MASTERARBEIT | MASTER'S THESIS

Titel | Title

Turing-Test der Übersetzung: Unterscheidung zwischen Human- und maschineller Übersetzung von Marketingtexten

verfasst von | submitted by

Sebastian Guttmann BA

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

Master of Arts (MA)

Wien | Vienna, 2026

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

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die Übersetzung von Marketingtexten nimmt in der internationalen Markenkommunikation eine zentrale Rolle ein (vgl. Stiegelbauer et al. 2024; Torresi 2021). Im Unterschied zu rein informativen Textsorten verfolgen Marketingtexte häufig eine persuasive Funktion und sollen Leser:innen gezielt ansprechen (vgl. Reiß 1971). Dadurch entstehen besondere Anforderungen an die Übersetzung, da neben der semantischen Genauigkeit auch Tonalität, Stil, Kreativität und adressatengerechte Formulierungen berücksichtigt werden müssen (vgl. Nord 1997; Reiß 1971).

Gleichzeitig werden maschinelle Übersetzungssysteme und Large Language Models (LLMs) zunehmend für Übersetzungszwecke eingesetzt (vgl. ELIS 2024). Diese Entwicklung betrifft nicht nur professionelle Übersetzer:innen, sondern auch Unternehmen, Auftraggeber:innen und private Nutzer:innen, die maschinell generierte Zieltexte erstellen, bearbeiten oder bewerten (vgl. ELIS 2025). Vor diesem Hintergrund stellt sich nicht nur die Frage, welche sprachliche Qualität maschinell erstellte Übersetzungen erreichen, sondern vor allem auch, wie diese von Rezipient:innen wahrgenommen werden. Besonders relevant ist dabei, ob Leser:innen erkennen können, ob ein Zieltext von einem Menschen oder maschinell erstellt wurde.

Während objektive Qualitätsbewertungen maschineller Übersetzungen wichtige Hinweise auf sprachliche Fehler oder Abweichungen liefern können, erfassen sie nur bedingt, wie Zieltexte tatsächlich auf Leser:innen wirken (vgl. Agrawal 2024; Kocmi et al. 2021). Gerade im Bereich der Marketingkommunikation ist jedoch die Wahrnehmung durch Rezipient:innen entscheidend, da diese mitbestimmt, ob ein Text als überzeugend, professionell und markenkonform wahrgenommen wird (vgl. Stiegelbauer et al. 2024; Torresi 2021).

Anknüpfend an diese Überlegungen untersucht die vorliegende Arbeit, inwieweit Humanübersetzungen und maschinell erstellte Übersetzungen von Marketingtexten aus dem Englischen ins Deutsche von Rezipient:innen korrekt als human oder maschinell übersetzt kategorisiert werden können. Bestehende Untersuchungen zur Wahrnehmung maschineller Übersetzung (z. B. Guerberof-Arenas & Toral 2022, 2024; Hassan et al. 2018; Toral & Way 2018) beziehen sich häufig auf informative oder literarische Textsorten oder auf nur sehr kurze Textausschnitte von wenigen Wörtern oder einzelnen Sätzen. Darüber hinaus beschränken sie sich häufig auf die Wahrnehmung von professionellen Übersetzer:innen (z. B. Aslan 2023; Delorme Benites et al. 2021; Guerberof-Arenas & Toral 2022). Für die vorliegende Arbeit ist jedoch

nicht nur die Einschätzung durch Sprachexpert:innen relevant, da auch Lai:innen zur Leser:innenschaft von Marketingtexten gehören und zugleich häufig als Nutzer:innen, Auftraggeber:innen oder Rezipient:innen maschineller Übersetzungen auftreten (vgl. ELIS 2025, 2026).

Die Untersuchung orientiert sich dabei am Grundgedanken des Turing-Tests (Turing 1950), bei dem geprüft wird, ob maschinelle Ausgaben von humanen Ausgaben unterschieden werden können. Übertragen auf den vorliegenden Untersuchungskontext steht somit nicht die objektive Messung von Übersetzungsqualität im Vordergrund, sondern die Wahrnehmung der Übersetzungsherkunft durch die Teilnehmer:innen.

Die Untersuchung basiert auf acht englischsprachigen Marketingtextausschnitten aus unterschiedlichen Branchen. Für jeden Textausschnitt wurden drei deutsche Zieltextvarianten erstellt, welche eine Humanübersetzung, eine Übersetzung mit einem maschinellen Übersetzungssystem sowie eine Übersetzung mit einem LLM umfassen. Konkret wurden hierfür DeepL und ChatGPT eingesetzt. Die Zieltexte wurden zwei Teilnehmer:innengruppen, Sprachexpert:innen und Lai:innen, vorgelegt. Als Sprachexpert:innen gelten Personen mit abgeschlossenem einschlägigem Sprachstudium oder mindestens dreijähriger Berufserfahrung in einem relevanten sprachbezogenen Bereich, etwa Übersetzen, Copywriting oder vergleichbaren Tätigkeiten. Den Teilnehmer:innen wurde in unterschiedlichen Sets jeweils eine der Übersetzungsvarianten vorgelegt und sie wurden angewiesen, zunächst einzuschätzen, ob es sich bei dem jeweiligen Zieltext um eine Humanübersetzung oder eine maschinelle Übersetzung handelt. Anschließend bewerteten sie die Texte anhand verschiedener Kategorien, darunter Flüssigkeit, Natürlichkeit und weitere wahrnehmungsbezogene Qualitätsmerkmale. Ausgehend von dieser Zielsetzung wurden folgende Forschungsfragen untersucht:

FF1: Inwieweit kategorisieren Lai:innen und Sprachexpert:innen Humanübersetzungen und von maschinellen Übersetzungssystemen und LLMs angefertigte Übersetzungen von Marketingtexten aus dem Englischen ins Deutsche korrekt?

FF2: Unterscheidet sich die Wahrnehmung maschinell erstellter Übersetzungen von maschinellen Übersetzungssystemen und LLMs?

FF3: Unterscheidet sich die Wahrnehmung maschineller Übersetzungen zwischen Sprachexpert:innen und Lai:innen?

Ausgehend von den Forschungsfragen werden leitende Grundannahmen formuliert, die der Strukturierung der Analyse dienen. Da vorangehende Studien zeigen, dass kreative Textsorten

weiterhin eine Herausforderung für die maschinelle Übersetzung darstellen können (vgl. Guerberof-Arenas & Toral 2022, 2024; Toral & Way 2018), wird angenommen, dass die Kreativitäts- bzw. Stilkomplexität der Textausschnitte die Klassifikation beeinflussen kann. Maschinell erstellte Übersetzungen könnten daher bei stärker kreativen oder stilistisch komplexen Texten eher als solche erkannt werden als bei weniger komplexen Textausschnitten. Im Hinblick auf das verwendete System wird angenommen, dass sich die Klassifikationsrate zwischen dem maschinellen Übersetzungssystem und dem LLM nicht grundlegend unterscheidet. Diese Annahme ist darauf zurückzuführen, dass heutige maschinelle Übersetzungssysteme zwar zu akkurateren, dafür ausgangstexttreueren und stilistisch oft starren Übersetzungen tendieren (vgl. Kamaluddin et al. 2024), während LLMs häufig flüssigere, dafür freiere und nicht immer akkurate Übersetzungen generieren (vgl. Ji et al. 2024; Sizov et al. 2024). Schließlich wird angenommen, dass Teilnehmer:innen mit sprachwissenschaftlicher Ausbildung oder mehrjähriger beruflicher Erfahrung im Bereich Textarbeit mehr Feingefühl und Sensibilität für Sprache im Allgemeinen entwickelt haben und sich intensiver mit den Texten beschäftigen (vgl. Kaspere et al. 2023) und dadurch andere Klassifikations- und Bewertungstendenzen aufweisen als Lai:innen.

Die Ergebnisse können Hinweise darauf liefern, wie Humanübersetzungen und maschinell erstellte Übersetzungen im Bereich der Marketingkommunikation wahrgenommen werden und ob die Einschätzung als human- bzw. maschinell übersetzter Text einen Einfluss auf die Qualitätsbewertung hat. Für die Translationswissenschaft ist dies relevant, weil dadurch die subjektive Wahrnehmung von Übersetzungsherkunft sowie der Zusammenhang zwischen geschätzter Herkunft und Qualität stärker in den Mittelpunkt rückt. Für professionelle Übersetzer:innen und Sprachdienstleister:innen könnte es von Interesse sein, ob und in welchen Bereichen maschinell generierte Zieltexte als menschlich wahrgenommen werden, was wiederum Hinweise auf mögliche Einsatzbereiche und Grenzen dieser Systeme im übersetzerischen Berufsalltag geben kann. Für Entwickler:innen von Übersetzungssystemen und LLMs können die Ergebnisse Hinweise darauf geben, welche sprachlichen Eigenschaften die Wahrnehmung von Übersetzungsqualität beeinflussen. Darüber hinaus können die Ergebnisse auch für Auftraggeber:innen relevant sein, die maschinelle Übersetzungssysteme oder LLMs für Marketingtexte einsetzen oder deren Einsatz erwägen.

Die vorliegende Arbeit ist wie folgt gegliedert. Kapitel 2 stellt zunächst den relevanten theoretischen Rahmen vor, bevor in Kapitel 3 der aktuelle Stand der Forschung näher erläutert wird. Anschließend wird in Kapitel 4 die Methodik beschrieben, bevor in Kapitel 5 und 6 die

Ergebnisse präsentiert und diskutiert werden. Abschließend werden in Kapitel 7 die zentralen Erkenntnisse zusammengefasst und ein Ausblick auf mögliche weitere Forschung gegeben.

2 Theoretische Grundlagen

Im folgenden Kapitel wird der theoretische Rahmen dieser Arbeit ausgeführt. Zunächst werden der Begriff der maschinellen Übersetzung sowie ihre historische Entwicklung beschrieben. Darauf aufbauend wird ein Überblick über die verschiedenen Architekturen der maschinellen Übersetzung sowie konkrete Anwendungen dieser vorgestellt. Anschließend werden Qualitätskonzepte sowie spezifische Bewertungsansätze von maschinellen Übersetzungen behandelt, bevor die Wahrnehmung durch Sprachexpert:innen und Lai:innen diskutiert wird. Abschließend wird der Turing-Test vorgestellt, der als theoretische Grundlage für den empirischen Teil dieser Arbeit eine zentrale Rolle einnimmt.

2.1 Maschinelle Übersetzung – Begriff und Abgrenzung

Seit den Anfängen der maschinellen Übersetzung in der ersten Hälfte des 20. Jahrhunderts wurden zahlreiche Versuche angestellt, den Begriff der Maschinellen Übersetzung (MÜ) zu definieren. Zunächst sollte erwähnt werden, dass insbesondere in der frühen Geschichte der MÜ zahlreiche Benennungen für denselben Begriff verwendet wurden. Neben maschinelle und computergestützte wurde auch automatische Übersetzung verwendet, was eine einheitliche Terminologie in der frühen Forschung erschwerte. Eine sehr bekannte Definition stammt von Hutchins (1995: 431): „The term 'machine translation' (MT) refers to computerized systems responsible for the production of translations with or without human assistance“. Nach dieser Definition ist der grundlegende Bestandteil einer maschinellen Übersetzung ein Computer, der Texte von einer Sprache in eine andere übersetzt. Diese Übersetzung kann entweder eigenständig oder mit der Unterstützung eines Menschen erfolgen. Oft verschwimmen insbesondere in Bezug auf die Rolle des Menschen jedoch die Grenzen zwischen tatsächlicher maschineller Übersetzung und computergestützter Übersetzung, weshalb Hutchins (1995: 431) seine Definition noch erweitert: „It excludes computer-based translation tools which support translators by providing access to online dictionaries, remote terminology databanks, transmission and reception of texts, etc.“. Während bei der computergestützten Übersetzung Programme wie Computer-Assisted Translation (CAT)-Tools oder Online-Glossare gemeint sind, die für die Unterstützung im Humanübersetzungsprozess konzipiert wurden, hat die maschinelle Übersetzung einen vollautomatisierten Übersetzungsprozess zum Ziel, bei dem der Maschinenoutput abgeschlossen ist, bevor er von Menschen bearbeitet wird, etwa in Form von Post-Editing (vgl. Hutchins 1995).

Kenny (2022: 32) beschreibt die maschinelle Übersetzung als automatische Produktion eines Zieltextes auf Basis eines Ausgangstextes. Kennys Definition überschneidet sich daher stark mit jener von Hutchins, ein prägnanter Unterschied ist jedoch, dass in Hutchins Definition explizit auch die potenzielle Rolle des Menschen in der maschinellen Übersetzung hervorgehoben wird, während Kennys Definition rein maschinenbasiert und dadurch vollautomatisch ist.

Hauenschild (2004: 756) kritisiert, dass unter dem Begriff Übersetzen im maschinellen Kontext häufig eigentlich Translation im Allgemeinen gemeint ist, da auch das maschinelle Dolmetschen oft synonym behandelt wird. Es wird daher hervorgehoben, dass in dieser Arbeit ausschließlich das Übersetzen, also die schriftliche Translation, behandelt wird. Die in diesem Kapitel dargestellten Definitionen bilden nur einen ausgewählten Ausschnitt der vielfältigen Begriffsbestimmungen maschineller Übersetzung ab. Sie erheben keinen Anspruch auf Vollständigkeit, sondern dienen als einführende begriffliche Grundlage für die weitere Arbeit.

Zusammenfassend kann gesagt werden, dass die verschiedenen angeführten Definitionen sehr ähnlich sind, jedoch je nach Ansatz mehr auf bestimmte Aspekte der maschinellen Übersetzung eingehen. Für diese Arbeit wird die maschinelle Übersetzung im Sinne von Hutchins als die vollautomatische Generierung eines Zieltextes aus einem schriftlichen Ausgangstext mittels Computersoftware verstanden. Etwaige menschliche Eingriffe erfolgen vor oder nach dem maschinellen Prozess und sind daher nicht Bestandteil der eigentlichen maschinellen Übersetzung.

2.2 Geschichtliche Entwicklung der maschinellen Übersetzung

Dieses Kapitel behandelt die zentralen Entwicklungen der maschinellen Übersetzung und zeigt, wie technologische Grenzen und Durchbrüche jeweils neue Paradigmen hervorgebracht haben. Die Darstellung bildet damit die Grundlage für das Verständnis zentraler Methoden und Ansätze moderner MÜ-Systeme.

Bei der historischen Darstellung der maschinellen Übersetzung ist eine Abgrenzung zu CAT-Tools und allgemeinen computerlinguistischen Entwicklungen notwendig, da sich technische Grundlagen und eigentliche MÜ-Forschung historisch oft überschneiden (vgl. Rothkegel 2004).

Als Anfangspunkt der modernen maschinellen Übersetzung werden häufig Dechiffriertechniken herangezogen, die im Zweiten Weltkrieg von England benutzt wurden, um deutsche Nachrichten automatisiert zu decodieren (vgl. Lennon 2014; Rothkegel 2004). Es handelt sich

hierbei zwar nicht um Übersetzung zwischen natürlichen Sprachen, die systematische Entschlüsselung von Codierungssprachen kann jedoch als kritischer Schritt hinsichtlich der Automatisierung sprachbezogener Aufgaben bezeichnet werden.

Ein Vorläufer der maschinellen Übersetzung sind die Übersetzungsmaschinen von Petr Smirnov-Trojanskij und Georges Artsrouni, welche 1933 unabhängig voneinander zum Patent angemeldet wurden (vgl. Lennon 2014; Rothkegel 2004; Schwartz 2017). Die Maschinen können als mechanische Wörterbücher beschrieben werden, die den Übersetzungsprozess unterstützen sollten, ohne selbstständig zu übersetzen. Auch hier handelt es sich noch nicht um eine MÜ-Anwendung, jedoch können die Maschinen als frühes Beispiel einer mechanisierten Übersetzungsanwendung für die Disziplin als wegweisend betrachtet werden. Darüber hinaus schlug Smirnov-Trojanskij ein Zwischensystem vor, mit Hilfe dessen menschliche Gedanken unabhängig von der Sprache dargestellt werden können. Dieser Ansatz erinnert bereits stark an spätere Interlingua-Modelle (vgl. Lennon 2014).

Mit der Entwicklung der ersten Computer erschlossen sich neue Möglichkeiten für die maschinelle Übersetzung. Der Mathematiker und Ingenieur Warren Weaver gilt als einer der Pioniere der MÜ in dieser Zeit. Zusammen mit Andrew Booth erarbeitete er aufbauend auf zeitgenössischer Computertechnologie erste konzeptuelle Ansätze und stützte sich auf die Annahme, dass kryptografische Verfahren zur Übersetzung natürlicher Sprachen verwendet werden können (vgl. Hutchins 2001a; Lennon 2014; Rothkegel 2004). Weavers (1949) 1949 erschienenes Memorandum gilt als Startpunkt der modernen MÜ-Forschung und trug maßgeblich zur Förderung und Forschungsaktivität in den USA in den folgenden Jahrzehnten bei. Bemerkenswert ist Weavers Einschätzung, dass die maschinelle Übersetzung von literarischen Texten potenziell zu herausfordernd gesehen werden kann. Technische Dokumente hingegen können maßgeblich von maschinellen Übersetzungsprozessen profitieren (vgl. Schwartz 2017).

1951 wurde für Yehoshua Bar-Hillel am Massachusetts Institute of Technology (MIT) die erste vollfinanzierte Forschungsposition für maschinelle Übersetzung eingerichtet und 1952 fand die erste MÜ-Konferenz in den USA statt (vgl. Hutchins 2001a). Da bereits weitgehend bekannt war, dass die maschinelle Übersetzung zumindest auf absehbare Zeit noch menschliche Vor- oder Nachbearbeitung benötigte, war das Hauptthema, wie dies am besten und effizientesten umgesetzt werden konnte. Zur Sicherung weiterer Förderungen wurde 1954 von IBM in Zusammenarbeit mit der Universität von Georgetown eine Übersetzungsdemonstration durchgeführt, die viel mediale Beachtung erhielt (vgl. Lennon 2014). Das Experiment wurde als Erfolg wahrgenommen, was zu hohen staatlichen, militärischen und privaten Förderungen in den

USA sowie zu den Gründungen erster Forschungsgruppen in der Sowjetunion führte, und markierte den Beginn des goldenen Zeitalters der maschinellen Übersetzung (vgl. Hutchins 2001a; Schwartz 2017).

In den 1950er- und frühen 1960er-Jahren florierte die Forschungstätigkeit im Bereich der MÜ. Weltweit entstanden zahlreiche Forschungsgruppen, wenngleich aufgrund der geopolitischen Lage der Fokus auf dem Sprachenpaar Englisch–Russisch lag. Zunehmend etablierte sich die RegelBasierte Maschinelle Übersetzung (RBMÜ) als dominierendes Paradigma (vgl. Hutchins 2001a). Die RBMÜ lässt sich in die drei grundlegenden Ansätze des direkten, des Transfer- und des Interlingua-Ansatzes unterteilen (vgl. Koehn 2010; Schwartz 2017). Die technischen Funktionsweisen dieser Ansätze werden im Kapitel 2.3 noch genauer ausgeführt.

Sehr schnell kristallisierten sich zwei Strömungen heraus. Zum einen gab es einen empirisch angelegten Ansatz, bei dem das Ziel nicht eine makellose automatische Übersetzung, sondern vielmehr eine brauchbare Rohübersetzung war (vgl. Hutchins 2001a). Für dieses Vorhaben eignete sich der direkte Ansatz sehr gut, da mit wenig Analyseaufwand schnell Output generiert werden konnte.

Parallel dazu entwickelten sich Forschungsgruppen, die theoriebasiert arbeiteten und der Linguistik eine bedeutende Rolle zusprachen (vgl. Hutchins 2001a). Der Fokus lag bei dieser Strömung auf syntaktischen und semantischen Analysen, wodurch sich sowohl Interlingua- als auch Transfersysteme als besonders interessant und geeignet für diese Untersuchungen erwiesen (vgl. Hutchins 2001a). Die Berücksichtigung der Linguistik in der Computerwissenschaft sollte später maßgeblich zur Entstehung der Disziplin der Computerlinguistik beitragen (vgl. Rothkegel 2004).

Anfang der 1960er-Jahre wurde der langsame Fortschritt der MÜ zunehmend kritisiert und viele noch zu Beginn der 1950er-Jahre optimistische Forscher:innen wie Yehoshua Bar-Hillel äußerten Zweifel daran, dass vollautomatische Übersetzungen in absehbarer Zeit erreichbar seien (vgl. Lennon 2014). Als Konsequenz wurde 1964 das Automatic Language Processing Advisory Committee (ALPAC) gegründet, das den aktuellen Stand der MÜ evaluieren sollte (vgl. Hutchins 2001a). Der 1966 veröffentlichte Bericht (Automatic Language Processing Advisory Committee (ALPAC) 1966) des Komitees kam zu dem Schluss, dass maschinelle Übersetzung weder ökonomisch noch qualitativ konkurrenzfähig sei, was einen generellen Stopp staatlicher Fördergelder in den USA für die kommenden zwei Jahrzehnte zur Folge hatte (vgl. Schwartz (2017).

Der Forschungsbetrieb kam nicht gänzlich zum Erliegen, verlagerte sich jedoch stärker auf Interlingua- und Transfermodelle. Während der Bericht den Optimismus in den USA deutlich schwächte, wurde in Europa nach wie vor die Relevanz der maschinellen Übersetzung gesehen. Zugleich gewann die Forschung rund um wissensbasierte KI-Systeme an Einfluss, insbesondere im Bereich der natürlichen Sprachverarbeitung. Diese Entwicklungen prägten die Forschung bis in die 1980er-Jahre und bereiteten indirekt den Boden für den späteren Übergang zu statistischen Methoden der MÜ (vgl. Hutchins 2001a).

Ende der 1980er- und Anfang der 1990er-Jahre arbeiteten Forschungsgruppen von IBM mit korpusbasierten Ansätzen. Diese versuchten, statistische Zusammenhänge von Wörtern, Phrasen oder Wortgruppen in einem Korpus zu errechnen und diese dann bei der maschinellen Zieltextgenerierung zu berücksichtigen (vgl. Hutchins 2005). Die vielversprechenden Ergebnisse dieser Versuche verschafften der statistischen maschinellen Übersetzung einen Durchbruch (vgl. Hutchins 2001a; Koehn 2010).

1997 konnte durch die Zusammenarbeit von SYSTRAN und der Suchmaschine AltaVista mit Babel Fish das erste kostenlose Online-Übersetzungstool veröffentlicht werden, was erstmals der breiten Öffentlichkeit freien Zugang zu maschineller Übersetzung bot (vgl. Schmalz 2018). Auch wenn dieses System noch auf regelbasierten Methoden beruhte, markierte es einen entscheidenden Schritt in Richtung alltäglicher Verfügbarkeit von MÜ-Anwendungen (vgl. Gaspari 2024).

