auf technischer Ebene in der vorliegenden Arbeit nicht näher eingegangen werden soll (vgl. Lu und Fantinuoli 2025: 5; Lu und Han 2023: 109; Son und Kim 2023: 3; Wang und Fantinuoli 2024: 328). Sie werden üblicherweise genutzt, da sie effektiv und einfach anzuwenden sind und rasch Ergebnisse liefern (vgl. Han und Lu 2021: 77; Lu und Han 2023: 109; Son und Kim 2023: 3). Ziel bei diesen Metriken ist es grundsätzlich, maschinell generierten Output mit menschlich erzeugten Referenztexten zu vergleichen (vgl. Fantinuoli 2025: 218; Han und Lu 2021: 77f; Sperber et al. 2025: 23342; Wang und Fantinuoli 2024: 328). Je näher die maschinell generierten Inhalte an den von professionellen Sprachmittler*innen erzielten Leistungen liegen, desto höher soll die erzielte Qualität sein (vgl. Han und Lu 2021: 77).

Durch LLM entstehen neue Möglichkeiten, komplexe und umfangreiche Daten zu analysieren und Muster aufzudecken und dabei komplexen sprachlichen Phänomenen Rechnung zu tragen, die auch bei der automatischen Bewertung von Dolmetschleistungen zur Anwendung kommen (vgl. Han und Lu 2021: 78; Jiang und Zhang 2025: 1f; Wang und Fantinuoli 2024: 328f). Diesen Entwicklungen ist das Aufkommen fortschrittlicher Methoden zur automatischen Beurteilung von Qualität ohne menschliche Referenztexte zu verdanken (vgl. Taylor 2009: 48; Wang und Fantinuoli 2024: 328). Die darauf basierenden Metriken beruhen auf Sprachmodellen, die zu diesem Zweck einem entsprechenden Training unterlaufen sind, um menschenähnliche Urteile ohne Rückgriff auf Referenztexte treffen zu können (vgl. Han und Lu 2021: 76, 78; Jiang und Zhang 2025: 3).

Diesen Ausführungen entsprechend und angesichts der inhärenten Grenzen menschlicher Beurteilung – etwa im Hinblick auf das von Menschen erzielbare Tempo – wurden Bestrebungen aufgenommen, die Funktionalität der vollautomatischen Beurteilung vermehrt auf menschliche Dolmetschungen auszuweiten (vgl. Han und Lu 2021: 82; Jiang und Zhang 2025: 1). Durch die maschinelle Bewertung sollen Beurteilungsprozesse beschleunigt, vereinfacht und günstiger gemacht werden, während zugleich die Anwendung subjektiver Kriterien und Vorstellungen vermieden werden soll (vgl. Fantinuoli 2025: 218; Han und Lu 2021: 82; Lu und Fantinuoli 2025: 5). Ferner sollen derartige Systeme, sofern sie einfach zugänglich und

ausreichend leistungsstark sind, auch Studierenden und aktiven Dolmetscher*innen einen Nutzen bringen, indem sie rasch Rückmeldungen und Informationen im Hinblick auf etwaiges Verbesserungspotenzial liefern (vgl. Wang und Fantinuoli 2024: 328).

Eine weitverbreitete automatische Bewertung von Dolmetschleistungen könnte jedoch durch eine möglicherweise daraus resultierende ständige Überwachung der allgemeinen Zufriedenheit von Dolmetscher*innen abträglich sein sowie in jenen Szenarien negative Folgen mit sich führen, in denen Entscheidungen über die Eignung für Dolmetschleistungen gänzlich Maschinen überlassen werden, während relevante Informationen vom betreffenden System nicht im erforderlichen Umfang erfasst und somit unzureichend berücksichtigt werden (vgl. Han und Lu 2021: 82; Wang und Fantinuoli 2024: 333). Darüber hinaus könnte die Zugänglichkeit zu solchen Systemen im Wege eines urheberrechtlichen Schutzes beschränkt sein (vgl. Han und Lu 2021: 81).

Diesen Ausführungen entsprechend kann die automatische Bewertung von Dolmetschqualität von Nutzen und wertvoll sein, wenn das Ziel verfolgt wird, rasch, systematisch und unter Nutzung bestimmter, womöglich beständiger Parameter Vergleichbarkeit zu schaffen. Allerdings wird bei den von Menschen angefertigten Dolmetschleistungen relativ zum Ausgangsmaterial häufig gezielt restrukturiert und strategisch auf Auslassungen zurückgegriffen, was nicht zwingenderweise Qualitätsdefiziten gleichzusetzen ist (vgl. Wang und Fantinuoli 2024: 328). Die resultierende schwächer ausgeprägte Übereinstimmung zwischen Ausgangsmaterial und Zieltext erschwert textbasierte Evaluationsansätze (vgl. Wang und Fantinuoli 2024: 328). Dies ist insofern problematisch, als der Großteil der maschinellen, automatischen Bewertung nach wie vor textbasiert erfolgt und folglich verschiedene, die Darbietung und Präsentationsebene betreffende und für die Wahrnehmung maßgebliche Merkmale – wie Prosodie, Pausenverhalten, Natürlichkeit – nur eingeschränkt berücksichtigt (vgl. Lu und Fantinuoli 2025: 5).

Da Menschen den Zieltext rezipieren und somit – anhand der ihnen individuell und unmittelbar greifbaren Kriterien – über die Qualität einer Dolmetschleistung urteilen, wurde die technologische Möglichkeit der maschinellen Bewertung von Dolmetschleistungen auf den vorangehenden Seiten zwar vorgestellt, jedoch in der vorliegenden Arbeit nicht zur Anwendung gebracht. Zur Beantwortung der dieser Arbeit zugrunde liegenden Fragestellungen wird eine rezipient*innenbasierte Beurteilung der Qualität herangezogen, da nur auf diesem Wege die relevanten Wahrnehmungsdimensionen der Menschen erfasst werden können, die maschinell

generierte Leistungen tatsächlich rezipieren. Wenngleich rein maschinell erzeugte Bewertungen von Dolmetschleistungen durchaus zur Optimierung von entsprechenden Systemen beitragen oder auch professionellen Sprachmittler*innen einen Nutzen bieten können, obliegt es letztlich stets den Nutzer*innen von Dolmetschleistungen – egal welchen Ursprungs – zu beurteilen, ob sie die ihnen vorgelegte Leistung als qualitativ hochwertig wahrgenommen haben oder nicht.

3 Methodik

In diesem Kapitel werden Forschungsdesign und Methodik der vorliegenden Arbeit behandelt. Zuerst werden die Zielsetzung der vorliegenden Arbeit erläutert sowie die zu beantwortenden Forschungsfragen und die zugrunde liegenden Hypothesen dargelegt. Im Anschluss daran wird ausgeführt, wie der zum Erhalt der erforderlichen Daten erforderliche Versuch gestaltet wird, auf welches Erhebungsinstrument zu diesem Zweck zurückgegriffen wird sowie mit welchem Ablauf und unter Einbindung welcher Versuchsteilnehmer*innen die Daten erhoben werden. Ferner wird erläutert, wie die Erhebung durchgeführt wurde und wie die auf diesem Wege erhaltenen Angaben, Daten und Informationen zur weiteren Verwendung in der vorliegenden Arbeit aufbereitet werden. Schließlich werden zur korrekten Einordnung in das Gefüge einschlägiger oder weiterführender wissenschaftlicher Werke die dieser Arbeit anhaftenden Limitationen und Grenzen aufgezeigt.

3.1 Zielsetzung, Forschungsfragen und Hypothesen

In den vorangehenden Kapiteln wurden die Funktionsweisen, Einsatzmöglichkeiten, Ausgestaltungsformen, Eigenheiten, Charakteristika, Probleme und Potenziale maschineller Dolmetschsysteme beleuchtet und die für das Verständnis der in der vorliegenden Arbeit zur Anwendung gebrachten Qualitätskriterien notwendigen theoretischen Grundlagen und Überlegungen behandelt. Wenngleich vollautomatisch erbrachte Dolmetschleistungen seit ihrer Konzeption eine lange Zeit von einer Reihe tiefgreifender technischer sowie weitgehend beständiger Limitationen geprägt waren, konnten in der Entwicklung maschineller Dolmetschsysteme beträchtliche Fortschritte verzeichnet und leistungsstarke Systeme hervorgebracht werden. Abseits der Implementierung in maschinellen Dolmetschsystemen kann zwar bereits heute dank jüngsten technologischen Fortschritten und Technologien von dem menschlichen Gegenstück kaum unterscheidbare Sprache künstlich erzeugt werden, unter den noch nicht restlos überwundenen Hürden und Herausforderungen im Kontext maschineller Dolmetschungen verbleiben jedoch nach wie vor insbesondere der Umgang mit Prosodie und eine möglichst menschlich und natürlich anmutende Sprachsynthese. Der Aspekt der Prosodie hat erhebliche Auswirkungen auf die Darbietung, Aussprache und Präsentation maschinell erstellter Dolmetschungen und somit auch auf ihre Wahrnehmung durch Rezipient*innen.

Nutzer*innen maschinell generierter Dolmetschungen haben eigene subjektive und zutiefst persönliche Erwartungshaltungen sowie Idealvorstellungen, anhand derer konsumierte Dolmetschleistungen gemessen und die entsprechenden Dolmetschsysteme bewertet werden.

Die vom Publikum empfundene und der Darbietung beigemessene Qualität wird weitestgehend von subjektiven Einschätzungen und Empfindungen geleitet. Diesen Überlegungen entsprechend und basierend auf der Hypothese, dass eine maschinell generierte und unzureichend natürlich klingende Sprache der wahrgenommenen Qualität abträglich ist, werden in der vorliegenden Arbeit primär die die Präsentationsebene betreffenden Aspekte der Sprachsynthese und ihr Einfluss sowie auserlesene weitere Qualitätsparameter anhand eines Experiments untersucht. Ziel der Arbeit ist es, zu analysieren und zu betrachten, wie die Qualität einer von einem Menschen erbrachten und einer maschinell generierten Dolmetschung derselben Ausgangsrede von Rezipient*innen bewertet wird, sowie etwaige Muster und Unterschiede aufzudecken, zu kontextualisieren und Schlüsse daraus zu ziehen. Die maschinelle Dolmetschung wurde hierbei anhand eines gegenwärtig handelsüblichen Dolmetschsystems erzeugt, um mit einer Momentaufnahme der aktuell maschinell erzielbaren Dolmetschleistungen die Qualitätswahrnehmungen zu erfassen.

Aus dieser Zielsetzung ergeben sich folgende Forschungsfragen:

1. Wie wird die Qualität einer mit einem maschinellen Dolmetschsystem erstellten Verdolmetschung von Rezipient*innen wahrgenommen, die den Ausgangstext nicht hören?
2. Wie wird die Qualität einer von einer professionellen Dolmetscherin dargebotenen Dolmetschung derselben Ausgangsrede von Rezipient*innen wahrgenommen, die den Ausgangstext nicht hören?
3. Welche Unterschiede ergeben sich bei der Qualitätswahrnehmung, wenn die von den Rezipient*innen wahrgenommene Qualität der maschinellen Verdolmetschung mit der wahrgenommenen Qualität der von der professionellen Dolmetscherin dargebotenen Verdolmetschung derselben Ausgangsrede verglichen werden?

Um Antworten auf diese Fragestellungen zu erhalten, wurde ein Experiment konzipiert und durchgeführt, in dessen Rahmen zwei Rezipient*innen-Gruppen jeweils entweder die maschinell generierte Verdolmetschung oder die von einem Menschen dargebotene Verdolmetschung zu hören bekamen, ohne über die Existenz zweier Gruppen, den zugrunde liegenden Forschungsgegenstand oder den Rückgriff auf maschinelle Inhalte informiert worden zu sein. Im Anschluss füllen die Teilnehmenden der beiden Gruppen einen – mit Ausnahme der jeweils eingebetteten Dolmetschleistung identen – Fragebogen zur wahrgenommenen Qualität aus. Die dieser Untersuchung zugrunde liegenden Hypothesen lauten wie folgt:

1. Eine nicht natürlich anmutende Stimme ist der Bewertung der Gesamtqualität der rezipierten Leistung abträglich;
2. die Unterschiede bei der wahrgenommenen Qualität der beiden Gruppen beziehen sich in erster Linie auf die primär die Präsentationsebene betreffenden Aspekte der Natürlichkeit, des Pausenverhaltens beziehungsweise des Sprechtempos, der Stimme, der Flüssigkeit und der erforderlichen Anstrengung beim Zuhören;
3. je sicherer sich die Teilnehmer*innen sind, dass es sich um eine maschinell erzeugte Leistung handelt, desto geringer die wahrgenommene Qualität.

An dieser Stelle ist festzuhalten, dass über inhaltliche Aspekte der für die vorliegende Arbeit herangezogenen maschinell generierten Dolmetschung sowie allgemein über die zum aktuellen Zeitpunkt maschinell erzielbaren Leistungen weder Aussagen getroffen noch Schlüsse gezogen werden. Inhaltliche Merkmale wie etwa fachliche und terminologische Korrektheit, Äquivalenz und Vollständigkeit, die bei Qualitätsbeurteilungen – insbesondere durch Gegenüberstellung von Ausgangs- und Zieltext – häufig beleuchtet werden, sind für die vorliegende Arbeit gegenstandslos. Stattdessen wird die Wahrnehmung von Qualität der zwei präsentierten Dolmetschleistungen ausschließlich auf Grundlage des ausgegebenen Outputs durch Erfragung der subjektiven Einschätzung der Rezipient*innen anhand standardisierter Likert-Items (siehe Kapitel 3.4) sowie ergänzender offener Fragen betrachtet.

3.2 Versuchsdesign und Teilnehmer*innen

Ziel ist es, die Wahrnehmung der Qualität von zwei gesonderten Dolmetschleistungen durch Rezipient*innen im Wege einer Befragung zu erfassen. Zu diesem Zweck wird die Gesamtheit der betreffenden Rezipient*innen randomisiert einer von zwei separaten Gruppen zugeteilt. Jeder der zwei Gruppen wird einer von zwei Fragebögen vorgelegt, die mit zwei kritischen Ausnahmen vollkommen identisch sind. Lediglich der in den jeweiligen Fragebogen eingebettete Audioclip der entsprechenden Dolmetschleistung sowie eine versteckte und für Rezipient*innen unsichtbare technische Variable zur eindeutigen Zuordnung der Befragungsergebnisse zum jeweiligen Fragebogen unterscheiden die beiden Erhebungsinstrumente. Folglich wird die Wahrnehmung der Qualität beider Dolmetschungen anhand identer Kriterien durch die Fragebögen erfragt, ungeachtet ihres menschlichen oder maschinellen Ursprungs. Die Unterschiede bei den Bewertungen können somit grundsätzlich – vorbehaltlich zufallsbedingter Unterschiede zwischen den Gruppen und entsprechender Antwortvariabilität – auf die jeweilige Dolmet-

schleistung zurückgeführt und ihnen zugeschrieben werden. Für die Untersuchung der vorliegenden Arbeit wurde die von einer professionellen Dolmetscherin angefertigte Simultandolmetschung herangezogen, die unter für die Ausübung des Dolmetschberufs üblichen Bedingungen unmittelbar bei der Darbietung des Ausgangstext angefertigt wurde. Auf Grundlage desselben Ausgangstextes wurde mithilfe eines gegenwärtig handelsüblichen maschinellen Dolmetschsystems die für die Analyse der vorliegenden Arbeit genutzte maschinelle Dolmetschung ex post erzeugt.

Die Teilnehmenden werden weder im Vorfeld noch im Laufe der Befragung darüber informiert, dass sie zur Befragung in Gruppen aufgeteilt werden, dass dementsprechend zwei unterschiedliche Versionen des Fragebogens existieren, dass maschinell generierte Inhalte eine Rolle spielen und dass der Zweck der Befragung darin liegt, die Grundlage für die Analyse der Qualitätswahrnehmung menschlich und maschinell erzeugter Dolmetschleistungen zu schaffen. Diese Informationen wurden den Teilnehmenden mit Vorbedacht und dem Ziel vorenthalten, Erwartungshaltungen, Voreingenommenheit und Vorwissen zu minimieren beziehungsweise zu vermeiden. Da keine Quelltextvergleiche angestellt werden, die Beurteilung ausschließlich auf dem Zieloutput beruht und dies für das Ziel der vorliegenden Arbeit unerheblich ist, wird den Rezipient*innen ferner die den Dolmetschleistungen zugrunde liegende Ausgangsrede nicht vorgelegt. Durch diese Gestaltungsentscheidungen und diesen Umgang mit Informationen sollen Störvariablen im Hinblick sowohl auf die zugrunde liegende Thematik als auch auf Ziel und Design der Befragung reduziert werden. Im Resultat soll ausschließlich die Wahrnehmung der Qualität der jeweiligen Dolmetschleistung im Zentrum stehen sowie erfasst und analysiert werden.

Um für die Befragung in Betracht kommen zu können, müssen die Teilnehmenden die Kriterien, über muttersprachliche Deutschkenntnisse zu verfügen und beruflich nicht als Dolmetscher*in tätig zu sein, erfüllen. Personen, die als Dolmetscher*innen tätig sind, werden von der vorliegenden Studie ausgenommen, da sie aufgrund ihres Fachwissens beziehungsweise ihrer Erfahrung sowohl einen anderen Zugang zu sowie Blick auf Dolmetschleistungen haben und die einzelnen Items des Fragebogens anders wahrnehmen könnten als Laien. Darüber hinaus könnten sie Ausrichtung und Zweck der Studie antizipieren und im Hinblick auf den Einsatz von maschinellen Dolmetschsystemen Vorbehalte hegen, die von der breiteren Bevölkerung anders gesehen beziehungsweise nicht geteilt werden. Da für die Analyse dennoch von Relevanz sein kann, in welchem Umfang die Teilnehmenden im Bereich des Dolmetschens Erfah-

rungen gesammelt haben, wird dies im Fragebogen erfragt. Alle Teilnehmer*innen werden direkt kontaktiert und auf diesem Wege rekrutiert. Alle Ausschlusskriterien werden im Zuge der Kontaktaufnahme abgefragt und verifiziert. Es wurde nicht mit Anreizen zur Teilnahme geworben. Für die Befragung wird eine Gesamtpopulation von insgesamt 50 und je 25 pro Gruppe herangezogen.

3.3 Material

Um das Forschungsziel der vorliegenden Arbeit zu erfüllen und etwaige Auswirkungen der den Dolmetschleistungen anhaftenden Unterschiede hinsichtlich der herangezogenen Qualitätsfaktoren auf die folglich beigemessene Qualität maschinell generierter Dolmetschungen zu analysieren, wurden eine maschinell erzeugte und eine menschengemachte Dolmetschung herangezogen, die auf einem Ausschnitt eines bei der Sitzung im Europäischen Parlament vom 23.1.2025 dargebotenen moderat fachlichen Beitrags beruhen. Dieser Ausgangstext wurde von Giuseppe Antoci, Mitglied des Europäischen Parlaments in der Fraktion Die Linke, in italienischer Sprache vorgetragen und behandelt das Thema Cyberkriminalität im Zusammenhang mit Kryptowerten. Auch wenn keine Quelltextanalyse vorgenommen wird, sei an dieser Stelle festgehalten, dass dieser Beitrag ein der Kommunikationssituation angemessenes, gehobeneres Register und bei einer Gesamtlänge von 1 Minute und 15 Sekunden ein durchschnittliches Sprechtempo von 132 Wörtern pro Minute aufwies. Dieses Ausgangsmaterial bietet ein realitätsnahes Setting für professionelle Dolmetschleistungen und durch anhaftende Faktoren wie Länge, Register, Fachlichkeit, Kontext und Kommunikationssituation eine geeignete Grundlage für die bei der vorliegenden Arbeit zum Tragen kommenden Dolmetschleistungen.

Bei der von einem Menschen angefertigten Dolmetschung handelt es sich um die bei der entsprechenden Sitzung in Echtzeit von einer professionellen Dolmetscherin erzeugte Simultandolmetschung, die unmittelbar von der Webseite des Europäischen Parlaments als Audiospur bezogen wurde. Zu der Qualifikation, der genauen Beschaffenheit der Dolmetschkabine, etwaigen eingesetzten technischen Mitteln, der Vorbereitung und anderen für die Humandolmetschung womöglich relevanten Faktoren können keine Aussagen getroffen werden. Die maschinelle Dolmetschung wurde ursprünglich als Videoclip ausgegeben, bei der sowohl die italienische Ausgangsrede als auch die wiedergegebene Dolmetschung in Echtzeit auf dem Bildschirm zum Mitlesen angegeben wurden. Die Videospur wurde vom Autor der vorliegenden Arbeit gänzlich entfernt, damit den Rezipient*innen keine unerwünschten zusätzlichen Informationen zur Verfügung stehen, die die Befragungsergebnisse verzerren könnten. Die Audiospuren beider Dolmetschleistungen wurden editiert, damit jeweils zu Beginn und Ende der

Audioclips keine potenziell missverständlichen oder störenden Pausen vorhanden sind und bei Wiedergabe im Zuge der Befragung die jeweilige Dolmetschung auf Knopfdruck unverzüglich abgespielt wird. Dadurch sollen vergleichbare Präsentationsbedingungen geschaffen und Verzerrungen und Artefakte an den betreffenden Stellen der Clips vermieden werden. Beide Audioclips wurden ohne Bevorzugung auf exakt dieselbe Weise bearbeitet und die zugrunde liegenden Dolmetschleistungen in unverfälschter Form bewahrt. Innerhalb der Clips wurden keine Eingriffe vorgenommen und etwaige Pausen während der Wiedergabe dementsprechend unverändert beibehalten. Der Audioclip der maschinell angefertigten Dolmetschung hat im Resultat eine Länge von 1 Minute und 15 Sekunden und jener der von einem Menschen erstellten Dolmetschung eine Länge von 1 Minute und 14 Sekunden. Die Lautstärke beider Clips ist funktional ident. Bei der Audiospur der von einer professionellen Dolmetscherin angefertigten Leistung konnten vom Autor der vorliegenden Arbeit keine klaren Artefakte festgestellt werden. Bei der Audiospur der maschinell realisierten Leistung war an zwei Stellen ein leises papierähnliches Rascheln zu vernehmen. Ferner wurden beide Audiospuren auf idente Weise in den Fragebogen eingebettet (siehe Kapitel 3.4). Im Endergebnis fanden die Teilnehmer*innen bei der Befragung als Grundlage für die Wahrnehmungsbewertung bei identer Benutzeroberfläche jeweils einen von zwei Audioclips vor. Im Kontrast zur professionell angefertigten Dolmetschleistung wurde die maschinelle Dolmetschung unter Nutzung einer männlich klingenden Stimme erzeugt.