Ab ca. 2000 konnte sich die statistische maschinelle Übersetzung endgültig durchsetzen. Mit der stetigen Kommerzialisierung des Internets waren nun immer mehr Textkorpora und auch Open-Source-Programme, mit denen grundlegende SMÜ-Modelle ausgeführt werden konnten, weitgehend verfügbar. Zudem etablierten sich Evaluationsmetriken wie die von Papineni et al. (2002) vorgestellte BLEU-Metrik (Bilingual Evaluation Understudy), die systematische Vergleiche ermöglichten. Dadurch wurde auch der rasche Fortschritt statistischer Modelle besser sichtbar (vgl. Hutchins 2014).

Zahlreiche Firmen mit Fokus auf SMÜ wurden gegründet und auch große Technologiekonzerne wie Google, Microsoft oder Yahoo! begannen, SMÜ-Ansätze zu adaptieren (vgl. Koehn 2010). Zudem trugen geopolitische Ereignisse wie die Anschläge auf das World Trade Center in 2001 zu neuen, millionenschweren Förderprogrammen im Feld der vollautomatisierten Übersetzung bei (vgl. Koehn 2010). Dies lässt sich dadurch erklären, dass geopolitische Krisen den institutionellen Bedarf an Technologien erhöhen, mit denen große Mengen fremdsprachiger Informationen trotz begrenzter personeller Übersetzungsressourcen schnell und möglichst akkurat übersetzt und nutzbar gemacht werden können (vgl. Apter 2009).

Die statistische maschinelle Übersetzung blieb bis zur Mitte der 2010er-Jahre die vorherrschende technologische Methode und trug maßgeblich dazu bei, die maschinelle Übersetzung massentauglich zu machen. Dies lag vor allem daran, dass SMÜ-Systeme im Vergleich zu regelbasierten Systemen schneller und kosteneffizienter für deutlich mehr Sprachenpaare verbessert werden konnten (vgl. Gaspari 2024).

Ab ca. 2014 etablierte sich das neue Paradigma der Neuronalen Maschinellen Übersetzung (NMÜ). Es entstand aus den Grenzen der statistischen Verfahren, deren Übersetzungen zwar zunehmend flüssig wirkten, jedoch häufig kontextarm oder fragmentiert blieben (vgl. Gaspari 2024).

Erste erfolgreiche Ansätze zu künstlichen neuronalen Strukturen reichen bis in die 1950er-Jahre zurück (vgl. Farley & Clark 1954). Für die maschinelle Übersetzung konnten neuronale Verfahren jedoch erst deutlich später breiter eingesetzt werden, da die dafür erforderliche Rechenleistung und Datenverfügbar lange Zeit begrenzt waren. Daher verzögerte sich die Entwicklung erster konkreter Formen noch bis in die 2010er-Jahre (vgl. Koehn 2020). Mit den Recurrent Continuous Translation Models schlugen Kalchbrenner und Blunsom (2013) zwei Modelle vor, die auf kontinuierlichen Repräsentationen basieren und als Theoriegrundlage zu späteren NMÜ-Architekturen gesehen werden können.

2014 folgte das Encoder-Decoder-Konzept, das erstmals ganze Sätze als Bedeutungseinheit verarbeitete (vgl. Cho et al. 2014; Sutskever et al. 2014), jedoch bei längeren Sätzen an Qualität verlor. Dieses Problem wurde mit der Einführung des Attention-Mechanismus (vgl. Bahdanau et al. 2014) wesentlich reduziert, der kontextabhängige Gewichtungen einzelner Wörter erlaubte und die Grundlage moderner NMÜ-Architekturen wurde.

In den folgenden Jahren entwickelte sich die NMÜ rasch weiter. Wichtige Meilensteine waren die zunehmende Kommerzialisierung der NMÜ, wie etwa durch Googles Übergang von SMÜ zu NMÜ 2016 (vgl. Wu et al. 2016) sowie die Transformer-Architektur (vgl. Vaswani et al. 2017). Diese Entwicklungen führten nicht nur zu einem massiven Qualitätssprung im Kontext des BLEU-Scores, sondern auch zu einer deutlichen Energie- und Kostenersparnis (vgl. Ciubotaru 2025; Vaswani et al. 2017).

Bis heute ist die NMÜ, insbesondere die Transformer-Architektur, die dominante maschinelle Übersetzungsarchitektur (vgl. Wang et al. 2021). Zugleich zeichnen sich neue Entwicklungen im Bereich neuronaler Modelle ab, die die maschinelle Übersetzung künftig weiter prägen dürften.

Seit Beginn der 2020er-Jahre haben Large Language Models (LLMs) die Entwicklung der maschinellen Übersetzung wesentlich beeinflusst. Dabei handelt es sich nicht um eigenständige Übersetzungsarchitekturen, sondern um großskalige neuronale Sprachmodelle, die für ein breites Spektrum sprachbezogener Aufgaben eingesetzt werden können, darunter auch für Übersetzungsaufgaben (vgl. Brown et al. 2020; Luo et al. 2025). LLMs werden mit umfangreichen Textdaten trainiert und sind darauf ausgerichtet, allgemeine sprachliche Muster zu erfassen und sprachliche Ausgaben zu generieren (vgl. Brown et al. 2020)

LLMs wurden daher nicht explizit als maschinelle Übersetzungstools entwickelt, sie können aber als solche eingesetzt werden. Darüber hinaus ermöglicht die dialogbasierte Nutzung, dass zusätzliche Kontextinformationen etwa durch Prompting bereitgestellt werden können. Dies kann als funktionaler Vorteil gegenüber klassischen MÜ-Systemen gesehen werden, deren Eingabekontext meist auf Satz- oder Dokumentebene begrenzt ist (vgl. Brown et al. 2020).

LLMs bieten meist auch den Vorteil von multimodalen Verarbeitungsfähigkeiten, d. h. mehr als eine Eingabemodalität kann verarbeitet werden, von Text über Bild bis hin zur gesprochenen Sprache. Dies eröffnet neue Möglichkeiten für die maschinelle Übersetzung. Durch die Option, die Eingabe in Form von Text, aber auch beispielsweise zusätzlich in Form von Bildern zu gestalten, erhält das LLM mehr Kontext für die Generierung des Zieltextes und sorgt so für akkuratere Übersetzungen (vgl. Zuo et al. 2025).

Abschließend lässt sich festhalten, dass die Entwicklung der maschinellen Übersetzung eng mit den jeweiligen technologischen Möglichkeiten ihrer Zeit verknüpft war. Von den frühen regelbasierten Verfahren über statistische Methoden bis hin zur neuronalen maschinellen Übersetzung markieren die einzelnen Paradigmenwechsel Reaktionen auf die Grenzen der jeweils vorherrschenden Ansätze. LLMs bilden keine Ablösung, sondern eine Weiterentwicklung des neuronalen Paradigmas. Im folgenden Kapitel werden die technischen Grundlagen der jeweiligen Ansätze näher untersucht.

2.3 Ansätze der maschinellen Übersetzung

Während in Kapitel 2.2 die Entwicklung der maschinellen Übersetzung als Disziplin beschrieben wurde, und im Zuge dessen eher auf geschichtliche und soziale Thematiken eingegangen wurde, liegt in diesem Kapitel der Fokus auf den technischen Aspekten und Funktionsweisen ausgewählter, für die maschinelle Übersetzung als Disziplin relevantesten Ansätzen.

2.3.1 Regelbasierte maschinelle Übersetzung

Die regelbasierte maschinelle Übersetzung basiert auf der Annahme, dass sich Bedeutungsäquivalente zwischen Ausgangs- und Zielsprache systematisch abbilden lassen. Bei diesem Ansatz kommt linguistischen Eigenheiten, wie beispielsweise Grammatik oder Syntaxregeln, besondere Bedeutung zu. Die RBMÜ wird in drei Ansätze unterteilt, und zwar in den direkten, den Transfer- und den Interlingua-Ansatz (vgl. Elhamayed & Nour 2025; Hutchins 2001a).

Der direkte Ansatz arbeitet sprachenpaarbasiert auf der Wortebene, wobei Analyse und syntaktische Neustrukturierungen auf ein Minimum beschränkt werden. Die Übersetzung erfolgt dann mithilfe zweisprachiger Wörterbücher und sehr grundlegender, auf linguistische Regeln bedachter Anpassungen an die Zielsprache (vgl. Elhamayed & Nour 2025). Dieses Verfahren zeigt eine hohe Fehleranfälligkeit bei strukturell sehr unterschiedlichen Sprachen (vgl. Hutchins & Somers 1992).

Der Transfer-Ansatz stellt ein komplexeres Verfahren dar. Der Ausgangstext wird zunächst grundsätzlich analysiert und auf seine Strukturen heruntergebrochen. Im Anschluss wird eine passende Struktur in der Zielsprache gesucht, bevor die Übersetzung generiert wird (vgl. Elhamayed & Nour 2025). Dieser Ansatz ist somit flexibler als das direkte Verfahren, da er über die Wortebene hinausgeht.

Beim Interlingua-Ansatz wird der Ausgangstext zunächst auf morphologische, syntaktische und semantische Eigenheiten analysiert und in eine sprachenunabhängige Bedeutungsrepräsentation, die sogenannte Interlingua, übertragen. In einem weiteren Schritt wird der Zieltext unabhängig vom Ausgangstext generiert (vgl. Elhamayed & Nour 2025). Dieser Ansatz geht also von umfassenden semantischen Repräsentationen zwischen Sprachen aus, was bisher jedoch nicht umgesetzt werden konnte (vgl. Hutchins & Somers 1992).

Historisch wichtige Anwendungen der RBMÜ sind beispielsweise SYSTRAN (vgl. Toma 1976) sowie METEO (vgl. Chandioux 1976) oder METAL (vgl. Bennett & Slocum 1985), die in sehr spezifischen Fachbereichen erfolgreich angewandt wurden (vgl. Hutchins 2001a). Trotz ihrer Einschränkungen hinsichtlich Skalierbarkeit und Robustheit gegenüber sprachlicher Variation bilden regelbasierte Methoden die Grundlage vieler späterer Systeme und prägen bis heute Teilkomponenten kommerzieller Übersetzungssoftware.

2.3.2 Statistische maschinelle Übersetzung

Die Statistische Maschinelle Übersetzung (SMÜ) ist ein datengetriebener Ansatz der maschinellen Übersetzung, bei dem mithilfe umfassender zweisprachiger Textkorpora Wahrscheinlichkeitsmodelle mit dem Ziel trainiert werden, Zieltextsequenzen zu bestimmen, die mit

höchster Wahrscheinlichkeit einer gegebenen Ausgangstextsequenz entsprechen (vgl. Hutchins 2005). Grundlegend ist dabei die Unterscheidung zwischen Übersetzungsmodell und Sprachmodell. Während das Übersetzungsmodell Wahrscheinlichkeiten zwischen ausgangs- und zielsprachlichen Einheiten abbildet, bewertet das Sprachmodell monolingual die Wahrscheinlichkeit zielsprachlicher Wortfolgen und damit deren Flüssigkeit bzw. Plausibilität in der Zielsprache (vgl. Brown et al. 1990; Koehn 2010). Im Rahmen dieses Kapitels wird vor allem auf zwei Arten der SMÜ näher eingegangen. Zum einen auf die wortbasierte SMÜ, die die Grundlage vieler früher Systeme bildet, zum anderen auf die phrasenbasierte SMÜ, die über viele Jahre hinweg der dominierende Ansatz der statistischen maschinellen Übersetzung war und in zahlreichen kommerziellen Anwendungen eingesetzt wurde (vgl. Koehn 2010).

Einer der wichtigsten technischen Ausgangspunkte der wortbasierten SMÜ sind die IBM-Modelle (vgl. Brown et al. 1993). Ausgangs- und Zieltext werden dabei auf Wortebene einander zugeordnet und miteinander verglichen. Die hierfür verwendeten Trainingsdaten müssen zunächst auf Satzebene aligniert sein, sodass jedem Ausgangssatz ein entsprechender Zielsatz zugeordnet werden kann (vgl. Koehn 2010). Mithilfe dieser Trainingsdaten können statistische Wahrscheinlichkeiten für die jeweilige mögliche Übersetzung eines Worts abgeleitet werden. Koehn (2010: 83–84) zieht hierfür das deutsche Wort *Haus* als Beispiel heran. In den meisten Fällen wird Haus im Englischen mit *house* übersetzt, in anderen mit *building*, in wieder anderen mit *shell*, etc. Wortbasierte Modelle berechnen auf Basis der Trainingsdaten, welche Übersetzung eines bestimmten Wortes statistisch am wahrscheinlichsten ist.

Die IBM-Modelle bilden eine wichtige Grundlage für die Modellierung von Wortzuordnungen und Übersetzungswahrscheinlichkeiten. Gleichzeitig zeigen sich jedoch die Grenzen rein wortbasierter Ansätze. Koehn (2010: 113–115) betont, dass die korrekte Wortanordnung oft sehr vom Kontext abhängt. Insbesondere bei Funktionswörtern oder bei Redewendungen führt die maschinelle Übersetzung trotz korrekter Zuordnung selten zum Erfolg, weil Bedeutung nicht aus einer rein wortbasierten Statistik erschlossen werden kann.

Eine Weiterentwicklung der wortbasierten SMÜ ist die phrasenbasierte SMÜ. Statt jeden Satz auf einzelne lexikalische Einheiten herunterzubrechen, werden hier zusammenhängende Wortsequenzen, sogenannte Phrasen, herangezogen, die statistisch aus einem Parallelkorpus abgeleitet sind. Dadurch können Kontext und idiomatische Bedeutungen besser erfasst werden, wodurch Ausgaben flüssiger und angemessener wirken. Schon mit dem Grundmodell der phrasenbasierten statistischen maschinellen Übersetzung lassen sich im Vergleich zum wortbasierten SMÜ-Ansatz erhebliche Verbesserungen der Übersetzungsqualität feststellen (vgl. Koehn 2010).

Die SMÜ löste die RBMÜ rasch ab und blieb noch bis zur Mitte der 2010er-Jahre die vorherrschende technologische Methode, vor allem, weil diese Systeme schneller und kostengünstiger trainiert werden konnten (vgl. Gaspari 2024). Es wurden im Wesentlichen nur ausreichend große, gut ausgerichtete Parallelkorpora benötigt, die vergleichsweise effizient trainiert und somit einsatzfähig gemacht werden konnten. Waren keine ausreichend große Parallelkorpora verfügbar, konnte man immer noch auf Englisch als Pivotsprache zurückgehen, da hierfür die meisten zweisprachigen Parallelkorpora vorliegen. Dies ermöglichte es kommerziellen Systemen wie Google Translate oder Bing, so eine große Anzahl an Sprachen anzubieten (vgl. Gaspari 2024).

Trotz der Entwicklungen im Bereich der maschinellen Übersetzung blieb der Output von phrasenbasierten SMÜ-Systemen bis zuletzt oft fehlerhaft, da die Übersetzungen nach wie vor sehr segmentbasiert waren. Es kam daher häufig zu offensichtlichen Mängeln, wie beispielsweise grammatischen Unstimmigkeiten (vgl. Gaspari 2024).

Google Translate gilt als eines der bekanntesten und einflussreichsten Beispiele für die statistische maschinelle Übersetzung. Bis 2016 basierte das System auf phrasenbasierter SMÜ (vgl. Wu et al. 2016) und trug maßgeblich zur Massenanwendung von maschineller Übersetzung bei (vgl. Kenny 2022), worauf in Kapitel 2.4.2 noch genauer eingegangen wird.

2.3.3 Neuronale maschinelle Übersetzung

Die NMÜ arbeitet nicht mehr mit einzelnen lexikalischen Elementen und Wahrscheinlichkeiten, sondern vielmehr mit ganzen sprachlichen Mustern, die selbstständig durch neuronale Netze auf Basis großer Datenmengen gelernt werden, wodurch Outputs deutlich korrekter und flüssiger werden. Zwar zieht die NMÜ wie die SMÜ Parallelkorpora zur Übersetzung heran, sie benötigt jedoch viel größere Datensätze, um erfolgreiche Ergebnisse zu liefern. Eine grundlegende Änderung zur SMÜ ist auch, dass die NMÜ Repräsentationen lernt, die nicht nur Beziehungen innerhalb von einzelnen Sprachen, sondern auch über mehrere Sprachen hinweg modelliert. Dadurch können ganze Sätze global statt nur in Segmenten verarbeitet werden und Sprachkombinationen übersetzt werden, die so nicht in den Trainingsdaten enthalten waren.

Im Unterschied zur klassischen SMÜ basiert die NMÜ auf neuronalen Netzwerken, die speziell für Übersetzungsaufgaben trainiert werden. Wörter und Kontexte werden dabei nicht als isolierte Einheiten behandelt, sondern in Form kontinuierlicher Vektorrepräsentationen verarbeitet, wodurch größere sprachliche Zusammenhänge modelliert werden können (vgl. Bahdanau et al. 2015; Forcada 2017; Gaspari 2024)

Diese Netzwerke wurden ursprünglich von der Funktionsweise des menschlichen Gehirns inspiriert. Sie können jedoch aufgrund ihrer geringeren Komplexität nicht mit biologischen Prozessen verglichen werden, sondern basieren auf einer Vielzahl von mathematischen Ansätzen, die aus in Schichten miteinander verbundenen Neuronen bestehen (vgl. Büttner 2021; Forcada 2017). Jedes Neuron verarbeitet Eingabesignale, gewichtet diese anhand erlernter Parameter und aktiviert weitere Neuronen je nach Relevanz. Wichtig hier zu beachten ist, dass die Signalverarbeitung nicht ausschließlich sukzessiv erfolgt, sondern in einer parallelen Vorgehensweise, weshalb einzelne Neuronen in großen Gruppen in Schichten organisiert werden (vgl. Goodfellow et al. 2016). Das neuronale Modell lernt über viele Trainingsiterationen hinweg eine interne eigene Darstellung sprachlicher Muster auf Basis der mehrfach verarbeiteten Parallelkorpora (vgl. Bahdanau et al. 2015).

NMÜ-Modelle bedienen sich oft des Encoder-Decoder-Ansatzes. Ein Encoder wandelt den Eingabesatz in der Ausgangssprache in eine komprimierte Vektordarstellung um, die dem Decoder übergeben wird. Der Decoder generiert anschließend auf Basis des komprimierten Vektors des Encoders Position für Position den Satz in der Zielsprache. Encoder und Decoder bestehen aus jeweils einem neuronalen Netzwerk mit vielen versteckten Schichten, deren Parameter im Training gemeinsam optimiert werden (vgl. Bahdanau et al. 2015). Ein NMÜ-System kann auch neue oder seltene Wörter verarbeiten, da es über Ähnlichkeiten in den Wortmustern und Darstellung von Sprache auf der Subwort-Ebene Bedeutung erschließen kann (vgl. Sennrich et al. 2015). Dies stellt eine grundlegende Innovation im Vergleich zu SMÜ-Systemen dar (vgl. Forcada 2017).

Heutige NMÜ-Systeme basieren überwiegend auf der Transformer-Architektur (vgl. Vaswani et al. 2017). Diese Architektur nutzt Self-Attention, einen Mechanismus, der Beziehungen zwischen den einzelnen Token eines Satzes modelliert und dabei auch weit voneinander entfernte Elemente direkt miteinander in Bezug setzen kann. Dadurch können größere Kontexte effizienter erfasst werden als in vorhergehenden Architekturen (vgl. Vaswani et al. 2017; Stahlberg 2020).

Beispiele für die NMÜ sind Weiterentwicklungen von Google Translate, SYSTRAN oder DeepL.

2.3.4 Generative maschinelle Übersetzung

Wie bereits in Kapitel 2.2 erwähnt, basiert die Technologie hinter generativen LLMs auf der Transformer-Architektur (vgl. Zhao et al. 2026). Es handelt sich also bei LLMs nicht um dezi-

dierte Übersetzungssysteme, da die maschinelle Übersetzung nur eine der zahlreichen Anwendungen für LLMs ist (vgl. Bommasani et al. 2021). Aufgrund der zunehmenden Verwendung von LLMs für maschinelle Übersetzungsleistungen und der Relevanz eines LLMs für die im Rahmen dieser Arbeit durchgeführte Empirie, wird im Folgenden näher auf ihre Funktionsweise eingegangen. Während klassische NMÜ-Systeme häufig satz- oder abschnittsbasiert arbeiten, können LLMs innerhalb ihrer jeweiligen Kontextlänge auch größere Textzusammenhänge verarbeiten und dadurch über einzelne Sätze hinausgehende Kontextinformationen in die Übersetzung einbeziehen (vgl. Wang et al. 2024). Die Länge des Eingabetextes bleibt jedoch durch die maximale Kontextlänge des jeweiligen Modells beschränkt (vgl. Pawar et al. 2024).

Wie bei neuronalen maschinellen Übersetzungssystemen stützen sich LLMs häufig auf das Encoder-Decoder-Prinzip, wenngleich moderne LLMs oft auch nur auf einem Decoder basieren, da diese stark auf Generierung der Ausgabe fokussiert sein sollen (vgl. Zhao et al. 2026). Weiters gibt es auch LLMs, die sich nur auf den Encoder stützen, da die detaillierte Analyse der Eingabe im Vordergrund steht (vgl. Devlin et al. 2019). Generative LLMs erzeugen Texte autoregressiv, das heißt Token für Token auf Basis der zuvor verarbeiteten Eingabe und der bereits erzeugten Ausgabe (vgl. Ciubotaru 2025).

Während NMÜ-Tools mit zweisprachigen, definierten Paralleltexten trainiert werden, erfolgt das Training von LLMs über massive, heterogene Datensätze, die viele unterschiedliche Sprachen, Textsorten oder auch multimodale Inhalte wie Bild- oder Audiodateien enthalten können. Die Trainingskorpora umfassen hier oft mehrere Milliarden Wörter (vgl. Brown et al. 2020; Zhao et al. 2026). Brown et al. (2020) beschreiben die Sprachverarbeitung mit dem Rosettastein-Prinzip. Im Transformerkontext bedeutet das, dass LLMs Muster verschiedener Sprachen in einem gemeinsamen Repräsentationsraum lernen (vgl. auch Schmalz 2019). Daraus folgt, dass LLMs nicht explizit für die maschinelle Übersetzung trainiert werden, sondern primär für die Vorhersage des wahrscheinlichsten nächsten Textelements. Die Qualitätssteigerungen in der LLM-generierten maschinellen Übersetzung können daher als Nebenprodukt der allgemeinen Sprachmodellierung bezeichnet werden.

In der translatorischen Praxis funktionieren LLMs daher etwas anders als vorhergehende Strukturen. Während bei neuronalen maschinellen Übersetzungstools oft nur ein Ausgangstext angegeben wird, brauchen LLMs zusätzliche, auf den Auftrag zugeschnittene Angaben, wie beispielsweise Informationen zu Textsorte, Zielgruppe oder Terminologie. Diese Anweisungen werden als Prompts bezeichnet. Es wird dabei häufig zwischen System Prompts und Task Prompts unterschieden. System Prompts sind vorgegebene Steuerungsanweisungen auf Systemebene, die für die Endnutzer:innen oft nicht zugänglich sind. Task Prompts hingegen

sind Prompts, die von Endnutzer:innen selbst angewendet werden können, um die für die jeweilige Aufgabenstellung bestmögliche Lösung zu finden (vgl. Zhang et al. 2026).

Gut ausgearbeitete Prompts bieten daher die Möglichkeit, sehr spezifische Aufträge zu geben, was den Output des LLMs zielgerichteter werden lässt (vgl. Sejnowski 2023). Da ein LLM verschiedenste Outputs generieren kann, muss zunächst das exakte Ziel der Ausgabe definiert werden. Ein grundlegender Prompt für einen Übersetzungsauftrag könnte beispielsweise sein: *Übersetzen Sie bitte den folgenden Ausgangstext ins Deutsche*, wenngleich dieser vermutlich aufgrund mangelnder Präzision keine qualitativ hochwertige Übersetzung liefern würde. Um den Prompt zu verbessern, könnten beispielsweise Angaben zur Zielgruppe oder zum Zweck der Übersetzung hinzugefügt werden (vgl. Yamada 2023). Ein etablierter Ansatz ist zudem das Persona Prompting, bei dem LLMs angewiesen werden, sich in eine bestimmte Rolle hineinzuversetzen, um Ausgaben zu generieren, die der Perspektive oder Expertise dieser Rolle entsprechen (vgl. Elnashar et al. 2023; Lutz et al. 2025). Für die Übersetzung eignet sich hierfür je nach Ausgangstext etwa die Rolle professioneller Fach- oder Marketingübersetzer:innen. Beim Few-Shot Prompting werden zusätzlich Beispiele hinzugefügt, wodurch sich das System bei der Outputgenerierung an bestehenden Texten, zum Beispiel Übersetzungen, orientieren kann (vgl. Sahoo et al. 2025).

Weitere Prompting-Ansätze, wie beispielsweise Chain-of-Thought Prompting (CoT) oder Self-Consistency Prompting zielen darauf ab, die Outputs nachvollziehbarer zu machen (vgl. Sahoo et al. 2025). Diese eignen sich jedoch nur bedingt für Übersetzungsaufgaben, da sie durch Verkomplizierung neue Fehlerquellen eröffnen können. Unabhängig von der Prompting-Strategie sind Halluzinationen ein häufig auftretendes Phänomen (vgl. Wang et al. 2024). Halluzinationen werden als Ausgaben definiert, die scheinbar korrekt sind, jedoch falsche Informationen enthalten und so schwerwiegende Folgen nach sich ziehen können (vgl. Luo et al. 2024). Im Kontext der maschinellen Übersetzung beziehen sich Halluzinationen insbesondere auf falsche oder im Ausgangstext nicht vorhandene Informationen (vgl. Guerreiro et al. 2023).

Eine Möglichkeit, die Zuverlässigkeit generativer Modelle zu erhöhen, ist etwa die Einbindung externer Informationen. Ein Beispiel hierfür ist die Retrieval Augmented Generation (RAG), bei der das System relevante Informationen aus einer externen Quelle abruft, anhand derer der originale Prompt verfeinert bzw. dem System zusätzlicher Kontext bereitgestellt werden soll. So soll die Wahrscheinlichkeit erhöht werden, dass das LLM nur akkurate, nachweis-

bare Informationen für die Outputgenerierung verwendet (vgl. Sahoo et al. 2025). Im Übersetzungskontext eignet sich diese Methode etwa bei terminologischen Vorgaben, indem das LLM angewiesen wird, auf ein Glossar zurückzugreifen (vgl. Kim et al. 2024).

Zusammenfassend lässt sich sagen, dass LLMs Anwender:innen durch das Prompting mehr Flexibilität geben. Im Kontext der maschinellen Übersetzung können durch Angabe zusätzlicher Informationen wie Stil, Register, Zielgruppe oder Zweck der Übersetzung natürlichere und idiomatischere Zieltexte generiert werden (vgl. Yamada 2023). Gleichzeitig bleibt die Qualität der Ausgabe von der Präzision des Prompts, den Modellgrenzen und möglichen Halluzinationen abhängig.

2.4 Kommerzielle Systeme im Vergleich

Im Folgenden werden mit SYSTRAN, Google Translate und DeepL drei der wichtigsten kommerziellen maschinellen Übersetzungssysteme vorgestellt. Mit ChatGPT wird auch ein LLM betrachtet, das zwar nicht primär als maschinelles Übersetzungssystem entwickelt wurde, in der Praxis jedoch häufig für Übersetzungszwecke eingesetzt wird (vgl. Bommasani et al. 2021).

2.4.1 SYSTRAN

SYSTRAN war einer der ersten kommerziell erfolgreichen Ansätze zur regelbasierten maschinellen Übersetzung. Das 1968 von Peter Toma gegründete Unternehmen entwickelte zunächst für das US-Militär und später für die Europäische Kommission ein rein regelbasiertes Übersetzungssystem, das über Jahrzehnte hinweg in verschiedenen Institutionen genutzt wurde und auch als Grundlage für andere kommerzielle Anwendungen wie Google Translate diente (vgl. Dugast et al. 2007; Systran o.J.a).

Einen größeren Bekanntheitsgrad erlangte das Unternehmen aufgrund seiner Mitwirkung bei der Apollo-Soyouz-Mission, bei der SYSTRAN verwendet wurde, um Anweisungen aus dem Russischen ins Englische zu übersetzen, bevor es 1978 mit Xerox SYSTRAN auch seinen ersten kommerziellen Anwendungsfall verzeichnete (vgl. Büttner 2021; Systran o.J.b)

Ein historischer Schritt ist der Launch von Babel Fish von AltaVista, später als Yahoo! bekannt, im Jahre 1997. Das Online-Tool basierte auf SYSTRAN und ging als erster kostenloser Online-Service für die maschinelle Übersetzung in die Geschichte ein (vgl. Schmalz 2018).

SYSTRAN wurde über die Jahre erfolgreich weiterentwickelt, integrierte nach ursprünglich regelbasierten Ansätzen später hybride RBMÜ-SMÜ-Verfahren und entwickelte schließlich auch neuronale MÜ-Systeme (vgl. Crego et al. 2016). Das System wurde bzw. wird

bei der NASA, der NATO, General Motors, Xerox, der Europäischen Kommission (bis 2010) und dem US-Verteidigungsministerium eingesetzt (vgl. Systran o.J.c).

2.4.2 Google Translate

Google Translate wurde 2006 veröffentlicht, wenngleich in sehr eingeschränkter Form, da zu Beginn mit Englisch–Arabisch lediglich ein Sprachenpaar angeboten wurde (vgl. Büttner 2021). Google Translate galt als eine der ersten öffentlich zugänglichen kommerziellen Anwendungen, die, nicht wie viele andere zeitgenössische Programme, auf der regelbasierten maschinellen Übersetzung, sondern auf der statistischen maschinellen Übersetzung basierte (vgl. Schmalz 2018).

Innerhalb von nur zehn Jahren stockte Google Translate auf 103 unterstützte Sprachen auf, konnte die Meilensteine von 500 Millionen Nutzer:innen und 100 Milliarden übersetzte Wörter pro Tag erreichen und entwickelte sich dadurch zur meistgenutzten Online-Übersetzungssoftware auf dem Markt (vgl. Kenny 2022).

Einen grundlegenden Wandel markierte das Jahr 2016 mit der Einführung des Google Neural Machine Translation (GNMT)-Systems (vgl. Wu et al. 2016). Dieses nutzte erstmals ein Encoder-Decoder-Modell auf Basis rekurrenter neuronaler Netze in Kombination mit Attention-Mechanismen (vgl. Wu et al. 2016). Ziel war es, ganze Sätze als zusammenhängende Einheiten zu verarbeiten, anstatt Phrasen isoliert zu behandeln. Das GNMT-System kann semantische Beziehungen über den gesamten Satz hinweg erfassen. Damit gelang ein deutlicher Qualitätssprung. Übersetzungen wirkten grammatikalisch kohärenter, idiomatischer und kontextsensitiver. Besonders hervorzuheben ist der entstandene Nebeneffekt der sogenannten Zero-Shot-Übersetzung, die es ermöglichte, Sprachen zu übersetzen, für die kein direkter Parallelkorpus vorlag, ein Effekt der stark kontextbasierten Repräsentationen innerhalb des Modells (vgl. Johnson et al. 2017). Allein im Jahr 2022 konnten durch diese Technik 24 neue Sprachen zu Google Translate hinzugefügt werden (vgl. Bapna et al. 2022). Wie in Kapitel 2.3.3 beschrieben, setzte die auch beim GNMT-System Google verwendete Transformer-Architektur neue Qualitätsstandards in der Welt der maschinellen Übersetzung.

Heute bietet Google Translate maschinelle Übersetzung für über 249 Sprachen an, von Weltsprachen wie Deutsch und Englisch bis zu beinahe ausgestorbenen Sprachen wie Manx (vgl. Bapna et al. 2022).

2.4.3 DeepL

DeepL ist ein deutsches Unternehmen, das auf der Übersetzungssuchmaschine Linguee basiert und 2017 der Öffentlichkeit vorgestellt wurde. Das Unternehmen verwendete neuronale Netze ursprünglich für die qualitative Bewertung von Übersetzungen bei Linguee und entwickelte daraufhin mit den dadurch gesammelten Daten eine maschinelle Übersetzungslösung (vgl. Büttner 2021).

DeepL basiert auf der von Google entwickelten Transformer-Architektur und nutzt davon abgeleitete Optimierungen, durch die das System neue Rekorde beim BLEU-Score aufstellen konnte. Dies ist überwiegend auf die qualitativ hochwertigen, unterschiedlichen Trainingsdaten zurückzuführen, die DeepL verwendet (vgl. Kamaluddin et al. 2024; Schneider 2017).

Ein grundlegender Unterschied zwischen Google und DeepL ist die Herangehensweise an das Produkt. Während Google das Ziel hat, 1.000 Sprachen zu unterstützen, war DeepLs Sprachenportfolio von Anfang an bedeutend geringer (vgl. Bapna et al. 2022). Dies deutet darauf hin, dass bei DeepL der Fokus auf Qualität statt auf Quantität gelegt wurde, was sich vor allem auch an Qualitätsbewertungen bemerkbar machte (vgl. Kamaluddin et al. 2024). Im November 2025 wurden viele neue Sprachen in das DeepL-Portfolio hinzugefügt. So verfügt das System nun über 96 Sprachen, wovon der Großteil sich jedoch noch in der Beta-Phase befindet (vgl. DeepL 2025). Laut dem European Language Industry Survey (ELIS) 2024 ist DeepL vor Google Translate das meistgenutzte MÜ-Tool (vgl. ELIS 2024).

2.4.4 ChatGPT

ChatGPT ist ein Large Language Model von OpenAI, das 2022 veröffentlicht wurde (vgl. Chatterji et al. 2025). Wie spätere Google-NMÜ-Versionen oder DeepL baut auch ChatGPT auf der Transformer-Architektur auf. Im Gegensatz zu traditionellen NMÜ-Tools nutzt ChatGPT aber ein Decoder-only-Modell.

Trotz der Ähnlichkeit im Aufbau unterscheidet sich ChatGPT maßgeblich von den anderen vorgestellten Systemen. Es handelt es sich bei ChatGPT nicht explizit um ein Übersetzungstool, da maschinelle Übersetzung nur als ein Anwendungsfall des LLMs gesehen werden kann. Aufgrund dessen unterscheidet sich ChatGPT auch in der Trainingsstruktur. Als LLM erfolgt das Training des Modells nicht sprachenpaarspezifisch, sondern über einen massiven, multilingualen Korpus hinweg. Wie bereits in Kapitel 2.3.4 erwähnt, muss für ChatGPT als LLM ein Prompt verwendet werden. Zwar bietet auch DeepL Einstellungen hinsichtlich Register und Stil, Prompting bietet jedoch größere stilistische Flexibilität (vgl. Chen 2024; OpenAI 2023).

Seit GPT-4 (vgl. OpenAI 2023) ist ChatGPT multimodal, da erstmals auch Bildinputs zusätzlich zu Textinputs verarbeitet werden und somit auch in der Outputgenerierung berücksichtigt werden können. Davon kann auch die maschinelle Übersetzung profitieren, da Kontext beispielsweise auch in Form einer Bilddatei zur Verfügung gestellt werden kann und somit Ambiguitäten beseitigt werden können (vgl. Zuo et al. 2025).

Aus translationswissenschaftlicher Perspektive ist interessant zu erwähnen, dass die Grenzen zwischen Übersetzung und Textproduktion zunehmend verschwinden. Während bei anderen kontemporären Tools wie DeepL oder Google Translate noch klar ein Übertragen des Ausgangstextes, inklusive Ordnung auf Satz- und Textebene, stattfindet, generiert ChatGPT neue Zieltexte und ahmt dadurch menschenähnliche Schreibweisen nach, sofern nicht anders im Prompt verlangt (vgl. Chen 2024).

Im ELIS 2024-Bericht (ELIS 2024) wird ChatGPT nicht explizit als MÜ-Tool geführt, jedoch als KI-Tool für eine Vielzahl von Textarbeiten, darunter auch zur Unterstützung im Übersetzungsprozess. So ist ChatGPT universal sowohl von freiberuflich tätigen Übersetzer:innen, Sprachenunternehmen als auch von Sprachenabteilungen innerhalb größerer Firmen die am häufigsten eingesetzte KI-Alternative für beispielsweise maschinelle Übersetzung, Qualitätsmessung, Content-Erstellung oder auch audiovisuelle Übersetzung (vgl. ELIS 2024). Aufgrund dieser Tendenz ist ChatGPT als LLM auch für die maschinelle Übersetzung praktisch relevant.

2.5 Übersetzungsqualität

Im folgenden Kapitel werden zentrale Qualitätskonzepte in der Translationswissenschaft und aus der Computerlinguistik kurz beleuchtet. Dies kann als relevant betrachtet werden, da nur qualitativ hochwertige Übersetzungen für einen Turing-Test und damit den empirischen Teil in Frage kommen. Dieses Kapitel bietet daher einen Überblick über die verschiedenen Qualitätsdimensionen und Evaluierungsmethoden, bevor in Kapitel 2.6 näher auf die Wahrnehmung von Qualität der maschinellen Übersetzung eingegangen wird.

2.5.1 Qualitätskonzepte in der Translationswissenschaft

Die Übersetzungsqualität ist ein viel diskutierter Begriff in der Translationswissenschaft, da sie sich maßgeblich je nach Ansatz und Fokus unterscheidet. Es wird nicht von einer immer greifenden Qualitätsmetrik ausgegangen, sondern vielmehr von einem mehrdimensionalen Konstrukt, das je nach theoretischem Ansatz, Textsorte und Zielsetzung unterschiedlich bewertet wird.

Frühere Überlegungen zur Übersetzungsqualität orientierten sich stark am Konzept der Äquivalenz, das bereits in den 1960er-Jahren von Ansätzen wie denen von Nida (1964) oder Catford (1965) geprägt wurde. Hohe Übersetzungsqualität wurde hier durch möglichst weitgehende Übereinstimmung zwischen Ausgangs- und Zieltext definiert. Auf dieser Grundlage entwickelte Reiß (1971) einen der ersten textlinguistischen Ansätze, der Übersetzungsqualität an Texttyp und textsortenspezifischen Normen ausrichtete und damit über rein formale Äquivalenzvorstellungen hinausging. Eine systematische Ausarbeitung der verschiedenen Dimensionen von Äquivalenz erfolgte später durch Koller (1979), dessen Äquivalenztypologie in Form von denotativ, konnotativ, textnormativ, pragmatisch und formal lange als wichtiger Bezugsrahmen für Übersetzungsqualität galt.

Ein weiterer Zugang zur Übersetzungsqualität wird im Rahmen der funktionalen Übersetzung diskutiert. Übersetzung ist eine zielgerichtete Handlung, bei der die Qualität eines Zieltextes im Rahmen primär über die Erfüllung des Zwecks einer Übersetzung definiert wird. Entspricht eine Übersetzung den Anforderungen des Zielkontextes, so gilt sie als qualitativ hochwertig (vgl. Nord 1997; Reiß & Vermeer 1984). Äquivalenz spielt hinsichtlich Qualität bei den funktionalen Ansätzen also eine kleinere Rolle, da der Zweck der Übersetzung im Vordergrund steht.

Fehlerbasierte Bewertungsmodelle bilden einen weiteren, insbesondere in der Praxis etablierten Zugang. House (1977, 2015) unterscheidet etwa zwischen overt/covert Fehlern, während Modelle wie das LISA QA Model oder SAE J2450 Übersetzungsfehler auf Basis einer fehlertypologischen Liste bewerten. Auf diese Ansätze aufbauend entstand das Multidimensional Quality Metrics (MQM) Framework, das als umfassende Sammlung translationsbezogener Qualitätskriterien gilt (vgl. Lommel et al. 2014). MQM bietet eine systematische Einteilung von mehr als 100 Fehlertypen, z. B. Bedeutungs-, Terminologie-, Grammatik- oder Konsistenzfehler, sowie deren Schweregrade, und wird neben der Bewertung von humanübersetzten Texten auch explizit zur Anwendung an maschinellen Übersetzungen vorgeschlagen (vgl. Lommel et al. 2014).

Diese klassischen Ansätze legen nahe, dass Übersetzungsqualität relativ ist, da Übersetzungen je nach Definition von Qualität, Normen, Konventionen, Funktion oder auch linguistischen Merkmalen unterschiedlich evaluiert werden können. Für die maschinelle Übersetzung bilden sie einen wichtigen theoretischen Hintergrund, auf dessen Basis die Bewertung maschineller Systemausgaben durchgeführt werden kann.

2.5.2 Qualität von maschineller Übersetzung

Da sich die Fehlertypologien von Humanübersetzungen und maschineller Übersetzung in bestimmten Bereichen unterscheiden können, ist eine direkte Übertragung der oben beschriebenen traditionellen Modelle zur Qualitätsevaluierung nur eingeschränkt möglich.

Bei der maschinellen Übersetzungbewertung wird häufig zwischen manueller und automatischer Qualitätskontrolle unterschieden. Erste Modelle, die spezifisch zur manuellen Qualitätskontrolle von maschineller Übersetzungsarbeit entwickelt wurden, kamen bereits in den 1960er-Jahren auf (vgl. Han 2018). Carrolls (1966) Ansatz konzentriert sich dabei auf die zwei Qualitätsdimensionen Inhaltstreue (fidelity) bzw. Genauigkeit (accuracy) und Verständlichkeit (intelligibility). Carroll basiert sein Modell auf die Satzebene, da er sich der Kohärenz- und Kontextprobleme damaliger maschineller Übersetzungsprogramme bewusst war. Als verständlich gilt ein Satz, wenn er sich natürlich und flüssig liest, als wäre er in der Ausgangssprache geschrieben worden. Die Inhaltstreue bzw. Genauigkeit ist dann gegeben, wenn ein Zieltext die Bedeutung eines Ausgangstextes nicht ändert, verdreht oder verfälscht (vgl. Carroll 1966).

In den 1990er-Jahren wurde im Rahmen der ARPA MT Initiative weiterer Fortschritt hinsichtlich der Evaluierung von maschinellen Übersetzungen gemacht. Die ARPA MT Initiative war Teil des Human Language Technologies Program des US-Verteidigungsministerium und zielte darauf ab, wegweisende Fortschritte im Bereich der maschinellen Übersetzungstechnologie zu machen (vgl. White et al. 1994). Die Qualitätsmodelle untersuchten maschinelle Übersetzungen, üblicherweise Fragmente, die teilweise sogar kürzer als ein Satz sind, in den Dimensionen Adäquatheit (adequacy), Flüssigkeit (fluency) und Verständlichkeit (comprehension) bzw. Informativität (informativeness). Die Adäquatheit ist bei diesem Ansatz dann gegeben, wenn die Informationen des Ausgangstextes auch im Zieltext vorhanden sind und im Rahmen der Informativität auch verstanden werden können. Eine Übersetzung gilt dann als flüssig, wenn sie gut, also natürlich, klingt, ohne dabei zu beurteilen, ob die Übersetzung korrekt ist oder nicht (vgl. White et al. 1994).

Weitere Ansätze zur manuellen Qualitätsmessung sind beispielsweise aufgabenorientierte Modelle (z. B. Doyon et al. 1999), bei der die Qualität für bestimmte Anwendungsfälle überprüft wird, Post-Editing-Modelle, bei der eine Humanübersetzung einer Post-Editing-MÜ gegenübergestellt wird (z. B. Snover et al. 2006) oder das Segment Ranking (z. B. Callison-Burch et al. 2007), bei dem verschiedene maschinelle Übersetzungen nach Qualität für bestimmte Ausgangstextsegmente angeordnet und bewertet werden (vgl. Han 2018).

Da die manuelle Qualitätskontrolle jedoch als sehr subjektiv, arbeitsaufwendig und daher teuer und zudem auch selten reproduzierbar gilt, setzte immer mehr Forschungsarbeit zu

automatischen Bewertungssystemen ein. Während es auch Bewertungssysteme gibt, die ohne humane Referenzübersetzung arbeiten, wurden überwiegend Metriken entwickelt, die die Ähnlichkeit zwischen Maschinenoutput und Referenzübersetzungen in Kategorien wie Überlappung von Wörtern und Wortsequenzen, Wortstellung oder die Anzahl notwendiger Bearbeitungsschritte (lexikalische Ähnlichkeit) oder Syntax, Semantik und weitere linguistische Parameter (linguistische Ähnlichkeit) messen (vgl. Han 2018). Praktische Beispiele zur automatisierten Qualitätsevaluierung sind beispielsweise die Metriken Word-Error-Rate-Metrik (WER) (Nießen et al. 2000), wo eine Überprüfung der Wörter in der MÜ vorgenommen wird, insbesondere hinsichtlich ihrer Position im Segment; Position-Independent Word Error-Rate (PER) (Tillmann et al. 1997), wo der Fokus ebenfalls auf der Wortebene liegt, jedoch unabhängig von der Position; und die Translation Edit Rate (TER) (Snover et al. 2006), wo insbesondere auf den Aufwand eingegangen wird, um eine Übersetzung vergleichbar mit der menschlichen Referenzübersetzung zu machen. Die Funktionsweise baut dabei auf WER auf, jedoch sind sogenannte Änderungen in der Anordnung der Wörter erlaubt (vgl. Han 2018; Snover et al. 2009).

Die vermutlich bekannteste automatische Evaluierungsmetrik für maschinelle Übersetzung ist die Bilingual Evaluation Understudy (BLEU)-Metrik (Papineni et al. 2002), die eine sehr hohe Korrelation zu menschlich durchgeführten Bewertungen aufweist. Im Vergleich zur Qualitätsbestimmung durch Menschen sticht diese Metrik insbesondere durch Effizienz, Kostenersparnis, Sprachunabhängigkeit und Reproduzierbarkeit hervor. Vor allem die Zeitersparnis kann als kritischer Punkt betrachtet werden, da monatelange Analysen hinsichtlich dem hohen Tempo der Entwicklungen im Bereich der maschinellen Übersetzung sich als problematisch erweisen könnten. Der Kern von BLEU ist die Idee, dass eine Übersetzung umso besser ist, je ähnlicher sie professionellen Referenzübersetzungen ist. Der BLEU-Score arbeitet auf Basis von n-Grammen, also einer Folge von aufeinanderfolgenden Einheiten wie Wörtern, und bewertet, wie viele n-grams einer maschinellen Übersetzung in Referenzübersetzungen vorkommen. Hervorzuheben ist dabei, dass die Bewertung auf Korpusebene und nicht wie bei anderen Metriken auf Satzebene stattfindet, um zufällige Varianzen einzelner Sätze auszugleichen (vgl. Papineni et al. 2002). BLEU kann auch nach über 20 Jahren noch als Industriestandard zur Bewertung von maschineller Übersetzungsqualität bezeichnet werden. Wie bereits in Unterpunkt 2.4.3 beschrieben, wird die Qualität von kommerziellen Anwendungen wie DeepL nach wie vor mit dem BLEU-Score erhoben, um Fortschritte zu messen und zu publizieren.