Sowohl die Ausgangsrede als auch die beiden Dolmetschungen sind in den Anhängen 1 bis 3 in transkribierter Form enthalten. Um für die Betrachtung relevante Faktoren wie Pausenverhalten oder Auffälligkeiten beim Sprechtempo zu berücksichtigen, wurden die orthografischen Transkriptionen der beiden Dolmetschleistungen um entsprechende Faktoren erweitert. Dies erfolgte in Anlehnung an GAT 2 nach Selting et al. (2009), deren Ansatz für die Zwecke der vorliegenden Arbeit jedoch adaptiert wurde. Kurze Pausen von weniger als einer Sekunde wurden mit „(.)“ markiert und längere Pausen unter genauer Angabe der gemessenen Dauer erfasst, zum Beispiel „(3,1)“. Weitere Angaben zu den zusätzlich genutzten Markierungen sind den betreffenden in den Anhängen 2 und 3 enthaltenen Transkriptionen beigelegt. Die Ausgangsrede wurde rein in orthografischer Form transkribiert, da sie selbst nicht Gegenstand der Qualitätsbewertung war, sondern lediglich als Ausgangsmaterial für die Erzeugung der beiden Dolmetschleistungen herangezogen wurde. Eine ähnliche Erweiterung der orthografischen Transkription der Ausgangsrede konnte darüber hinaus nicht vorgenommen werden, da die Aufnahme zum Zeitpunkt der finalen Bearbeitung der vorliegenden Arbeit von der Plattform

des Europäischen Parlaments entfernt wurde. Den bisherigen theoretischen Ausführungen entsprechend liegt der Zweck der Transkriptionen nicht darin, Quelltextvergleiche und -analysen anzufertigen oder zu ermöglichen, sondern darin, für Transparenz, Nachvollziehbarkeit und Replizierbarkeit – insbesondere im Hinblick auf mögliche weiterführende Forschung – zu sorgen. Im Anschluss an die Erläuterungen zum Versuchsdesign, wird im Folgenden näher auf das Erhebungsinstrument eingegangen.

3.4 Fragebogen

Für die Erhebung der für die vorliegende Arbeit maßgeblichen subjektiven Qualitätswahrnehmungen und weiteren Informationen zu den Teilnehmenden wird ein online abrufbarer standardisierter Fragebogen genutzt, um ihre Urteile und Bewertungen in vergleichbarer und systematischer Weise unter Berücksichtigung verschiedener Qualitätsparameter im Wege einer über beide Gruppen funktional identischen Messung zu erfassen. Mehrere Faktoren haben zur Entscheidung beigetragen, einer webbasierten online-Umsetzung Vorzug vor einer Befragung vor Ort anhand eines physischen Fragebogens zu geben. Im Kern dieser Entscheidung steht die technische Realisierbarkeit zweier nahezu identischer Fragebögen, bei denen die Audioclips jeweils nahtlos eingebettet werden können. Im Zuge der Gestaltung des Erhebungsinstruments können die Audioclips permanent an einen der Fragebögen gekoppelt werden, wodurch Verwechslungen beziehungsweise inkorrekte Zuweisungen von Antworten einer Person zum betreffenden Fragebogen technisch ausgeschlossen sind. Die technische Einbettung der dafür erforderlichen Variable ist für Teilnehmende nicht ersichtlich.

Darüber hinaus können die Fragebögen den Teilnehmenden über die Online-Lösung einfach, unkompliziert, ohne Zeitverzögerung und unabhängig von räumlichen Distanzen zugänglich gemacht werden. Es steht den Teilnehmer*innen ferner gänzlich frei, einen für sie opportunen Zeitpunkt für die Nutzung des ihnen zugewiesenen Links zur Beantwortung des Fragebogens zu wählen, was der Rücklaufquote zuträglich sein könnte. Auf diesem Wege kann außerdem sichergestellt werden, dass die Fragebögen allen Rezipient*innen über beide Gruppen hinweg in zwei Varianten mit funktional identischem Rahmen – mit den oben dargelegten essenziellen Unterschieden – und identischer Umfrageumgebung in standardisierter Weise präsentiert werden. Die Form dieser Präsentation kann bei der Gestaltung des Fragebogens in hohem Maße kontrolliert und gesteuert werden.

Bei der Nutzung von Fragebögen ist zu beachten, dass Rezipient*innen dazu tendieren, die Form betreffende Aspekte – wie Flüssigkeit, Stimme, Rhythmus – höher zu bewerten, was

sich nach De Gregoris (2015: 57f) unter anderem wiederum auf die Wahrnehmung der Kohärenz und allgemeinen Qualität der Dolmetschleistung auswirkt. Ferner könnten Fragebögen die kognitiven Ressourcen der Rezipient*innen belasten, da es sich schwierig gestalten kann, bei einer ausgedehnten und umfangreichen Liste von Fragen zu Qualitätskriterien diese voneinander abzugrenzen, da sie zwar individuell dargeboten, inhaltlich jedoch verwandt sein könnten (vgl. Han und Lu 2021: 73). Dies könnte wiederum dazu führen, dass Rezipient*innen konzeptuell verwandte Fragen ähnlich oder gar ident behandeln und bewerten (vgl. Han und Lu 2021: 73). Folglich könnten Fragen, die entweder inhaltlich starke Überschneidungen aufweisen oder auf ähnliche Konzepte abzielen, redundant und dem insgesamt gewonnenen Umfang an Einblicken abträglich sein (vgl. Han und Lu 2021: 73). Daher wird bei dem für die vorliegende Arbeit entworfenen Fragebogen versucht, Anzahl und Umfang der Fragen auf ein zweckdienliches Maß zu beschränken, dabei konzeptuelle Überschneidungen zu minimieren und dennoch ein ausreichend und angemessen umfangreiches Spektrum an Qualitätskriterien zu erfragen.

Neben Umfang und Anzahl der Fragen ist ferner zu beachten, wie den Rezipient*innen Antwortmöglichkeiten dargeboten werden. Die Ausgestaltung von Bewertungsskalen – sprich die Anzahl der Abstufungen und die Ankerformulierung – kann die Interpretierbarkeit der Ergebnisse beeinflussen und ist daher Teil der methodischen Qualität des Designs (Pöchhacker 2005: 147f). Die bei Fragebögen zur Anwendung gelangende Formulierung der Fragen hat folglich erhebliche Auswirkungen darauf, wie die Rezipient*innen auf die Fragen reagieren, sie verstehen und schließlich beantworten. Die Art und Weise, wie Fragen gestellt und präsentiert werden, kann die Entscheidungen der Rezipient*innen ferner lenken und etwaige Vorurteile bedienen oder ihnen entgegenwirken (vgl. Kurz 2001: 397). Da sich die Formulierung von Fragebögen auf die Erhebung von Qualitätswahrnehmungen auswirken kann (vgl. Pöchhacker 2005: 156), wurde bei der Entwicklung der Skalen besonderes Augenmerk auf präzise Formulierungen und kohärente Abgrenzungen gelegt. Dabei wurden die Fragen sprachlich möglichst deutlich gefasst, begrifflich voneinander abgegrenzt und in einem Pretest auf Verständlichkeit geprüft. Durch die im Resultat zur Anwendung gebrachte Formulierung der Fragen sollen die entsprechenden Skalen möglichst eindeutig auf die ausgewählten Qualitätsdimensionen ausgerichtet sein.

Darüber hinaus wurde besonders auf konsistente Ankerformulierungen und terminologische Genauigkeit und Einheitlichkeit geachtet, um etwaigen Formulierungsartefakten vorzubeugen. Ferner wurde versucht, die Fragen laienverständlich zu formulieren und diese gleich-

zeitig auf eine Weise zu präzisieren, dass semantische Überlappungen zwischen den verschiedenen Kriterien auf ein Minimum reduziert werden. Ein weiterer Faktor, der eine klare Formulierung des Erhebungsinstruments ex ante voraussetzt, ist der Umstand, dass online durchgeführte Befragungen ohne Beteiligung des Autors vor Ort stattfinden, der die Teilnehmenden durch die Fragebögen führen und etwaige Unklarheiten oder Probleme ad hoc beheben könnte (vgl. Bethlehem 2010: 165).

Um über den gesamten Erhebungsprozess sicherzustellen, dass die im Fragebogen erhaltenen Informationen stets in der zur Erfüllung der Zielsetzungen erforderlichen Reihenfolge präsentiert und verfrühte Angaben, die zu Verzerrungen oder Voreingenommenheit führen könnten, vermieden werden, wurde die Befragung auf insgesamt vier mit relevanten Inhalten befüllte Seiten aufgeteilt. Gelangen Teilnehmende nach vollständigem Abschluss des Fragebogens zur fünften und letzten Seite, so werden die Eingaben ordnungsgemäß gespeichert und bei der Analyse der vorliegenden Arbeit berücksichtigt.

Nach der Einwilligung zur Teilnahme und Bestätigung der Erfüllung der Ausschlusskriterien erhielten Teilnehmende randomisiert einen Link, der sie zur Startseite eines der beiden Fragebögen führt. Den vorangehenden Ausführungen entsprechend sind die beiden Startseiten vollkommen ident. Diese Startseite dient dazu, den Teilnehmenden allgemeine Hinweise und eine grundsätzliche Orientierung zur Befragung zu bieten sowie grundlegende Einwilligungen zur Teilnahme und zur Datenverarbeitung einzuholen, ohne vorab zu weitgehende Angaben zu machen. Unter dem bewusst vage formulierten Titel „Befragung zu einer Dolmetschaufnahme“ werden klare Anweisungen gegeben, was für die Teilnehmenden konkret zu tun sein wird, sowie eine Information im zweckdienlichen Umfang darüber, welche Inhalte und Aufgaben die Teilnehmenden erwarten werden. Die Teilnehmenden werden darüber informiert, dass ihnen eine Dolmetschung einer kurzen Rede vorgelegt werden wird und sie anschließend Fragen dazu beantworten werden. Ihnen wird darüber hinaus mitgeteilt, dass mit der Befragung ein rein wissenschaftlicher Zweck verfolgt und die Teilnahme circa 10 Minuten in Anspruch nehmen wird. Hierbei werden die Teilnehmenden darauf hingewiesen, dass vorab nicht alle Details zum Untersuchungszweck genannt werden können, um die Ergebnisse nicht zu beeinflussen. Den obigen Ausführungen entsprechend enthält die Startseite keine Hinweise im Hinblick auf die Aufteilung in Gruppen, das Vorhandensein zweier Versionen des Fragebogens, die Rolle maschinell generierter Inhalte und tiefergehende Facetten des Zwecks der Befragung. Die Teilnehmenden werden ferner darum gebeten, bei der Rezeption nach Möglichkeit für eine ruhige

Umgebung zu sorgen und Kopfhörer zu verwenden. Es wird explizit dazu aufgerufen, die Aufnahme vollständig anzuhören und im Anschluss die Fragen und Punkte nach persönlichem Eindruck zu beantworten. Schließlich werden die Teilnehmenden darüber in Kenntnis gesetzt, dass sie durch Klicken der mit der Aufschrift „Weiter“ gekennzeichneten Schaltfläche ihre freiwillige Teilnahme sowie die Kenntnisnahme der Datenschutzinformationen bestätigen. Eine Datenschutzerklärung ist auf jeder einzelnen Seite des Fragebogens ersichtlich und aufrufbar.

Seite zwei des Erhebungsinstruments ist ausschließlich der Wiedergabe des Audioclips gewidmet, um Ablenkungen zu minimieren und sicherzustellen, dass der Audioclip tatsächlich in voller Länge angehört wird. Unter dem Schriftzug: „Bitte hören Sie sich den folgenden Clip einmal vollständig an“ finden die Teilnehmenden die eingebettete Funktion zur Wiedergabe der Audiospur. Teilnehmende müssen die Wiedergabe selbst starten und können diese in der Folge nach Belieben pausieren und wieder aufnehmen. Außerdem steht es ihnen frei, die Audiospur herunterzuladen. Die Möglichkeit, die Audiospuren zu beziehen, ist unabänderlich der technischen Umsetzung des Fragebogens geschuldet.

Die daran anschließende dritte Seite führt die Teilnehmenden zum Herzstück des Erhebungsinstruments. Zur standardisierten Erfassung ihrer Qualitätsurteile erwarten sie dort zwölf Likert-Items sowie zwei offene Fragen. Likert-Items beruhen auf dem nach Likert (1932) etablierten Ansatz, subjektive Urteile zu erfassen, indem Teilnehmende ihren Grad der Zustimmung oder Ablehnung zu gegebenen Aussagen anhand einer geordneten Reihe von Kategorien angeben, etwa von „strongly approve“ über „approve“, „undecided“, „disapprove“ zu „strongly disapprove“ (vgl. Boone und Boone 2012: 1; Likert 1932: 14f). Dabei wird häufig eine numerische Kodierung – etwa 1 bis 5 – eingesetzt (vgl. Boone und Boone 2012: 1). Analog dazu wird bei den zwölf im Fragebogen verwendeten an das Werk Likerts angelehnten Items jeweils eine fünfstufige Zustimmungsskala als Antwortmöglichkeiten gegeben. Wie auf der entsprechenden Seite des Fragebogens bei jedem Item jeweils dargelegt wird, entspricht „1“ dabei „Ich stimme gar nicht zu“ und „5“ entspricht „Ich stimme voll und ganz zu“. Es wurden lediglich die mit den jeweiligen Polen – sprich 1 und 5 – assoziierten sprachlichen Gegenstücke angegeben, wodurch die Abstufungen 2, 3 und 4 ohne explizite Benennungszuweisung in den Fragebogen eingebettet wurden. Diese Skala ist für die Teilnehmenden leicht verständlich und ermöglicht eine standardisierte Bewertung der jeweiligen Urteile.

Die Items und offenen Fragen des Fragebogens wurden basierend auf den in Kapitel 2.1 identifizierten relevanten Qualitätskriterien formuliert. Da Qualität in der Dolmetschwissenschaft den oben dargelegten theoretischen Überlegungen entsprechend als mehrdimensionales sowie kontext- und situationsabhängiges Konzept verstanden wird, werden mit dem Fragebogen neben der globalen Qualitätsbewertung auch einzelne Aspekte wahrgenommener Qualität bewertet. Dazu zählen etwa wahrgenommene Kohärenz, Vertrauenswürdigkeit, sprachliche Korrektheit, Flüssigkeit, Natürlichkeit, Stimmqualität, Angemessenheit des Sprechtempos und Professionalität.

Die formulierten in den Fragebogen eingebetteten und mit der fünfstufigen Zustimmungsskala versehenen Items lauten wie folgt:

- Ich konnte der Dolmetschung inhaltlich gut folgen.
- Die Dolmetschung war kohärent.
- Die Dolmetschung war sprachlich (Grammatik, Wortwahl, Satzbau) korrekt.
- Die Vortragsweise der Dolmetschung war flüssig.
- Die Formulierungen wirkten natürlich.
- Die Stimme war angenehm zu hören.
- Pausen/Verzögerungen haben mich beim Zuhören gestört.
- Das Sprechtempo war angenehm.
- Es war anstrengend, der Dolmetschung zu folgen.
- Die Dolmetschung wirkte auf mich professionell.
- Die Dolmetschung wirkte auf mich vertrauenswürdig.
- Die Dolmetschung war insgesamt von hoher Qualität.

Ergänzend dazu werden im Anschluss die beiden folgenden offenen Fragen gestellt:

- Gab es etwas, das Sie an der Dolmetschung besonders gut fanden?
- Gab es etwas, das Sie an der Dolmetschung besonders schlecht fanden?

Zu beachten ist, dass sich die fünfstufige Zustimmungsskala, die von „Ich stimme gar nicht zu“ zu „Ich stimme voll und ganz zu“ reicht, bei den beiden Items „Pausen/Verzögerungen haben

mich beim Zuhören gestört“ und „Es war anstrengend, der Dolmetschung zu folgen“ relativ zu den übrigen, positiv formulierten Items implizit von einer positiven in eine negative Wahrnehmung invertieren. Ferner müssen die Teilnehmenden jeden Punkt – mit Ausnahme der offenen Fragen – mit einer Antwort versehen, um die Seite abschließen und auf die nächste Seite gelangen zu können. Die Beantwortung der offenen Fragen ist fakultativ. Sie dienen dazu, die quantitativen Urteile zu kontextualisieren und konkrete Aspekte zu erfassen, die Teilnehmende als besonders gut oder schlecht empfunden haben.

Sobald die Fragen abgeschlossen und bestätigt wurden, gelangen Teilnehmende auf Seite 4, auf der weitere Angaben zur Person wie Alter, Bildungsgrad und das mit der eigenen Person identifizierte Geschlecht erfragt werden. Dabei wird auch die zentrale Frage gestellt, ob nach Ansicht der Teilnehmenden die Dolmetschleistung von einer Maschine angefertigt worden sein könnte. Dies erfolgt erst auf dieser an den Großteil der Fragen anschließenden Folgeseite, damit nicht unbeabsichtigt durch Scrollen ein Hinweis auf diese Frage erhalten werden kann und so Verzerrungen bei der Beantwortung der Items auf Seite 3 vermieden werden.

Wie bei den zwölf Likert-Items der Vorseite wurde eine fünfstufige Zustimmungsskala auch bei der Frage nach der Einschätzung im Hinblick auf den möglichen maschinellen Ursprung der Dolmetschleistung (1 = Nein, sicher nicht; 5 = Ja, ganz sicher) als auch bei der an dieser Stelle platzierten Frage nach der Erfahrung mit Dolmetschen (1 = Keine Erfahrung; 5 = Sehr viel Erfahrung) zur Anwendung gebracht. Sobald alle Antworten eingetragen und bestätigt sind, gelangen die Teilnehmenden schließlich auf die letzte Seite des Fragebogens, auf der ihre Teilnahme und Angaben bestätigt werden. Erst wenn diese Seite erreicht wird, gilt der Fragebogen als vollständig ausgefüllt und können die Ergebnisse berücksichtigt werden. Die Qualität beider Dolmetschleistungen und die Wahrnehmung durch die Rezipient*innen werden diesen Ausführungen entsprechend durch die beiden Fragebögen im Lichte derselben Kriterien abgefragt, ungeachtet von deren maschinellem oder menschlichem Ursprung. Indem jeder Gruppe somit bei gleichbleibendem Rahmen und Kontext je eine Dolmetschleistung isoliert als unterscheidender Faktor vorgelegt wird, soll hohe Vergleichbarkeit erzielt werden.

3.5 Durchführung

Die Erhebung wurde im Zeitraum von 10.2.2026 bis zum 23.5.2026 über insgesamt 102 Tage durchgeführt. Für die technische Online-Umsetzung des Erhebungsinstruments wurde das Tool

„LimeSurvey“¹ herangezogen. Nach Erstellung und Aktivierung der beiden Fragebögen wurden permanente Links erstellt, die theoretisch endlos zum Starten neuer Befragungsdurchgänge genutzt werden können. Den Teilnehmenden wurde nach Einwilligung und Bestätigung des Nichtvorhandenseins der Ausschlusskriterien randomisiert je einer der beiden Zugangslinks übermittelt.

Die im Vorfeld angestrebte Zahl von 50 Teilnehmenden zu je 25 pro Gruppe konnte erreicht werden. Um sicherzustellen, dass beide Gruppen bei Abschluss der Befragung mit einer identen Anzahl von Personen befüllt sind, wurde der letzten und 50. Teilnehmerin kein zufälliger Link übermittelt, sondern der jener Gruppe, bei der noch ein*e Teilnehmer*in gefehlt hat. Dem Feedback einer nicht quantifizierten Teilmenge von Teilnehmenden zufolge hat die Durchführung der Befragung weniger als die veranschlagten 10 Minuten in Anspruch genommen, die als Richtwert auf der Startseite der Befragung angegeben wurden. Es werden ausschließlich die abgegebenen Antworten erfasst, die als Datensatz automatisch jeweils mit einer fortlaufenden Identifikationsnummer versehen werden. Um eine umfassende Erfassung, Betrachtung und Auslegung der in den Anhängen 5 bis 8 enthaltenen Auflistungen der Antwortsätze im Hinblick auf die Qualitätsurteile und die getrennt dargestellten offenen Fragen zu ermöglichen, wurde den Datensätzen auf Grundlage der automatisch generierten fortlaufenden Identifikationsnummer für die Darstellung in der vorliegenden Arbeit eine übersichtliche Auswertungskennung zugewiesen. Zusätzlich wurden das Einreichungsdatum und das Datum der letzten Aktivität automatisch gespeichert sowie ob die betreffende Befragung erfolgreich abgeschlossen wurde. Im Rahmen der Befragung wurden keine Namen, E-Mail-Adressen oder sonstigen Angaben erhoben oder gespeichert, auf deren Grundlage die Teilnehmenden unmittelbar identifiziert werden können.

Die Teilnehmenden wurden hinterher über jene Aspekte informiert, die ihnen ex ante vorenthalten wurden, darunter dass sie zur Befragung in Gruppen aufgeteilt wurden, dass dementsprechend zwei unterschiedliche Versionen des Fragebogens existieren, dass maschinell generierte Inhalte eine Rolle spielen und dass der Zweck der Befragung darin liegt, die Grundlage für die Analyse der Qualitätswahrnehmung menschlich und maschinell erzeugter Dolmettschleistungen zu schaffen. Nach Erhalt aller Antwortsätze wurden diese schließlich für die anschließende Aufbereitung exportiert.

¹ <https://www.limesurvey.org/de>

3.6 Datenaufbereitung

Die von den Rezipient*innen eingeholten Bewertungen werden anhand einer fünfstufigen Zustimmungsskala erfasst. Für die Darstellung und Veranschaulichung der Ergebnisse und Antworten zu jedem Item werden Mittelwert und Standardabweichung angegeben, um deren Lage auf der Zustimmungsskala und die anhaftende Streuung der erhaltenen Bewertungen anzugeben. Dem Mittelwert – im Folgenden auch „Durchschnitt“ – ist die durchschnittliche Zustimmung zu entnehmen, während aus der ebenso errechneten Standardabweichung hervorgeht, in welchem Umfang die individuellen Bewertungsurteile von dem erfassten Durchschnitt abweichen. Die Standardabweichung wird in Excel anhand der Funktion STABW.S als Stichproben-Standardabweichung berechnet.

Es sei festgehalten, dass die Abstände zwischen den individuellen Stufen der für die vorliegende Arbeit genutzten Zustimmungsskala von den Rezipient*innen nicht zwingend als gleich groß betrachtet beziehungsweise ausgelegt werden könnten, wodurch zugrunde liegende Verteilungsmuster durch Mittelwerte verzerrt oder verdeckt werden können (vgl. Pöchlacker 2005: 147f). Daher werden ergänzend jeweils der Median und der Modus – sprich der am häufigsten gewählte Wert – sowie die prozentuellen Verteilungen der Antworten angegeben. Auf diesem Wege sollen Homo- beziehungsweise Heterogenität bei den erhaltenen Antworten und vorhandene Muster besser erfasst und ersichtlich gemacht werden.

Zusätzlich werden die erfassten Unterschiede bei der Wahrnehmung anhand der standardisierten Effektstärke d nach Cohen (1988) und dem darauf aufbauenden Werk von Lakens (2013) dargelegt. Durch Zugrundelegung dieses Wertes sollen Unterschiede standardisiert und relativ zur Streuung der Antworten veranschaulicht werden (vgl. Cohen 1988: 67; Lakens 2013: 3f). Im Resultat geht aus der ermittelten Effektstärke d hervor, welche Items am stärksten zu den Unterschieden zwischen den Gruppen beitragen (vgl. Lakens 2013: 3). Je stärker die standardisierte Differenz ausgehend von den Mittelwerten der beiden Gruppen in Bezug auf ein Item ausgeprägt ist, desto stärker trägt der betreffende Aspekt zum Gesamtunterschied bei den erfassten Wahrnehmungen bei (vgl. Lakens 2013: 3f). Dabei wird der Effekt nach Cohen (1988: 25f) als klein kategorisiert, wenn d bei 0,2 liegt, mittel bei 0,5 und groß bei 0,8.