Ein weiteres etabliertes System zur automatischen Qualitätsbestimmung ist METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee & Lavie 2005). ME-

TEOR ist eine Weiterentwicklung von BLEU und setzt direkt bei den Schwachstellen der Metrik an. Es wird der fehlende Recall-Anteil, also die Evaluierung ob Teile ausgelassen wurde, ergänzt, ebenso wie die fehlende explizite Wort-zu-Wort-Zuordnung, die geringe Sensitivität gegenüber Wortreihenfolge und die Tatsache, dass BLEU auf Satzebene oft unzuverlässige Werte produziert. Dies wird unter anderem durch den Fokus auf Unigramme, also auf einzelne Einheiten, statt auf n-Gramme erzielt (vgl. Banerjee & Lavie 2005). Insgesamt erzielt METEOR bessere Korrelationen mit menschlichen Bewertungen, gilt jedoch als komplexer, weshalb BLEU nach wie vor häufiger zur Anwendung kommt.

Weitere neuere Ansätze sind beispielsweise der BERTScore (Zhang et al. 2020), der mit kontextuellen Embeddings arbeitet, und COMET (Crosslingual Optimized Metric for Evaluation of Translation) (Rei et al. 2020), ein Ansatz, mit dem versucht wird, menschliche Bewertungen vorherzusagen.

Während diese Metriken für standardisierte Texte gut funktionieren, stoßen sie bei kreativen Textsorten an Grenzen, da stilistische, funktionale oder wirkungsbezogene Aspekte schwer messbar sind.

2.5.3 MÜ-Qualität von kreativen Texten

Kreative Textsorten, darunter insbesondere Marketing-, Werbe- und appellative Kommunikation, stellen maschinelle Übersetzungssysteme vor besondere Herausforderungen. Appellative Textsorten, wie etwa Werbe- oder Marketingtexte, zeichnen sich durch eine besonders starke stilistische Gestaltung, persuasive Funktion und Wirkungsorientierung aus (vgl. Reiß 1971). Nord (1997: 41–43) betont zudem, dass solche Texte häufig kulturell gebunden sind und ihre kommunikative Wirkung maßgeblich von kulturellen Erwartungen und Konventionen im Zielkontext abhängt. Qualität wird in diesem Bereich nicht allein an korrekter Bedeutungsübertragung gemessen, sondern an der adressatengerechten Wirkung des Zieltexts. Die Besonderheit bei der Übersetzung von kreativen Textsorten ist, dass im Übersetzungsprozess häufig Änderungen vorgenommen werden müssen, um den Zweck im Zieltext zu erfüllen. Es muss also etwas Neues entstehen, was einen Aspekt darstellt, der von maschinellen Systemen schwer umsetzbar ist (vgl. Guerberof-Arenas & Toral 2022).

Auf linguistischer Ebene ergibt sich eine zentrale Schwierigkeit aus der figurativen und kreativen Natur sprachlicher Mittel in appellativen Textsorten, also u. a. Marketingtexten. Metaphern, Wortspiele, idiomatische Formulierungen oder kulturelle Anspielungen werden nicht immer erkannt und auch statistisch nicht erwartet. Bedeutungsnuancen gehen verloren, stilistische Besonderheiten werden geglättet und kreative Effekte durch generische Formulierungen

ersetzt (vgl. Guerberof-Arenas & Toral 2022). In der Forschung wird immer wieder darauf hingewiesen, dass maschinelle Systeme bei literarischen und anderen stark kreativen Textsorten systematisch schlechter abschneiden, was sich grundsätzlich auch auf Marketingkommunikation übertragen lässt (vgl. Guerberof-Arenas & Toral 2022). Werbe- und Marketingtexte zeichnen sich durch eine besonders komplexe sprachliche Gestaltung aus, da sie nicht nur informieren, sondern gezielt Emotionen ansprechen und eine bestimmte Wirkung beim Publikum erzeugen sollen (vgl. Stiegelbauer et al. 2024; Torresi 2021). Übersetzer:innen müssen dabei sensibel mit kulturellen Unterschieden umgehen und sorgfältig darauf achten, dass zentrale persuasive Elemente nicht abgeschwächt oder verfälscht werden. Eine gelungene Übersetzung im Werbekontext trägt wesentlich dazu bei, die Markenidentität zu bewahren und zugleich die Attraktivität des Produkts oder der Dienstleistung für neue Zielmärkte zu steigern. Damit bildet sie einen entscheidenden Bestandteil internationaler Marketingstrategien (vgl. Stiegelbauer et al. 2024).

Setzt man dies in den Kontext der maschinellen Übersetzung, wird deutlich, dass aktuelle Systeme zwar in der Lage sind, flüssige und oberflächlich kohärente Zieltexte zu erzeugen, jedoch häufig Schwierigkeiten haben, die stilistische Feinheit, die kulturelle Einbettung und die persuasive Intention solcher Texte adäquat abzubilden (vgl. Majji et al. 2025). Die genannten Besonderheiten sind daher nicht nur theoretisch relevant, sondern haben auch direkte Implikationen für die vorliegende Untersuchung. Kreative Marketingtexte eignen sich daher besonders gut, um Unterschiede in der Wahrnehmung maschinell und human übersetzter Texte sichtbar zu machen. Da die MÜ bei kreativen Texten zwar hohe sprachliche Flüssigkeit, aber häufig Probleme mit kulturellen oder emotionalen Nuancen aufweist, kann angenommen werden, dass Sprachexpert:innen aufgrund ihrer Ausbildung und Erfahrung diese Abweichungen eher erkennen als Lai:innen; eine Annahme, die wahrnehmungsbasierte Experimente besonders interessant macht.

2.6 Wahrnehmung der maschinellen Übersetzung

Während in Kapitel 2.5 die Qualität maschineller Übersetzung anhand theoretischer Modelle und Evaluierungsmetriken beschrieben wurde, kann angenommen werden, dass diese Ansätze insbesondere bei kreativen Textsorten nur begrenzt erfassen, wie Übersetzungen tatsächlich von den Rezipient:innen beurteilt werden. Für das Ziel dieser Arbeit ist umso relevanter, wie Übersetzungen von Rezipient:innen wahrgenommen und interpretiert werden. Dieses Kapitel bietet zunächst einen Überblick über die historische Entwicklung der Wahrnehmung und geht anschließend auf die Besonderheiten und Unterschiede in der Wahrnehmung der maschinellen

Übersetzung von Sprachexpert:innen und Lai:innen hinsichtlich Erwartungen, Rezeptionen und Nutzungsmuster ein.

2.6.1 Historische Aspekte der Wahrnehmung

Die Wahrnehmung maschineller Übersetzung (MÜ) hat sich seit den ersten experimentellen Systemen der 1950er Jahre mehrfach gewandelt und ist eng an technologische Paradigmenwechsel geknüpft.

Mit Weavers Memorandum 1949 und ersten Anwendungen wurde förmlich eine Euphoriewelle ausgelöst, die durch klaren Optimismus geprägt war und wie in Kapitel 2.2 beschrieben durch hohe Förderungen bestätigt wurde. Die maschinelle Übersetzung wurde als regelbasiertes, lösbares Problem wahrgenommen. Diese Phase sollte bis 1964 anhalten, als immer offensichtlicher wurde, dass damalige maschinelle Übersetzungsprogramme zwar nützlich waren, um Informationen aus dem Ausgangstext zu erhalten, jedoch bei weitem nicht imstande waren, menschlicher Übersetzungsqualität nahe zu kommen. Der ALPAC-Report brachte in den 1960er-Jahren ein klares Negativmomentum für die Wahrnehmung von maschineller Übersetzung.

> Although widely condemned at the time as biased and short-sighted, its influence was profound, bringing a virtual end to MT research in the USA for over a decade. In addition, it put an end to the widespread perception of MT as a leading area of research in the investigation of computers and natural language; from this time on, for example, researchers in information retrieval dropped all interest in MT and indeed in computational linguistics as a whole. (Hutchins 2001a: 6–7)

Der Bericht schadete der Wahrnehmung der maschinellen Übersetzung demnach immens und sorgte dafür, dass die Forschung stark reduziert wurde. Dennoch zeigte die Veröffentlichung von SYSTRAN und die Anwendung bei der Europäischen Kommission in den 1970er Jahren, dass die maschinelle Übersetzung nicht aufgegeben wurde. Sie wurde nun nicht mehr als den Menschen ersetzende Maschine, sondern vielmehr als dem Menschen dienende Architektur gesehen (vgl. Hutchins 2014). Mit der RBMÜ kamen erste kostenlose Übersetzungstools auf den Übersetzungsmarkt, was erstmals auch Lai:innen einen unkomplizierten Zugang zu MÜ-Anwendungen ermöglichte und einen neuen Markt erschloss (vgl. Kenny 2018).

Mit der statistischen maschinellen Übersetzung verschob sich die Wahrnehmung erneut. Texte wurden nun als flüssiger im Vergleich zu Outputs von regelbasierten Architekturen wahrgenommen und dank der zunehmenden Anzahl an Online-Tools wie Google Translate wurde der Zugang zur maschinellen Übersetzung leichter (vgl. Gaspari & Hutchins 2007; Sreelekha 2017).

Der Paradigmenwechsel zur neuronalen maschinellen Übersetzung läutete ein weiteres Kapitel für die Wahrnehmung der MÜ ein. Outputs waren nun viel korrekter, klangen noch flüssiger, idiomatischer und dadurch auch humaner (vgl. Forcada 2017; Gaspari 2024). Dies war ein wichtiger Schritt hin zu einem kontextsensitiven System; Bahdanau et al. (2014: 1) sprechen sogar von einem Netzwerk, „das Sätze liest". Studien wie Hassan et al. (2018) beanspruchten sogar Human Parity, also die Gleichwertigkeit mit Humanübersetzung. Diese Aussage erhielt viel mediale Aufmerksamkeit und wurde vielerorts kommentiert, was auch Einfluss auf die öffentliche Präsenz und Wahrnehmung nahm (vgl. Kenny 2018).

Gleichzeitig entstand ein Diskurs über mögliche Bedrohungen des Berufsstands (vgl. Li 2023; Moorkens 2018) und auch terminologisch fand ein Wandel statt. Der Begriff übersetzen wurde zunehmend von generieren abgelöst (vgl. z. B. Forcada 2017; Koehn 2020), was eine klare Unterscheidung zwischen Mensch und Maschine nahelegt.

Mit der Veröffentlichung großer Sprachmodelle (LLMs) erfuhr die Wahrnehmung von maschineller Übersetzung einen weiteren grundlegenden Wandel. LLMs haben in der Wahrnehmung „revolutioniert, wie Maschinen die menschliche Sprache verstehen, verarbeiten und generieren" (Ciubotaru 2025: 380), was die Debatte der Human Parity einmal mehr verschärft.

2.6.2 Wahrnehmung durch Sprachexpert:innen

Zu Beginn wurde der maschinellen Übersetzung von professionellen Übersetzer:innen eher kritisch entgegengesehen, da erste konkrete Anwendungen in den 1950er-Jahren vielversprechend waren und daher auch medial und fördertechnisch stark unterstützt wurden (vgl. Hutchins 2001b). Schnell stellte sich jedoch heraus, dass die maschinelle Übersetzung noch weit weg davon war, menschliche Qualität zu erreichen und sehr offensichtliche Fehler machte. Diese Erkenntnis prägte die Wahrnehmung der Übersetzer:innen hinsichtlich der Qualität nachhaltig. Trotz der stetigen Errungenschaften und dem stetigen technologischen Fortschritt hielt sich der schlechte Ruf aufgrund früher qualitativ minderwertiger Ergebnisse über lange Zeit hartnäckig. Maschinelle Übersetzung wurde daher schon bald nicht mehr als Bedrohung angesehen (vgl. Aslan 2023; Kenny 2018).

In den 2000er-Jahren, als die statistische MÜ den aktuellen Stand der Technik darstellte, zeigten Studien, dass Übersetzer:innen sich zwar noch primär auf Translation Memories stützen, gleichzeitig aber das Potenzial von qualitativ hochwertigen MÜ-Systemen erkannten, was eine Änderung in der Wahrnehmung auch unter Sprachexpert:innen nahelegt (vgl. Li 2023). Post-Editing von maschinengenerierten Übersetzungen wurde immer verbreiteter (vgl. Kenny 2018), wenngleich professionelle Übersetzer:innen dessen Effizienz unterschätzten oder die

Tätigkeit als nicht sonderlich hilfreich wahrnahmen. Ein überwiegender Teil, insbesondere im kreativen Raum, bevorzugte daher, ohne maschinelle Hilfe zu übersetzen (vgl. Moorkens et al. 2018).

Mit der Einführung von neuronalen maschinellen Übersetzungssystemen änderte sich die professionelle Wahrnehmung maßgeblich, was starken Einfluss auf das klassische Rollenbild des:der Übersetzer:in nahm. Zwar zeichnete sich bereits in den späten Jahren der SMÜ ein Trend zum Post-Editing von MÜ-Outputs ab, doch vor allem mit der NMÜ wurde ein klarer Shift vom reinen Übersetzen zum Post-Editing in der translatorischen Berufspraxis immer deutlicher (vgl. Kenny 2018). Gleichzeitig wirft die NMÜ bis heute neue Probleme für das Post-Editing auf. Die Fehler von SMÜ-Systemen waren häufig grammatischer Natur und daher klar erkennbar. NMÜ-Outputs hingegen wirken formal oft korrekt, weisen jedoch inhaltliche Fehler oder Ungereimtheiten auf. Revisor:innen erhalten dadurch ein falsches Gefühl der Korrektheit und müssen viel konzentrierter arbeiten (vgl. Gaspari 2024).

Eine weitere wichtige Wahrnehmungskomponente betrifft die Übersetzer:innenausbildung. Studien (vgl. z. B. Aslan 2023 & Delorme Benites et al. 2021) ergaben, dass bis zu 90 % der Studierenden maschinelle Übersetzungstool verwenden und weiterempfehlen, sich jedoch nicht sicher in deren Verwendung fühlen. Auch Chen (2024) hebt die Wichtigkeit der Eingliederung des Umgangs mit generativen KI-Tools in die formelle Übersetzer:innenausbildung hervor und betont, dass das Know-How im Umgang mit ihnen künftig einen klaren Wettbewerbsvorteil bilden kann. Zwar wird der Umgang mit Technologien wie CAT-Tools schon lange vorausgesetzt, mit dem Aufkommen von LLMs erweitert sich das erwartete Skillset von Übersetzer:innen jedoch noch weiter. Neben den linguistischen Kenntnissen wird häufig auch erwartet, dass Übersetzer:innen Beratungsdienstleistungen zu KI-Tools anbieten und tiefgehendes Wissen über den Umgang mit diesen aufweisen, diese trainieren und schließlich auch Qualitätskontrolle durchführen können (vgl. Chen 2024). Mit dem Fortschritt wird es dadurch auch immer schwieriger, klare Grenzen bei den Aufgabenbereichen von Übersetzer:innen zu ziehen.

Aktuelle Branchendaten unterstreichen, dass viele Übersetzer:innen den technologischen Wandel zunehmend skeptisch betrachten. Der European Language Industry Survey (ELIS) 2024 zeigt eine überwiegend negative Einstellung, insbesondere unter freiberuflichen Sprachdienstleister:innen.

> This negative sentiment of independent professionals continues to be fuelled by a perceived lack of fair remuneration linked to the rise of post-editing which replaces higher paid and more rewarding traditional human translation. Many participants fear that widespread acceptance of AI by the general public will legitimize the use of MT and thereby accelerate this evolution. (ELIS 2024: 4)

Professionelle Übersetzer:innen müssen aufgrund der mittlerweile branchenweiten Erwartung des Post-Editings finanzielle Einbußen hinnehmen, da dieses schlechter als eine Neuübersetzung bezahlt wird. Im Gegensatz zu den Studien von 2018, also vor der kommerziellen Einführung von generativen KI-Systemen, ist die Angst groß, von der KI ersetzt zu werden. Zwar scheint den befragten professionellen Übersetzer:innen laut des Berichts (ELIS 2024) klar, dass menschliche Arbeit nach wie vor qualitativ hochwertiger ist, jedoch zweifeln sie an, dass die Industrie und Endverbraucher:innen das auch erkennen.

2.6.3 Wahrnehmung durch Lai:innen

Zwar wurde im Laufe der Jahrzehnte medial viel über die maschinelle Übersetzung berichtet, doch wirklich greifbar wurde sie im nicht-professionellen Kontext erst mit der Verfügbarkeit des ersten Online-MÜ-Tools Babel Fish im Jahr 1998 (vgl. Schmalz 2019). Von den Medien wurde Babel Fish als wegweisende technologische Innovation dargestellt, jedoch war der Output des regelbasierten Systems nur sehr begrenzt nutzbar. Trotz großem Fehlerpotenzial wurde Babel Fish 2001 bereits für rund 1,3 Millionen Übersetzungen täglich genutzt (vgl. Schmalz 2019). Dies war ein frühes Indiz dafür, dass viele Nutzer:innen für grundlegendes Textverständnis bereit waren, qualitative Einschränkungen zu akzeptieren.

Die Veröffentlichung kostenloser Online-Übersetzungssysteme veränderte die Rolle der maschinellen Übersetzung im Alltag nachhaltig. Kenny (2018: 62–63) hebt hervor, dass für die Zwecke mancher Nutzer:innen eine qualitativ minderwertige Übersetzung reicht, insbesondere, wenn es nur darum geht, die Quintessenz eines Ausgangstext zu verstehen. In diesen Fällen kann davon ausgegangen werden, dass präferiert auf kostenlose Online-Tools zurückgegriffen wurde, statt professionelle Übersetzer:innen zu beauftragen.

Mit der Einführung von Googles SMÜ-Architektur in den 2000er-Jahren stieg nicht nur die Übersetzungsqualität, sondern vor allem die Alltagspräsenz. Nutzer:innen mussten nun nicht mehr aktiv nach einem Übersetzungstool suchen, sondern es war für viele nun direkt vom Startbildschirm ihres Browsers verfügbar. Diese Tendenz verfestigte sich insbesondere im Laufe der letzten Jahre noch weiter. Lai:innen kommen in beinahe jeder Anwendung in Kontakt mit der maschinellen Übersetzung (vgl. Bowker 2021). Browser bieten bereits in der Suchleiste maschinelle Übersetzung fremdsprachiger Webseiten auf Knopfdruck und beliebte Social-Media-Apps integrieren einen Übersetzungsbutton unter jedem fremdsprachigen Post als festen Bestandteil der Benutzeroberfläche (vgl. Abbildung 1 und 2).

Abbildung 1: Automatische Übersetzungsfunktion auf Instagram (eigener Screenshot)

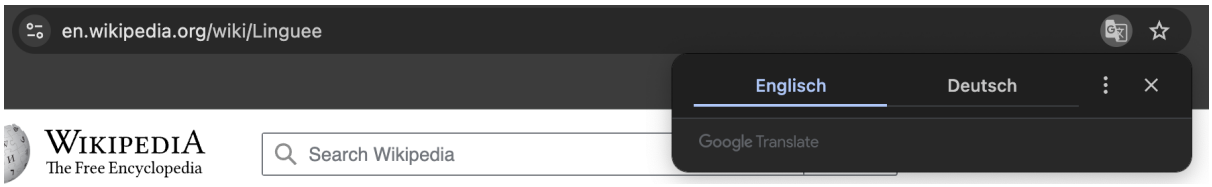

Abbildung 2: Automatische Übersetzungsfunktion im Google Chrome-Browser (eigener Screenshot)

Darüber hinaus werden Videos in zahlreichen Settings wie Instagram oder YouTube teilweise schon ohne explizite Nutzer:innennachfrage automatisch mittels KI untertitelt (vgl. Aslan 2023). Meta geht hier sogar noch einen Schritt weiter und bietet Contentcreator:innen mit einem Publikum von über 1.000 Nutzern die Möglichkeit, ihre Videos automatisch mittels KI zu übersetzen und die Videos zu dubben. Die KI ahmt dabei die Stimme der Creator:innen nach und es kann auch festgelegt werden, dass die Mundbewegungen mittels Videobearbeitung angepasst werden (vgl. Perez 2025). Es stellt sich die Frage, inwiefern Nutzer:innen erkennen, dass es sich dabei nicht um Originalinhalte, sondern um eine automatisch übersetzte und modifizierte Version handelt.

Für Lai:innen stehen häufig praktische Vorteile im Vordergrund. Dank maschineller Übersetzungsintegrationen erhalten Lai:innen Zugriff auf Inhalte, die sie sonst nicht konsumieren könnten (vgl. Aslan 2023). Neben dem Freizeitgebrauch eröffnet die maschinelle Übersetzung auch Lösungsansätze zu gesellschaftlich relevanten Problemfeldern, etwa die Kommunikation im Migrations- oder Gesundheitskontext. Dort ermöglicht sie eine grundlegende Verständigung zwischen Menschen, die über keine gemeinsame Lingua Franca verfügen. Dies geschieht sowohl institutionell, beispielsweise durch Programme wie MaTIAS (vgl. Macken et al. 2024), als auch privat über Endgeräte (vgl. Johnson 2016). Für Lai:innen steht damit in

bestimmten Fällen primär der praktische Nutzen im Vordergrund, nicht die absolute Genauigkeit.

Mit dem Aufkommen von LLMs ändert sich die Wahrnehmung von Lai:innen zusehends. Wie Branchendaten nahelegen, werden LLMs aufgrund der breiten medialen Berichterstattung und der kostenlosen Verfügbarkeit verstärkt für Übersetzungen verwendet und als vermeintlich gleichwertige Alternative zu professionellen Dienstleistungen genutzt (vgl. ELIS 2024). Gleichzeitig birgt die hohe sprachliche Natürlichkeit ein Wahrnehmungsrisiko. Wie in Kapitel 2.3.4 erläutert, sind LLMs anfällig für Halluzinationen, also für die Erzeugung plausibel klingender, aber inhaltlich unzutreffender Informationen. Während davon ausgegangen werden kann, dass Sprachexpert:innen solche Fehler meist erkennen, könnten Lai:innen, die nicht mit der Verwendung eines LLMs als Übersetzungssystem vertraut sind, die hohe Fluency als inhaltliche Korrektheit fehlinterpretieren (vgl. Xiao et al. 2025).

Zusammenfassend zeigt sich, dass Sprachexpert:innen und Lai:innen die maschinelle Übersetzung aus unterschiedlichen Perspektiven wahrnehmen. Während Sprachexpert:innen technologische Entwicklungen häufig mit ökonomischen und beruflichen Risiken verbinden, profitieren Lai:innen vor allem von der zunehmenden Zugänglichkeit, Geschwindigkeit und Nützlichkeit. Es steht bei Lai:innen also eher die funktionale Verständlichkeit und weniger die inhaltliche Präzision im Vordergrund. Dieser Faktor spielt für die Erkennung maschineller Übersetzung bzw. für die wahrgenommene Qualität im Rahmen dieser Arbeit eine zentrale Rolle.

2.7 Der Turing-Test

Der Turing-Test (vgl. Turing 1950) dient seit den frühen Anfängen der KI-Forschung als Maßstab dafür, ob maschineller Sprachoutput vergleichbar mit dem eines Menschen ist. Turings Konzept wird heute in zahlreichen Bereichen der Computer- und Sprachwissenschaft als theoretischer Bezugsrahmen herangezogen, darunter auch in der maschinellen Übersetzung. Im Folgenden werden die Grundidee und Entwicklungen, die Übertragung des Turing-Tests auf den Bereich der maschinellen Übersetzung sowie zentrale Kritikpunkte dargestellt.

2.7.1 Ursprünge und Entwicklung des Turing-Tests

Der Turing-Test findet seinen Ursprung im 1950 veröffentlichten Artikel von Turing (1950). Turing stellt eine scheinbar einfache Frage: „Can machines think?“ (Turing 1950: 433). Diese setzt er in den Kontext des Imitation Games, bei dem ein Mann (A), eine Frau (B) und ein:e Fragesteller:in (C) räumlich voneinander getrennt sind. Mittels schriftlicher Fragen muss C nun

herausfinden, wer der Mann und wer die Frau ist. Turing stellt sich nun die Frage, was wäre, wenn A durch eine Maschine ersetzt werden würde und ob C feststellen könnte, dass A in Wahrheit eine Maschine ist. Im Folgenden stellt er sich die Frage, ob die Maschine den:die Fragesteller:in oft genug davon überzeugen kann, ein Mensch zu sein. Es soll demnach überprüft werden, ob ein digitaler Computer intelligent genug sein kann, um Menschen täuschen zu können. Turing definiert keine klaren Metriken, wann das Imitation Game als bestanden gilt, prognostiziert jedoch:

> I believe that in about fifty years' time it will be possible, to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning. (Turing 1950: 442)

Daraus lässt sich ableiten, dass Turing kein fehlerfreies Bestehen des Imitation Game voraussetzt. Vielmehr beschreibt er eine Situation, in der ein:e durchschnittliche:r Fragesteller:in die Maschine nach fünf Minuten Befragung nur noch mit einer Wahrscheinlichkeit von höchstens 70 % korrekt identifizieren kann. Die daraus häufig abgeleitete 30-%-Marke bezeichnet somit die Wahrscheinlichkeit einer Fehlidentifikation, nicht den Anteil getäuschter Beobachter:innen, und wird in der Diskussion um den Turing-Test häufig als historischer Richtwert herangezogen (vgl. Shah & Warwick 2010).

Erste Versuche, wenngleich nicht immer explizit und formal als Turing-Tests bezeichnet, wurden bereits in den 1960er-Jahren erfolglos durchgeführt. Bekannte Beispiele hierfür sind ELIZA (Weizenbaum 1966), ein Computer, der eine Psychiaterin imitiert, und PARRY (Colby et al. 1972), ein anderer Computer, der hingegen vorgibt, ein Patient zu sein, der an paranoider Schizophrenie leidet. PARRY war so erstellt, dass er immer zum Thema passende, aber dennoch sehr allgemeine Antworten gab, und schaffte es in einem 1979 durchgeführten Experiment, Psychiater:innen in 50 % der Testfälle so zu verunsichern, dass sie sogar häufiger dachten, der menschliche Patient sei ein Computer als umgekehrt (vgl. Warwick & Shah 2016). Nach jahrzehntelanger Forschung und unzufriedenstellenden Ergebnissen gab es dann in den 1980er-Jahren einen Wandel in der Wissenschaft. Statt der Frage nachzugehen, ob eine Maschine, die den Test besteht, intelligent sei, fragte man sich nun eher, ob eine Maschine überhaupt je in der Lage sei, den Test zu bestehen (vgl. French 2000; Mauldin 1994).

1991 wurde der Loebner Preis eingeführt, ein jährlicher Turing-Test-Wettbewerb mit wechselnden spezifischen Themen (vgl. Shieber 1994). Die Programmierer:innen der ersten Maschine, die mehr als 30 % einer Jury in Themengebieten wie etwa Eishockey oder Haustiere davon überzeugen können, dass sie ein Mensch ist, gewannen einen monetären Preis von

100.000 US-Dollar (vgl. Shieber 1994). Der Preis wurde jedoch mit dem Jahr 2019 aufgrund mangelnder Fördermittel ausgesetzt (vgl. Wakefield 2019).

Ein wichtiger Meilenstein für den Turing-Test ist Eugene Goostman, ein Chatbot, der einen 13-jährigen, englischsprachigen Jungen aus der Ukraine simuliert (vgl. Warwick & Shah 2016). Technisch gesehen war Eugene Goostman im Vergleich zu älteren Programmen wie ELIZA oder PARRY ein späteres und stärker auf Wettbewerbssituationen zugeschnittenes Dialogsystem. Das System konnte 33 % der Jury davon überzeugen, dass es sich bei Eugene Goostman um einen Menschen handelte (vgl. Warwick & Shah 2016).

Eugene Goostmans Bestehen des Turing-Tests ließ jedoch die Wogen hochschlagen und genoss eine starke Medienberichterstattung. Oft wurde kritisiert, dass der Bot den Turing-Test nur bestehen konnte, weil er durch viele Tricks täuscht, etwa dadurch, dass er einen 13-Jährigen imitiert, gebrochenes Englisch spricht und wie ältere Systeme Fragen tendenziell ausweicht, statt sie direkt zu beantworten (vgl. Neufeld & Finnestad 2020).

Mit dem Aufkommen und der breiten Verfügbarkeit von LLMs ist der Dialog mit Maschinen alltäglicher geworden. Moderne LLMs können in dialogischen Situationen häufig überzeugende sprachliche Performanz zeigen (vgl. Jones & Bergen 2025). Zugleich wird in der Forschung hervorgehoben, dass diese Performanz nicht automatisch mit Bedeutungsverstehen oder Intelligenz gleichzusetzen ist (vgl. Bender & Koller 2020). Vielmehr stützen sich LLMs auf große Trainingskorpora, ohne Bedeutung im menschlichen Sinn zu verarbeiten oder über eine direkte Verbindung zur realen physischen Welt zu verfügen (vgl. Bender & Koller 2020; Sejnowski 2023).

Insgesamt zeigt die historische Entwicklung des Turing-Tests, dass sich Turings Gedankenexperiment im Laufe der Jahrzehnte zu einem flexibel ausgelegten praktischen Bezugsrahmen entwickelt hat. Diese Perspektive ist auch für seine Adaption auf maschinelle Übersetzung zentral. Im folgenden Unterkapitel wird daher die Relevanz des Turing-Tests in der Translationswissenschaft beleuchtet.

2.7.2 Der Turing-Test in der Translationswissenschaft

Die Übertragung des Turing-Tests auf die maschinelle Übersetzung erfolgt meist in einer rein produktorientierten Form. Während Turings ursprüngliches Imitation Game sprachliche Performanz in einem Dialog fokussiert, stellt sich im Übersetzungskontext die Frage, ob Leser:innen erkennen können, ob ein Zieltext von einem Menschen oder von einem maschinellen System stammt. Der Untersuchungsgegenstand verschiebt sich also auf den Zieltext. Der Turing-

Test fungiert damit als indikatorisches Verfahren zur Erfassung der wahrgenommenen Natürlichkeit, Kohärenz und Stilangemessenheit einer Übersetzung. Ein Beispiel für eine solche Herangehensweise findet sich etwa bei Popel et al. (2020); diese Studie wird in Abschnitt 3 ausführlicher diskutiert.

In der KI-Forschung und damit auch in Bezug auf die maschinelle Übersetzung gilt häufig die Gleichwertigkeit mit menschlicher Leistung als das ultimative Ziel. Der Turing-Test wird daher oftmals so abgeleitet, dass untersucht wird, ob ein Text menschlicher Qualität entspricht. Dies kann etwa durch Befragungen festgestellt werden. Die Idee ist demnach sehr ähnlich, da wie beim Turing-Test festgestellt werden soll, ob die Qualität der Übersetzung mit der eines Menschen gleichzusetzen ist. Ein wichtiges Beispiel sind die Untersuchungen von Hassan et al. (2018), die behaupteten, dass der Output ihres NMÜ-Systems mit den Übersetzungsleistungen eines Menschen gleichzusetzen ist. Die Durchführungsweise ist jedoch etwas anders, da die Übersetzungsqualität nicht anhand der Rezeption eines menschlichen Zielpublikums gemessen, sondern mittels Bewertungsmethoden festgestellt wird. Der Fokus hier liegt demnach auf der statistischen Ununterscheidbarkeit der Zieltexte, was jedoch nicht bedeutet, dass sie von Rezipient:innen auch so wahrgenommen werden und dass anhand eines Turing-Tests Human Parity erzielt worden wäre (vgl. Chatzikoumi 2020).

Ein weiteres wichtiges Beispiel sind die Untersuchungen von Wu et al. (2016), bei denen die Verbesserung der Übersetzungsqualität eines zum damaligen Zeitpunkt neuartigen NMÜ-Systems von Google belegt werden sollte. Mehrere hunderte Textproben auf Satzbasis, die nach Zufallsprinzip von Wikipedia und News-Webseiten entnommen wurden, wurden durch das NMÜ-System übersetzt und anschließend sowohl nach BLEU-Score als auch von Proband:innen nach einem Skalenschema bewertet. Ergebnisse zeigten, dass bei manchen Textproben Human Parity erreicht wurde. Zu beachten gilt jedoch, dass es sich um informative Textsorten handelt und dass die Autor:innen selbst anmerken, dass manche Zieltexte aufgrund der randomisierten Art der Textprobe ohne Kontext kaum bewertbar sind (vgl. Wu et al. 2016). Es handelt sich hierbei zwar nicht um einen klassischen Turing-Test, man sieht jedoch klar die Überschneidungen zwischen der Thematik der Human Parity und der Qualitätsbewertung in der Translationswissenschaft, wie in Kapitel 2.5 näher ausgeführt.

Als Ausnahme zu den weit verbreiteten, rein ausgangstextfokussierten Ansätzen zum Turing-Test in der Translationswissenschaft beschreibt Melby (2018) einen Translation Turing Test (TTT), der für den gesamten Übersetzungsprozess ausgelegt ist. Wie bei Alan Turings ursprünglicher Idee beginnt der Turing-Test mit einem Dialog, jedoch findet hier eine Änderung von zufälligen Personas zu A/B, also professionelle:r Übersetzungsprojektmanger:in bzw.

imitierende Maschine, und C als Kund:in statt. Statt eines willkürlichen Gesprächs sprechen die Parteien in einem vom:von der Kund:in initiierten Gespräch über den Übersetzungsauftrag, also beispielsweise dessen Spezifikationen, Ablauf oder Preis. Sobald ein Konsens erreicht wurde, wird der Dialog pausiert und die Abwicklung des Projekts selbst findet statt. Sollte der:die Projektmanager:in ein Mensch sein, findet er:sie eine:n Übersetzer:in, der:die Übersetzung wie mit dem:der Kund:in besprochen abwickelt, sollte es sich aber um eine Maschine handeln, führt sie die Übersetzung selbst durch. Unabhängig davon, wie lange der Prozess tatsächlich dauert, wird der Dialog erst wieder aufgegriffen, sobald die Übersetzung durchgeführt wurde. Die Übersetzung wird nun an den:die Auftraggeber:in übermittelt, der:die nun ausreichend Zeit bekommt, um einzuschätzen, ob er:sie mit einem Menschen oder einer Maschine kommuniziert hat. Wenn Kund:innen in mehr als 30 % der Fälle falsch liegen, gilt der TTT als bestanden. Dieser Ansatz kann als äußerst anspruchsvoll gesehen werden, da er eine Erweiterung des klassischen Turing-Tests ist. Bereits diesen zu bestehen, hat sich in der Vergangenheit als überaus schwierig erwiesen. Sollte die Maschine den Dialog jedoch meistern, müssen auch das Projektmanagement und die Übersetzung den Anforderungen standhalten. Da letztere völlig ohne menschlichen Einfluss und somit ohne Post-Editing erfolgen muss, gilt dieses Gesamtszenario als sehr unwahrscheinlich.

Wenngleich der Turing-Test in der KI-Forschung lange als Grundstock galt, gibt es nach wie vor geteilte Meinungen zum Test. Im folgenden Kapitel wird daher auf die Rezeption und Kritik des Gedankenexperiments und seinen Ausläufern eingegangen.

2.7.3 Rezeption und Kritik

Seit der Veröffentlichung von Turings Ideen (1950) lassen Turings Überlegungen die Wogen hochschlagen. Vielmals wird Turings Artikel auch als meistdiskutierter Artikel in den Computerwissenschaften bezeichnet. Während manche Turings Ansatz als unbrauchbar oder sogar als Behinderung für den wissenschaftlichen Fortschritt in der KI-Wissenschaft bezeichnen, sehen bzw. sahen andere in ihm den Grundstein der KI-Forschung (vgl. French 2000; Saygin et al. 2000).

Frühe Kritiken setzten bereits bei der zentralen Frage des Turing-Tests an, welche die Fähigkeit der Imitation eines Menschen durch eine Maschine adressiert. So war für manche, insbesondere zu Zeiten der Veröffentlichung des Turing-Tests, die Idee, dies überhaupt zu hinterfragen, bereits problematisch, da der Mensch als etwas Besonderes bzw. Überlegenes galt (vgl. Saygin et al. 2000). Dies geht auch Hand in Hand mit der Tatsache, dass Maschinen kein

Bewusstsein haben, demnach auch keine Gefühle oder Emotionen wie Erfolge oder Enttäuschung wahrnehmen und Kreativität aufweisen können (vgl. Saygin et al. 2000). French (2000: 2) sieht hier jedoch auch ein weitverbreitetes Missverständnis als Reibungspunkt, da der Test nie dafür intendiert war, Menschen, die beim Test scheiterten und eine Maschine fälschlicherweise als Mensch identifizierten, als unintelligent zu bezeichnen, sondern vielmehr umgekehrt, Maschinen als intelligent genug darzustellen.

Häufig zitierte Kritikpunkte des Turing-Tests sind die Sprachzentriertheit und damit die Gleichstellung der menschlichen Intelligenz mit der Fähigkeit einer Maschine, eine menschliche Konversation zu imitieren. Ein Gedankenexperiment zu diesem Thema führt Searle (1980) durch. Er argumentiert, dass ein System auch dann keine semantische Kompetenz besitzen muss, wenn es korrekt erscheinende sprachliche Antworten generiert. Für Searle beweist reine Performanz keine Intentionalität, kein Verständnis und keine Bedeutungszuschreibung. Searle adressiert Turings Frage nach der Fähigkeit von Maschinen zu denken direkt und beantwortet sie schlichtweg mit nein. Die Menschen, die Maschinen programmieren, können denken, die Maschinen aber nicht. Um zu denken, fehlen Maschinen biologische Merkmale wie ein Nervensystem (vgl. Searle 1980).

Auch Block (1981) kritisierte dies und legte mit seinem hypothetischen System Blockhead nahe, dass eine Maschine von Menschen sehr gut programmiert werden kann. Statt tatsächlich intelligent wie ein Mensch zu sein und die Inputs rational zu verstehen und zu verarbeiten, würde sie vielmehr auf den umfangreichen Speicher zurückgreifen, der vorab implementiert wurde, und so menschliches Verhalten imitieren. Blocks Hypothese erinnert stark an die Grundprinzipien heutiger LLMs, die auf einem riesigen Trainingskorpus trainiert werden und auf Basis der daraus gelernten Muster und internen Repräsentationen Antworten generieren, die mit der Sprachproduktion von Menschen vergleichbar ist.

Ein weiterer Kritikpunkt ist, dass es einen Unterschied macht, wer tatsächlich an einem Turing-Test teilnimmt. Im Rahmen des Loebner Preises wurden beispielsweise bewusst Leute aus der allgemeinen Bevölkerung als Preisrichter ausgewählt, insbesondere mit dem Augenmerk darauf, keine bedeutende Expertise im Bereich der Computerwissenschaften zu haben (vgl. Mauldin 1994; Shieber 1994). Bei einem Turing-Test mit Fokus auf maschineller Übersetzung kann daher davon ausgegangen werden, dass Sprachexpert:innen die sprachlichen Muster der Maschine eher erkennen als Lai:innen.

Zusammengefasst bleibt der Turing-Test dennoch ein wertvolles und viel verwendetes, wenngleich begrenzt belastbares Werkzeug zur Untersuchung der Wahrnehmung maschineller

Übersetzungen. Wie oben beschrieben liegt seine Stärke weniger in der Erfassung von Intelligenz, sondern in der Analyse menschlicher Wahrnehmung in Bezug auf die maschinelle Sprachproduktion. Dennoch lässt sich feststellen, wie zeitlos und flexibel auslegbar Turings Überlegungen sind, weshalb sie als Untersuchungsgrundlage für den empirischen Teil dieser Arbeit gewählt wurden.

3 Aktueller Forschungsstand

Während Kapitel 2 die theoretischen und qualitativen Grundlagen der Masterarbeit darlegt, widmet sich Kapitel 3 der empirischen Forschungslage zu den für diese Arbeit zentralen Fragen der Erkennbarkeit, Wahrnehmung und Bewertung maschineller Übersetzungen.

Eine vielzitierte Publikation von Hassan et al. (2018) legt nahe, dass neuronale maschinelle Übersetzungssysteme bereits ein Qualitätsniveau erreichen können, das mit Humanübersetzungen vergleichbar ist. In einer Folgestudie (Läubli et al. 2018) wurde festgestellt, dass bei maschinell generierten Übersetzungen auf Satzebene tatsächlich kaum statistische Unterschiede in der Bevorzugung festgestellt werden konnten, es aber eine klare Tendenz zur Humanübersetzung auf Dokumentebene gab. Die Studie verdeutlicht damit nicht nur Unterschiede in der tatsächlichen Qualität, sondern vor allem in der Wahrnehmung. Während einzelne Sätze maschinell übersetzter Texte als durchaus menschenähnlich eingestuft werden, erkennen Leser:innen auf Dokumentebene deutliche Brüche in Kohärenz und Wirkung. Diese Ergebnisse legen nahe, dass Qualitätsunterschiede zwischen Human- und maschinell generierten Übersetzungen nicht ausschließlich textintern, sondern auch kontextabhängig wahrgenommen werden.

Aktuelle Studien (vgl. Ghassemiazghandi 2024; Kamaluddin et al. 2024; Sizov et al. 2024) zeigen, dass NMÜ-Übersetzungen, darunter insbesondere DeepL-Übersetzungen, in den letzten Jahren immer höhere BLEU-Scores erreichen und in automatischen Evaluationsmetriken teilweise höhere Werte als Humanübersetzungen erzielen, während von ChatGPT geschriebene Essays höhere Qualitätsbewertungen als von Menschen verfasste Texte erzielen. NMÜ-Systeme tendieren dabei zu akkurateren Übersetzungen als LLMs, während sich LLMs besser vom Ausgangstext lösen können. Dies führt dazu, dass LLMs im Sprachenpaar Deutsch–Englisch Humanübersetzer:innen in Bezug auf den Sprachgebrauch, insbesondere bei Terminologie und Modalverben, besser nachahmen konnten, wenngleich sie insgesamt jedoch stilistisch weniger variantenreich bleiben und daher nur begrenzt kreative Lösungen anbieten (vgl. Daems et al. 2025; Sizov 2024). Darüber hinaus halluzinieren LLMs manchmal, was im Falle der Anwendung als Übersetzungstool als eine inkorrekte, vollständige Loslösung vom Ausgangstext beschrieben werden kann (vgl. Ji et al. 2024). Dies ist insbesondere für kreative Übersetzungen relevant. Während zu ausgangstexttreue Übersetzungen unnatürlich klingen könnten, könnten Halluzinationen zu Bedeutungsverschiebungen oder Brüchen bei der Tonalität bzw. dem Markenstil führen.

Während maschinelle Übersetzungstools in automatischen Evaluationsmetriken teilweise sehr gute Ergebnisse erzielen, zeigen mehrere Studien, dass diese Bewertungen nicht

zwangsläufig mit der von Menschen wahrgenommenen Übersetzungsqualität übereinstimmen (vgl. Agrawal 2024; Kocmi et al. 2021).

Humanübersetzungen erzielen in aktuellen Studien höhere Kreativitätsrankings als post-editierte oder reine MÜ und werden insgesamt besser bewertet (vgl. Guerberof-Arenas & Toral 2022). Humanübersetzer:innen lösen sich besser vom Ausgangstext und tendieren zu neuartigen, flexiblen Übersetzungslösungen, während maschinelle Übersetzungen sehr nah am Ausgangstext bleiben. Selbst wenn dieselben Übersetzer:innen mit Post-Editing arbeiten, neigen sie zu weniger kreativen Lösungen und einer höheren Fehleranzahl, was auf den Wandel von Urheber:in zu Lektor:in zurückzuführen ist (vgl. Daems et al. 2025; Guerberof-Arenas & Toral 2022).

Zu beachten sind zudem Unterschiede zwischen verschiedenen Textsorten. Mehrere Studien weisen darauf hin, dass die maschinelle Übersetzung insbesondere bei kreativen Textsorten nach wie vor auf Grenzen stößt. Toral und Way (2018) und Guerberof-Arenas und Toral (2022, 2024) zeigen, dass maschinelle Übersetzungssysteme bei literarischen oder stark persuasiven Texten weiterhin deutliche Einschränkungen aufweisen. Während maschinelle Übersetzungstools bei strukturellen Kategorien wie der durchschnittlichen Satzlänge mit Humanübersetzungen vergleichbare Ergebnisse erzielen konnten, schnitten sie bei linguistischen Merkmalen wie der Terminologie sowie kreativen Übersetzungslösungen deutlich schlechter ab.

Darüber hinaus weisen Guerberof-Arenas und Toral (2024) darauf hin, dass die Rezeption maschineller Übersetzungen stark vom Sprachenpaar abhängen kann. In ihrer Studie nahmen Proband:innen Unterschiede zwischen Human- und Maschinenübersetzungen im Sprachenpaar Englisch–Katalanisch deutlich stärker wahr als im Sprachenpaar Englisch–Niederländisch. Dies könnte darauf zurückgeführt werden, dass niederländische Leser:innen es gewohnt sind, Literatur auch in der Originalsprache Englisch zu konsumieren, was eine Präferenz zu einer wörtlicheren, ausgangstexttreueren Übersetzung nahelegt.