Die Effektstärke d wird in der vorliegenden Arbeit wie in der nachstehenden Gleichung beschrieben pro Item berechnet, wodurch resultierende positive Werte höhere Bewertungen seitens der die menschengemachte Dolmetschung rezipierenden Teilnehmenden bedeuten (vgl. Cohen 1988: 66f; Lakens 2013: 3).

$$d = \frac{M_{\text{menschliche Dolmetschung}} - M_{\text{maschinelle Dolmetschung}}}{SD_{\text{gepoolt}}}$$

Die in dieser Gleichung angeführte gemeinsame Standardabweichung (gepoolte SD) wird hierbei auf der Grundlage der gruppenübergreifend erfassten Abweichungen errechnet und zeigt, in welchem Ausmaß die Antworten beider Gruppen typischerweise voneinander abweichen (vgl. Lakens 2013: 3f). Dies dient dazu, die Mittelwertunterschiede für die Berechnung der Effektstärke d auf eine einheitliche und vergleichbare Skala zu setzen (vgl. Cohen 1988: 67; Lakens 2013: 3). Die gepoolte SD wird dabei wie folgt berechnet (vgl. Cohen 1988: 67; Lakens 2013: 3):

$$SD_{\text{gepoolt}} = \sqrt{\frac{(n_{\text{menschl. Dolm.}} - 1)SD_{\text{menschl. Dolm.}}^2 + (n_{\text{masch. Dolm.}} - 1)SD_{\text{masch. Dolm.}}^2}{n_{\text{menschl. Dolm.}} + n_{\text{masch. Dolm.}} - 2}}$$

Da der Wert n der Teilmenge der entsprechenden Gruppe entspricht und dieser Wert je Gruppe 25 beträgt, lässt sich diese Gleichung zur besseren Verständlichkeit wie folgt vereinfachen:

$$SD_{\text{gepoolt}} = \sqrt{\frac{SD_{\text{menschliche Dolmetschung}}^2 + SD_{\text{maschinelle Dolmetschung}}^2}{2}}$$

Um die Hypothesen zu den Zusammenhängen zwischen Stimme beziehungsweise Natürlichkeit und der wahrgenommenen Gesamtqualität sowie zwischen der Einschätzung, ob es sich bei der rezipierten Leistung um eine maschinell generierte Dolmetschung handelt, und der wahrgenommenen Gesamtqualität zu beleuchten, werden darüber hinaus in Kapitel 4.5 die Korrelationen r zwischen den betreffenden Angaben der Teilnehmenden errechnet, um den zugrunde liegenden Zusammenhang der Werte zu quantifizieren. Ein positiver errechneter Wert entspricht hierbei einem positiven linearen Zusammenhang – je höher Wert A desto höher Wert B – und ein negativer Wert wiederum einen entsprechend umgekehrten negativen Zusammenhang – je höher Wert A desto geringer Wert B (vgl. Schober et al. 2018: 1763). Der Größe des errechneten Werts r – der zwischen -1 und +1 liegt – kann entnommen werden, wie ausgeprägt dieser Zusammenhang ist – je höher der Wert, desto stärker ist dieser Zusammenhang bei den entsprechenden Werten (vgl. Schober et al. 2018: 1763). Der lineare Korrelationskoeffizient r wird hierbei in Excel durch Nutzung der Funktion KORREL auf Grundlage der Antworten der Teilnehmenden errechnet. Rückschlüsse auf eine etwaige anhaftende Kausalität werden aus den errechneten Korrelationswerten nicht abgeleitet (vgl. Schober et al. 2018: 1767).

Ferner werden die erhaltenen Antworten bei den beiden Aussagen „Pausen/Verzögerungen haben mich beim Zuhören gestört“ und „Es war anstrengend, der Dolmetschung zu folgen“ bei bestimmten Darstellungen und Auslegungen invertiert. Inversion bedeutet in diesem Zusammenhang konkret, dass die erhaltenen Antworten so behandelt werden, als wären die beiden Pole umgedreht (demnach entspricht „5“ „Ich stimme gar nicht zu“ und „1“ „Ich stimme voll und ganz zu“). Dies ist erforderlich, da nur bei diesen beiden Items eine höhere Zustimmung mit einer negativeren Wahrnehmung assoziiert ist. Zur Berechnung der Effektstärke d wird kategorisch eine solche Inversion vorgenommen. Werden im Folgenden weitere entsprechende Inversionen vorgenommen, so wird dies jeweils an der betreffenden Stelle angemerkt.

Die oben dargelegten zur Anwendung gebrachten Metriken werden bei der Darstellung und Behandlung der Ergebnisse zwecks Übersichtlichkeit wie folgt abgekürzt: Mittelwert (M), Standardabweichung (SD), Median (Mdn), Modus (Mo), Cohen's d (d), Korrelation (r) und die Differenz der Mittelwerte (ΔM). Aus praktischen Gründen und um Übersichtlichkeit zu bewahren, werden die beiden Gruppen, auf die die Teilnehmenden zu gleichen Teilen aufgeteilt wurden, ferner wie folgt bezeichnet: Die Gruppe, der die maschinell angefertigte Dolmetschung vorgelegt wurde, wird im Folgenden AD-Gruppe genannt (automatische Dolmetschung) und die andere wird als MD-Gruppe bezeichnet (menschliche Dolmetschung).

3.7 Limitationen und Grenzen

Nachdem auf den vorangehenden Seiten Erhebung, Methodik und Design der Erhebung eingehend beleuchtet wurden, sollen im Folgenden Limitationen und Grenzen der Erhebungsmethode eingeräumt und behandelt werden, um Aussagen und Ergebnisse der vorliegenden Arbeit – insbesondere im Hinblick auf eine etwaige weiterführende Forschung – in einen angemessenen Rahmen zu fassen. Auch wenn alle Teilnehmenden zwar gezielt für die Befragung rekrutiert wurden, kann nach Übermittlung eines der Zugangslinks nicht nachvollzogen oder kontrolliert werden, wie dieser verwendet wird. Beispielsweise kann dieser entgegen der abgegebenen Zusage nicht zur Teilnahme verwendet oder auch weitergegeben und von einer anderen, nicht autorisierten Person genutzt werden. Außerdem können Befragungen gestartet und dennoch nicht abgeschlossen werden. Im Resultat entsteht dadurch eine Form der Selbstselektion, in deren Folge die tatsächlich realisierte Stichprobe nicht gänzlich der Kontrolle des Autors unterliegt (vgl. Bethlehem 2010: 163). In diesem Zusammenhang sei auch darauf hingewiesen, dass es sich bei den Teilnehmenden um eine Gelegenheitsstichprobe handelt (vgl. Etikan et al. 2016: 1f; Springer 2026). Es wurden Personen ausgewählt, die ad hoc zur Verfügung standen

und grundsätzlich in der Lage sind, auf das online eingebettete Erhebungsinstrument zuzugreifen. Die Ergebnisse der vorliegenden Arbeit müssen folglich als Ausdruck der im Zuge der ihr zugrunde liegenden Erhebung individuell realisierten Stichprobe ausgelegt werden.

Die Aussagekraft des Gruppenvergleiches innerhalb der realisierten Stichprobe bleibt davon zwar unberührt, auf die Gesamtbevölkerung generalisierbare Rückschlüsse sind in der Folge jedoch nur unter Vorbehalt möglich. Die Generalisierbarkeit ist auch im Hinblick auf die Ausgangsrede und die darauf basierenden maschinell erzeugten beziehungsweise menschengemachten Dolmetschleistungen eingeschränkt. Jede Rede bringt eigene Herausforderungen mit sich und Dolmetschleistungen variieren kontext- und situationsabhängig von Person zu Person und von System zu System. Es wurde lediglich eine einzelne Dolmetschleistung eines einzelnen Dolmetschsystems herangezogen, wodurch nur beschränkt Rückschlüsse auf andere Systeme beziehungsweise die Allgemeinheit der derzeit verfügbaren Dolmetschsysteme gezogen werden kann. Diese Überlegungen treffen analog auch auf die von der professionellen Dolmetscherin erzeugte Leistung zu.

Darüber hinaus können die Hörsituation und -umgebung, die zum Zugriff auf das Erhebungsinstrument genutzten Endgeräte und die Einhaltung der Aufforderung, Kopfhörer zu nutzen, nicht kontrolliert werden. Wie bereits erwähnt wurden ferner beide Clips minimal editiert, um gehaltlose Pausen zu Beginn und Ende zu entfernen und zwei Audiospuren zu erhalten, die die Ausgangsrede dennoch zur Gänze abdecken und inhaltlich die gewollte Grundlage für die Wahrnehmungsbewertung liefern. Zwar werden dadurch Verzerrungen vermieden und die Vergleichbarkeit erhöht, dennoch wird dieser Umstand zwecks Vollständigkeit und Transparenz als Eingriff festgehalten.

4 Ergebnisse

Die nach der vorstehend beschriebenen Methodik erhobenen Daten werden im Folgenden im Lichte des Forschungsziels im Detail dargelegt und ausgewertet. Zunächst wird durch Veranschaulichung und Analyse der demographischen Angaben der Proband*innen die erfasste Stichprobe beleuchtet. Um der anschließenden Betrachtung einen Rahmen zu geben, wird festgehalten, in welchem Umfang die Teilnehmenden über Erfahrung im Zusammenhang mit Simultandolmetschen verfügen sowie die von den Proband*innen angegebenen Eindrücke, ob die ihnen vorgelegte Dolmetschleistung menschlichen oder maschinellen Ursprungs ist. Die von den Teilnehmenden erhaltenen Angaben zu den Likert-Items und etwaige Antworten auf die offenen Fragen werden im Anschluss daran festgehalten, übersichtlich dargestellt und ausgewertet. Da die Wahrnehmung einer maschinell erzeugten Dolmetschleistung im primären Forschungsfokus der vorliegenden Arbeit liegt, wird im Anschluss an die Betrachtung der demographischen Angaben mit der Auswertung der von der AD-Gruppe erhaltenen Ergebnisse begonnen. Im darauffolgenden Kapitel 4.3 wird auf die Ergebnisse der MD-Gruppe eingegangen. Dabei erkannte Unterschiede zwischen den beiden Gruppen werden in der Folge aufgedeckt und thematisiert sowie die Forschungsfragen und die Hypothesen eingehend behandelt.

4.1 Datengrundlage und Stichprobe

Insgesamt wurden 50 Fragebögen vollständig abgeschlossen, sachgemäß bestätigt und können bei der vorgenommenen Analyse berücksichtigt werden. Je 25 davon entfallen auf die AD-Gruppe und die MD-Gruppe. Nicht zur Gänze beziehungsweise ordnungsgemäß abgeschlossen wurden 3 Fragebögen der MD-Gruppe und 9 der AD-Gruppe, die folglich nicht in die Analyse einbezogen wurden. Keine Person wurde aufgrund der Nichterfüllung der erforderlichen Kriterien abgelehnt oder ausgenommen. In der nachstehenden Tabelle sind die demographischen Angaben der Teilnehmenden (Alter, Bildungsgrad, Geschlecht) zwecks Übersicht über die Stichprobe wiedergegeben. Die Prozentangaben beziehen sich jeweils auf die betreffende Spalte (AD, MD oder Gesamt). In den Anhängen 5 bis 8 sind Aufstellungen aller Angaben der Teilnehmenden, einschließlich detaillierterer demographischer Angaben, enthalten, denen weitere Einzelheiten entnommen werden können, von deren Angabe an dieser Stelle zwecks Übersichtlichkeit abgesehen wurde.

Tabelle 1: Demographische Angaben der Stichprobe

	AD	MD	Gesamt
Teilnehmerzahl	25 (100 %)	25 (100 %)	50 (100 %)
Alter			
18 bis 29 Jahre	5 (20 %)	7 (28 %)	12 (24 %)
30 bis 39 Jahre	13 (52 %)	13 (52 %)	26 (52 %)
40 bis 49 Jahre	4 (16 %)	1 (4 %)	5 (10 %)
50 bis 59 Jahre	1 (4 %)	2 (8 %)	3 (6 %)
60+ Jahre	2 (8 %)	2 (8 %)	4 (8 %)
Höchster Bildungsabschluss			
Lehrabschluss	2 (8 %)	3 (12 %)	5 (10 %)
Meisterprüfung	1 (4 %)	0 (0 %)	1 (2 %)
Matura	5 (20 %)	7 (28 %)	12 (24 %)
Bachelor	6 (24 %)	6 (24 %)	12 (24 %)
Master/Magister	11 (44 %)	9 (36 %)	20 (40 %)
Doktor	0 (0 %)	0 (0 %)	0 (0 %)
Geschlecht			
männlich	10 (40 %)	9 (36 %)	19 (38 %)
weiblich	15 (60 %)	16 (64 %)	31 (62 %)
divers	0 (0 %)	0 (0 %)	0 (0 %)
keine Angabe	0 (0 %)	0 (0 %)	0 (0 %)

Das Durchschnittsalter der Teilnehmenden der AD-Gruppe beträgt 36,8 Jahre, der Median 34 Jahre und der Modus 32 Jahre. Zwecks Übersichtlichkeit wurden zur Zusammenfassung der Altersgruppen Kategorien definiert, in die die Teilnehmenden gereiht wurden. Knapp mehr als die Hälfte der Teilnehmenden der AD-Gruppe (13 Personen bzw. 52 %) fällt in die Kategorie der 30- bis 39-Jährigen. Sieben Personen sind älter als 40 Jahre (insgesamt 28 % der AD-Gruppe) und 5 Personen (20 %) fallen in die Gruppe der 18- bis 29-Jährigen. Die jüngste Person unter den Teilnehmenden der AD-Gruppe hat ein Alter von 26 Jahren angegeben und die beiden ältesten ein Alter von 62.

Fast die Hälfte aller Teilnehmenden dieser Gruppe (11 Personen bzw. 44 %) verfügt über einen Magister- oder Masterabschluss. Zusammen mit den Personen, die über einen Bachelorabschluss verfügen, zeigt sich folglich mit rund zwei Drittel der Teilnehmenden (17 Personen bzw. 68 %) ein hoher Akademikeranteil. 60 % identifizieren sich als weiblich und 40 % als männlich. Unter den Teilnehmenden der AD-Gruppe befindet sich ferner die gruppenübergreifend einzige Personen, die eine Meisterprüfung absolviert hat.

Das Durchschnittsalter der Teilnehmenden der MD-Gruppe beträgt 35,4 Jahre, der Median 31 Jahre und der Modus 27 Jahre. Auch in dieser Gruppe entfallen 52 % (13 Personen) auf die Kategorie der 30- bis 39-Jährigen. 28 % (7 Personen) entfallen auf die 18- bis 29-Jährigen und 5 Personen (20 %) auf die verbleibenden drei gestaffelten Kategorien der Personen in einem Alter von 40 Jahren oder mehr. Die beiden jüngsten Teilnehmenden haben ein Alter von 26 angegeben, und die beiden ältesten – exakt wie in der AD-Gruppe – ein Alter von 62.

Auch in der MD-Gruppe verfügt der Großteil (15 Personen bzw. 60 %) über einen Hochschulabschluss, wovon 6 (24 % der MD-Gruppe) über einen Bachelorabschluss und 9 (36 % der MD-Gruppe) über einen Master- oder Magisterabschluss verfügen. Unter den verbleibenden 10 Personen haben 7 (28 %) einen Maturaabschluss und 3 (12 %) einen Lehrabschluss. 36 % der Teilnehmenden dieser Gruppe identifizieren sich als männlich und 64 % als weiblich.

Diesen Zahlen entsprechend sind unter den Teilnehmenden gruppenübergreifend in vielerlei Hinsicht Ähnlichkeiten und Parallelen zu erkennen. Sowohl im Hinblick auf die Geschlechteraufteilung, das Bildungsniveau als auch die Altersstruktur sind die Teilnehmenden auf die betreffenden Kategorien ähnlich aufgeteilt. Über beide Gruppen hinweg befinden sich je 13 Personen (52 %) in der Kategorie der 30- bis 39-Jährigen.

Insgesamt ist etwa ein Viertel (24 %) aller Personen älter als 40 (7 Personen bei AD, 5 bei MD) und 14 % sind über 50 (3 Personen bei AD, 4 bei MD). Allgemein ist in der Stichprobe ein hohes Bildungsniveau zu erkennen. 88 % aller Teilnehmenden verfügen über einen Maturaabschluss und 64 % der Gesamtmenge darüber hinaus auch über einen Hochschulabschluss – 24 % aller Teilnehmenden können einen Bachelor- und 40 % einen Master- beziehungsweise Magisterabschluss vorweisen. Insgesamt verfügen 6 Personen (12 % der gesamten Stichprobe) über keinen Maturaabschluss, wovon je drei auf eine der zwei Gruppen entfallen – zwei Lehrabschlüsse und eine Meisterprüfung in der AD-Gruppe sowie 3 Lehrabschlüsse in der MD-Gruppe. Über beide Gruppen übergreifend identifizieren sich 62 % der Teilnehmenden als weiblich und 38 % als männlich. Keine der teilnehmenden Personen verfügt über einen Dokortitel. Kein*e Teilnehmer*in hat sich im Hinblick auf das Geschlecht als divers identifiziert oder die Angabe hierzu verweigert.

Im Hinblick auf die Angaben der Teilnehmenden über den Umfang an Erfahrung, den sie im Bereich des Simultandolmetschens gesammelt haben, lassen sich ebenso große Ähnlichkeiten zwischen den Gruppen erkennen, wie aus der nachstehenden Abbildung hervorgeht.

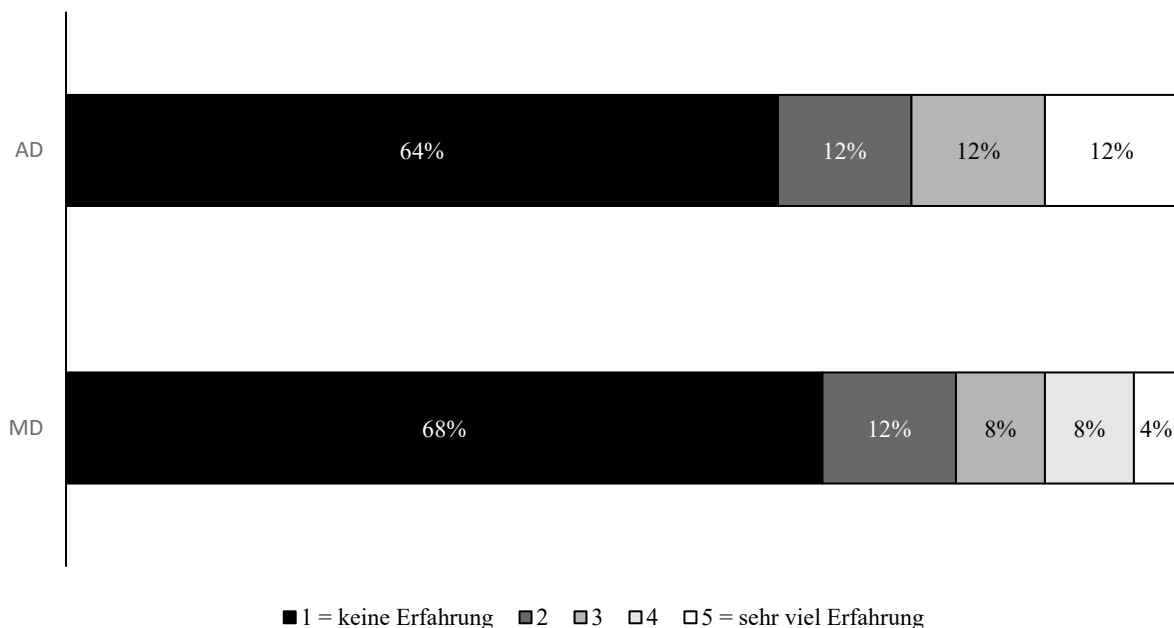


Abbildung 1: Angabe der Erfahrung mit Simultandolmetschen in Prozent

Über beide Gruppen hinweg zeigt sich, dass der Großteil der Teilnehmer*innen – 16 der AD-Gruppe und 17 der MD-Gruppe – in dieser Hinsicht über keinerlei Erfahrungen verfügt. Unter den Teilnehmenden der AD-Gruppe haben je 3 Personen die Antwortmöglichkeiten 2, 3 und 5 gewählt. Dahingegen haben in der MD-Gruppe 3 Personen die Antwortmöglichkeit 2, je 2 Personen haben die Antwortmöglichkeiten 3 und 4 und eine Person die Antwortmöglichkeit 5 gewählt. Der von den Teilnehmenden gesammelte Erfahrungsschatz im Bereich Simultandolmetschen ist somit über beide Gruppen hinweg mit leichten Variationen vergleichbar. Insgesamt befinden sich in der Stichprobe gruppenübergreifend 33 Personen, die über keine Erfahrungen verfügen, und 4 Personen, die über sehr viel Erfahrung verfügen.

Bevor im Folgenden näher auf die Qualitätskriterien und die diesbezüglich erhaltenen Antworten der Teilnehmenden auf die betreffenden Likert-Items und die offenen Fragen eingegangen wird, soll nun zuerst die grundsätzliche Frage behandelt werden, wie die Teilnehmenden eingeschätzt haben, ob es sich ihrer Ansicht nach bei der ihnen jeweils vorgelegten Dolmetschleistung um eine maschinell erzeugte Dolmetschung handelt. Dies soll als Referenz und Rahmen für die Betrachtung der in der Folge angegebenen und analysierten Daten dienen. Die entsprechenden Ergebnisse sind in der nachstehenden Abbildung 2 festgehalten.

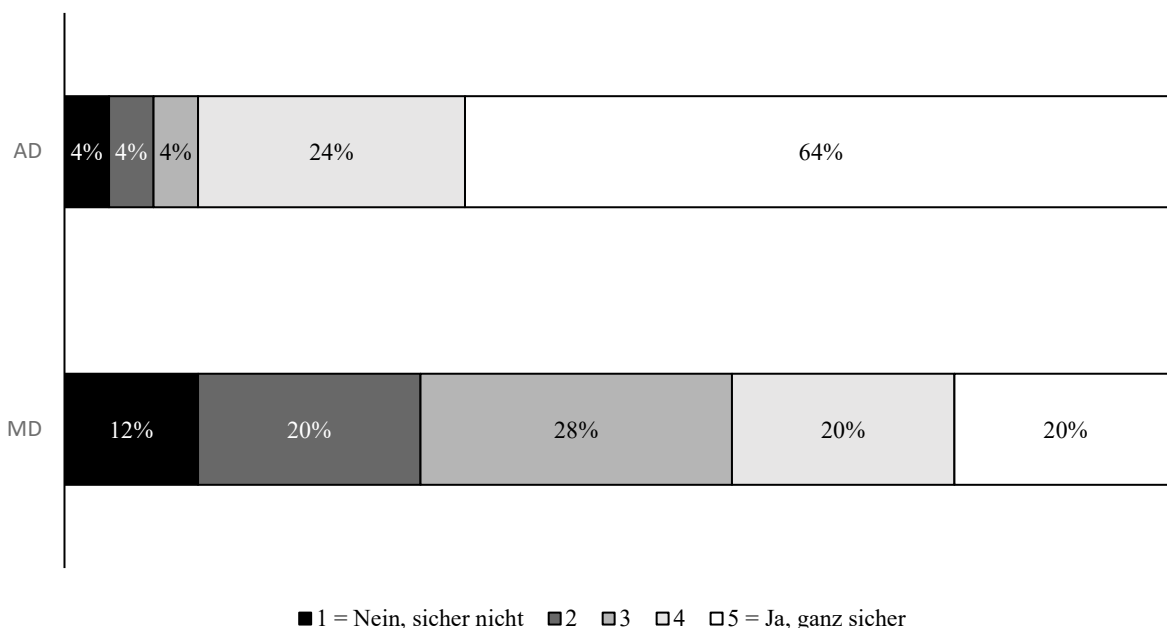


Abbildung 2: Einschätzung, ob Dolmetschung maschinell generiert wurde, in Prozent

Die Mehrheit der Rezipient*innen der maschinell erzeugten Dolmetschung (16 Personen bzw. 64 % der AD-Gruppe) waren sich sehr sicher, dass die Dolmetschung von einer Maschine erzeugt worden ist. 6 Personen (24 %) haben Antwortmöglichkeit 4 gewählt und der Leistung somit auch eher einen maschinellen Ursprung beigemessen. Die verbleibenden drei Antwortmöglichkeiten wurden je von einer Person ausgewählt. Demzufolge war sich ein*e Rezipient*in der maschinell erzeugten Dolmetschung gänzlich sicher, dass die rezipierte Leistung keinen maschinellen Ursprung hat.