Diese Ergebnisse verdeutlichen, dass Qualitätsunterschiede zwischen Human- und maschinellen Übersetzungen nicht ausschließlich auf den technischen Entwicklungsstand der Systeme zurückzuführen sind, sondern auch von der jeweiligen Textsorte sowie vom sprachlichen und kulturellen Kontext abhängen.

Hiebl und Gromann (2024) vergleichen Humanübersetzungen mit maschinellen Übersetzungen verschiedener MÜ-Tools mithilfe des Best-Worst-Scalings (BWS). Teilnehmende wählten dabei jeweils die beste und schlechteste Übersetzung aus und bewerteten sie zusätzlich mit einer Likert-Skala. Die Ergebnisse zeigen, dass Humanübersetzungen insgesamt am häu-

figsten als beste Option gewählt wurden, während DeepL direkt dahinter die zweitbesten Resultate lieferte. Die Ergebnisse zeigen auch, dass bereits kleine Unterschiede in Wortwahl oder Stil einen erheblichen Einfluss auf die Wahrnehmung von Übersetzungsqualität haben können.

In einer Studie von Kaspere et al. (2023) zur Wahrnehmung und Akzeptanz maschineller Übersetzungen konnten klare Unterschiede zwischen professionellen Sprachdienstleister:innen, also Übersetzer:innen, Copywriter:innen und Lektor:innen, und Lai:innen festgestellt werden. Beiden Zielgruppen wurden mit Google Translate aus dem Englischen ins Litauische übersetzte Texte vorgelegt und mithilfe von Eye-Tracking wurde untersucht, inwiefern sich Augenbewegungen sowie die Dauer der Textbetrachtung zwischen den Gruppen unterschieden. Alle Proband:innen wussten, dass es sich um maschinenübersetzte Textproben handelt.

Die Ergebnisse zeigen, dass sich professionelle Sprachdienstleister:innen länger und genauer mit den vorgelegten Textproben auseinandersetzen. Zudem wurde festgestellt, dass die Qualität und Gebrauchstauglichkeit von Lai:innen höher eingeschätzt wurde als von professionellen Sprachdienstleister:innen. Dies lässt sich vermutlich darauf zurückführen, dass professionelle Sprachdienstleister:innen darauf trainiert sind zu erkennen, welche Merkmale einen qualitativ hochwertigen Text ausmachen (vgl. Kaspere et al. 2023).

Für die vorliegende Arbeit sind diese Erkenntnisse insbesondere im Hinblick auf die Frage interessant, ob sich Unterschiede in der Wahrnehmung von Sprachexpert:innen und Lai:innen feststellen lassen. Zum einen kann davon ausgegangen werden, dass ein längeres Auseinandersetzen Einfluss auf die Wahrnehmung haben könnte. Zum anderen könnte eine erhöhte Sensibilität für qualitative Merkmale dazu führen, dass Texte im Rahmen der Bewertung kritischer auf potenzielle Fehler oder stilistische Ungereimtheiten untersucht werden.

Neben der Bewertung der Übersetzungsqualität beschäftigt sich ein wachsender Teil der Forschung auch mit der Frage, inwieweit Leser:innen KI-generierte Texte überhaupt erkennen (z. B. Doru et al. 2025; Milička et al. 2025). Gehrmann et al. (2019) stellten fest, dass Proband:innen in etwa 54 % der Fälle, und damit nur knapp über dem Zufallswert, richtig zuordnen konnten, ob es sich bei dem Probetext um einen KI-generierten oder um einen von Menschen verfassten Text handelt. Mehrere Studien zeigen zudem, dass die Textsorte eine große Rolle bei der Erkennbarkeit von KI-generierten Texten spielen kann. Bei stark fachspezifischen Textsorten wie medizinischen oder geisteswissenschaftlichen Essays liegt die richtige Erkennungsrate weit über dem Zufallswert (vgl. Doru et al. 2025). Darüber hinaus zeigen Studien, dass Proband:innen je nach Textsorte unterschiedliche Vorannahmen über die Urheberschaft eines Textes haben. So erwarten sie bei stark informativen oder wissenschaftlichen Texten eher, dass diese von einem KI-Tool generiert wurden, während sie bei interaktiven oder

erzählenden Textsorten die Autor:innenschaft tendenziell eher Menschen zuschreiben (vgl. Milička et al. 2025). Diese Ergebnisse legen nahe, dass das Bewusstsein darüber, ob ein Text von einem Menschen oder einer Maschine erstellt wurde, die Wahrnehmung und Bewertung eines Textes beeinflussen kann. Im translationswissenschaftlichen Kontext ist dabei jedoch zu berücksichtigen, dass nicht nur allgemeine Zuschreibungen zu Mensch oder Maschine relevant sind, sondern auch spezifische Erwartungen an Richtigkeit, Verlässlichkeit und Angemessenheit einer Übersetzung eine zentrale Rolle einnehmen. Diese Annahme wird von Asscher und Glikson (2023) explizit in einem Übersetzungssetting nachgewiesen. Sobald Teilnehmende wussten, dass es sich um eine maschinelle Übersetzung handelte, bewerteten sie diese niedriger hinsichtlich ihrer Genauigkeit und Zuverlässigkeit. Auch Qiu (2025) stellte in einer Untersuchung zu von DeepL rohübersetzten Untertiteln fest, dass Proband:innen, die die Untertitel als maschinell übersetzt erkannten, angaben, weniger Interesse an dem Film zu haben. Dennoch brach kein:e Teilnehmer:in den Film ab. Dies legt nahe, dass die Wahrnehmung von maschinell übersetzten Texten einen Einfluss auf die Rezeption hat, wenngleich sie diese nicht verhindert.

Klassische und abgewandelte Formen des Turing-Tests werden auch mehr als 50 Jahre nach seiner Einführung noch regelmäßig durchgeführt. Jones und Bergen (2024a, 2024b, 2025) führten Turing-Tests mit verschiedenen LLMs durch. In ihrer Studie führten Proband:innen Konversationen mit den KI-Tools in einer klassischen Nachrichten-App-Situation unter Anwendung verschiedener Konfigurationen durch. In ihrem ersten Versuch zeigte ChatGPT-4.0 die stärkste Leistung und konnte bereits bis zu 50 % aller Proband:innen davon überzeugen, dass sie mit einem Menschen Nachrichten austauschen. In der Folgestudie (vgl. Jones & Bergen 2025) wurden stärkere KI-Architekturen verwendet, darunter auch GPT-4.5, die GPT-Version, die auch im Rahmen dieser Arbeit verwendet wird. ChatGPT-4.5 wurde mit 73 % unter Anwendung von Persona Prompting am häufigsten als Mensch wahrgenommen, während der ungepromptete Bot immerhin bis zu 42 % aller Menschen davon überzeugen konnte, ein Mensch zu sein. Diese Resultate legen nahe, dass der sprachliche Output von ChatGPT-4.5 in vielen Fällen als menschlich wahrgenommen wird und damit auch, dass LLMs den Turing-Test bestehen können. Diese Ergebnisse sind besonders relevant für Studien zur KI-Erkennung und zeigen darüber hinaus, dass Persona Prompting die Wahrnehmung stark beeinflussen kann.

Rathi et al. (2025) untersuchten, wie gut ChatGPT bei Inverted und Displaced Turing-Tests abschneidet. Bei einem Inverted Turing-Test übernimmt ein KI-System die Rolle des:der Fragesteller:in. Bei einem Displaced Turing-Test hingegen bewerten Personen vorab erstellte Inhalte und müssen einschätzen, ob diese von Menschen oder KI-Systemen stammen (vgl. Rathi et al. 2025). Die Ergebnisse zeigen, dass GPT-4 sowohl in Displaced als auch in Inverted

Turing-Tests häufiger als Mensch wahrgenommen wurde als tatsächliche menschliche Teilnehmer:innen (vgl. Rathi et al. 2025). Dies bestätigt frühere Forschungsergebnisse, nach denen KI-Systeme häufiger fälschlicherweise als Mensch wahrgenommen werden als umgekehrt (vgl. auch Chein et al. 2024).

Trotz der regen Forschungsaktivität zur Wahrnehmung und Erkennbarkeit von maschinell übersetzten Texten gibt es bislang nur wenige explizite Untersuchungen darüber, inwieweit Leser:innen zwischen maschinell generierten und humanübersetzten Texten unterscheiden können. Vor allem fehlen Studien, die sich auf kreative bzw. appellative Textsorten stützen, welche die Expertise der Leser:innen miteinbeziehen und spezialisierte Übersetzungssysteme wie DeepL und LLMs wie ChatGPT direkt miteinander vergleichen. Aufgrund dieser Forschungslücke untersucht die vorliegende Arbeit, inwiefern Leser:innen von DeepL und ChatGPT generierte und humanübersetzte kreative Texte unterscheiden können und welchen Einfluss ein sprachwissenschaftlicher Hintergrund oder Berufserfahrung im Umgang mit Texten auf diese Einschätzung hat.

4 Methodik

Im folgenden Kapitel wird die Forschungsmethodik der vorliegenden Arbeit erläutert. Zur Beantwortung der Forschungsfragen wurden Daten mittels eines Online-Fragebogens erhoben, der über SoSci Survey (2026) aufgesetzt und durchgeführt wurde. Neben den Zielgruppen werden der Aufbau des Fragebogens, die Aufgabenstellungen für die Teilnehmer:innen sowie die Vorgehensweise bei der Datenanalyse vorgestellt. Darüber hinaus werden die Auswahl der Textausschnitte sowie deren Besonderheiten erläutert.

4.1 Zielgruppen

Da der Forschungsfokus auf einem Turing-Test-ähnlichen Experiment mit ausschließlich dem deutschen Zieltext als Bewertungsmaterial lag, war ein hohes Deutschniveau der Teilnehmenden Voraussetzung. Im Rahmen dieser Untersuchung galt daher eine Mindestanforderung von C1 gemäß dem Europäischem Referenzrahmen.

> Kann ein breites Spektrum anspruchsvoller, längerer Texte verstehen und auch implizite Bedeutungen erfassen. Kann sich spontan und fließend ausdrücken, ohne öfter deutlich erkennbar nach Worten suchen zu müssen. Kann die Sprache im gesellschaftlichen und beruflichen Leben oder in Ausbildung und Studium wirksam und flexibel gebrauchen. Kann sich klar, strukturiert und ausführlich zu komplexen Sachverhalten äußern und dabei verschiedene Mittel zur Textverknüpfung angemessen verwenden. (Europäischer Referenzrahmen o.J.)

Dieses Niveau wurde als angemessen betrachtet, da Teilnehmende auf diesem Niveau zum einen ein sehr starkes allgemeines Textverständnis aufweisen und zum anderen auch implizite Bedeutungen verstehen, was als wesentlich für eine fundierte Beurteilung der Texte im Rahmen des Fragebogens angesehen wurde. Darüber hinaus gilt C1 für viele sprachbezogene Bachelorstudien als Mindestniveau, wie zum Beispiel für Transkulturelle Kommunikation an der Universität Wien (vgl. Universität Wien o.J.), was wiederum als besonders relevant für die Expert:innengruppe erschien.

Um herauszufinden, ob es einen Unterschied in der Wahrnehmung von maschineller Übersetzungsqualität zwischen Lai:innen und Sprachexpert:innen gibt, wurden zwei Zielgruppen gewählt. Als Sprachexpert:innen wurden zum einen Teilnehmende definiert, die über ein abgeschlossenes relevantes Sprachstudium verfügten. Als relevant wurden in diesem Kontext abgeschlossene Bachelor- oder Bakkalaureatsstudien oder höhere Abschlüsse in einem sprach- oder translationswissenschaftlichen Fach betrachtet. Beispiele hierfür sind Translationswissenschaft, Deutsch als Fremd- und Zweitsprache, Linguistik oder Germanistik. Zum anderen wur-

den auch Teilnehmende zur Expert:innengruppe gezählt, die über mindestens drei Jahre einschlägige berufliche Erfahrung im Umgang mit der deutschen Sprache verfügten, was der Mindeststudienzeit für die meisten relevanten Bachelorstudien entspricht. Jedenfalls als einschlägige Berufserfahrung wurden Copywriting, Übersetzung, Lokalisierung und Lektorat erachtet. Für die Einstufung als Sprachexpert:in reichte es, eine der beiden Voraussetzungen zu erfüllen. Diese Kriterien wurden gewählt, da sie eine erhöhte metasprachliche Sensibilität und Erfahrung in der Analyse sprachlicher Strukturen voraussetzen und somit die sprachliche Auseinandersetzung von Lai:innen überschreiten.

Als Lai:innen wurden Teilnehmende klassifiziert, die weder ein abgeschlossenes Studium in einem sprach- oder translationswissenschaftlichen Fach noch einschlägige berufliche Erfahrung im professionellen Umgang mit der deutschen Sprache verfügten. Diese Kriterien wurden gewählt, um eine klare Abgrenzung zur Gruppe der Sprachexpert:innen zu gewährleisten.

4.2 Untersuchungsmaterial und Übersetzungen

Im Folgenden werden zunächst die Kriterien erläutert, nach denen die Textausschnitte für die Untersuchung ausgewählt wurden. Anschließend wird der Übersetzungsprozess dargestellt und die einzelnen Textausschnitte hinsichtlich ihrer sprachlichen, stilistischen und übersetzungsrelevanten Eigenschaften charakterisiert, wobei auch Besonderheiten der jeweiligen Übersetzungen berücksichtigt werden.

4.2.1 Auswahlkriterien der Textausschnitte

Da im Rahmen dieser Arbeit der Schwerpunkt auf der Übersetzung von appellativen bzw. kreativen Texten lag, wurden für die Untersuchung Textausschnitte von Webseiten verschiedener Marken herangezogen. Diese Textsorte wurde für die vorliegende Untersuchung als besonders relevant erachtet, da sie häufig eine hohe Dichte an evaluativen, emotionalen und stilistisch konnotierten Elementen aufweist, die für maschinelle Übersetzungssysteme eine besondere Herausforderung darstellen können.

Die Textausschnitte wurden aus dem Englischen ins Deutsche übersetzt und umfassten zwischen 80 und 130 Wörter. Diese Länge wurde als angemessen erachtet, um aussagekräftige Ergebnisse nicht nur für einzelne Sätze, sondern auch für zusammenhängende Textabschnitte zu erhalten, ohne dabei durch zu lange Bewertungsaufgaben Ermüdungseffekte im Rahmen des Fragebogens zu riskieren.

Darüber hinaus wurde darauf geachtet, Texte aus verschiedenen Branchen auszuwählen, um mögliche Verzerrungen durch branchenspezifische Ausdrucksmuster zu vermeiden und eine größere Bandbreite an stilistischen Variationen abzudecken.

4.2.2 Erstellung der Übersetzungen

Die maschinellen Übersetzungen wurden im Zeitraum vom 27.01.2026 bis zum 02.02.2026 mit DeepL (2026) und ChatGPT (OpenAI 2026) erstellt. Für beide Systeme wurde jeweils die Pro-Version für die Generierung der deutschen Zieltexte verwendet. Bei DeepL konnten im Vergleich zur kostenlosen Version Voreinstellungen getroffen werden, wie beispielsweise der gewünschte Formalitätsgrad. In der vorliegenden Untersuchung wurden die Einstellungen jedoch auf automatisch belassen, um das vom System gewählte Formalitätsniveau nicht zu beeinflussen.

Bei ChatGPT bietet die Pro-Version im Vergleich zur kostenlosen Version mehr Transparenz bezüglich des eingesetzten Sprachmodells. Dadurch konnte sichergestellt werden, dass zum Zeitpunkt der Generierung das neueste GPT (Version 5.2) zur Anwendung kam.

Wie bereits in Kapitel 2.4.4 beschrieben, benötigt ChatGPT zur Generierung einer Übersetzung einen Prompt. Da es sich bei ChatGPT nicht primär um ein Übersetzungssystem handelt, kommt dem Prompting eine zentrale Bedeutung zu. Sejnowski (2023) zeigt, dass detailreiche und zielgerichtete Prompts zu qualitativ hochwertigeren ChatGPT-Outputs führen. Dies legt nahe, dass das Prompting bei der Generierung hochwertiger Zieltexte eine entscheidende Rolle spielen kann.

Für die Gestaltung des Prompts existieren, wie in Kapitel 2.3.4 beschrieben, zahlreiche Strategien. Es wurde entschieden, einen Prompt für alle Textausschnitte zu verwenden, um etwaige Verzerrungen zwischen den einzelnen Texten zu vermeiden. Auf Basis der theoretischen Grundlagen wurden verschiedene Promptformulierungen getestet, wobei sich zeigte, dass eine Balance zwischen ausreichend konkreten Anweisungen und einer zu starken inhaltlichen Festlegung von zentraler Bedeutung ist. Der erwähnte Basisprompt *Übersetzen Sie bitte den folgenden Ausgangstext ins Deutsche* stellte sich schnell als ungeeignet heraus, da er zu wenige explizite Anweisungen enthielt und zu sehr wörtlichen Übersetzungen der Textausschnitte führte. Im Gegensatz dazu erwies sich zu spezifisches Prompting, etwa Angaben zum Formalitätsgrad oder Markenton, ebenfalls als ungeeignet, da die Textausschnitte aufgrund der sehr unterschiedlichen Branchen auch sehr unterschiedliche Anforderungen aufweisen.

Am geeignetsten erwies sich ein Persona Prompt (vgl. Elnashar et al. 2023; Lutz et al. 2025), bei dem dem KI-System die Rolle eines:einer erfahrenen Übersetzer:in mit der Spezialisierung auf Marketingkommunikation zugewiesen wurde. Ziel dieses Prompts war es, das Sprachmodell zu Übersetzungsentscheidungen zu veranlassen, die sich an denen professioneller menschlicher Übersetzer:innen orientieren. Darüber hinaus wurde besonderer Wert auf die Beibehaltung der Intention und des Markentons gelegt, was im Kontext der Marketingübersetzung als besonders wichtig erachtet wurde. Zudem wurde das System angewiesen, keine Erklärungen hinzuzufügen (vgl. Abbildung 3).

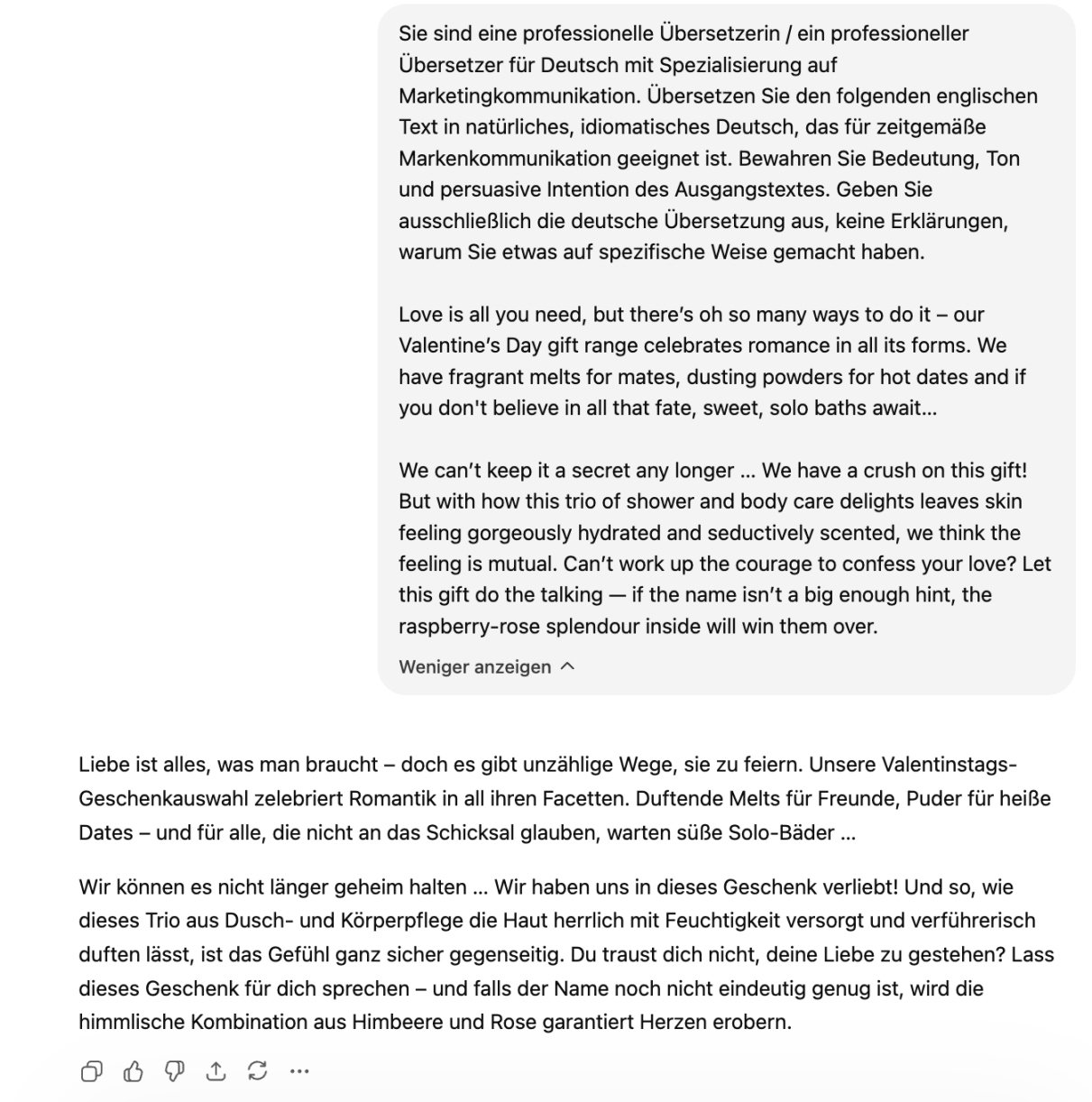

Abbildung 3: Prompt in ChatGPT mit generierter Übersetzung von Textausschnitt 5 (Lush) (eigener Screenshot)

Für die Humanübersetzung wurde eine auf Marketingkommunikation spezialisierte Übersetzerin mit mehr als zehn Jahren Berufserfahrung ausgewählt. Es wurde bewusst nur eine Übersetzerin einbezogen, um mögliche Verzerrungen aufgrund qualitativer Unterschiede zwischen mehreren Übersetzer:innen zu vermeiden. Die Übersetzerin erhielt Anweisungen, die inhaltlich mit dem Prompt vergleichbar waren, und wurde darüber informiert, dass ihre Übersetzungen im Rahmen der vorliegenden Studie verwendet werden. Die Textausschnitte wurden der Übersetzerin in einem Microsoft-Word-Dokument zur Verfügung gestellt und als Auftragszeitraum wurde eine Woche festgelegt.

Sowohl im Prompt als auch in der Anleitung für die Humanübersetzerin wurde bewusst darauf verzichtet, kreative Übersetzungsentscheidungen zu unterbinden, um auch transkreative Lösungen zu ermöglichen. Dies wurde als notwendig erachtet, um ein realistisches Übersetzungsszenario nachzubilden, da solche Strategien insbesondere bei der Übersetzung von kreativen Textsorten häufig zur Anwendung kommen.

Abschließend wurden sämtliche Übersetzungen einheitlich formatiert, um visuelle Unterschiede zu vermeiden. Dies war erforderlich, da DeepL häufig den gesamten Zieltext fett hervorhob, während ChatGPT in einem Beispiel einen zusätzlichen Zeilenumbruch zwischen Textabschnitten erzeugte und in einem anderen Textausschnitt einzelne Textpassagen ebenfalls fett formatierte. Sprachlich wurden die Übersetzungen nicht verändert.

4.2.3 Charakterisierung der Textausschnitte und ihrer Übersetzungen

Die ausgewählten Textausschnitte stammen aus unterschiedlichen Branchen, wie etwa der Automobil-, Sport- oder Lebensmittelindustrie, und unterscheiden sich hinsichtlich der Dichte stilistischer und kreativer Elemente. Während einige davon näher an informativen Produktbeschreibungen bleiben, weisen andere für kreative Textsorten typische Elemente wie evaluierende Adjektive, Reimschemata oder Wortspiele auf (vgl. Reiß 1971). Im Folgenden werden sowohl die Textausschnitte als auch ihre Übersetzungen hinsichtlich sprachlicher Besonderheiten, daraus resultierender Übersetzungsherausforderungen sowie stilistischer Unterschiede untersucht. Die vollständigen Textausschnitte sowie ihre Übersetzungen sind im Anhang A1 dokumentiert. Insgesamt wurden die Textausschnitte so gewählt, dass sie einen graduellen Anstieg stilistischer und kreativer Komplexität aufweisen. Während die Textausschnitte 1–4 stärker informationsorientiert und strukturell konventionell bleiben, zeigen die Textausschnitte 5–8 eine höhere Dichte an komplexen kreativen und stilistischen Mitteln. Diese Unterschiede werden in der Diskussion im Hinblick auf ihre Auswirkungen auf die Erkennbarkeit maschineller Übersetzungen weiter aufgegriffen. Um Redundanzen zu vermeiden, werden bei späteren

Textausschnitten insbesondere jene Merkmale hervorgehoben, die neue oder besonders relevante Unterschiede zwischen den Übersetzungen zeigen.

4.2.3.1 Textausschnitt 1: Mazda

Der erste Textausschnitt stammt vom japanischen Autohersteller Mazda (vgl. Mazda 2026). Wie bei einigen der anderen Textausschnitte werden hier produktbeschreibende, informative Elemente mit persuasiver Sprache kombiniert. Es wird ein neues Automodell vorgestellt, wobei mit *Japanese craftsmanship* oder *In Japan* insbesondere die japanische Herkunft und die damit implizierte hohe Produktqualität betont wird. Der Textausschnitt enthält zahlreiche evaluierende Adjektive wie *breathtaking* sowie bildhafte Formulierungen wie *empty space is seen as a canvas*, die auf eine emotionale Ansprache abzielen. Dennoch bleibt der Fokus stark auf dem Produkt sowie auf seinen Funktionen und Vorteilen.

Linguistisch kann dieser Textausschnitt als relativ einfach eingestuft werden. Darüber hinaus enthält er viele standardisierte Begriffe aus der Automobilbranche, weshalb er diesbezüglich kaum Schwierigkeiten aufwerfen dürfte. Die zahlreichen evaluierenden und emotionalen Adjektive könnten jedoch herausfordernd sein, da sie an die deutschsprachige Zielkultur angepasst werden müssen und daher nicht wörtlich übersetzt werden können.

Die maschinell generierten Übersetzungen bleiben dem englischen Ausgangstext insgesamt sehr treu. DeepL hält sich dabei weitgehend an den Satzbau des Ausgangstextes, während ChatGPT diesen teilweise verändert, etwa durch die Verwendung von Doppelpunkten, um den Text lebendiger zu gestalten. Die Humanübersetzerin restrukturiert die Sätze hingegen deutlicher, was typisch für Transkreation im Marketingbereich ist, bei der insbesondere in appellativen Texten häufig die Wirkung gegenüber der formalen Nähe zum Ausgangstext priorisiert wird (vgl. Błaszkowska 2021).

Beide maschinellen Übersetzungen bleiben auch hinsichtlich Terminologie und Bildsprache sehr nah am Ausgangstext, wodurch sie stellenweise technisch und wenig idiomatisch wirken. So wird *canvas* beispielsweise als *Leinwand* übersetzt, obwohl hier konzeptuell eher ein leerer Raum gemeint ist. Die Humanübersetzerin verwendet hingegen mit *ein unbeschriebenes Blatt* eine deutlich freiere Lösung, die die intendierte Wirkung des Ausgangstextes akkurater einfängt.

Zudem weisen sowohl DeepL als auch ChatGPT ungenaue Übersetzungen auf. So wird beispielsweise *hit the open road* mit *auf die Straße fahren* (DeepL) bzw. mit *eine Fahrt auf offener Straße* (ChatGPT) übersetzt, obwohl sinngemäß der Aufbruch zu einem Abenteuer oder einer Reise gemeint ist. Die Humanübersetzerin distanziert sich hier stärker vom Ausgangstext,

indem sie den Satz umstrukturiert und mit *Ganz gleich, wohin Ihr Weg Sie führt* beginnt, was insgesamt für ein flüssigeres Leseerlebnis sorgt.

4.2.3.2 Textausschnitt 2: Atomic

Der zweite Textausschnitt stammt vom österreichischen Skihersteller Atomic (vgl. Atomic 2026). Konkret wurden zwei Absätze einer Produktbeschreibung zweier Skimodelle herangezogen. Statt das Produkt selbst im Detail zu beschreiben, inszeniert der Text ein narrativ geprägtes Verwendungssetting rund um mehrere Skitage in Nordamerika. Darüber hinaus arbeitet der Text stark mit bildhafter Sprache, die auf die Zielgruppe der Marke ausgerichtet ist. Dies zeigt sich anhand von Motiven wie dem Skifahren im Neuschnee, der Liebe zum Skisport und dem Abschluss des Skitages in einer warmen Hütte.

Sprachlich weist der Textausschnitt zahlreiche evaluierende Adjektive und Phrasen wie *state of the art* oder *intrepid* sowie idiomatische Ausdrücke wie *biggest personalities* auf, die die Leser:innen von der Überlegenheit des Produkts überzeugen sollen. Die verwendete Terminologie ist dabei mit Begriffen wie *powder*, *hit the hill* oder *lines* häufig sehr zielgruppenspezifisch.

Eine besondere Herausforderung dieses Ausschnitts liegt darin, die bildhafte und emotionale Wirkung des Ausgangstextes sowie dessen narrative Elemente angemessen in den Zieltext zu übertragen. Zusätzlich enthält der Ausgangstext einen sehr langen Satz, der für einen natürlichen Lesefluss im Deutschen vermutlich aufgeteilt werden müsste. Dies stellt potenziell eine Herausforderung für die maschinellen Übersetzungssysteme dar, die tendenziell eher nah am Ausgangstext bleiben.

Auch in diesem Textausschnitt lassen sich klare Unterschiede zwischen den maschinellen und der Humanübersetzung erkennen. Der von DeepL generierte Text weicht kaum von der Struktur des englischen Originals ab, wodurch die Sätze lang, starr und unnatürlich wirken. ChatGPT nimmt auch hier eine Mittelposition ein, indem Strukturelemente wie Gedankenstriche hinzugefügt werden, während die Humanübersetzerin klare Umstrukturierungen oder Gruppierungen von Informationen bevorzugt. Die Transkreation in der Humanübersetzung führt zu einem natürlich klingenden und flüssigen Zieltext, wobei einzelne Informationen ausgelassen werden, um den Markenton des englischen Originals beizubehalten.

Auch auf terminologischer Ebene zeigen sich deutliche Unterschiede. Insbesondere die zuvor genannten, stark kulturell- und zielgruppenspezifischen Begriffe wurden von den maschinellen Übersetzungssystemen nur unzureichend erfasst, wodurch die Übersetzungen unna-

türlich oder sogar repetitiv klingen. Ein Beispiel hierfür ist die Kombination von *North American powder* und *windblown snow*. Sowohl DeepL als auch ChatGPT geben beide Begriffe mit *Schnee* wieder, DeepL etwa mit *des nordamerikanischen Pulverschnees, des vom Wind verwehten Schnees*. Die Humanübersetzerin löst diese Passage mit *der endlosen Vielfalt nordamerikanischer Schneelandschaften*. Zwar werden Details aus dem Ausgangstext weggelassen, der Text klingt dafür flüssig und elegant. Der Begriff *everyday skiers* wurde von DeepL und ChatGPT als *alltägliche Skifahrer* und *Skifahrer aus dem Alltag* übersetzt und hat damit einen leicht abwertenden Charakter. Die Humanübersetzerin wählte mit *echte Skifahrer:innen* einen viel positiveren und gegenderten Zugang, der dem Markenton entspricht.

Hervorzuheben ist jedoch, dass der Humanübersetzerin mit der Übersetzung von *skier* als *Skier* ein Fehler unterlaufen ist, der im Kontext der intendierten Zielgruppe auffällig sein dürfte.

4.2.3.3 Textausschnitt 3: Hisense

Der dritte Textausschnitt stammt vom chinesischen Elektronikhersteller Hisense (vgl. Hisense 2026). Der Text beschreibt die TV-Produktpalette des Unternehmens und zielt darauf ab, von den Vorteilen der Geräte zu überzeugen.

Charakteristisch ist hier insbesondere die direkte Ansprache der Leser:innen, die eine dialogische, emotional geprägte Werbesprache erzeugt. Aufgrund der direkten Ansprache muss eine übersetzungsrelevante Entscheidung hinsichtlich der Formalität getroffen werden. Während das englische *you* Formalität nur implizit andeutet, muss in der Übersetzung mit *du* oder *Sie* explizit entweder ein informelles oder formelles Register gewählt werden. Diese Entscheidung kann neben der sprachlichen Wahrnehmung auch die Markenwahrnehmung beeinflussen. Darüber hinaus enthält der Text zahlreiche evaluierende und werbetypische Ausdrücke wie *high-end performance*, *immersive features*, *cutting-edge* oder *premium viewing experiences*, die persuasiv wirken sollen. Die Herausforderung besteht darin, diese werbliche Tonalität im Zieltext beizubehalten, ohne dass dieser übersetzt oder unnatürlich wirkt.

In diesem Textbeispiel zeigen sich erneut deutliche Unterschiede zwischen den Übersetzungen. DeepL bleibt auch hier dem Ausgangstext in seiner Struktur treu. Auch ChatGPT orientiert sich stark am Original, abgesehen von kleinen Änderungen wie der Einfügung eines Doppelpunkts. Die Humanübersetzerin hält sich zwar ebenfalls überwiegend an die Ausgangstextstruktur, verwendet jedoch zusätzliche Elemente wie Doppelpunkte oder Gedankenstriche und fasst Informationen aufgrund von Wiederholungen zusammen, wodurch sowohl einzelne Sätze als auch der gesamte Text verkürzt werden.

Der von DeepL generierte Zieltext wirkt insgesamt trocken und verliert größtenteils seine persuasive und emotionale Wirkung. Dies ist zum einen auf den stark am Englischen orientierten Satzbau zurückzuführen, der im Deutschen ungewohnt und monoton klingt. Zum anderen wurden sehr emotionale, evaluierende englische Formulierungen wie *TVs that are hard not to fall in love with* mit *Fernseher, die Sie begeistern werden* abgeschwächt, wodurch die intendierte Wirkung verloren geht und der Text eher an eine sachliche Produktbeschreibung erinnert. Auch ChatGPT weist ähnliche Probleme auf, wenngleich der Text insgesamt flüssiger und überzeugender wirkt. Im Gegensatz zu DeepL übersetzt ChatGPT jedoch die zuvor erwähnte Phrase wörtlich mit *ein TV-Portfolio, in das man sich kaum nicht verlieben kann*. Diese doppelte Verneinung klingt im Deutschen sehr unnatürlich. Die Humanübersetzerin fasst diesen Satz mit einigen der Informationen im Zieltext zu *Hisense bringt Qualität zum Verlieben direkt in dein Wohnzimmer* zusammen. Zwar weicht diese Lösung stark vom Ausgangstext ab, erhält jedoch die intendierte Wirkung.

Eine weitere idiomatische Schwierigkeit zeigt die Textpassage *Whether you're into movies, video games, or just catching the big game, Hisense has a TV that's right for you*. *The big game* ist im Englischen geläufig und wird auch ohne explizite Erklärung mit einem großen Sportereignis in Verbindung gesetzt. Im Deutschen hingegen wirkt *das große Spiel* von DeepL unvollständig. ChatGPT erkennt aus dem Kontext heraus, dass es sich um Sport handelt und übersetzt die Phrase angemessen mit *das große Sportereignis*, während die Humanübersetzerin mit *Sport-Highlights zum Mitfiebern* ein zusätzliches emotionales Element einführt, das den Text deutlich werblicher wirken lässt.

Sowohl DeepL als auch ChatGPT wählten eine formelle Kund:innenansprache, was dafür sorgt, dass die Texte distanzierter wirken. Darüber hinaus nutzen beide Systeme den Markennamen, um vom Unternehmen zu sprechen. Die Humanübersetzerin wählt hingegen eine informelle Ansprache, wodurch der Text persönlicher wirkt, und verwendet *wir* statt den Markennamen, was den Text noch zugänglicher wirken lässt. Auf der offiziellen Webseite des Unternehmens werden sowohl formelle als auch informelle Formen der Kund:innenansprache verwendet, teilweise sogar auf einer einzigen Seite. Es lassen sich daher keine Rückschlüsse auf die korrekte Verwendung ziehen.

4.2.3.4 Textausschnitt 4: Puma

Der vierte Textausschnitt stammt von der Webseite des deutschen Sportartikelherstellers Puma (vgl. Puma 2026a, 2026b). Im Vergleich zu den vorangegangenen Textausschnitten handelt es sich hier um eine Übersicht über verschiedene Kollektionen. Da die einzelnen Beschreibungen

sehr kurz sind, wurden die Ausgangstexte von zwei unterschiedlichen Unterseiten des Unternehmens kombiniert, um die vorgegebene Ausgangstextlänge von mindestens 80 Wörtern zu erreichen.

Die Beschreibungen sind durch prägnantes Marketing-Copywriting gekennzeichnet. Die jeweiligen Kollektionen werden in ein bis zwei kurzen Sätzen vorgestellt, wobei stark mit Bildsprache wie *intersection of sport, style, and creativity* oder *The street is your track, the city is your field* gearbeitet wird. Auch hier finden sich mit *bold* oder *vibrant* evaluierende Adjektive, die persuasiv wirken sollen. Darüber hinaus adressieren die Kollektionsbeschreibungen unterschiedliche Zielgruppen. Während die erste und vierte Kollektion als Lifestyle-Produkte den Fokus auf das Design legen, richtet sich die zweite mithilfe urbaner Bildsprache an Läufer:innen in der Großstadt. Die dritte Kollektion wiederum betont mit Stabilität und Widerstandsfähigkeit zwei für den Skatesport sehr wichtige Eigenschaften.

Eine Herausforderung für die Übersetzung besteht darin, die emotionale Wirkung und Bildsprache in nur wenigen Worten überzeugend zu erfassen. Während die vorhergehenden Textausschnitte aus längeren Absätzen bestehen, weisen diese Kollektionsbeschreibungen stark elliptische Strukturen mit einer Länge von jeweils maximal zwei kurzen Sätzen auf. Der adäquaten Übertragung der stilistischen Elemente kommt daher eine besondere Bedeutung zu, damit die persuasive Wirkung im Zieltext erhalten bleibt.

Auch bei diesem Textausschnitt lassen sich ähnliche Muster wie in den vorangegangenen Beispielen beobachten. DeepL repliziert den Satzbau des Ausgangstextes, ChatGPT löst sich teilweise davon und die Humanübersetzerin geht überwiegend transkreativ vor. DeepL wählt etwa mit *Obermaterial* einen sehr konservativen, technischen Zugang zur Fachterminologie, der für die moderne Sportbranche unpassend ist und die Texte viel zu formell wirken lässt. ChatGPT verwendet einen modernen Sprachgebrauch mit zahlreichen Anglizismen, was der Branche entspricht. Die Humanübersetzerin entfernt sich sehr weit vom Ausgangstext und ersetzt oder lässt viele Begriffe aus dem Englischen aus. Ein Beispiel hierfür ist der Begriff *streetwear staple*. DeepL übersetzt ihn mit *Streetwear-Klassiker*, ChatGPT mit *fester Bestandteil der Streetwear*, während die Humanübersetzerin ihn mit *mit urbanem, modernem Lifestyle* ersetzt. DeepL begeht hier auch einige Übersetzungsfehler. *Elevate the brand* wird zu *hebt die Marke hervor* und *on and off the board* wird zu *auf dem Board und außerhalb*, was in beiden Fällen nicht der intendierten Wirkung entspricht.

Insgesamt zeigt sich, dass die Humanübersetzerin in diesem Textbeispiel einen stark transkreativen Ansatz verfolgt, um dem Markenton gerecht zu werden. Die Länge des Zieltextes entspricht dabei in etwa jener des Ausgangstextes, während die maschinellen Übersetzungen

deutlich länger ausfallen. Die Transkreation sorgt für einen besonders flüssigen und natürlichen Zieltext.

4.2.3.5 Textausschnitt 5: Lush

In diesem Textausschnitt des britischen Kosmetikkonzerns Lush wird ein Valentinstagsgeschenksset vorgestellt (vgl. Lush 2026a, 2026b). Um die erforderliche Mindestlänge zu erreichen, wurden auch hier Texte aus zwei unterschiedlichen Produkten des Valentinstagssortiments kombiniert. Dieser Textausschnitt stellt eine besondere Herausforderung für die Übersetzung dar, da er zahlreiche kreative und bildliche Produktbeschreibungen enthält, wie beispielsweise *fragrant melts* oder *dusting powders*. Zusätzlich werden literarische Stilmittel wie Alliterationen, etwa *melts for mates,* sowie mit *mates… dates*; *fate… await* ein lockeres Reimschema am Ende des ersten Abschnitts verwendet. Darüber hinaus wird das Geschenk mit Ausdrücken wie *we have a crush on this gift*, *we think this feeling is mutual* oder *let the gift do the talking* personifiziert.

Ein direkter Vergleich der Übersetzungen mit dem Ausgangstext würde die fehlende Übertragung literarischer Stilmittel sofort sichtbar machen. Da die Teilnehmenden im Rahmen des Turing-Tests jedoch keinen Zugriff auf den englischen Ausgangstext haben, ist es nicht zwingend erforderlich, diese Stilmittel im Zieltext beizubehalten. Entscheidend ist vielmehr, die spielerische und humorvolle Tonalität dieses Textausschnitts auch in der Übersetzung zu erhalten.

Sowohl DeepL als auch ChatGPT gelingt es nicht vollständig, die Wirkung des Ausgangstextes ins Deutsche zu übertragen. Die literarischen Stilmittel gehen darüber hinaus auch in der Humanübersetzung überwiegend verloren. Der von DeepL generierte Zieltext wirkt stilistisch sehr trocken und erinnert stellenweise mehr an eine Auflistung als an einen persuasiven Marketingtext. Ein Beispiel hierfür ist *Wir haben duftende Melts für Freunde, Puder für heiße Dates, und wenn Sie nicht an das Schicksal glauben, erwarten Sie süße Solo-Bäder....* Das System wählt dabei sogar das formelle *Sie* als Ansprache, was weder zur Marke noch zum Textinhalt passt. ChatGPT trifft mit *du* und ausweichenden Formulierungen wie *für alle, die* den Markenton besser. Der Text wirkt auch flüssiger und vermeidet mit *Duftende Melts für Freunde, Puder für heiße Dates – und für alle, die nicht an das Schicksal glauben, warten süße Solo-Bäder …* die bei DeepL beobachteten Probleme. Die Humanübersetzerin ersetzt die literarischen Stilmittel durch sehr emotional aufgeladene Begriffe, beispielsweise *schockverliebt*, sowie durch besonders starke Bildsprache, um die Wirkung des Ausgangstextes zu erhalten. Sie baut zusätzlich *Self Care* explizit ein, welches im englischen Original nur durch den Begriff

sweet solo baths mitschwingt. Dies kann als klassische transkreative Strategie gesehen werden (vgl. Błaszkowska 2021), die dazu beiträgt, dass die Übersetzung noch persuasiver wirkt. Auch das physische Produkt *fragrant melts* wird in der Humanübersetzung zu *schmelzende Düfte*. Dies ist terminologisch zwar ungenau, erzeugt jedoch ein sehr starkes Bild.

Die Übersetzung von *the raspberry-rose splendour inside will win them over* veranschaulicht sehr gut die verschiedenen Herangehensweisen. DeepL übersetzt diesen Satzteil mit *... wird die Himbeer-Rosen-Pracht darin Ihr Gegenüber überzeugen*. *Himbeer-Rosen-Pracht* liest sich sehr technisch und wirkt holprig, *das Gegenüber* kann als zu neutral im Kontext eines Valentinstagsgeschenks eingestuft werden. Die von ChatGPT generierte Version, *... wird die himmlische Kombination aus Himbeere und Rose garantiert Herzen erobern*, zeigt eine viel flüssigere Herangehensweise. *Himmlisch* ist ein stark evaluierendes Adjektiv, das zum Kontext passt und auch die abstraktere Formulierung *wird garantiert Herzen erobern* wirkt gelungen. Die Humanübersetzerin geht mit *wird der Mix aus Himbeere und Rose jedes Herz im Sturm erobern* einen Schritt weiter und erzeugt ein noch kräftigeres Bild.

4.2.3.6 Textausschnitt 6: Oatly

Der sechste Textausschnitt stammt von einer Produktseite des schwedischen Lebensmittelunternehmens Oatly (vgl. Oatly 2026). Insgesamt enthält der Textausschnitt eine hohe Dichte an Fachvokabular aus dem Bereich der Ernährungswissenschaft, das auf die gesundheitlichen Eigenschaften des Produkts verweist, wie beispielsweise *unsaturated fat*, *betaglucans* oder *GMOs*. Trotz dieser Fachterminologie kann der Text insgesamt als sprachlich relativ einfach eingestuft werden.

Charakteristisch für den Text ist ein stark ironischer und humorvoller Unterton, der durch verschiedene Elemente weiter verstärkt wird. Dazu zählen unter anderem metasprachliche Kommentare sowie bewusste Abschweifungen wie *big, scientific word for soluble fiber from oats* oder *we're talking about less fat. Or karma. Well, whatever we're talking about, it's the good kind*. Eine Herausforderung bei der Übersetzung könnte darin bestehen, diesen ironischen Ton im Zieltext angemessen wiederzugeben, ohne dabei den Textfluss zu stören.