Dahingegen zeigt sich unter den Teilnehmenden der MD-Gruppe ein grundlegend anderes Bild. Wie Abbildung 2 zu entnehmen ist, sind die Einschätzungen der Teilnehmenden dieser Gruppe wesentlich gleichmäßiger auf die Möglichkeiten aufgeteilt als die der anderen Gruppe. Mit 7 Personen (28 %) entschieden sich die meisten Teilnehmenden für die mittlere Antwortmöglichkeit. Für Antwortmöglichkeiten 2, 3 und 5 entschieden sich je 5 Personen (20

%) und 3 Personen für Antwortmöglichkeit 1. Die festgestellte Aufteilung der Antworten deutet darauf hin, dass sich die Personen allgemeinhin unsicher waren, ob die ihnen vorgelegte Dolmetschleistung menschlichen oder maschinellen Ursprungs ist. Hervorzuheben ist in diesem Zusammenhang insbesondere, dass die Zahl der Teilnehmenden der MD-Gruppe, die mit der größtmöglichen Sicherheit angegeben haben, dass es sich bei der Dolmetschung menschlichen Ursprungs tatsächlich um eine maschinell erzeugte Leistung handelt, die Zahl jener Personen übersteigt, die sich ganz sicher waren, dass diese Dolmetschung nicht maschinell erzeugt wurde.

4.2 Wahrnehmung der maschinell erzeugten Dolmetschung

Nachdem nun die demographischen Angaben in Kontext gesetzt und durch Darlegung der Angaben zur gesammelten Erfahrung im Bereich des Simultandolmetschens und zur Einschätzung des Ursprungs der rezipierten Dolmetschleistung ein Rahmen etabliert wurde, soll im Folgenden näher auf die Antworten der Teilnehmenden der AD-Gruppe eingegangen werden. Zunächst werden die Antworten auf die offenen Fragen betrachtet, da sie zwecks Verständnisses und Einordnung der Antwortentscheidungen der Rezipient*innen mit den resultierenden Antwortmustern bei den Likert-Items in Relation gesetzt werden müssen. Dazu werden die von den Teilnehmenden der AD-Gruppe erhaltenen Antworten auf die offenen Fragen nach Möglichkeit gesammelt auf die geschlossenen Likert-Items rückbezogen, um die heterogenen, nicht standardisierten Eingaben nachvollziehbar zu machen.

Individuelle positive Einschätzungen wurden mit der Frage „Gab es etwas, das Sie an der Dolmetschung besonders gut fanden?“ und negative Einschätzungen mit der Frage „Gab es etwas, das Sie an der Dolmetschung besonders schlecht fanden?“ erfragt. Die Antworten einiger Proband*innen, die grundsätzlich mit einer schlichten Verneinung gleichzusetzen sind – es wurde wiederholt mit einem bloßen „Nein“ oder funktional Gleichwertigem geantwortet – werden gemeinsam mit den leer gelassenen Feldern bei beiden Befragungsgruppen grundsätzlich so ausgelegt, dass in der Wahrnehmung der betreffenden Rezipient*in jeweils keine nennenswerten negativen beziehungsweise positiven Faktoren zu erkennen waren, die über die bereits erfragten Kategorien hinausgehen. Da den Fragen jeweils bewusst ein offener Rahmen gegeben wurde, konnten die Rezipient*innen nicht nur eine, sondern auch mehrere Beobachtungen in diesem Feld teilen. Wird in einer Antwort auf mehrere Aspekte Bezug genommen, so werden die jeweiligen Aussagen dementsprechend individuell pro ermitteltem Aspekt berücksichtigt. Allgemeine Muster sowie jene Antworten, bei denen keine Entsprechung ermittelt werden konnte, werden im Hinblick auf die AD-Gruppe unmittelbar nachstehend beleuchtet.

17 Personen der AD-Gruppe ließen das Antwortfeld zur positiv formulierten Frage leer oder verneinten die Frage und 8 Personen versahen das Feld mit einer Antwort. Bei den Antworten auf die negativ formulierte Frage waren 16 Antworten und 9 funktionale Verneinungen zu verzeichnen. Wenngleich die Anzahl an erkannten negativen Auffälligkeiten bereits in dieser Hinsicht deutlich überwiegt, verschiebt sich dieses Verhältnis durch die Anzahl der diesen Antworten entnommenen individuell genannten Aussagen noch weiter ins Negative. Positive Rückmeldungen gab es je eine im Hinblick auf die Flüssigkeit, das Sprechtempo, die sprachliche Korrektheit und die Fähigkeit, der Dolmetschung inhaltlich folgen zu können. Die Stimme wurde dreimal positiv erwähnt. Eine Antwort enthielt eine positiv anmutende Vermutung im Hinblick auf eine mögliche attraktive Preisgestaltung für Kunden, die in der vorliegenden Arbeit nicht weiter berücksichtigt wird.

Während den positiven Antworten je eine einzige Aussage entnommen werden konnte, wurde in den negativ formulierten Aussagen häufig mehr als ein Aspekt zur Sprache gebracht. Einmal negativ erwähnt wurde die Korrektheit der Dolmetschung, je zweimal ihre Flüssigkeit und Kohärenz, je dreimal die Natürlichkeit und die Stimme, viermal die – mangelhaft ausgeprägte – Möglichkeit, der Dolmetschung inhaltlich folgen zu können, und je siebenmal das Tempo und das Pausenverhalten. Die Aussage, die Dolmetschung wirkte sehr maschinell, konnte keiner Kategorie eindeutig zugewiesen werden. Dem Verweis auf ein komisches Bauchgefühl, in dessen Rahmen die Dolmetschung als unrealistisch und unglaubwürdig bezeichnet wurde, kann ebenso kein eindeutig identifizierbarer Bezug auf eine der herangezogenen Kategorien entnommen werden, jedoch ein grundsätzlich negatives Urteil über die Gesamtqualität und Vertrauenswürdigkeit. Jene Antworten der AD-Gruppe, die mit den Likert-Items funktional gleichgesetzt werden können, werden nachstehend bei deren Betrachtung in größerem Detail dargelegt. Alle Antworten der AD-Gruppe auf die offenen Fragen sind in Anhang 8 enthalten.

Die nachstehenden Tabellen 2 und 3 bieten eine Übersicht über die Antworten der AD-Gruppe auf die zwölf erfragten Likert-Items. Im Fragebogen selbst war die Aussage „Die Dolmetschung war sprachlich korrekt“ zur besseren Verständlichkeit mit dem Hinweis „(Grammatik, Wortwahl, Satzbau)“ versehen, der nachstehend zwecks Übersichtlichkeit entfernt wurde. Angegeben werden die Ergebnisse der Befragung anhand der Parameter Mittelwert (M), Standardabweichung (SD), Median (Mdn) und Modus (Mo) (siehe Kapitel 3.6). In den betreffenden Tabellen werden die Antwortmöglichkeiten zwecks leichter Nachvollziehbarkeit des Befra-

gungshergangs in derselben Reihenfolge angegeben wie im Erhebungsinstrument. Eine vollständige Übersicht aller erfassten Informationen, deren Angabe an dieser Stelle eine sinnvolle und nachvollziehbare Betrachtung erschweren würde und daher nicht zweckdienlich erscheint, ist in den Anhängen 5 bis 8 enthalten. Wie in Kapitel 3.4 erläutert wurde, entspricht die Antwortmöglichkeit 5 der größtmöglichen Zustimmung zur betreffenden Aussage und Antwortmöglichkeit 1 der geringstmöglichen Zustimmung.

Tabelle 2: Zusammenfassung der Antworten der AD-Gruppe in Zahlen

AD	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>Mo</i>
Ich konnte der Dolmetschung inhaltlich gut folgen.	3,24	1,05	3	3/4
Die Dolmetschung war kohärent.	2,68	1,07	2	2
Die Dolmetschung war sprachlich korrekt.	3,36	1,22	3	3/4
Die Vortragsweise der Dolmetschung war flüssig.	2,24	1,13	2	2
Die Formulierungen wirkten natürlich.	1,92	1,11	2	1
Die Stimme war angenehm zu hören.	2,36	1,15	2	2
Pausen/Verzögerungen haben mich beim Zuhören gestört.	4,16	1,34	5	5
Das Sprechtempo war angenehm.	2,48	1,08	2	2
Es war anstrengend, der Dolmetschung zu folgen.	3,68	1,22	4	4
Die Dolmetschung wirkte auf mich professionell.	2,24	1,30	2	1
Die Dolmetschung wirkte auf mich vertrauenswürdig.	2,40	1,19	2	2
Die Dolmetschung war insgesamt von hoher Qualität.	2,16	1,11	2	2

Zur besseren Veranschaulichung wurden die Befragungsergebnisse ferner in der folgenden Tabelle im Detail nach der jeweiligen Verteilung auf die verfügbaren Antwortmöglichkeiten dargestellt. Die von 1 bis 5 reichenden Spaltenüberschriften entsprechen den Antwortmöglichkeiten (1 = „Ich stimme gar nicht zu“; „5“ = „Ich stimme voll und ganz zu“). Sie werden als prozentuelle Anteile an den Antworten der AD-Gruppe angegeben.

Tabelle 3: Verteilung der Antworten der AD-Gruppe in Prozent

AD	Verteilung (in %)				
	1	2	3	4	5
Ich konnte der Dolmetschung inhaltlich gut folgen.	8	12	36	36	8
Die Dolmetschung war kohärent.	8	44	28	12	8
Die Dolmetschung war sprachlich korrekt.	8	16	28	28	20
Die Vortragsweise der Dolmetschung war flüssig.	28	40	16	12	4
Die Formulierungen wirkten natürlich.	44	36	8	8	4
Die Stimme war angenehm zu hören.	24	36	28	4	8
Pausen/Verzögerungen haben mich beim Zuhören gestört.	12	0	8	20	60
Das Sprechtempo war angenehm.	16	40	32	4	8
Es war anstrengend, der Dolmetschung zu folgen.	8	8	20	36	28
Die Dolmetschung wirkte auf mich professionell.	36	32	12	12	8
Die Dolmetschung wirkte auf mich vertrauenswürdig.	24	36	24	8	8
Die Dolmetschung war insgesamt von hoher Qualität.	28	48	8	12	4

Da im Zentrum der vorliegenden Arbeit der insgesamt gewonnene Eindruck der zugrunde liegenden Qualität der Dolmetschleistungen liegt, wird im Folgenden von der Reihenfolge des Erhebungsinstruments abgewichen. Zuallererst wird unter Nutzung der erfassten Parameter der Gesamtqualität, Professionalität und Vertrauenswürdigkeit der zentrale Aspekt der allgemein erkannten Qualitätswahrnehmung beleuchtet, um den folgenden Ausführungen einen angemessenen Rahmen und eine faktenbasierte Grundlage zu geben.

Unmittelbar an den Antworten auf das darauf abzielende Likert-Item gemessen wurde die maschinell erzeugte Dolmetschleistung insgesamt tendenziell negativ bewertet. Die wahrgenommene Gesamtqualität liegt demnach im Durchschnitt bei 2,16 (SD = 1,11) und der Median sowie der Modus jeweils bei 2. Beinahe die Hälfte aller Teilnehmenden dieser Gruppe (12 Personen bzw. 48 %) stellten der Dolmetschung mit einer Bewertung von 2 ein eher negatives Urteil aus. Eine Person wählte jedoch die bestmögliche Antwortmöglichkeit aus und schrieb der rezipierten Dolmetschung somit die höchstmögliche Qualität zu. Etwas mehr als ein Viertel aller Teilnehmenden der AD-Gruppe gaben der maschinell erzeugten Leistung im Hinblick auf die erzielte Qualität die negativste Bewertung.

Geringfügig positiver fielen die Ergebnisse im Hinblick auf die wahrgenommene Vertrauenswürdigkeit (M = 2,40; SD = 1,19; Mdn = 2; Mo = 2) und Professionalität (M = 2,24; SD = 1,30; Mdn = 2; Mo = 1) aus. Dementsprechend wurde bei der Professionalität die negativste Auswahlmöglichkeit am häufigsten von den Teilnehmenden gewählt. Je zwei Personen

stellten mit Blick auf Professionalität und Vertrauenswürdigkeit das bestmögliche Urteil aus. Der Standardabweichung und der dargestellten Verteilung ist bei diesen drei Aspekten zu entnehmen, dass die Bewertungen zwischen den Teilnehmenden durchaus in erheblichem Maße variieren. Im Gesamteindruck ist die maschinell erzeugte Dolmetschleistung im Lichte der Professionalität und Vertrauenswürdigkeit von den Rezipient*innen eher negativ bewertet worden.

Anschließend an den Gesamteindruck, die Professionalität und die allgemeine Vertrauenswürdigkeit werden nun die vorwiegend die Darbietung und Präsentationsebene betreffenden Aspekte beleuchtet. Die Flüssigkeit der rezipierten Dolmetschleistung wurde dem Gesamteindruck ähnlich negativ wahrgenommen. Im Schnitt wurde dieser Punkt mit einer Bewertung von 2,24 ($SD = 1,13$) versehen. Vergleichbar negativ fielen Modus und Median mit einem Wert von jeweils 2 aus. Eine Person stellte der maschinell generierten Dolmetschleistung jedoch das bestmögliche Urteil hinsichtlich ihrer Flüssigkeit aus. Bei den offenen Fragen wurde dieser Punkt je einmal positiv und einmal negativ angemerkt. Positiv vermerkt wurde das Fehlen von Hesitationen und bemängelt wurde die abgehackte Sprechweise.

Als eine der negativsten unter den Kategorien wurde mit einer durchschnittlichen Bewertung von 1,92 ($SD = 1,12$; $Mdn = 2$) die Natürlichkeit der Dolmetschung beurteilt. Fast die Hälfte (11 Personen bzw. 44 %) der Teilnehmenden der AD-Gruppe stellten der Natürlichkeit die schlechteste Note aus, wodurch der Modus folglich ebenfalls 1 beträgt. Insgesamt wurde die maschinelle Leistung somit als überaus unnatürlich aufgefasst. Im Rahmen der drei diesbezüglich erhaltenen Antworten auf die offenen Fragen wurden im Hinblick auf die Natürlichkeit unter anderem die unnatürliche Sprachmelodie, die Formulierung der Sätze und die ungewöhnliche Aussprache bemängelt. Positive Äußerungen wurden in diesem Zusammenhang keine getätigt.

Der ebenso primär die Präsentationsebene betreffende Aspekt der Stimme wurde in vergleichbarem Maße negativ aufgenommen und erhielt im Schnitt einen Wert von 2,36 ($SD = 1,15$). Nur zwei Personen (8 %) nahmen die Stimme sehr positiv auf und eine (4 %) eher positiv. Der Großteil der Teilnehmenden nahm diesen Aspekt jedoch negativ wahr – 6 Personen (24 %) sehr negativ und 9 Personen (36 %) eher negativ. Bemerkenswerterweise wurde dieser Aspekt im Zuge der Beantwortung der offenen Fragen dreimal positiv zur Sprache gebracht. Dabei wurde die Stimme als angenehm, ruhig beziehungsweise deutlich bezeichnet. Bei den drei Antworten auf die offene Frage nach negativen Aspekten wurde die Stimme jeweils ohne Nennung von Einzelheiten bemängelt.

Das Sprechtempo wurde mit einem Mittelwert von 2,48 ($SD = 1,08$; $Mdn = 2$; $Mo = 2$) ebenso überwiegend negativ bewertet. Obwohl auch dieser Aspekt von zwei Personen sehr positiv und einer Person eher positiv bewertet wurde, wurde am häufigsten (10 Personen bzw. 40 %) eine eher negative Bewertung abgegeben. Knapp ein Drittel der Proband*innen der MD-Gruppe (8 Personen bzw. 32 %) entschieden sich bei ihrer Antwort für die neutrale Mitte. Einer positiven Erwähnung bei den offenen Fragen stehen sieben negative entgegen. Einmal wurde das Tempo explizit als zu schnell bezeichnet und sechsmal wurde unmittelbar auf die Variation im Sprechtempo als problematischen Faktor verwiesen. Dies hat es aus Sicht zweier Rezipient*innen ferner erschwert, der Dolmetschung zu folgen.

Als der größte Kritikpunkt der Proband*innen hat sich das Pausenverhalten herausgestellt. Wie auch die Aussage zur empfundenen Anstrengung ist dieser Punkt hinsichtlich der Zustimmungsskala relativ zu den anderen Punkten invers – je höher der abgegebene Wert, desto negativer die Beurteilung. Der Großteil aller Teilnehmenden (15 Personen bzw. 60 %) stimmten voll und ganz zu, dass das Pausenverhalten beziehungsweise die Verzögerungen sie beim Zuhören störten, wodurch folglich auch Median und Modus bei der negativsten Zahl 5 liegen. Durch Inversion der erhaltenen Angaben ($M = 4,16$; $SD = 1,34$; $Mdn = 5$; $Mo = 5$) können mit den anderen Angaben vergleichbarere Werte ($M = 1,84$; $SD = 1,34$; $Mdn = 1$; $Mo = 1$) erhalten werden. Eine entsprechende Darstellung, in der die invertierten Werte aufgenommen wurden, ist Anhang 4 zu entnehmen. Dieses Bild wird durch diverse Rückmeldungen bei den offenen Fragen in Hinblick auf das Pausenverhalten bestätigt, das dort – wie das Sprechtempo – siebenmal bemängelt wurde. Dabei wurden die Pausen nicht nur allgemein, sondern im Speziellen auch in Bezug auf deren Länge und Platzierung als störend genannt. Verzögerungen wurden dabei einmal explizit als negativer Faktor erwähnt. Beachtenswert ist, dass drei Personen (12 %) diesem Urteil widersprechen und bei ihrer Beantwortung des Likert-Items angaben, dass sie diesen Aspekt überhaupt nicht als störend aufgefasst haben. Die relativ hohe Standardabweichung von 1,32 ist unter anderem diesem Umstand geschuldet. Keine Person beantwortete die offenen Fragen mit einem positiven Bezug in dieser Hinsicht.

Die bei der Rezeption empfundene Anstrengung wurde von den Rezipient*innen mit einem Durchschnitt von 3,68 ($SD = 1,22$) bewertet. Sowohl Median als auch Modus liegen bei dieser Kategorie bei 4. Der zur Vergleichbarkeit invertierte Durchschnittswert beträgt 2,32 ($Mdn = 2$; $Mo = 2$). Je 2 Personen (je 8 %) gaben an, dass die Rezeption für sie mit geringen beziehungsweise keinen Anstrengungen verbunden war. Kein*e Teilnehmer*in nahm bei der

Beantwortung der offenen Fragen auf diesen Aspekt Bezug. Im Resultat wurde die bei der Rezeption wahrgenommene Anstrengung allgemein hin eher hoch eingeschätzt.

Im Hinblick darauf, wie einfach oder schwierig es war, der Dolmetschung inhaltlich zu folgen, zeigt sich ein ausgewogenes und neutrales Bild. Der Mittelwert beträgt hierbei 3,24 (SD = 1,05). Jeweils 9 Personen (36 %) wählten die Antwortmöglichkeiten 3 und 4 und jeweils 2 Personen (je 8 %) stellten diesem Aspekt das bestmögliche beziehungsweise schlechtmögliche Urteil aus. Bei den Antworten auf die offenen Fragen kam diese Kategorie einmal im Hinblick auf die empfundene Klarheit positiv zur Sprache sowie viermal negativ. Dabei verwiesen die betreffenden Rezipient*innen auf die eingesetzten Pausen, die abgehackte Sprechweise, die Geschwindigkeitsänderungen und die ungewöhnliche Aussprache als erschwerende Faktoren.

Die maschinell erzeugte Dolmetschleistung wurde von den Rezipient*innen im Allgemeinen mit einem Durchschnittswert von 2,68 (SD = 1,07; Mdn = 2; Mo = 2) als eher nicht kohärent wahrgenommen. Knapp die Hälfte der Rezipient*innen (11 Personen bzw. 44 %) bewertete die Kohärenz eher negativ. Je zwei Personen (je 8 %) maßen diesem Aspekt die bestmögliche beziehungsweise die schlechtmögliche Ausprägung bei. Im Rahmen der zwei negativen Rückmeldungen zu den offenen Fragen wurde unter anderem darauf verwiesen, dass die Dolmetschung nicht immer Sinn ergeben habe.

Hinsichtlich der wahrgenommenen sprachlichen Korrektheit zeigt sich ein ähnliches Bild. Mit einer durchschnittlichen Bewertung von 3,36 wurde dieser Aspekt insgesamt weder als besonders positiv noch als besonders negativ angesehen. Der vergleichsweise hohen Standardabweichung (1,22) entsprechend wurden der sprachlichen Korrektheit mitunter auch klarer polarisierte Urteile ausgesprochen. 5 Personen (20 %) haben voll und ganz zugestimmt, dass die Dolmetschung im Hinblick auf Grammatik, Wortwahl und Satzbau korrekt war, während 2 Personen (8 %) dieser Aussage vollkommen widersprochen haben. Im Rahmen der offenen Fragen wurde dieser Aspekt einmal mit dem Hinweis, die Terminologie sei naheliegend, vorsichtig positiv genannt und einmal unter Verweis auf die unvollständigen Sätze als negativ.

Diesen Ausführungen kann entnommen werden, dass die maschinell erzeugte Dolmetschleistung in den erfragten Kategorien insgesamt recht negativ wahrgenommen wurde. Insbesondere die Natürlichkeit, das Pausenverhalten und das unmittelbar abgefragte Gesamturteil wurden von den Rezipient*innen klar negativ beurteilt. Den Standardabweichungen zufolge waren die Ergebnisse dabei jeweils recht heterogen, wodurch die erfassten Durchschnittswerte

nicht ohne Weiteres auf die gesamte Stichprobe generalisiert werden können. Dabei hat eine nicht unerhebliche Anzahl an Personen häufig gänzlich unterschiedliche Meinungen vertreten. Die sprachliche Korrektheit wurde zwar relativ zu den anderen Kategorien am positivsten beurteilt, hat jedoch mit einem Mittelwert von 3,36 trotzdem ein lediglich in der Mitte der Skala zu verortendes Ergebnis erzielt. Im Hinblick darauf, wie gut die Proband*innen der Dolmetschung folgen konnten, wurde relativ zu den anderen Faktoren ebenso ein eher positives, jedoch in absoluten Zahlen im Schnitt ($M = 3,24$) ein gleichermaßen durchzogenes Urteil ausgestellt. Flüssigkeit, Tempo, Stimme, Natürlichkeit, Professionalität und Vertrauenswürdigkeit wurden allesamt eher negativ bewertet. Relativ neutral wurde neben der bei der Rezeption empfundenen Anstrengung auch der Aspekt gesehen, wie leicht es ist, der Dolmetschung inhaltlich zu folgen. Anschließend an die Beleuchtung der Resultate der AD-Gruppe werden im Folgenden die Antworten der MD-Gruppe analog dazu und in ähnlicher Manier dargestellt und behandelt.

4.3 Wahrnehmung der von einem Menschen angefertigten Dolmetschung

Wie bereits im Rahmen der Darbietung der Wahrnehmung der AD-Gruppe wird zunächst auf die Antworten auf die offenen Fragen eingegangen. Im Hinblick auf die offenen Fragen ist vorab festzuhalten, dass die Proband*innen in der MD-Gruppe mehrmals vom vorgeschlagenen, bewusst jeweils positiv beziehungsweise negativ formulierten Rahmen der beiden Fragen abgewichen sind. Beispielsweise wurde bei der Frage nach positiven Aspekten einmal das stockende Tempo bemängelt und einmal die missverständliche und – aufgrund des eigentlich positiv formulierten Rahmens – nicht eindeutig auszulegende Aussage getätigt: „Nein, wirkte wie KI“. Auch bei der negativ formulierten offenen Frage wurde von einer Teilnehmer*in die Vermutung geäußert, dass es sich bei der von einer professionellen Dolmetscherin angefertigten Leistung womöglich um ein KI-generiertes Produkt handeln könnte.

Die positiv formulierte offene Frage wurde von 6 Personen mit einer gehaltvollen Antwort versehen und 19 Personen gaben entweder keine Antwort oder eine einer Verneinung gleichwertige Antwort. Die negativ formulierte offene Frage wurde wiederum von 14 Personen beantwortet und 11 Personen haben das entsprechende Feld unausgefüllt gelassen. Als positive Aspekte wurden je zweimal das Tempo, die Stimme und die sprachliche Korrektheit angeführt. Als negative Faktoren vermerkt wurden einmal die Stimme, dreimal das Sprechtempo, fünfmal die Flüssigkeit und sechsmal das Pausenverhalten. Lediglich einmal wurde eine Antwort abgegeben, die weder mit der gestellten Frage noch mit dem Forschungsgegenstand der vorliegenden Arbeit inhaltliche Überschneidungen aufweist und folglich außer Acht gelassen wird (für eine vollständige Auflistung aller Antworten auf die offenen Fragen siehe Anhänge 7 und 8).

Die genannten von den Teilnehmenden der MD-Gruppe abgegebenen Antworten, die mit den Likert-Items funktional gleichgesetzt werden können, werden nachstehend im Rahmen der Betrachtung der betreffenden Kategorien näher erläutert. Analog zu der im Hinblick auf die AD-Gruppe gebotenen Übersicht ist auch für die Resultate der MD-Gruppe im Folgenden eine zusammenfassende Tabelle unter Angabe derselben Parameter und in derselben Reihenfolge der Items angegeben, aus der die wichtigsten Informationen zu den Ergebnissen dieser Gruppe entnommen werden können.

Tabelle 4: Zusammenfassung der Antworten der MD-Gruppe in Zahlen

MD	M	SD	Mdn	Mo
Ich konnte der Dolmetschung inhaltlich gut folgen.	4,00	0,87	4	4
Die Dolmetschung war kohärent.	3,60	0,91	4	4
Die Dolmetschung war sprachlich korrekt.	4,12	0,73	4	4
Die Vortragsweise der Dolmetschung war flüssig.	2,24	0,97	2	2
Die Formulierungen wirkten natürlich.	2,76	1,23	3	3
Die Stimme war angenehm zu hören.	2,84	1,07	3	3
Pausen/Verzögerungen haben mich beim Zuhören gestört.	4,04	1,06	4	5
Das Sprechtempo war angenehm.	2,84	1,03	3	3
Es war anstrengend, der Dolmetschung zu folgen.	3,28	1,34	4	4
Die Dolmetschung wirkte auf mich professionell.	2,96	1,17	3	2
Die Dolmetschung wirkte auf mich vertrauenswürdig.	3,16	1,18	3	3
Die Dolmetschung war insgesamt von hoher Qualität.	2,88	1,01	3	3

Zwecks Veranschaulichung werden die Ergebnisse in der nachstehenden Tabelle auch nach der Verteilung der gewählten Antworten dargestellt. Die Verteilung wird als prozentueller Anteil an den Antworten der MD-Gruppe angegeben.

Tabelle 5: Verteilung der Antworten der MD-Gruppe in Prozent

MD	Verteilung (in %)				
	1	2	3	4	5
Ich konnte der Dolmetschung inhaltlich gut folgen.	0	4	24	40	32
Die Dolmetschung war kohärent.	0	12	32	40	16
Die Dolmetschung war sprachlich korrekt.	0	4	8	60	28
Die Vortragsweise der Dolmetschung war flüssig.	24	40	24	12	0
Die Formulierungen wirkten natürlich.	20	20	32	20	8
Die Stimme war angenehm zu hören.	16	16	36	32	0
Pausen/Verzögerungen haben mich beim Zuhören gestört.	0	12	16	28	44
Das Sprechtempo war angenehm.	8	28	44	12	8
Es war anstrengend, der Dolmetschung zu folgen.	12	20	16	32	20
Die Dolmetschung wirkte auf mich professionell.	8	32	28	20	12
Die Dolmetschung wirkte auf mich vertrauenswürdig.	4	28	36	12	20
Die Dolmetschung war insgesamt von hoher Qualität.	12	20	36	32	0

Im Gesamteindruck wurde die von einer professionellen Dolmetscherin angefertigte Dolmetschleistung von den Rezipient*innen überwiegend im mittleren Bereich der Skala bewertet. Die Bewertung der Gesamtqualität liegt im Schnitt bei 2,88 (SD = 1,01; Mo = 3), und auch bei den Urteilen zur Vertrauenswürdigkeit (M = 3,16; SD = 1,18; Mo = 3) und Professionalität (M = 2,96; SD = 1,17; Mo = 2) zeigen die Zahlen ein ähnliches Bild. Der bei den beiden letztgenannten Parametern erfassten Standardabweichung zufolge, sind die Bewertungen nicht gänzlich homogen ausgefallen, sondern variierten zwischen den Teilnehmenden in nicht unerheblichem Maße. Bei allen drei Aussagen wurde im Median mit der Antwortmöglichkeit 3 die Mitte gewählt. Bemerkenswerterweise gab keine einzige Person der MD-Gruppe an, voll und ganz zuzustimmen, dass die von einem Menschen angefertigte Dolmetschung ihres Erachtens insgesamt von hoher Qualität ist. Drei Personen (12 %) haben angegeben, überhaupt nicht zuzustimmen, dass ihnen eine Dolmetschung hoher Qualität vorgelegt worden sei. Bei der Beantwortung der offenen Fragen wurde weder auf den Gesamteindruck noch auf die Professionalität oder Vertrauenswürdigkeit direkt Bezug genommen. Im Resultat wurde die menschengemachte Dolmetschung unter dem Gesichtspunkt dieser drei Aspekte insgesamt weder besonders positiv noch besonderes negativ beurteilt.

Die Flüssigkeit der Vortragsweise wurde dahingegen vergleichsweise kritisch gesehen. Im Schnitt wurde dieses Item mit 2,24 (SD = 0,97) beurteilt. Median und Modus fielen mit jeweils 2 ähnlich negativ aus. Keine einzige Person bewertete die Flüssigkeit durch Auswahl

der Antwortmöglichkeit 5 als besonders positiv, und nur zwei Personen (12 %) gaben mit Auswahlmöglichkeit 4 ein eher positives Urteil ab. Sechs Personen und damit knapp ein Viertel der Teilnehmenden der MD-Gruppe nahmen die Flüssigkeit sehr negativ wahr. Im Rahmen der Beantwortung der offenen Fragen wurde fünfmal negativ auf die Flüssigkeit der dargebotenen Leistung verwiesen. Dabei wurden unter anderem die abgehackte Sprechweise, Füllwörter, Hesitationen und ein erkanntes Stottern bemängelt.

Die Natürlichkeit der Vortragsweise wurde ähnlich, jedoch etwas weniger kritisch beurteilt, wobei höhere Abweichungen zwischen den Teilnehmenden zu erkennen sind. Nur zwei Personen (8 %) schätzten die Vortragsweise der Dolmetschung als sehr flüssig ein, gegenüber fünf Personen (20 %), die deren Natürlichkeit gänzlich negativ bewerteten. Bei einer dieser Zahlen entsprechend verständlicherweise etwas höheren Standardabweichung von 1,23 wurde bei diesem Aspekt im Schnitt ein Wert von 2,76 erzielt. Modus und Median liegen jeweils bei 3, wodurch veranschaulicht wird, dass die Wahrnehmung der Natürlichkeit insgesamt durchgezogen ist. Im Zuge der offenen Fragen wurden keine Äußerungen zur Natürlichkeit getätigt.

Ähnliche Ergebnisse lieferte die Erhebung zur Einschätzung der Stimme. Der Mittelwert bei diesem Item beträgt 2,84 bei einer Standardabweichung von 1,07, wobei Median und Modus ebenfalls mit einem Wert von exakt 3 in der Mitte der Skala anzusiedeln sind. Es ist festzuhalten, dass auch der Aspekt der Stimme bei der Beantwortung des betreffenden Items von keiner einzigen Person sehr positiv bewertet wurde. Positiv angemerkt wurden jedoch die Tonalität und die Tonlage im Rahmen von zwei Antworten auf die offenen Fragen. Einmal wurde die Stimme ohne weitere Details als störend angegeben.

Wie bei den übrigen bereits dargelegten primär die Präsentationsebene betreffenden Aspekten wurden auch im Hinblick auf das Sprechtempo überwiegend in der Mitte der Skala anzusiedelnde Bewertungen abgegeben. Der Mittelwert beträgt wie bei den zur Wahrnehmung der Stimme abgegebenen Antworten exakt 2,84 und der Median und der Modus betragen wie bei den Aspekten der Natürlichkeit und der Stimme ebenso exakt 3. Auch die Standardabweichung ist mit 1,03 der bei dem Aspekt der Stimme erfassten Standardabweichung überaus ähnlich. Polarisierter sind die Antworten auf die offenen Fragen ausgefallen. Das Sprechtempo wurde zweimal als positiver Faktor angeführt und dreimal – unter Verweis auf irritierende Verzögerungen und ein unzureichendes Tempo – als negativer.

Der Aspekt der Pausen beziehungsweise Verzögerungen wurde von den Proband*innen der MD-Gruppe überaus kritisch bewertet. Im Durchschnitt wurde eine Bewertung von 4,04

(SD = 1,06) abgegeben, die nach der aufgrund der negativen Formulierung der Aussage erforderlichen Inversion einem Wert von 1,96 entspricht. Der Modus beträgt 5, womit die Rezipient*innen folglich am häufigsten angaben, dass das Pausenverhalten und die Verzögerungen als überaus störend empfunden wurden. Der Modus ist damit in der MD-Gruppe der negativste unter allen Items – der durch Inversion erhaltene Moduswert von 1 wurde bei keinem anderen Item in der MD-Gruppe erzielt. Keine Teilnehmer*in gab an, dass Pausen beziehungsweise Verzögerungen überhaupt nicht störend waren. Der Median ist mit einem Wert von 4 (2 nach Inversion) geringfügig positiver ausgefallen als der Modus. Die Pausen wurden ferner sechsmal im Rahmen der Beantwortung der offenen Fragen als störend angeführt, wobei insbesondere Hesitationen, ein vernommenes Stottern sowie Anzahl und Länge der Pausen bemängelt wurden.

Die Proband*innen konnten der menschengemachten Dolmetschung insgesamt sehr gut folgen. Mit einem Mittelwert, Median und Modus von jeweils exakt 4 und einer vergleichsweise geringen Standardabweichung von 0,87 wurde dieser Aspekt allgemein sehr positiv bewertet. Keine einzige Person gab an, der Dolmetschung inhaltlich gar nicht folgen zu können, eine Person (4 %) behauptete, der Dolmetschung eher nicht folgen zu können, und beinahe ein Drittel der Teilnehmenden (8 Personen bzw. 32 %) wählten bei diesem Aspekt den bestmöglichen Wert aus. Es wurden in dieser Hinsicht keine Äußerungen bei der Beantwortung der offenen Fragen verzeichnet.

Die Kohärenz der Dolmetschung wurde tendenziell positiv bewertet. Mit einem Mittelwert von 3,60 und einer Standardabweichung von 0,91 zählt dieser Aspekt zu den insgesamt eher als positiv wahrgenommenen Aspekten. Die positive Beurteilung wird ferner durch den Median und den Modus untermauert, die beide jeweils 4 betragen. Keine einzige Person beurteilte die Kohärenz als sehr negativ, gegenüber vier Personen (16 %), die sie als besonders positiv einstufen. Auch auf diesen Aspekt wurde bei den Antworten der offenen Fragen nicht Bezug genommen.

Der Parameter, der von den Teilnehmenden der MD-Gruppe am höchsten bewertet wurde, ist die sprachliche Korrektheit mit einem Mittelwert von 4,12 und einer bemerkenswert geringen Standardabweichung von 0,73. 60 % aller Teilnehmenden der MD-Gruppe (15 Personen) gaben an, dass die rezipierte Dolmetschleistung sprachlich eher korrekt sei. Kein*e Teilnehmer*in gab an, dass sie sprachlich gänzlich inkorrekt gewesen sei, und lediglich einer Person zufolge war die Dolmetschung sprachlich eher inkorrekt. Zwei Personen (8 %) entschieden

sich für die Mitte und sieben Personen (28 %) waren der Auffassung, dass die Dolmetschung sprachlich gänzlich korrekt gewesen sei. Entsprechend betragen Median und Modus jeweils 4. Im Zuge zweier positiver Äußerungen zu den offenen Fragen wurden Korrektheit und Genauigkeit als besonders positive Faktoren herausgestellt.

Insgesamt wurde die von einer professionellen Dolmetscherin erbrachte Leistung von den Rezipient*innen insbesondere im Hinblick auf ihre Kohärenz, sprachliche Korrektheit und in Bezug darauf positiv gesehen, wie leicht es war, der Dolmetschung inhaltlich zu folgen. Ein markant negatives Urteil konnte dahingegen bei den Aspekten der Flüssigkeit und der bei der Darbietung eingesetzten Pausen und wahrgenommenen Hesitationen festgestellt werden. Die verbleibenden Kategorien der Natürlichkeit, der Stimmqualität, des Sprechtempos, der bei der Rezeption empfundenen Anstrengung sowie der beigemessenen Professionalität und Vertrauenswürdigkeit erhielten allgemein eher durchgezogene Bewertungen. Im Rahmen des unmittelbar darauf abzielenden Items wurde die Gesamtqualität von den Teilnehmenden recht neutral, jedoch mit leichter negativer Tendenz beurteilt. Im Anschluss an die detaillierte Aufschlüsselung der Wahrnehmungen der beiden Gruppen werden die erfassten Ergebnisse im folgenden Unterkapitel zueinander in Beziehung gesetzt, um Muster aufzudecken und Rückschlüsse über die Unterschiede in der wahrgenommenen Qualität ziehen zu können.

4.4 Gruppenübergreifende Gegenüberstellung der Ergebnisse

In diesem Abschnitt werden die erfassten Ergebnisse der beiden Gruppen einem direkten Vergleich unterzogen, um etwaige Unterschiede, Regelmäßigkeiten und Muster aufzuzeigen. Dies erfolgt insbesondere im Lichte der Forschungsfrage, welche Unterschiede sich durch einen solchen Vergleich ergeben, und der Hypothese, dass sich mögliche Unterschiede bei der wahrgenommenen Qualität in erster Linie auf die Präsentationsebene betreffende Aspekte beziehen.

Zunächst wird anhand der errechneten Mittelwerte eine grundlegende Übersicht über die erfassten Ergebnisse und deren Abweichungen zwischen den Gruppen gegeben. Dabei werden die beiden negativ formulierten Aussagen – betreffend die Pausen und Verzögerungen sowie die bei der Rezeption empfundene Anstrengung (siehe Kapitel 3.4) – im vorliegenden Kapitel durchgehend funktional invertiert, wodurch bei allen Items höhere Werte einer positiven Ausprägung entsprechen und niedrige Werte einer negativen. Bei den zur Veranschaulichung erstellten Tabellen wurde die Frage nach der Einschätzung zum maschinellen Ursprung paraphrasiert, um die optische Nachvollziehbarkeit der Darstellung zu bewahren. Primär werden

die erfassten Ergebnisse anhand des Mittelwerts und dessen Differenz (ΔM) zwischen den Gruppen angegeben.

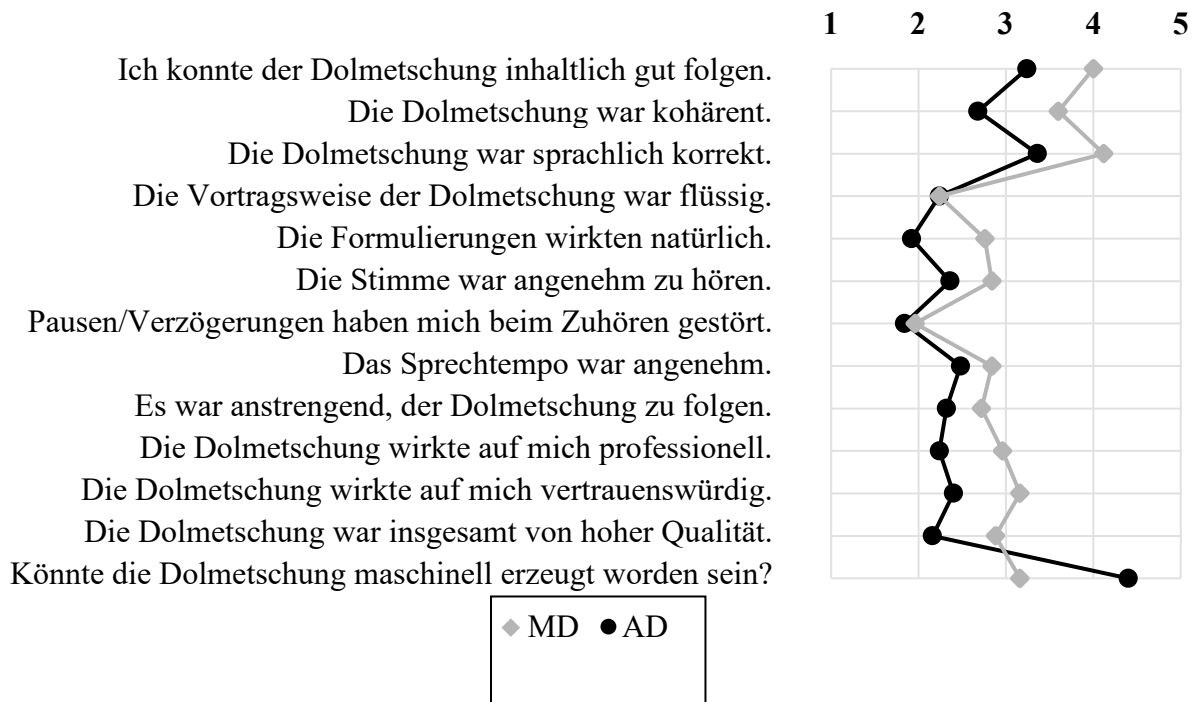


Abbildung 3: Darstellung der Mittelwertunterschiede

Im Hinblick auf das von den Rezipient*innen abgegebene Gesamturteil wurde keiner der beiden Dolmetschungen im Mittel eine hohe Qualität zugesprochen. Die maschinelle Dolmetschung wurde mit einem Durchschnittswert von 2,16 merklich schlechter bewertet als die Dolmetschung menschlichen Ursprungs, bei der ein Mittelwert von 2,88 erzielt wurde. Die Standardabweichungen sind hierbei überaus ähnlich ausgefallen ($AD = 1,11$; $MD = 1,01$). Beinahe die Hälfte aller Teilnehmenden der AD-Gruppe (48 %) hat der maschinellen Dolmetschung mit Antwortmöglichkeit 2 eine eher schlechte Qualität beigemessen. Dahingegen waren die Antworten der MD-Gruppe weniger stark auf eine Option konzentriert – der größte Anteil entfiel mit 33 % auf Antwortmöglichkeit 3. Nur eine einzige der AD-Gruppe zugehörige Person und keine Teilnehmer*in der MD-Gruppe stimmte voll und ganz zu, dass die jeweils rezipierte Dolmetschung insgesamt von hoher Qualität war. Dieser Aussage stimmten 7 Teilnehmer*innen der AD-Gruppe und 3 der MD-Gruppe überhaupt nicht zu.

Tabelle 6: Gruppenvergleich betreffend Professionalität, Vertrauenswürdigkeit und Qualität

	MD		AD		ΔM
	M	SD	M	SD	
Die Dolmetschung wirkte auf mich professionell.	2,96	1,17	2,24	1,30	0,72
Die Dolmetschung wirkte auf mich vertrauenswürdig.	3,16	1,18	2,40	1,19	0,76
Die Dolmetschung war insgesamt von hoher Qualität.	2,88	1,01	2,16	1,11	0,72

Bei den Aspekten der Professionalität und Vertrauenswürdigkeit zeigt sich ein ähnliches Bild. Den von den Teilnehmenden der AD-Gruppe erhaltenen Angaben zufolge wurden beide Aspekte eher negativ wahrgenommen (siehe Tabelle 6). Die Urteile der Teilnehmenden der MD-Gruppe liegen im Schnitt nahe der Mitte der Skala und damit deutlich über den Ergebnissen der AD-Gruppe. Werden alle drei Aspekte pro Gruppe isoliert betrachtet, so wird deutlich, dass sie jeweils ähnlich ausgeprägt sind. Im Hinblick auf das Gesamturteil, die beigemessene Professionalität und Vertrauenswürdigkeit schnitt die menschliche Dolmetschung klar – jedoch nicht in erheblichem Umfang – besser ab.

Tabelle 7: Gruppenvergleich im Hinblick auf die Präsentationsebene

	MD		AD		ΔM
	M	SD	M	SD	
Die Vortragsweise der Dolmetschung war flüssig.	2,24	0,97	2,24	1,13	0,00
Die Formulierungen wirkten natürlich.	2,76	1,23	1,92	1,12	0,84
Die Stimme war angenehm zu hören.	2,84	1,07	2,36	1,15	0,48
Pausen/Verzögerungen haben mich beim Zuhören gestört.	1,96	1,06	1,84	1,34	0,12
Das Sprechtempo war angenehm.	2,84	1,03	2,48	1,08	0,36
Es war anstrengend, der Dolmetschung zu folgen.	2,72	1,34	2,32	1,22	0,40

Im Folgenden werden die primär die Präsentationsebene betreffenden Aspekte in den Fokus gerückt. Der Aspekt der Flüssigkeit wurde von beiden Gruppen überaus ähnlich bewertet. Der Mittelwert liegt jeweils exakt bei 2,24, wobei bei der AD-Gruppe stärkere Abweichungen zu verzeichnen sind. Sehr ähnlich wurden von beiden Gruppen ferner das Pausenverhalten und die Verzögerungen beurteilt. Durch Umkehrung der Bewertung liegt der dabei erzielte Mittelwert der MD-Gruppe bei 1,96 und jener der AD-Gruppe bei 1,84. Dementsprechend erhielt das von der AD-Gruppe wahrgenommene Pausenverhalten unter allen Kategorien gruppenübergreifend im Schnitt die negativste Bewertung. Kein*e einzige*r Teilnehmer*in der MD-Gruppe gab an, dass die Pausen sie*ihn überhaupt nicht gestört haben, gegenüber drei Personen der AD-Gruppe. Die hohe Übereinstimmung zwischen den Gruppen in Hinblick auf die Wahrnehmung

der gesetzten Pausen und die Verzögerungen zeigt sich auch an der Anzahl der abgegebenen Bemerkungen im Zuge der Beantwortung der offenen Fragen – 6 Teilnehmer*innen der MD-Gruppe und 7 Teilnehmer*innen der AD-Gruppe hinterließen einen kritischen Kommentar zu diesem Aspekt. Insgesamt beurteilen beide Gruppen diesen Aspekt als überaus negativ.

Ein weiterer von der AD-Gruppe überaus negativ bewerteter Aspekt ist die Natürlichkeit mit einem Durchschnittswert von 1,92. Die MD-Gruppe hat die von ihnen rezipierte Dolmetschleistung mit einem Mittelwert von 2,76 ebenso eher als unnatürlich eingestuft, zugleich jedoch auch klar natürlicher als die AD-Gruppe. Folglich ist die Natürlichkeit jener der die Präsentationsebene betreffenden Faktoren, bei dem der auf dem Mittelwert beruhende Unterschied zwischen den Gruppen am deutlichsten ausgeprägt ist.

Ähnliches trifft auch – wenngleich in geringerem Ausmaß – auf die Wahrnehmung der Stimme (AD = 2,36; MD = 2,84), des Sprechtempos (AD = 2,48; MD = 2,84) und der bei der Rezeption empfundenen Anstrengung (invertiert: AD = 2,32; MD = 2,72) zu: beide Gruppen beurteilten die jeweils rezipierte Leistung im Hinblick auf diese Faktoren im Durchschnitt eher negativ, die MD-Gruppe jedoch markant negativer. Im Zuge der Beantwortung der offenen Fragen wurde der Aspekt der Stimme von den Rezipient*innen der MD-Gruppe zweimal positiv und einmal negativ erwähnt und von der AD-Gruppe je dreimal positiv und negativ und fand somit gruppenübergreifend in ambivalenter Weise Erwähnung. Gruppenübergreifend wurde kein einziger primär die Präsentationsebene betreffender Aspekt im Schnitt eher positiv bewertet. Zudem wurde kein einziger Aspekt von der AD-Gruppe positiver bewertet als von der MD-Gruppe.

Tabelle 8: Gruppenvergleich betreffend primär inhaltlich-strukturelle Aspekte

	MD		AD		AM
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Ich konnte der Dolmetschung inhaltlich gut folgen.	4,00	0,87	3,24	1,05	0,76
Die Dolmetschung war kohärent.	3,60	0,91	2,68	1,07	0,92
Die Dolmetschung war sprachlich korrekt.	4,12	0,73	3,36	1,22	0,76

Auch bei den inhaltlich-strukturellen Aspekten schnitt die maschinell generierte Dolmetschung im Schnitt schlechter ab als die von einer professionellen Dolmetscherin angefertigte Leistung. Bei den drei darunter subsumierten Kategorien erhielten beide Dolmetschungen wesentlich bessere Bewertungen als bei den die Darbietung betreffenden Aspekten (siehe Tabelle 8). Hervorzuheben ist, dass die Rezipient*innen der MD-Gruppe der Dolmetschung im Schnitt gut folgen

konnten ($M = 4,00$) und die rezipierte Leistung als sprachlich korrekt empfanden ($M = 4,12$). Diese beiden Aspekte wurden zwar auch von den Rezipient*innen der maschinellen Leistung eher positiv empfunden, jedoch – wie der Tabelle zu entnehmen ist – in auffällig geringerem Ausmaß. Dabei wurde die Kohärenz der maschinellen Dolmetschung sogar eher negativ empfunden ($M = 2,68$). Bemerkenswert ist ferner, dass kein*e einzige*r Teilnehmer*in der MD-Gruppe bei einer dieser drei Kategorien die negativste Auswahlmöglichkeit auswählte, während dies bei der AD-Gruppe jeweils 2 Personen taten. Die für diese Kategorien von der MD-Gruppe erfassten Standardabweichungen sind darüber hinaus deutlich geringer ausgefallen als die von der AD-Gruppe erfassten Werte, was auf eine heterogenere Verteilung der Angaben hindeutet. Die von einem Menschen angefertigte Dolmetschung stellte sich den abgegebenen Urteilen entsprechend im Resultat somit durchschnittlich als klar kohärenter, korrekter und inhaltlich nachvollziehbarer heraus.

Insgesamt lassen sich durch den Vergleich zwischen den von den Teilnehmenden der beiden Gruppen erhaltenen Angaben gruppenübergreifend ähnliche Muster sowie eine in nahezu allen Kategorien klar positivere Bewertung der von einer professionellen Dolmetscherin angefertigten Leistung erkennen. Im Hinblick auf die primär die Präsentationsebene betreffenden Faktoren zeigt der Vergleich, dass beide Leistungen nicht positiv wahrgenommen wurden, jedoch die maschinelle Leistung bei allen Faktoren – mit Ausnahme der Flüssigkeit – im Schnitt klar negativer aufgefasst wurde. Ferner wurden ausschließlich die sprachliche Korrektheit ($M = 3,36$) und der Faktor, wie leicht es ist, der Dolmetschung inhaltlich zu folgen, ($M = 3,24$) bei der AD-Leistung im Schnitt mit leichter positiver Tendenz, jedoch in absoluten Zahlen trotzdem mittelmäßig bewertet. Diese beiden Aspekte wurden von der MD-Gruppe mit einem Mittelwert von 4,12 im Hinblick auf die sprachliche Korrektheit und 4,00 in Bezug darauf, wie leicht der Dolmetschung inhaltlich gefolgt werden konnte, insgesamt positiv gesehen. Ferner wurde die Kohärenz der professionell angefertigten Dolmetschleistung von deren Rezipient*innen mit einem Durchschnittswert von 3,6 und damit ebenso mit leichter positiver Tendenz bewertet. Die Vertrauenswürdigkeit ($M = 3,16$) wurde von der MD-Gruppe ferner als mittelmäßig bewertet. Gruppen- und damit dolmetschleistungsübergreifend wurden ausnahmslos alle anderen Faktoren – wenngleich in unterschiedlichem Ausmaß – negativ oder eher negativ bewertet.

An diese auf den Mittelwert und entsprechende Unterschiede fokussierte Betrachtung anschließend werden die Unterschiede bei der Wahrnehmung nun anhand der Effektstärke dargelegt, um zugrunde liegende Unterschiede relativ zur Streuung der Antworten zu veranschaulichen. Aus der nachstehenden Tabelle 9 geht hervor, welche Items am stärksten zu den

Unterschieden zwischen den Gruppen beitragen. Je höher der Wert, desto stärker trägt der betreffende Aspekt zum Gesamtunterschied bei den erfassten Wahrnehmungen bei. Die Kategorien werden zur besseren Veranschaulichung in absteigender Größe angegeben. Für Erläuterungen zur Berechnung der Effektstärke d siehe Kapitel 3.6.

Tabelle 9: Reihung der Gruppenunterschiede nach Effektstärke d

	d	0	0,5	1
Die Dolmetschung war kohärent.	0,93			
Ich konnte der Dolmetschung inhaltlich gut folgen.	0,79			
Die Dolmetschung war sprachlich korrekt.	0,76			
Die Formulierungen wirkten natürlich.	0,71			
Die Dolmetschung war insgesamt von hoher Qualität.	0,68			
Die Dolmetschung wirkte auf mich vertrauenswürdig.	0,64			
Die Dolmetschung wirkte auf mich professionell.	0,58			
Die Stimme war angenehm zu hören.	0,43			
Das Sprechtempo war angenehm.	0,34			
Es war anstrengend, der Dolmetschung zu folgen.	0,31			
Pausen/Verzögerungen haben mich beim Zuhören gestört.	0,10			
Die Vortragsweise der Dolmetschung war flüssig.	0,00			

Durch Anwendung der Effektgröße d wird die anhand der Mittelwerte gewonnene Beobachtung bestätigt, dass sich die Unterschiede bei der Wahrnehmung entgegen der im Vorfeld aufgestellten Hypothese weniger auf die primär die Präsentationsebene betreffenden Aspekte konzentrieren als auf Kohärenz, sprachliche Korrektheit und die Beschaffenheit der Dolmetschungen, inhaltlich nachvollziehbar zu sein, wobei sich die Kohärenz als jener Faktor herausstellte, der am stärksten zu den erfassten Unterschieden zwischen den Gruppen beitrug. Daneben trug auch die wahrgenommene Natürlichkeit – bei der die Mittelwerte zwischen den Gruppen am umfangreichsten voneinander abweichen – markant zu den insgesamt erfassten Unterschieden zwischen den Gruppen bei. Allgemein hin divergiert die Effektstärke d aufgrund der ähnlichen Streuung bei den Antworten der Gruppen nicht erheblich von den Mittelwertunterschieden. Die geringsten Unterschiede und damit die größten Übereinstimmungen sind bei der wahrgenommenen Flüssigkeit und dem Pausenverhalten beziehungsweise Verzögerungen zu erkennen. Ähnlichkeiten sind bei den Bewertungen der Gruppen auch in Bezug darauf festzustellen, wie anstrengend es war, der jeweiligen Dolmetschung zu folgen, und wie angenehm das Sprechtempo war.

Diesen Ausführungen entsprechend wurden alle Kategorien von den Rezipient*innen der von einer professionellen Dolmetscherin angefertigten Leistung mit einer Ausnahme – dem Aspekt der Flüssigkeit, der im Schnitt ident beurteilt wurde – positiver gesehen als von den Rezipient*innen der maschinell generierten Dolmetschung. Die Unterschiede zwischen den beiden Gruppen sind dabei nicht bei den primär auf der Präsentationsebene anzusiedelnden Aspekten zu verorten. Hauptunterschiede betreffen gemessen an der Effektstärke die Kohärenz, sprachliche Korrektheit und den Aspekt, wie einfach es ist, der Leistung inhaltlich zu folgen. Diese Faktoren entsprechen zugleich auch jenen Aspekten, die bei beiden Gruppen jeweils relativ zum Rest der von der betreffenden Gruppe erhaltenen Angaben am positivsten aufgefasst wurden. Bei ausschließlicher Betrachtung der Mittelwertunterschiede konnte zudem die Natürlichkeit als ein weiterer die Gruppen erheblich unterscheidender Faktor erkannt werden. Im Resultat folgen beide Leistungen – gemessen an der durch die Items erfragten Beobachtungen und Wahrnehmungen – einem überaus ähnlichen Muster, wobei die menschliche Leistung als im Durchschnitt deutlich besser wahrgenommen wurde.

4.5 Betrachtung von Korrelationen im Hinblick auf die Hypothesen

Für das Forschungsziel der vorliegenden Arbeit wurde unter anderem die Hypothese aufgestellt, dass die Gesamtqualität der rezipierten Leistung aufgrund einer etwaigen nicht natürlich anmutenden Stimme niedriger bewertet wird. Um die dementsprechend vermutete Verbindung zwischen den betreffenden Faktoren der Natürlichkeit beziehungsweise der Stimmqualität und der angegebenen Gesamtbewertung zu betrachten, wurde für die entsprechenden Faktoren deren Korrelation r errechnet. Zur Berechnung der Korrelation r siehe Kapitel 3.6.

Auf Grundlage der von der AD-Gruppe abgegebenen Angaben wurde im Hinblick auf die Korrelation zwischen Stimme und Qualität ein Wert von 0,74 ermittelt. Demzufolge geht mit einem höheren Wert bei der Stimme und einer entsprechend positiveren Wahrnehmung auch ein deutlich positiveres Urteil für die Gesamtqualität der rezipierten Leistung einher. Dieser Effekt trifft gleichermaßen auf gegenteilige Situationen zu: je negativer der Aspekt der Stimme wahrgenommen wird, desto geringer fällt auch die der Leistung insgesamt zugesprochene Qualität aus.

Dem in Bezug auf die Korrelation zwischen Natürlichkeit und Qualität errechneten Wert von 0,79 zufolge entspricht auch bei der Natürlichkeit ein höherer erzielter Wert klar einem höheren Wert bei der angegebenen Gesamtqualität. Im Resultat zeigt sich, dass beide Aspekte

in erheblichem Maße an das erfragte Globalurteil gekoppelt sind. Je höher Stimme und Natürlichkeit von den Rezipient*innen der maschinell erzeugten Leistung bewertet werden, desto positiver wurde auch die Qualitätswahrnehmung beurteilt. Dieser Effekt wird anhand der nachstehenden Tabelle und der darin enthaltenen Kombinationen der betreffenden Aussagen in Bezug auf Stimme beziehungsweise Natürlichkeit in der von der Teilnehmer*in jeweils gewählten Ausprägung von 1 bis 5 und des von derselben Person abgegebenen Gesamturteils der entsprechend anhaftenden Qualität für beide Gruppen veranschaulicht.

Stimme (y) / Qualität (x)						Natürlichkeit (y) / Qualität (x)					
AD			MD			AD			MD		
	1	2	3	4	5		1	2	3	4	5
5				1	1	5					1
4		1				4				2	
3	1	2	2	2		3		1	1		
2	2	7				2	2	5	1	1	
1	4	2				1	5	6			

Tabelle 10: Antwortverteilungen von Stimme bzw. Natürlichkeit in Relation zu Gesamtqualität

Auch bei den entsprechenden Korrelationen der MD-Gruppe zeigt sich ein positives Verhältnis zwischen den genannten Aspekten. Im Hinblick auf die Korrelation zwischen Stimme und Qualität beträgt der errechnete Wert 0,29. Demzufolge wurde auch bei der von einer professionellen Dolmetscherin angefertigten Leistung die Qualität höher bewertet, je positiver die Stimme aufgenommen wurde, wenngleich in weniger ausgeprägtem Umfang als von der AD-Gruppe. Die für die MD-Gruppe errechnete Korrelation zwischen Natürlichkeit und Qualität beträgt dahingegen 0,61, wodurch auch zwischen diesen beiden Faktoren eine Korrelation in erheblicher Ausprägung festgestellt werden konnte.

Im Resultat sind die erkannten Korrelationseffekte bei der maschinell erzeugten Leistung insbesondere im Hinblick auf den Aspekt der Stimme stärker hervorgetreten als bei der von einer professionellen Dolmetscherin angefertigten Dolmetschung. Die Korrelation zwischen Stimme und Qualität ist zwar auch bei der MD-Gruppe positiv, jedoch in vergleichsweise moderatem Umfang, weswegen andere Faktoren das Gesamturteil dieser Gruppe scheinbar stärker getragen haben. Die durch den Rückgriff auf Korrelationswerte gewonnenen Resultate sind vereinbar mit der Hypothese, dass eine geringere Natürlichkeit und negativ wahrgenommene Stimmqualität mit schlechter aufgefasster Gesamtqualität einhergehen.

Im Zuge der Darlegung des Forschungsziels der vorliegenden Arbeit wurde ferner die Hypothese aufgestellt, dass die Qualität der jeweils rezipierten Leistung umso negativer wahrgenommen wird, je sicherer sich die Teilnehmenden sind, dass es sich um eine maschinell erzeugte Leistung handelt. Um diese Hypothese näher zu betrachten, wurde die Korrelation zwischen dieser Einschätzung und dem abgegebenen Urteil zur Gesamtqualität errechnet. Die Verteilungen der von den Rezipient*innen erhaltenen Antworten zum entsprechenden Likert-Item sind in nachstehender Tabelle 11 festgehalten, um den Zusammenhang zwischen den Antworten zu den beiden betreffenden Kategorien zu veranschaulichen.

Einschätzung zu maschinellem Ursprung (y) / Qualität (x)													
AD							MD						
	1	2	3	4	5		1	2	3	4	5		
5	7	8		1		5	2	1		2			
4		3	2	1		4	1	1	2	1			
3				1		3			4	3			
2		1				2		2	2	1			
1					1	1		1	1	1			

Tabelle 11: Antwortverteilung der Einschätzung zum maschinellen Ursprung relativ zur wahrgenommenen Qualität

Aus den von den Teilnehmenden der AD-Gruppe erhaltenen Antworten betreffend die Einschätzung zum maschinellen Ursprung und zur wahrgenommenen Gesamtqualität wurde eine Korrelation von -0,64 errechnet. Diesem Wert kann entnommen werden, dass die Qualität insgesamt umso negativer wahrgenommen wurde, je höher die Möglichkeit eingeschätzt wurde, dass es sich bei der rezipierten Dolmetschung um eine maschinell generierte Leistung handelt. Aus den von den Rezipient*innen der menschlich angefertigten Dolmetschung erhaltenen Angaben ergibt sich eine Korrelation in Höhe von -0,20, wodurch sich derselbe Effekt in schwächerer Ausprägung auch bei der MD-Gruppe erkennen lässt. Im Resultat besteht somit – auch gruppenübergreifend – ein klares negatives Verhältnis zwischen diesen beiden Aspekten. Wird der jeweils rezipierten Leistung im Rahmen des der vorliegenden Arbeit zugrunde liegenden Experiments ein maschineller Ursprung zugesprochen, so wird der entsprechenden Dolmetschung zugleich geringere Qualität beigemessen.

Im Hinblick auf die oben dargelegten Korrelationswerte und die dadurch aufgedeckten Regelmäßigkeiten sei festgehalten, dass ihnen keine Aufschlüsse über etwaige Kausalitäten entnommen werden können. Die Korrelationswerte geben Auskunft über die Richtung und Stärke der Zusammenhänge, jedoch nicht über deren exakte Auslöser. Auch Drittvariablen oder andere Faktoren könnten zu den erkannten Zusammenhängen beigetragen oder geführt haben

oder gar beide Werte zugleich beeinflusst haben. Aus den erkannten Korrelationswerten kann lediglich abgeleitet werden, dass sich die betreffenden Wahrnehmungen der Teilnehmenden gemeinsam je nach Ausprägung verändern.

5 Diskussion und Schlussfolgerungen

Ziel der vorliegenden Arbeit war es, zu erfassen, wie die Qualität einer von einem gegenwärtig handelsüblichen Dolmetschsystem angefertigten Dolmetschleistung von Personen, die nicht als Dolmetscher*innen tätig sind, wahrgenommen wird. Zugleich wurde betrachtet, wie eine von einer professionellen Dolmetscherin auf Grundlage derselben Ausgangsrede angefertigte Dolmetschung von ebenfalls nicht als Dolmetscher*innen tätigen Personen wahrgenommen wird, sowie die von diesen beiden Gruppen erfassten Qualitätsurteile miteinander verglichen und anhaftende Unterschiede, Muster und Besonderheiten aufgedeckt. Im Mittelpunkt stand dabei die Wahrnehmung von Qualität durch die Rezipient*innen nach ihren individuellen Urteilen ausschließlich anhand der beiden konkret dargebotenen Leistungen.

Bei der Untersuchung wurde im Rahmen von drei im Vorfeld definierten Hypothesen davon ausgegangen, dass eine nicht natürlich anmutende Stimme der Gesamtqualität der rezipierten Leistung abträglich ist, dass sich die zwischen den Qualitätsurteilen der beiden Gruppen erfassten Unterschiede in erster Linie auf die primär die Präsentationsebene betreffenden Aspekte der Natürlichkeit, des Pausenverhaltens beziehungsweise des Sprechtempos, der Stimme, der Flüssigkeit und der erforderlichen Anstrengung beim Zuhören beziehen und dass die Qualität umso negativer wahrgenommen wird, je sicherer sich die Teilnehmer*innen sind, dass es sich bei der ihnen vorliegenden Dolmetschung um eine maschinell erzeugte Leistung handelt.

Zur Erreichung des Forschungsziels wurde ein Experiment durchgeführt, in dessen Rahmen den Teilnehmenden eine Dolmetschung zur Rezeption vorgelegt wurde und ihre Qualitätsurteile im Wege einer Befragung erfasst wurden. Für die zu diesem Zweck herangezogene Stichprobe wurden 50 nicht als Dolmetscher*innen tätige Personen rekrutiert, die randomisiert einer von zwei Gruppen zu je 25 Personen zugeordnet wurden. Beiden Gruppen wurde je einer von zwei Fragebögen vorgelegt, die mit einer bedeutenden Ausnahme – der jeweils eingebetteten, entweder maschinell generierten oder menschengemachten Dolmetschleistung – vollkommen ident waren. Die Erfassung der Qualitätsurteile erfolgte im Rahmen eines Online-Fragebogens anhand von zwölf Likert-Items unter Verwendung einer fünfstufigen Zustimmungsskala sowie zwei ergänzenden offenen Fragen. Zusätzlich – und im Rahmen des Fragebogens technisch von den auf die Wahrnehmung der Qualität abzielenden Items getrennt – wurden ferner demografische Angaben erfasst sowie die Vorerfahrung mit Simultandolmetschen und die individuelle Einschätzung, ob die jeweils rezipierte Leistung maschinellen Ursprungs ist. Im Zuge der Befragung wurde sowohl unmittelbar die der rezipierten Leistung beigemessene Gesamtqualität festgehalten als auch elf weitere Qualitätskriterien zur Beantwortung vorgelegt

(für eine übersichtliche Zusammenfassung der Fragen und Antworten siehe Anhänge 4 bis 8). Zur mehrdimensionalen Betrachtung der Ergebnisse wurden jeweils der Mittelwert, die Antwortverteilungen, der Median, der Modus, die standardisierte Effektstärke d nach Cohen (1988) und vereinzelt der Korrelationskoeffizient r errechnet und für die Analyse herangezogen. Infolge der Auswertung der ermittelten Daten lassen sich die dieser Arbeit zugrunde liegenden Forschungsfragen wie folgt beantworten:

Im Rahmen der erfassten Stichprobe wurde die maschinelle Dolmetschung von ihren Rezipient*innen sowohl in der unmittelbar abgefragten Gesamtwahrnehmung der Qualität als auch hinsichtlich des Großteils der einzelnen Qualitätsdimensionen überwiegend negativ bewertet. Ausschließlich im Hinblick auf die sprachliche Korrektheit sowie darauf, ob der Dolmetschung inhaltlich gefolgt werden konnte, wurden im Schnitt eher positive Bewertungen abgegeben. Am negativsten haben die Rezipient*innen der maschinell erzeugten Leistung das Pausenverhalten beziehungsweise die Verzögerungen bewertet, gefolgt von der beigemessenen Natürlichkeit. Im Kontrast dazu wurde die von einem Menschen angefertigte Dolmetschung sowohl im Hinblick auf das Gesamturteil als auch beinahe alle einzelnen Qualitätskategorien – mit einer Ausnahme – deutlich positiver wahrgenommen als die maschinelle Leistung. Einzig der Aspekt der Flüssigkeit wurde im Schnitt von beiden Gruppen ident bewertet. Insgesamt wurde jedoch auch der von einer professionellen Dolmetscherin angefertigten Leistung von ihren Rezipient*innen eine mittelmäßige Qualität beigemessen. Dabei waren das Pausenverhalten beziehungsweise die Verzögerungen – wie auch bei der maschinell erzeugten Dolmetschung – der am stärksten bemängelte Aspekt, gefolgt von der Flüssigkeit. Am positivsten wurden – ebenso wie bei der maschinell angefertigten Dolmetschung – die sprachliche Korrektheit und die inhaltliche Nachvollziehbarkeit der professionell angefertigten Leistung wahrgenommen. Im Resultat wurde beiden Leistungen von den betreffenden Rezipient*innen eher geringe Qualität beigemessen, wobei die menschlich erzeugte Dolmetschung jedoch im Schnitt deutlich positiver bewertet wurde.

Beachtenswerterweise eint beide Gruppen, dass am Pausenverhalten beziehungsweise an den Verzögerungen jeweils – bei ähnlicher Ausprägung – die größte Kritik geübt und der Aspekt der Flüssigkeit von beiden Gruppen im Schnitt ident beurteilt wurde. Dieser Umstand legt vor dem Hintergrund der dennoch erzielten höheren Gesamtqualität der Humandolmetschung im Rahmen der erfassten Stichprobe die Vermutung nahe, dass diese beiden Faktoren – wenngleich sie gewiss in das Gesamtbild eingeflossen sind – für die Unterschiede in der Gesamtwahrnehmung nicht ausschlaggebend waren.

Entgegen der ursprünglich aufgestellten Hypothese bezogen sich die markantesten Unterschiede zwischen den abgegebenen Bewertungen – je nach herangezogener Metrik – in erster Linie auf die wahrgenommene Kohärenz, Natürlichkeit und inhaltliche Nachvollziehbarkeit. Die Hypothese, dass die Unterschiede in erster Linie durch die primär die Präsentation und Darbietung betreffenden Aspekte der Natürlichkeit, des Pausenverhaltens beziehungsweise des Sprechtempos, der Stimme, der Flüssigkeit und der erforderlichen Anstrengung beim Zuhören getragen werden, konnte im Rahmen der erfassten Stichprobe und der erhaltenen Angaben nicht bestätigt werden. Sowohl anhand der errechneten Mittelwertunterschiede als auch durch Betrachtung der standardisierten Effektstärke d stellte sich der Aspekt der Kohärenz als wichtigstes und ausschlaggebendstes Unterscheidungsmerkmal zwischen den beiden herangezogenen Leistungen heraus. Im Hinblick auf die der vorliegenden Arbeit zugrunde liegenden Dolmetschungen scheint die Auffassung, dass insbesondere Stimme und Sprachfluss Unterschiede in der beigemessenen Qualität bedingen, den Angaben der Rezipient*innen entsprechend folglich zu eng gefasst.

Im Zuge der Gegenüberstellung der Ergebnisse wurde ferner aufgedeckt, dass zwei Gruppen von jeweils drei Qualitätsaspekten gruppenübergreifend im Hinblick auf die durchschnittlich erzielten Werte überaus ähnlich wahrgenommen wurden. Je Gruppe wurden einerseits die Aspekte der sprachlichen Korrektheit, der Kohärenz und der inhaltlichen Nachvollziehbarkeit ähnlich bewertet und andererseits wurde auch den Aspekten der anhaftenden Professionalität, Vertrauenswürdigkeit und der Gesamtqualität jeweils ähnliche Urteile ausgesprochen. Dies könnte im Rahmen der erfassten Stichprobe jeweils auf Korrelationen bei den drei betreffenden Aspekten hindeuten.

Zur Beantwortung und Betrachtung der beiden weiteren Hypothesen wurden entsprechende Korrelationen zwischen den betreffenden Qualitätsaspekten erfasst. Den in Bezug auf die Stimmqualität beziehungsweise Natürlichkeit und die wahrgenommene Gesamtqualität errechneten Korrelationswerten konnte jeweils ein positiver linearer Zusammenhang zwischen diesen Faktoren entnommen werden. Demnach geht mit höher bewerteter Natürlichkeit beziehungsweise Stimmqualität jeweils auch ein positiveres Gesamturteil einher – und umgekehrt. Diese Korrelationen konnten zwar gruppenübergreifend festgestellt werden, im Hinblick auf die Beziehung zwischen Stimme und Gesamtqualität bei den Rezipient*innen der menschengemachten Dolmetschung fällt die entsprechend erfasste Korrelation jedoch deutlich geringer

aus. Dies könnte im Rahmen der erfassten Stichprobe darauf schließen lassen, dass die Stimmqualität bei Leistungen maschinellen Ursprungs, die als solche erkannt werden, kritischer beurteilt wird.

Darüber hinaus wurde analog dazu auch der Korrelationskoeffizient zwischen der Einschätzung der Teilnehmenden, ob ihnen eine maschinell generierte Dolmetschung vorgelegt wurde, und der wahrgenommenen Gesamtqualität errechnet und dabei ein negativer linearer Zusammenhang erkannt. Demzufolge wurde der rezipierten Leistung eine umso geringere Qualität beigemessen, je sicherer sich die Rezipient*innen waren, dass sie eine maschinelle Leistung rezipiert haben. Auch dieser Zusammenhang ist gruppenübergreifend zu beobachten, konnte jedoch bei den Rezipient*innen der maschinell generierten Dolmetschung in ausgeprägterer Form festgestellt werden.

Eine weitere im Rahmen der Untersuchung erhaltene, im Vorfeld jedoch nicht angestrebte Erkenntnis liegt darin, dass die Anzahl der Rezipient*innen der professionell angefertigten Dolmetschung, die sich sicher waren, dass es sich bei der ihnen vorgelegten Leistung um ein maschinell generiertes Produkt handelt, die Anzahl der Teilnehmenden dieser Gruppe übersteigt, die sich sicher waren, dass die rezipierte Leistung nicht maschinellen Ursprungs ist. Dies könnte einerseits an der nach Auffassung beider Gruppen im Schnitt mangelhaften Flüssigkeit oder dem viel bemängelten Pausenverhalten liegen, andererseits jedoch auch an einer möglichen hohen Erwartung der Allgemeinbevölkerung an maschinelle Dolmetschsysteme, wonach sie Leistungen, die der von der professionellen Dolmetscherin angefertigten gleichwertig sind, auch Maschinen zutrauen und daher nicht ohne Weiteres zwischen Mensch und Maschine unterscheiden können.

Aus diesen Ergebnissen lassen sich verschiedene Ansätze für weiterführende Forschung ableiten. So könnten beispielsweise im Rahmen von Replikationsstudien Dolmetschungen derselben Rede verschiedener Dolmetscher*innen der maschinellen Dolmetschung gegenübergestellt werden, um angesichts des Umstandes, dass die maschinell generierte Dolmetschung mit einer männlich anmutenden Stimme versehen wurde, potenzielle genderbasierte Biases zu erforschen. Ebenso könnten Dolmetschungen verschiedener maschineller Dolmetschsysteme mit unterschiedlichen Sprachsyntheseprogrammen einerseits einander gegenübergestellt werden, andererseits jedoch auch fortschrittlichen Syntheseprogrammen, mit denen sich von einem Menschen kaum unterscheidbare Leistungen erzeugen lassen, wodurch weiter herausgefiltert

werden könnte, in welchem Umfang sich die Natürlichkeit der Sprache auf die wahrgenommene Qualität auswirkt. Erforschenswert wäre ferner ein realitätsnäheres Szenario, bei dem die Rezipient*innen nicht nur die Dolmetschung hören, sondern auch die Ausgangsrede. Dadurch könnten sie – bei Verständnis der Ausgangssprache – auch die Vollständigkeit der Verdolmetschung einschätzen, was als weiterer Faktor in ihre Gesamtbewertung einfließen könnte.

Bibliographie

- Amini, Mansour; Ibrahim-González, Noraini und Ayob, Leelany (2013). Quality of Interpreting from Users' Perspectives. *International Journal of English and Education* 2(1), 89-98.
- Barfuß, Elisabeth (2017). *Skype Translator. Reaktionen des maschinellen Dolmetschsystems auf problemauslösende Faktoren in der Spracherkennung im Englischen*. Masterarbeit, Universität Wien.
- Batliner, Anton; Buckow, Jan; Niemann, Heinrich; Nöth, Elmar und Warnke, Volker (2000). The Prosody Module. In: Wolfgang, Wahlster (Hg.) *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin/Heidelberg: Springer, 106-121.
- Bethlehem, Jelke (2010). Selection Bias in Web Surveys. *International Statistical Review* 78 (2), 161-188. <https://doi.org/10.1111/j.1751-5823.2010.00112.x>
- Black, Alan W. und Lenzo, Kevin A. (2004). Multilingual Text-to-Speech Synthesis. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* 3, 761-764. <https://doi.org/10.1109/ICASSP.2004.1326656>
- Boone, Harry N. und Boone, Deborah A. (2012). Analyzing Likert Data. *Journal of Extension* 50 (2), Article 48. <https://doi.org/10.34068/joe.50.02.48>
- Burkhardt, Felix (2001). *Simulation emotionaler Sprechweise mit Sprachsynthesystemen*. Aachen: Shaker Verlag.
- Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat, Susanne; Klabunde Ralf und Langer, Hagen (2001). *Computerlinguistik und Sprachtechnologie – Eine Einführung*. Heidelberg: Spektrum Akademischer Verlag.
- Chan, Cedric und Kuang, Jianjing (2024). Exploring the Accuracy of Prosodic Encodings in State-of-the-Art Text-to-Speech Models. In: *Speech Prosody 2024*, 27-31. <https://doi.org/10.21437/SpeechProsody.2024-6>
- Chang, Janie (2011). *Speech Recognition Leaps Forward*. <https://www.microsoft.com/en-us/research/blog/speech-recognition-leaps-forward/> (Stand: 28.04.2020).
- Christodoulides, George und Lenglet, Cédric (2014). Prosodic Correlates of Perceived Quality and Fluency in Simultaneous Interpreting. In: Campbell, Nick; Gibbon, Dafydd

- und Hirst, Daniel (Hg.) *Speech Prosody. Proceedings of the 7th International Conference on Speech Prosody*, 1002-1006. https://www.isca-archive.org/speechprosody_2014/christodoulides14_speechprosody.pdf (Stand: 31.05.2026).
- Crumpton, Joe und Bethel, Cindy L. (2016). A Survey of Using Vocal Prosody to Convey Emotion in Robot Speech. *International Journal of Social Robotics* 8 (2), 271-285.
- Cohen, Jacob (1988): *Statistical Power Analysis for the Behavioral Sciences*. 2. Aufl. New York: Lawrence Erlbaum Associates.
- Cooper, Erica; Huang, Wen-Chin; Tsao, Yu; Wang, Hsin-Min; Toda, Tomoki und Yamagishi, Junichi (2024). A Review on Subjective and Objective Evaluation of Synthetic Speech. *Acoustical Science and Technology* 45 (4), 161-183.
<https://doi.org/10.1250/ast.e24.12>
- Dutoit, Thierry (1997). *An Introduction to Text-to-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers.
- Etikan, Ilker; Musa, Sulaiman A. und Alkassim, Rukayya S. (2016). Comparison of Convenience Sampling and Purposive Sampling. *American Journal of Theoretical and Applied Statistics* 5 (1), 1-4. <https://doi.org/10.11648/j.ajtas.20160501.11>
- Euler, Stephan (2006). *Grundkurs Spracherkennung. Vom Sprachsignal zum Dialog; Grundlagen und Anwendung verstehen; mit praktischen Übungen*. Wiesbaden: Vieweg.
- Fantinuoli, Claudio (2018). Interpreting and Technology: The Upcoming Technological Turn. In: Fantinuoli, Claudio (Hg.) *Interpreting and Technology*. Berlin: Language Science Press, 1-12. <https://doi.org/10.5281/zenodo.1493289>
- Fantinuoli, Claudio (2025). Machine Interpreting. In: Davitti, Elena; Korybski, Tomasz und Braun, Sabine (Hg.) *The Routledge Handbook of Interpreting, Technology and AI*. London: Routledge, 209-226. <https://doi.org/10.4324/9781003053248-16>
- Fügen, Christian; Waibel, Alex und Kolss, Muntsin (2007). Simultaneous Translation of Lectures and Speeches. *Machine Translation* 21 (4), 209-252.
- Fünfer, Sarah (2013). *Mensch oder Maschine? Dolmetscher und maschinelles Dolmetschsystem im Vergleich*. Berlin: Frank & Timme.

- Galdino, Julio Cesar; Matos, Ariadne Nascimento; Svartman, Flaviane Romani Fernandes und Aluísio, Sandra Maria (2025). The Evaluation of Prosody in Speech Synthesis: a Systematic Review. *Journal of the Brazilian Computer Society* 31 (1), 465-486. <https://doi.org/10.5753/jbcs.2025.5468>
- Glück, Helmut (2010). *Metzler Lexikon Sprache*. 4. Auflage. Stuttgart: Metzler.
- Grbić, Nadja (2008). Constructing Interpreting Quality. *Interpreting* 10 (2), 232-257.
- Han, Chao und Lu, Xiaolei (2021). Interpreting Quality Assessment Re-Imagined: The Synergy Between Human and Machine Scoring, *Interpreting and Society* 1 (1), 70-90. <https://doi.org/10.1177/27523810211033670>
- Hamon, Olivier; Mostefa, Djamel und Choukri, Khalid (2007). End-to-End Evaluation of a Speech-to-Speech Translation System in TC-STAR. In: Maegaard, Bente (Hg.) *Proceedings of Machine Translation Summit XI: Papers*. <https://aclanthology.org/2007.mtsummit-papers.30/> (Stand: 05.06.2026).
- Härtel, Johannes (2016). Vollautomatisches Dolmetschen – Möglichkeiten und Grenzen. *Lebende Sprachen* 61 (1), 67-116.
- Horváth, Ildikó (2021). Speech Translation vs. Interpreting. *Language Studies and Modern Humanities* 3 (2), 174-187. <https://doi.org/10.33910/2686-830X-2021-3-2-174-187>
- Hübener, Kai (1997). *Akustische Segmentierung zur Erkennung gesprochener Sprache*. Dissertation, Universität Hamburg.
- Joshi, Prem Lal und Marthandan, Govindan (2018). The Hype of Big Data Analytics and Auditors. *Emerging Markets Journal*, 8 (2), 1-4. <https://doi.org/10.5195/emaj.2018.153>
- Jiang, Zhaokun und Zhang, Ziyin (2025). From Black Box to Transparency: Enhancing Automated Interpreting Assessment with Explainable AI in College Classrooms. *Research Methods in Applied Linguistics* 4 (3). <https://doi.org/10.1016/j.rmal.2025.100237>
- Kade, Otto (1968). *Zufall und Gesetzmäßigkeit in der Übersetzung*. Leipzig: Verl. Enzyklopädie.
- Kalina, Sylvia (2002). Quality in Interpreting and Its Prerequisites: A Framework for a Comprehensive View. In: Garzone, Giuliana und Viezzi, Maurizio (Hg.) *Interpreting in the 21st Century: Challenges and Opportunities*. Amsterdam: Benjamins, 121-130.

- Klein, Dominik; Tran-Gia, Phuoc und Hartmann, Matthias (2013). Big Data. *Informatik-Spektrum* 36 (3), 319-323.
- Klie, Leonard (2015). Skype Translator Emerges. *Speech Technology* 20 (1), 13.
- Koehn, Philipp (2010). *Statistical Machine Translation*. Cambridge: Cambridge University Press.
- Kurz, Ingrid (2001). Conference Interpreting: Quality in the Ears of the User. *Meta* 46 (2), 394-409.
- Kutz, Wladimir (2012). *Dolmetschkompetenz: was muss der Dolmetscher wissen und können?*. Europäischer Universitätsverlag.
- Lakens, Daniël (2013): Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-Tests and ANOVAs. *Frontiers in Psychology* 4, Article 863. <https://doi.org/10.3389/fpsyg.2013.00863>
- Lam, Perry; Zhang, Huayun; Chen, Nancy F.; Sisman, Berrak und Herremans, Dorien (2024): PRESENT: Zero-Shot Text-to-Prosody Control. *IEEE Signal Processing Letters* 32. 776-780. <https://doi.org/10.1109/LSP.2025.3528359>
- Li, Jiahuan; Zhou, Hao; Huang, Shujian; Cheng, Shanbo und Chen, Jiajun (2024). Eliciting the Translation Ability of Large Language Models via Multilingual Finetuning with Translation Instructions. *Transactions of the Association for Computational Linguistics* 12, 576-592. https://doi.org/10.1162/tacl_a_00655
- Likert, Rensis (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology* 22 (140), 1-55.
- Liu, Zhao-Ci; Ling, Zhen-Hua; Hu, Ya-Jun; Pan, Jia; Wang, Jin-Wei und Wu, Yun-Di (2023): Speech Synthesis with Self-Supervisedly Learnt Prosodic Representations. In: *Proceedings of Interspeech 2023*, 7-11. <https://doi.org/10.21437/Interspeech.2023-1292>
- Lu, Xiaolei und Han, Chao (2023). Automatic Assessment of Spoken-Language Interpreting Based on Machine-Translation Evaluation Metrics. A Multi-Scenario Exploratory Study. *Interpreting* 25 (1), 109-143. <https://doi.org/10.1075/intp.00076.lu>

- Lu, Xinchao und Fantinuoli, Claudio (2025). Machine and Computer-Assisted Interpreting: Innovations in and Implications for Interpreting Practice, Pedagogy and Research. *Linguistica Antverpiensia* 24, 1-22. <https://doi.org/10.52034/lans-tts.v24i.869>
- Mack, Gabriele (2002). Die Beurteilung professioneller Dolmetschleistungen. In: Best, Joanna und Kalina, Sylvia (Hg.) *Übersetzen und Dolmetschen: eine Orientierungshilfe*. Tübingen: Francke, 110-118.
- Marz, Nathan; Warren, James und Lorenzen, Knut, (2016). *Big data: Entwicklung und Programmierung von Systemen für große Datenmengen und Einsatz der Lambda-Architektur*. Frechen: MITP.
- Neururer, Daniel; Dellwo, Volker und Stadelmann, Thilo (2024). Deep Neural Networks for Automatic Speaker Recognition do not Learn Supra-Segmental Temporal Features. *Pattern Recognition Letters* 181, 64-69. <https://doi.org/10.1016/j.patrec.2024.03.016>
- Nöth, Elmar (1991). *Prosodische Information in der automatischen Spracherkennung: Berechnung und Anwendung*. Tübingen: Niemeyer.
- Olive, Joseph; van Santen, Jan; Möbius, Bernd und Shih, Chilin (1998). Synthesis. In: Sproat, Richard W. (Hg.) *Multilingual Text-to-Speech Synthesis: The Bell Labs approach*. Boston: Kluwer. 191-228.
- Peinl, René (2021). *Sprachsynthese - State-of-the-Art in englischer und deutscher Sprache*. <https://doi.org/10.48550/arxiv.2106.06230>
- Petrick, Rico (2009). *Robuste Spracherkennung unter raumakustischen Umgebungsbedingungen*. Dissertation, Technische Universität Dresden.
- Pfister, Beat und Kaufmann, Tobias (2008). *Sprachverarbeitung. Grundlagen und Methoden der Sprachsynthese und Spracherkennung*. Berlin/Heidelberg: Springer.
- Pöchhacker, Franz (2005). Quality Research Revisited. *The Interpreters' Newsletter* 13, 143-166.
- Pöchhacker, Franz (2012). Qualität, die man versteht. Ein funktional-kognitiver Ansatz. In: Ahrens, Barbara; Albl-Mikasa, Michaela und Sasse, Claudia (Hg.) *Dolmetschqualität in Praxis, Lehre und Forschung. Festschrift für Sylvia Kalina*. Tübingen: Narr Francke Attempto Verlag, 19-31.

- Pöchhacker, Franz (2024). Is Machine Interpreting Interpreting?. *Translation Spaces*.
<https://doi.org/10.1075/ts.23028.poc>
- Ramlow, Markus (2009). *Die maschinelle Simulierbarkeit des Humanübersetzens. Evaluation von Mensch-Maschine-Interaktion und der Translatqualität der Technik*. Berlin: Frank & Timme.
- Ren, Yi; Tan, Xu; Qin, Tao; Zhao, Sheng; Zhao, Zhou und Liu, Tie-Yan (2019). Almost Un-supervised Text to Speech and Automatic Speech Recognition. In: Chaudhuri, Kamalika und Salakhutdinov, Ruslan (Hg.) *Proceedings of the 36th International Conference on Machine Learning*. Proceedings of Machine Learning Research 97, 5410-5419. <https://doi.org/10.48550/arXiv.1905.06791>
- Rennert, Sylvi (2010). The Impact of Fluency on the Subjective Assessment of Interpreting Quality. *The Interpreters' Newsletter* 15, 101-115. <http://hdl.handle.net/10077/4752> (Stand: 03.06.2026).
- Sasu, David (2025). *Leveraging and Probing Speech Prosody to Improve Spoken Language Processing*. Dissertation, IT University of Copenhagen. https://pure.itu.dk/ws/portal-files/portal/113177735/David_Sasu_Temp_Thesis.pdf (Stand 22.05.2026).
- Selting, Margret; Auer, Peter; Barth-Weingarten, Dagmar; Bergmann, Jörg; Bergmann, Pia; Birkner, Karin; Couper-Kuhlen, Elizabeth; Deppermann, Arnulf; Gilles, Peter; Günthner, Susanne; Hartung, Martin; Kern, Friederike; Mertzlufft, Christine; Meyer, Christian; Morek, Miriam; Oberzaucher, Frank; Peters, Jörg; Quasthoff, Uta; Schütte, Wilfried; Stukenbrock, Anja und Uhmann, Susanne (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 10, 353-402.
- Schober, Patrick; Boer, Christa und Schwarte, Lothar A. (2018): Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia* 126 (5), 1763-1768.
<https://doi.org/10.1213/ANE.0000000000002864>
- Son, Jungha und Kim, Boyoung (2023). Translation Performance from the User's Perspective of Large Language Models and Neural Machine Translation Systems. *Information* 14 (10), 574. <https://doi.org/10.3390/info14100574>
- Sperber, Matthias; de Seyssel, Maureen; Bao, Jiajun und Paulik, Matthias (2025). Toward Machine Interpreting: Lessons from Human Interpreting Studies. In: *Proceedings of*

- the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, 23338-23353. <https://doi.org/10.48550/arXiv.2508.07964>
- Springer Link Lehrbuch für Psychologie (2026). Gelegenheitsstichprobe oder Ad-hoc-Stichprobe. <https://www.lehrbuch-psychologie.springernature.com/glossar/gelegenheitsstichprobe-oder-ad-hoc-stichprobe> (Stand: 31.05.2026).
- Taylor, Paul (2009). *Text-to-speech synthesis*. Cambridge: Cambridge University Press.
- Tran, The Anh; Adnan, Azmat; Yihao, Wu und Siong, Chng Eng (2025). TS-VAD+: Modularized Target-Speaker Voice Activity Detection for Robust Speaker Diarization. In: *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 630-635.
<https://doi.org/10.1109/APSIPAASC65261.2025.11249081>
- Waibel, Alex und Fügen, Christian (2008). Spoken Language Translation. *IEEE Signal Processing Magazine* 25 (3), 71-79.
- Wang, Xiaoman und Fantinuoli, Claudio (2024). Exploring the Correlation between Human and Machine Evaluation of Simultaneous Speech Translation. In: Scarton, Carolina; Prescott, Charlotte; Bayliss, Chris; Oakley, Chris; Wright, Joanna; Wrigley, Stuart; Song, Xingyi; Gow-Smith, Edward; Bawden, Rachel; Sánchez-Cartagena, Víctor M.; Cadwell, Patrick; Lapshinova-Koltunski, Ekaterina; Cabarrao, Vera; Chatzitheodorou, Konstantinos; Nurminen, Mary; Kanojia, Diptesh und Moniz, Helena (Hg.) *Proceedings of the 25th Annual Conference of the European Association for Machine Translation* 1, 327-336. <https://aclanthology.org/2024.eamt-1.28/> (Stand 04.06.2026).
- Wang, Xiaoman und Wang, Binhua (2025). Advancing Automatic Assessment of Target-Language Quality in Interpreter Training with Large Language Models: Insights from Explainable AI. *The Interpreter and Translator Trainer* 19 (3-4), 465-485.
<https://doi.org/10.1080/1750399X.2025.2533015>
- Wu, Shao-Chuan (2010). *Assessing Simultaneous Interpreting: A Study on Test Reliability and Examiners' Assessment Behaviour*. Dissertation, Newcastle University.

- Xie, Tianxin; Rong, Yan; Zhang, Pengfei; Wang, Wenwu und Liu, Li (2025). Towards Controllable Speech Synthesis in the Era of Large Language Models. A Systematic Survey. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 764-791. <https://doi.org/10.18653/v1/2025.emnlp-main.40>
- Zeng, Zhen; Wang, Jianzong; Cheng, Ning und Xiao, Jing (2020). Prosody Learning Mechanism for Speech Synthesis System Without Text Length Limit. *Interspeech 2020*, 4422-4426. <https://doi.org/10.21437/Interspeech.2020-2053>
- Zwischenberger, Cornelia (2010). Quality Criteria in Simultaneous Interpreting: an International vs. a National View. *The Interpreters' Newsletter* 15, 127-142. https://www.researchgate.net/publication/302189394_Quality_criteria_in_simultaneous_interpreting_An_international_vs_a_national_view (Stand 03.06.2026).
- Zwischenberger, Cornelia (2013). *Qualität und Rollenbilder beim simultanen Konferenzdolmetschen*. Berlin: Frank & Timme.

Anhänge

Anhang 1: Ausgangsrede, orthografische Transkription

„Grazie presidente! Buongiorno Commissario Brunner. Europol ci segnala un incremento nell'uso criminale delle criptovalute, nel riciclaggio di denaro e per la richiesta dei riscatti dopo gli attacchi informatici. La Blockchain facilita trasferimenti rapidi di capitali a livello globale, offrendo ai criminali un vantaggio significativo. Nel campo della cybercriminalità emergono tecniche avanzatissime che richiedono competenze elevate degli investigatori. Tali competenze necessitano di personale adeguatamente formato. Inoltre, cresce l'uso di criptovalute ancorate al valore delle materie prime apprezzate dai capi criminali per la loro stabilità e facile comprensione. La mancanza di strumenti adeguati per il tracciamento delle criptovalute in alcuni Stati Membri sta facendo aumentare le richieste di supporto investigativo ad Europol. Di fronte a queste sfide e conoscendo la sua grande sensibilità è essenziale un impegno coordinato per sviluppare standard globali e condivisi al fine di combattere efficacemente l'uso illegale di questa tecnologia. Grazie.“

Anhang 2: Von einem Menschen angefertigte Dolmetschung, erweiterte orthografische Transkription

„Vielen Dank, Frau Präsidentin! Guten Tag, Herr Kommissar Brunner. (1,3) Es werden die Kryptowerte immer mehr (.) kriminell genutzt für Geldwäsche, für (1,4) IT-Angriffe und Lösegeld(.)for(.)derungen. Es (.) bieten diese Werte den (.) Kriminellen (.) Vorteile. Es entstehen da (.) fortschrittlichste Techniken, wo hohe Kompetenzen erforderlich sind. Und dies erfordert adäquat (.) geschultes Personal. (2,6) Also, äh (1,4) die (.) krimi(.)nellen (.) Kreise nutzen dies ganz gezielt. Wir (.) brauchen adäquate Instrumente. [auffällig schnell gesprochen] In einigen Mitglied[/auffällig schnell gesprochen]staaten (.) gibt es (.) immer mehr (.) Nach(.)fra(.)ge und auch Anfragen bei Europol. Das sind hoch(.)sensible Fragen. Wir brauchen hier koordinierte [auffällig schnell gesprochen] Anstrengungen [/auffällig schnell gesprochen] zur Entwicklung (.) glo(.)baler (1,3) Standards zur Bekämpfung der illegalen Nutzung dieser Technologien.“

Zeichen	Bedeutung
(.)	Hörbare Pause von weniger als einer Sekunde
(#,##)	Gemessene Pause von mehr als einer Sekunde, in Sekunden
[auffällig schnell gesprochen] ... [auffällig schnell gesprochen]	Passage, die auffällig schnell ausgesprochen wurde

Anhang 3: Maschinell angefertigte Dolmetschung, erweiterte orthografische Transkription

„Vielen Dank, Herr Präsident. Guten Morgen, Herr Bruner. (2,9) Europol meldet eine Zunahme der kriminellen Nutzung von Kryptowährungen, (.) der Geldwäsche und der Forderung nach Lösegeldern nach Cyberangriffen. (6,3) Blockchange erleichtert schnelle Kapitaltransfer weltweit, was Kriminellen einen erheblichen Vorteil im Bereich der Cyberkriminalität verschafft. [Geräusch: papierähnliches Rascheln] [auffällig schnell gesprochen] Es entstehen sehr fortgeschrittene Techniken [/auffällig schnell gesprochen]. [Geräusch: papierähnliches Rascheln] (1,6) Dies erfordert hohe Fähigkeiten der Ermittler. (.) Diese Fähigkeiten erfordern ausreichend geschultes Personal. (.) Darüber hinaus nimmt die Verwendung von Kryptowährungen, (.) die an den Wert von Rohstoffen gekoppelt sind (.) [auffällig schnell gesprochen] und von kriminellen Bossen für ihre Stabilität geschätzt werden [/auffällig schnell gesprochen], zu. [auffällig schnell gesprochen] Es ist leicht zu verstehen [/auffällig schnell gesprochen]. Der Mangel an geeigneten Tools zur Verfolgung von Kryptowährungen. (5,7) In einigen Mitgliedstaaten er sucht Europol zunehmend um Unterstützung bei den Ermittlungen. (2,4) Angesichts dieser Herausforderungen und im Wissen um ihre große Sensibilität ist eine koordinierte Anstrengung unerlässlich, (.) Entwicklung globaler und gemeinsamer Standards erforderlich, um die illegale Nutzung dieser Technologie wirksam zu bekämpfen.“

Zeichen	Bedeutung
(.)	Hörbare Pause von weniger als einer Sekunde
(#,##)	Gemessene Pause von mehr als einer Sekunde, in Sekunden
[Geräusch: ...]	Hörbares auffälliges nichtsprachliches Geräusch
[auffällig schnell gesprochen] ... [/auffällig schnell gesprochen]	Passage, die auffällig schnell ausgesprochen wurde

Anhang 4: Gegenüberstellung der Gruppen anhand M, SD und ΔM

Die Aussagen „Pausen/Verzögerungen haben mich beim Zuhören gestört.“ und „Es war anstrengend, der Dolmetschung zu folgen.“ wurden invertiert – siehe Kapitel 3.6.

	MD		AD		ΔM
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Ich konnte der Dolmetschung inhaltlich gut folgen.	4,00	0,87	3,24	1,05	0,76
Die Dolmetschung war kohärent.	3,60	0,91	2,68	1,07	0,92
Die Dolmetschung war sprachlich korrekt.	4,12	0,73	3,36	1,22	0,76
Die Vortragsweise der Dolmetschung war flüssig.	2,24	0,97	2,24	1,13	0,00
Die Formulierungen wirkten natürlich.	2,76	1,23	1,92	1,11	0,84
Die Stimme war angenehm zu hören.	2,84	1,07	2,36	1,15	0,48
Pausen/Verzögerungen haben mich beim Zuhören gestört.	1,96	1,06	1,84	1,34	0,12
Das Sprechtempo war angenehm.	2,84	1,03	2,48	1,08	0,36
Es war anstrengend, der Dolmetschung zu folgen.	2,72	1,34	2,32	1,22	0,40
Die Dolmetschung wirkte auf mich professionell.	2,96	1,17	2,24	1,30	0,72
Die Dolmetschung wirkte auf mich vertrauenswürdig.	3,16	1,18	2,40	1,19	0,76
Die Dolmetschung war insgesamt von hoher Qualität.	2,88	1,01	2,16	1,11	0,72

Anhang 5: Antworten der MD-Gruppe auf den Fragebogen

Teilnehmer*in	Punkt des Fragebogens nach untenstehender Legende																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
MD01	3	4	5	1	4	3	4	1	4	3	4	2	1	5	w	35	Master/Magister
MD02	5	4	4	4	3	4	4	5	1	4	5	4	5	1	w	62	Bachelor
MD03	4	4	5	2	3	3	5	3	4	3	3	3	3	2	w	27	Bachelor
MD04	4	3	4	1	1	3	5	2	5	2	2	2	2	1	w	57	Matura
MD05	5	5	5	3	5	4	3	4	2	4	5	3	1	3	w	31	Bachelor
MD06	4	4	4	4	5	4	2	4	2	5	5	4	1	1	w	27	Matura
MD07	5	3	4	2	3	2	5	2	4	1	3	3	3	1	m	30	Master/Magister
MD08	4	4	4	2	4	2	4	3	1	3	3	4	2	1	m	33	Master/Magister
MD09	4	3	4	3	3	4	5	4	4	4	3	4	3	1	w	31	Matura
MD10	5	5	4	3	4	3	3	3	2	5	5	4	3	1	m	31	Master/Magister
MD11	2	2	3	1	1	1	5	2	4	1	2	1	5	2	w	30	Master/Magister
MD12	5	5	5	3	3	3	4	3	2	3	3	3	2	1	w	31	Matura
MD13	5	5	4	1	2	1	5	1	4	2	2	3	4	1	m	36	Bachelor
MD14	3	3	4	2	1	2	5	3	4	2	3	2	5	1	m	27	Lehrabschluss
MD15	4	3	4	2	1	4	2	3	5	3	2	1	5	1	m	62	Lehrabschluss
MD16	5	4	5	4	4	4	3	5	1	5	5	4	4	1	w	59	Lehrabschluss
MD17	3	3	3	2	2	4	5	2	4	2	2	3	2	4	w	26	Master/Magister
MD18	3	2	2	1	1	4	5	3	5	2	4	2	2	4	m	27	Bachelor
MD19	4	4	5	3	2	1	4	2	3	2	2	3	4	2	w	26	Matura
MD20	5	3	4	2	3	3	5	2	2	2	2	2	4	1	m	28	Bachelor
MD21	4	4	4	3	4	3	2	3	3	3	3	3	3	3	w	33	Master/Magister
MD22	3	3	5	2	3	3	4	3	5	4	4	4	3	1	w	30	Master/Magister
MD23	3	2	4	1	2	1	5	2	5	2	1	1	4	1	w	40	Master/Magister
MD24	4	4	4	2	2	2	3	3	3	4	3	4	5	1	m	32	Matura
MD25	4	4	4	2	3	3	4	3	3	3	3	3	3	1	w	35	Matura

Für die vorstehende Tabelle gilt folgende Legende:

Punkt	Item des Fragebogens
1	Ich konnte der Dolmetschung inhaltlich gut folgen.
2	Die Dolmetschung war kohärent.
3	Die Dolmetschung war sprachlich (Grammatik, Wortwahl, Satzbau) korrekt.
4	Die Vortragsweise der Dolmetschung war flüssig.
5	Die Formulierungen wirkten natürlich.
6	Die Stimme war angenehm zu hören.
7	Pausen/Verzögerungen haben mich beim Zuhören gestört.
8	Das Sprechtempo war angenehm.
9	Es war anstrengend, der Dolmetschung zu folgen.
10	Die Dolmetschung wirkte auf mich professionell.
11	Die Dolmetschung wirkte auf mich vertrauenswürdig.
12	Die Dolmetschung war insgesamt von hoher Qualität.
13	Glauben Sie, die Dolmetschung könnte von einer Maschine erzeugt worden sein?
14	Wie viel Erfahrung haben Sie mit dem Thema Simultandolmetschen?
15	Mit welchem Geschlecht identifizieren Sie sich?
16	Wie alt sind Sie?
17	Was ist Ihr höchster Bildungsabschluss?

Anhang 6: Antworten der AD-Gruppe auf den Fragebogen

Teilnehmer*in	Punkt des Fragebogens nach untenstehender Legende																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
AD01	4	4	5	4	3	3	1	3	1	4	4	3	4	5	w	35	Master/Magister
AD02	2	2	1	1	1	2	3	2	4	2	1	2	5	2	w	29	Matura
AD03	4	2	3	2	1	3	5	3	4	1	3	2	4	1	m	33	Matura
AD04	4	2	2	2	1	3	5	2	3	1	3	1	5	1	m	41	Master/Magister
AD05	1	2	5	2	3	2	5	2	5	2	1	2	5	1	w	27	Matura
AD06	4	3	3	1	2	1	5	2	4	4	2	2	2	5	w	32	Master/Magister
AD07	4	4	3	2	1	1	5	1	5	2	3	1	5	1	w	36	Bachelor
AD08	3	2	3	2	1	2	5	2	4	2	3	2	5	1	w	26	Bachelor
AD09	3	2	4	1	1	2	5	1	5	1	1	1	5	3	w	30	Master/Magister
AD10	2	2	3	3	1	2	5	3	4	2	2	1	5	2	w	46	Master/Magister
AD11	2	1	2	1	2	1	5	1	5	1	2	1	5	3	m	29	Master/Magister
AD12	4	4	3	4	2	2	4	3	3	2	2	2	5	1	m	35	Bachelor
AD13	3	3	4	2	2	3	4	3	3	4	4	4	4	2	m	34	Bachelor
AD14	3	3	4	3	2	1	1	2	2	3	2	2	4	1	w	34	Bachelor
AD15	4	5	5	4	4	3	3	3	2	3	3	4	5	1	m	27	Lehrabschluss
AD16	5	5	5	5	5	5	1	5	1	5	5	5	1	1	w	59	Lehrabschluss
AD17	4	2	4	2	1	2	4	3	4	1	2	2	5	1	m	32	Master/Magister
AD18	4	3	4	2	2	3	5	2	4	2	2	2	5	1	w	30	Matura
AD19	3	2	1	1	1	4	5	2	4	1	2	2	5	3	w	32	Master/Magister
AD20	3	3	5	3	4	5	5	5	3	5	5	4	3	1	m	43	Master/Magister
AD21	1	1	3	1	2	1	5	1	5	1	1	1	5	1	m	40	Master/Magister
AD22	3	3	4	3	2	3	4	2	3	3	2	3	4	1	w	62	Bachelor
AD23	5	2	2	2	1	1	5	2	5	1	1	1	5	1	m	62	Meisterprüfung
AD24	3	3	4	2	2	2	4	3	4	2	3	2	4	1	w	36	Matura
AD25	3	2	2	1	1	2	5	4	5	1	1	2	5	5	w	30	Master/Magister

Für die vorstehende Tabelle gilt folgende Legende:

Punkt	Item des Fragebogens
1	Ich konnte der Dolmetschung inhaltlich gut folgen.
2	Die Dolmetschung war kohärent.
3	Die Dolmetschung war sprachlich (Grammatik, Wortwahl, Satzbau) korrekt.
4	Die Vortragsweise der Dolmetschung war flüssig.
5	Die Formulierungen wirkten natürlich.
6	Die Stimme war angenehm zu hören.
7	Pausen/Verzögerungen haben mich beim Zuhören gestört.
8	Das Sprechtempo war angenehm.
9	Es war anstrengend, der Dolmetschung zu folgen.
10	Die Dolmetschung wirkte auf mich professionell.
11	Die Dolmetschung wirkte auf mich vertrauenswürdig.
12	Die Dolmetschung war insgesamt von hoher Qualität.
13	Glauben Sie, die Dolmetschung könnte von einer Maschine erzeugt worden sein?
14	Wie viel Erfahrung haben Sie mit dem Thema Simultandolmetschen?
15	Mit welchem Geschlecht identifizieren Sie sich?
16	Wie alt sind Sie?
17	Was ist Ihr höchster Bildungsabschluss?

Anhang 7: Antworten der MD-Gruppe auf die offenen Fragen

Teilnehmer*in	Gab es etwas, das Sie an der Dolmetschung besonders gut fanden?
MD01	
MD02	Das Tempo war angenehm
MD03	
MD04	
MD05	
MD06	
MD07	
MD08	
MD09	Die Tonalität der Stimme war sehr angenehm.
MD10	Obwohl das Tempo teils stockend wirkte, würde ich das auf die vorsichtige Wortwahl der Dolmetscherin zurückführen, wodurch die Dolmetschung insgesamt als sehr genau und bedacht beschrieben werden kann.
MD11	Nein, wirkte wie KI
MD12	
MD13	
MD14	
MD15	Nein
MD16	
MD17	Die Sätze wurden trotz langer Pausen dann meistens doch korrekt beendet.
MD18	langsame Geschwindigkeit
MD19	
MD20	
MD21	
MD22	Tonlage
MD23	
MD24	
MD25	Nein, gab es nicht

Teilnehmer*in	Gab es etwas, das Sie an der Dolmetschung besonders schlecht fanden?
MD01	Die Hesitationen und Pausen.
MD02	Die Verzögerungen waren irritierend
MD03	
MD04	abgehackte Sprechweise
MD05	Häufiges Ähm dazwischen
MD06	
MD07	Pausen, stottern
MD08	
MD09	
MD10	Für nicht geübte Ohren kann das langsamere Tempo zu einem Bruch im roten Faden führen.
MD11	Zu lange Pausen in den Sätzen, das störte beim zuhören
MD12	
MD13	
MD14	
MD15	Meine erste Dolmetschung!
MD16	
MD17	Die Pausen waren manchmal zu lange, was es schwieriger machte, dem Inhalt zu folgen.
MD18	Es fehlte die Phonetik
MD19	
MD20	nicht flüssig und zu langsam
MD21	
MD22	Zu viele Pausen im Satz
MD23	Mein erster Gedanke war: ist das KI generiert?
MD24	
MD25	die Pausen

Anhang 8: Antworten der AD-Gruppe auf die offenen Fragen

Teilnehmer*in	Gab es etwas, das Sie an der Dolmetschung besonders gut fanden?
AD01	Dass es keine Hesitationen und Füllwörter gab.
AD02	
AD03	Ruhige Stimme, deutliche Aussprache.
AD04	
AD05	
AD06	
AD07	
AD08	
AD09	
AD10	
AD11	Als Hörer nicht. Als Kunde könnte ich mir vorstellen, dass die Preisgestaltung wirtschaftlich attraktiv ist.
AD12	Klare wörter
AD13	Sprach tempo war gut, in der annahme das es eine Live dolmetschung war.
AD14	
AD15	
AD16	
AD17	Die Stimme war an sich angenehm
AD18	
AD19	Stimm Lage war angenehm
AD20	
AD21	
AD22	Nein
AD23	
AD24	Gab es nicht
AD25	Die Terminologie klang naheliegend.

Teilnehmer*in	Gab es etwas, das Sie an der Dolmetschung besonders schlecht fanden?
AD01	Wortwahl und teilweise zu schnell.
AD02	Komisches Bauchgefühl! Unrealistisch - unglaubwürdig
AD03	Pausenplatzierung ist seltsam. Konnte schlecht dem Fluss folgen. Übersetztes ergab nicht immer Sinn in der ganzen Story.
AD04	
AD05	
AD06	
AD07	Die Pausen und das abgehackte Sprechen
AD08	Verzögerungen
AD09	
AD10	
AD11	Die hohen Schwankungen bei der Sprechgeschwindigkeit, einige kurze Sätze oder Teilsätze, die sich nicht in die Struktur der Rede einfügen und deren Aussage im Kontext sich daher nicht erschloss, und die - auch für maschinell erzeugte Stimmen - schlechte und unnatürliche Sprachmelodie.
AD12	Die Stimme und Formulierung der Sätze
AD13	Die Aussprache von einigen Wörtern war etwas ungewöhnlich, wodurch ich minimal länger gebraucht habe um den Inhalt initial zu verstehen. Aber die Aussprache war konsistent wodurch spätere Wiedererkennung der gleichen Wörter OK war.
AD14	
AD15	
AD16	
AD17	Unterschiedliche Geschwindigkeit der Dolmetschung
AD18	Das Tempo war mal sehr schnell und dann wieder total langsam, das hat es sehr schwierig gemacht den Faden nicht zu verlieren in den langen Pausen.
AD19	es wirkte sehr maschinell
AD20	
AD21	Geschwindigkeitsveränderungen, Pausen, abgehackt Sprache, inhaltlich schwer zu folgen
AD22	Tempowechsel
AD23	Die Pausen und die Stimme
AD24	Sprechtempovariation
AD25	Die langen Pausen und die unvollständigen Sätze.

Abstract Deutsch

Die vorliegende Arbeit untersucht die rezipient*innenbasierte Qualitätswahrnehmung einer von einem gegenwärtig handelsüblichen Dolmetschsystem erzeugten und einer von einer professionellen Dolmetscherin angefertigten Dolmetschleistung derselben Ausgangsrede durch Personen, die nicht als Dolmetscher*innen tätig sind. Im Mittelpunkt steht die Frage, ob und in Hinblick auf welche Qualitätsdimensionen sich die Bewertung der beiden Dolmetschungen unterscheidet. Dazu wurde eine Online-Befragung mit zwei Gruppen durchgeführt. Die Teilnehmenden rezipierten pro Gruppe entweder die maschinell generierte oder die von einem Menschen angefertigte Dolmetschleistung und bewerteten diese anschließend anhand von Likert-Items und einer Zustimmungsskala sowie offenen Fragen. Die Ergebnisse zeigen, dass beiden Leistungen zwar insgesamt mittelmäßige Qualität beigemessen wurde, die von einem Menschen angefertigte Dolmetschung jedoch über nahezu alle Qualitätskategorien hinweg positiver wahrgenommen wurde. Die markantesten Unterschiede zwischen den Bewertungen bezogen sich primär auf die wahrgenommene Kohärenz, Natürlichkeit und inhaltliche Nachvollziehbarkeit. Pausenverhalten beziehungsweise die Verzögerungen wurden gruppenübergreifend am stärksten bemängelt und der Aspekt der Flüssigkeit wurde ident beurteilt. Ferner wurden gruppenübergreifende positive Korrelationen – wenngleich in unterschiedlichem Ausmaß – zwischen den Aspekten der Natürlichkeit beziehungsweise Stimmqualität und dem Gesamturteil festgestellt. Gleichmaßen zeigte sich, dass der rezipierten Leistung eine umso geringere Qualität beigemessen wurde, je sicherer sich die Rezipient*innen waren, dass sie eine maschinelle Leistung rezipiert haben.

Abstract English

This paper explores the perception of quality of an interpretation generated by a currently available machine interpreting system and an interpretation produced by a professional interpreter based on the same source material – each as perceived by recipients who are not working as interpreters themselves. The focus lies on assessing whether quality perceptions differ between the two interpretations and, if so, in terms of which quality dimensions. For this purpose, an online survey was conducted using two groups. The participants of each group received either the machine- or human-generated interpretation which they subsequently evaluated based on Likert Items and a Likert Scale as well as open questions. The findings reveal that both interpretations were deemed to be of mediocre quality, yet the human-generated interpretation achieved superior scores across almost all quality criteria. The most prominent differences between the assessments primarily relate to perceived coherence, naturalness and whether recipients were able to comprehend the underlying content. Across both groups, the use of pauses and delays in the delivery were criticized the most while fluidity was perceived identically. Further, positive correlations were identified among both groups, albeit each to a different extent, between the overall perceived quality on the one hand and naturalness or quality of voice on the other hand. Similarly, this study revealed that greater certainty among recipients that they had listened to a machine-generated interpretation correlates with lower perceived quality.