Die DeepL-Übersetzung wirkt sehr formell. Nicht nur die Ansprache mit *Sie* lässt den Text distanziert wirken, sondern auch die humorvollen, ironischen Anspielungen können insgesamt nicht adäquat übertragen werden. Der von ChatGPT generierte Text liest sich erneut deutlich flüssiger und kreativer, trifft mit *du* auch das Register besser, kann aber die Anspielungen ebenfalls nur bedingt wiedergeben. Die Humanübersetzerin greift mit *Haferflocken (nur*

flüssiger) oder *gutes Wort, um damit anzugeben* auch in diesem Beispiel auf Transkreationen und Einschübe zurück, um den ironisch-humorvollen Ton beizubehalten.

4.2.3.7 Textausschnitt 7: Innocent

Der siebte Textausschnitt stammt vom britischen Getränkeunternehmen Innocent und stellt einen Smoothie aus dessen Produktpalette vor (vgl. Innocent 2026). Der Text zeichnet sich durch dynamisches und kreatives Copywriting aus und enthält einige für die Übersetzung herausfordernde Elemente wie kulturspezifische Abkürzungen oder Wortspiele mit dem englischen Begriff *defence* darunter auch sprachliche Neukreationen wie *de-fence*. Neben der direkten Ansprache treten auch Interjektionen wie *Phew* auf, die zum Kontext passend übersetzt werden müssen.

DeepL übersetzt auch in diesem Fall sehr wörtlich, wodurch die Übersetzung aufgrund der Wortneuschöpfungen und Anspielungen zusammenhangslos wirkt. ChatGPT versucht mit *„ent-zäunen“*, die Wortneuschöpfung durch Anführungszeichen und einen Bindestrich zu übertragen, was im Deutschen jedoch nicht funktioniert, da sich *Abwehr* und *Zaun* im Gegensatz zum englischen *fence* und *defence* nicht überlappen. Die Humanübersetzerin setzt mit *Manche bauen Zäune, um sich zu schützen (was wahnsinnig viel Arbeit macht), andere trinken einfach unseren vitaminhaltigen Super-Smoothie* wieder auf eine Transkreation und überträgt somit die humorvolle Komponente des Ausgangstextes.

4.2.3.8 Textausschnitt 8: Casper

Der letzte Textausschnitt beschreibt ein Hundebett der US-amerikanischen Tierpflegemarke Casper (vgl. Casper 2026). Der Text verbindet typische Fachterminologie aus Produktbeschreibungen mit zwei an das Produkt angepassten idiomatischen Phrasen und Wortspielen. Die erste Phrase, *Let sleeping dogs lie, on foam*, hat im Deutschen mit *Schlafende Hunde soll man nicht wecken* ein Äquivalent mit Hundebezug, das jedoch auf einen anderen Begriff aufbaut, weshalb eine wörtliche Übersetzung nicht zielführend ist. Die zweite Phrase, *Dogs will dig it, and scratch it*, kann als für die Übersetzung noch schwieriger eingestuft werden, da sie neben *dig it* (etwas mögen, gut finden) auch auf die Gewohnheit von Hunden anspielt, zu graben.

Die Herausforderung für die Übersetzung besteht darin, die Phrasen nicht nur als idiomatische Einheiten zu erkennen, sondern sie auch so zu übertragen, dass sie für das Produkt funktionieren. Eine wörtliche Übersetzung im Deutschen ist daher nicht angemessen.

DeepL ignoriert beide Wortspiele und übersetzt sie wörtlich. *Liegen lassen* ist zudem auch negativ konnotiert, wodurch die intendierte Wirkung nicht erreicht wird. ChatGPT erkennt

das erste Sprichwort und adaptiert dessen Struktur. Auffällig ist, dass es dieses dann mit *Schlafende Hunde soll man ruhen lassen – auf Schaumstoff* jedoch wörtlich übersetzt, wodurch es potenziell nicht erkannt wird. Die implizite Bedeutung der zweiten Phrase geht auch hier gänzlich verloren. Die Humanübersetzerin übersetzt das erste Sprichwort mit *Schlafende Hunde soll man nicht wecken – aber weich betten* adäquat und findet eine stimmige Lösung für den Zusatz. Sie erkennt auch, dass es keine Entsprechung der zweiten idiomatischen Phrase im Deutschen gibt und wählt stattdessen mit *Da wird der Hund in der Pfanne verrückt (vor Freude)* ein zweites deutschsprachiges Sprichwort mit Hundebezug, das in diesem Kontext besonders gut funktioniert. Darüber hinaus baut sie mit *Pfote drauf* oder *Zweibeiner* auch weitere Anspielungen mit Hundebezug ein, die ebenfalls gut zur Produktrhetorik passen.

Zusammenfassend zeigt die Analyse der Textausschnitte, dass sowohl DeepL als auch ChatGPT in der Lage sind, Marketingtexte mit stark informationsorientierten Inhalten verständlich und weitgehend korrekt ins Deutsche zu übertragen. DeepL tendiert dazu, sich stark am Satzbau und an der lexikalischen Struktur des Ausgangstextes zu orientieren, während ChatGPT insbesondere bei Texten mit narrativen Elementen oder direkter Leser:innenansprache tendenziell flüssigere und natürlichere Übersetzungen generiert. Dies wird vor allem anhand des gewählten Registers ersichtlich. DeepL bleibt bei allen Textbeispielen auf einer formellen Ebene, während ChatGPT die Ansprache je nach Branche des Ausgangstextes häufig angemessen wählt. Bei Textausschnitten, die komplexe sprachliche Mittel, Humor, Ironie, Sprichwörter oder Wortneuschöpfungen enthalten, stoßen beide maschinellen Übersetzungssysteme an ihre Grenzen. Hier zeigt sich, dass formale Äquivalenz nicht zwangsläufig zu funktionaler Äquivalenz führt, was insbesondere bei DeepL häufig zu beobachten ist, da zu starke Ausgangstexttreue bei kreativen Texten im Deutschen schnell unidiomatisch wirken kann. Es kann daher angenommen werden, dass die von ChatGPT generierten Zieltexte häufiger als humanübersetzt eingeschätzt werden als jene von DeepL.

Die Humanübersetzungen können hingegen insgesamt als die adäquatesten und zielkulturgerechtesten Zieltexte betrachtet werden. Auffallend ist, dass die Übersetzerin insbesondere bei Texten mit stark kreativen oder humorvollen Elementen transkreative Übersetzungsstrategien anwendet und der intendierten Wirkung des Ausgangstextes mehr Gewicht als der formalen Äquivalenz zum Ausgangstext zuschreibt. Diese Beobachtung steht im Einklang mit den in Kapitel 3 ausgeführten Grenzen der maschinellen Übersetzung, da vor allem bei kreativen Textsorten ist davon auszugehen, dass maschinelle Übersetzungssysteme dem Menschen weiterhin häufig unterlegen sind.

4.3 Erstellung der Umfrage

Die Umfrage umfasste insgesamt 52 Fragen, davon entfielen 16 auf Angaben zur Person und die Nutzung von maschinellen Übersetzungstools bzw. Large Language Models für Übersetzungszwecke, 32 auf die Bewertung der Textausschnitte im Rahmen der Hauptbefragung sowie vier Fragen auf den Abschlussteil. Abhängig von den Antworten im soziodemografischen Teil des Fragebogens verringerte sich die Anzahl der Fragen. Die genaue Logik sowie der gesamte Fragebogen sind in Anhang A2 dargestellt.

Der Fragebogen startete mit einer einleitenden Infoseite, auf der die Teilnehmenden begrüßt und über den Ablauf sowie die Datenverarbeitung informiert wurden. Die Umfrage wurde anonym durchgeführt und erfasste ausschließlich Daten, die für die Untersuchung relevant waren. Mit dem Anklicken von Weiter bestätigten die Proband:innen ihre freiwillige Teilnahme an der Studie. Ein Abbruch war jederzeit möglich.

Der soziodemografische Abschnitt des Fragebogens bestand überwiegend aus Single- und Multiple-Choice-Fragen. Bei einzelnen Fragen hatten Teilnehmende zudem die Möglichkeit, zusätzliche Informationen oder Erklärungen bereitzustellen.

Zunächst wurden Informationen zum Alter erhoben. Als zentrale Frage zur Qualifikation zur Teilnahme an der Studie wurde anschließend erfragt, ob Deutsch die Erstsprache des:der Teilnehmenden ist. Teilnehmende, die Deutsch als Erstsprache angaben, wurden direkt zur Teilnahme zugelassen. Im Falle einer Verneinung folgten Fragen zur Erstsprache und zu den Deutschkenntnissen. Die Sprachkompetenz wurde über Selbsteinschätzung erhoben, da die Studie als Online-Befragung konzipiert war und keine standardisierte externe Sprachtestung vorsah. Wie in Kapitel 4.1 beschrieben, galt C1 als Mindestniveau zur Teilnahme an der Studie. Diese Voraussetzung wurde bereits im Einladungsschreiben sowie auf der Willkommensseite des Fragebogens bekanntgegeben. Teilnehmende, die ein Niveau von A1/A2 oder B1/B2 angaben, wurden von der Befragung ausgeschlossen.

Zur Klassifikation der Teilnehmenden als Lai:innen oder Sprachexpert:innen wurden im Fragebogen gezielt Angaben zu Studium und Beruf erhoben. Die Zuordnung erfolgte auf Basis der in Kapitel 4.1 festgelegten Kriterien. Explizit zur Auswahl stehende und jedenfalls für die Expert:innenrolle geltende Studien waren: Transkulturelle Kommunikation/Translation, Linguistik, Deutsche Philologie/Germanistik und Deutsch als Zweitsprache/Fremdsprache. Darüber hinaus bestand die Möglichkeit, mittels eines freien Eingabefeldes weitere Studienabschlüsse anzugeben, die individuell auf ihre fachliche Relevanz für den Expert:innenstatus geprüft wurden.

In Bezug auf die Berufserfahrung der Teilnehmenden standen folgende Bereiche der Textarbeit explizit zur Auswahl und galten jedenfalls für die Expert:innenrolle: Übersetzung, Lokalisierung, Copywriting und Lektorat. Auch hier konnten weitere Berufsfelder angegeben werden, die anschließend individuell auf ihre Relevanz geprüft wurden.

Im darauffolgenden Block des Fragebogens wurde die Nutzung von maschineller Übersetzung hinsichtlich Häufigkeit, Zweck und verwendeter Tools erfragt. Teilnehmende, die angaben, noch nie maschinelle Übersetzungstools verwendet zu haben, wurden nach den Gründen dafür gefragt.

Vor dem darauffolgenden Hauptexperiment wurde eine kurze Überleitung eingeblendet. Diese informierte die Teilnehmenden darüber, dass sie anschließend mehrere Textausschnitte sehen und zu jedem Text Fragen beantworten würden. Diese Überleitung diente der klaren Strukturierung des Fragebogens und dem Übergang in den experimentellen Teil. Den Teilnehmenden wurde nun eines von vier Sets in zufälliger Reihenfolge zugeteilt. Diese zufällige Zuteilung legte fest, welche Übersetzungsvariante, also Humanübersetzung, DeepL oder ChatGPT, den Teilnehmenden in den einzelnen Textbeispielen angezeigt wurde. Die zufällige Reihenfolge erfolgte automatisiert über einen Zufallsgenerator und war für die Teilnehmenden nicht sichtbar.

Der Anteil von maschinellen und Humanübersetzungen betrug jeweils vier Textausschnitte und war daher ausgewogen. Es wurde entschieden, die Befragung in vier Sets aufzuteilen, damit jeder Textausschnitt abhängig vom Set jeweils zumindest einmal maschinell und einmal humanübersetzt vorkam. Dadurch konnten in der nachfolgenden Analyse Rückschlüsse hinsichtlich des Niveaus an stilistischen und kreativen Mitteln der einzelnen Textausschnitte gezogen werden. Die maschinellen Übersetzungen wurden dabei stets entweder von DeepL oder ChatGPT angefertigt, wobei bei jedem Set konsequent immer nur eines von beiden Systemen zur Übersetzung herangezogen wurde. Dieser Aufbau wurde gewählt, um stichhaltige Aussagen zum Abschneiden und zur Bewertungsqualität des jeweils verwendeten Systems treffen zu können. Tabelle 1 verdeutlicht die angewandte Logik der Textverteilung in den vier Sets. Die Kürzel UP01 bis UP08 bezeichnen dabei die jeweiligen Textausschnitte, die zur besseren Nachvollziehbarkeit zusätzlich mit dem dazugehörigen Markennamen versehen sind. Die vollständigen englischen Originaltexte sowie alle Übersetzungsvarianten sind in Anhang A1 aufgeführt.

Tabelle 1: Verteilung der Textausschnitte und Systeme in den Befragungssets

Set	**Verwendetes System**	**Maschinell übersetzte Textausschnitte**	**Humanübersetzte Textausschnitte**
Set A	DeepL	UP05 (Lush), UP06 (Oatly), UP07 (Innocent), UP08 (Casper)	UP01 (Mazda), UP02 (Atomic), UP03 (Hisense), UP04 (Puma)
Set B	DeepL	UP01 (Mazda), UP02 (Atomic), UP03 (Hisense), UP04 (Puma)	UP05 (Lush), UP06 (Oatly), UP07 (Innocent), UP08 (Casper)
Set C	ChatGPT	UP05 (Lush), UP06 (Oatly), UP07 (Innocent), UP08 (Casper)	UP01 (Mazda), UP02 (Atomic), UP03 (Hisense), UP04 (Puma)
Set D	ChatGPT	UP01 (Mazda), UP02 (Atomic), UP03 (Hisense), UP04 (Puma)	UP05 (Lush), UP06 (Oatly), UP07 (Innocent), UP08 (Casper)

Die Teilnehmenden sahen jeweils nur eine Variante pro Text. Es handelte sich somit um ein Between-Subjects-Design auf Set-Ebene (vgl. Pyka & Furchheim 2017), bei dem jede Person pro Textausschnitt jeweils nur eine Übersetzungsvariante sah. Zusätzlich zur Set-Verteilung war auch die Reihenfolge der Textausschnitte zufällig gewählt, sodass diese nicht von ihrer übersetzerischen Schwierigkeit beeinflusst wurde, sondern zufällig erfolgte.

Die Teilnehmenden erhielten nun acht Übersetzungen von Marketingtexten verschiedener Unternehmenswebseiten, ohne dabei den englischen Originaltext zu sehen. Zu jedem Textausschnitt beantworteten die Teilnehmenden zunächst eine Einschätzungsfrage, ob es sich um eine maschinelle oder eine Humanübersetzung handelte (Single-Choice). Anschließend bewerteten sie die sprachlichen bzw. stilistischen Merkmale, die zu dieser Einschätzung geführt haben, auf einer Likert-Skala sowie die Sicherheit ihrer Einschätzung ebenfalls mittels Skala. Bei den stilistischen Merkmalen mussten Teilnehmende auf einer fünfstufigen Likert-Skala (vgl. Gritsch 2012; Likert 1932) von 1 (= gar nicht) bis 5 (= sehr) angeben, wie flüssig und natürlich der Text klingt, wie stimmig die Wortwahl, der Satzbau sowie der Text insgesamt ist und wie gut der Stil und Ton zum Texttyp passt. Auch die Angabe zur Sicherheit der eigenen Einschätzung wurde mittels einer Skala von 1 (= sehr unsicher) bis 5 (= sehr sicher) erhoben. Zum Abschluss eines jeden Textausschnitts konnten die Teilnehmenden mittels einer offenen Frage

optional einen Kommentar hinterlassen (vgl. Abbildung 4).

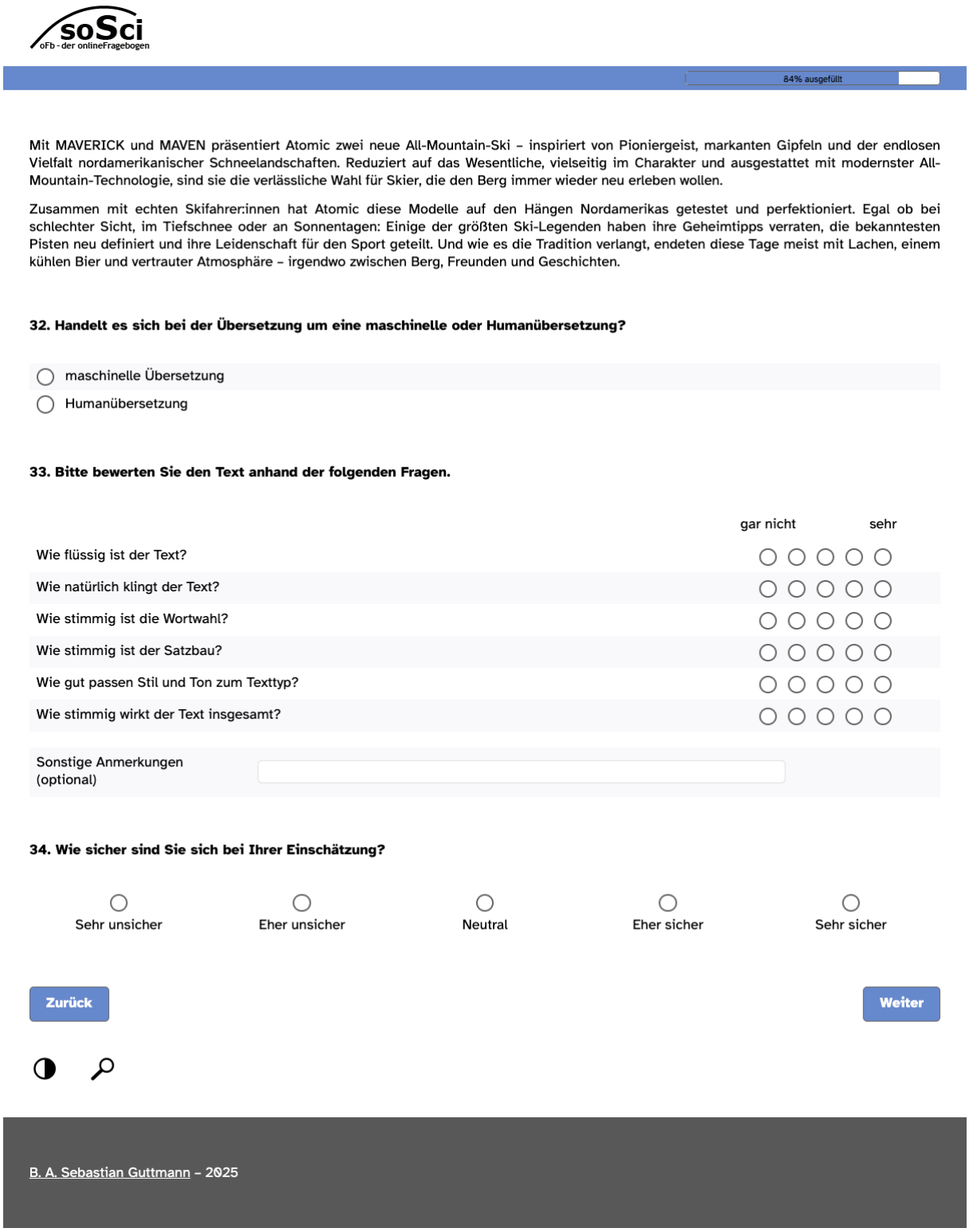

Abbildung 4: Aufbau der Bewertungsseite im Online-Fragebogen am Beispiel des Atomic-Textausschnitts (eigener Screenshot)

Am Ende des Fragebogens folgten mehrere Abschlussfragen zur Selbsteinschätzung in Form von Skalen und offenen Fragen, um den Teilnehmenden die Möglichkeit zu geben, weitere Gedanken, Überlegungen oder Zweifel zu teilen. Den Abschluss bildeten eine kurze Danksagung und die Information, dass die Teilnehmenden erfolgreich teilgenommen haben und die Seite nun schließen können.

4.4 Durchführung der Erhebung

Nach der Konzeption der Umfrage wurde dieser zunächst in einer Erstversion auf SoSci Survey erstellt. Anschließend wurde er zur inhaltlichen und formalen Überprüfung mit der Betreuerin geteilt. Das erhaltene Feedback wurde danach in einer überarbeiteten Version implementiert. Konkret wurden einige Formulierungen geändert, die Bewertung der stilistischen Eigenschaften der Textausschnitte von einem Multiple-Choice-Format auf eine Skalenbewertung umgestellt sowie eine zusätzliche offene Frage für Feedback zum Fragebogen ergänzt.

Im nächsten Schritt wurde ein Pretest unter realen Bedingungen mit sechs Proband:innen durchgeführt, darunter drei Lai:innen und drei Sprachexpert:innen. Zwei der Teilnehmer:innen verfügten über Erfahrung im Bereich UX-Research und konnten dadurch methodischen Input zur Umsetzung der Umfrageerstellung und Erhebung von Nutzer:innendaten geben. Darüber hinaus wurde darauf geachtet, dass der Fragebogen sowohl auf Desktop- als auch auf mobilen Geräten getestet wurde, um potenzielle Unterschiede in der Darstellung und der Bedienbarkeit feststellen zu können. Zudem wurden noch weitere Ungereimtheiten im Gesamtlayout festgestellt, insbesondere in der einheitlichen Darstellung von Fragebeschreibungen und der Formatierung der Textausschnitte. Zusätzlich wurden einzelne Antwortmöglichkeiten zur Frage zum höchsten erreichten Studienabschluss überarbeitet und präzisiert. Eine zentrale Anpassung betraf die Logik der zufälligen Zuteilung der Bewertungsaufgabe zu den Textausschnitten. Da es jederzeit möglich war, auch nach dem Start der Bewertungsaufgabe zurück zu den soziodemografischen Fragen zu navigieren, wurde ein Fehler im PHP-Code behoben, der eine erneute zufällige Zuteilung nach dem Zurücknavigieren ausgelöst hätte. Diese Korrektur war notwendig, um eine saubere Datenerhebung sowie eine stichhaltige Auswertung der Daten sicherzustellen.

Vor Beginn der Hauptbefragung wurden noch vier Testdurchgänge sowohl in der mobilen Version als auch der Desktopversion unter realen Testbedingungen durchgeführt, um sicherzustellen, dass die zufällige Zuteilung und Datenerhebung auf beiden Plattformen funktionierten. Diese dabei generierten Datensätze dienten ausschließlich zu Testzwecken und wurden vor Beginn der eigentlichen Datenerhebung vollständig gelöscht.

Die finale Datenerhebung begann am 16. Februar 2026. Die Proband:innen wurden persönlich, per E-Mail sowie über Online-Plattformen wie LinkedIn oder Facebook sowie Vereine bzw. Verbände rekrutiert. Ziel der Datenerhebung waren mindestens 60 abgeschlossene Datensätze in einer ausgewogenen Verteilung von 50 % Sprachexpert:innen und 50 % Lai:innen. Um dieses Verhältnis zu erreichen, wurden insbesondere Sprachexpert:innen gezielt angeschrieben. Diese Verteilung wurde durch regelmäßige Zwischenanalysen überprüft, um im Falle einer sehr ungleichmäßigen Verteilung gezielt nachrekrutieren zu können.

Die Daten wurden regelmäßig im Hinblick auf Ausfallraten und technische Auffälligkeiten kontrolliert. Es zeigte sich jedoch keine erhöhte Ausfall- bzw. Abbruchrate bei einzelnen Fragen. Nachdem genügend Teilnehmende akquiriert werden konnten, wurde die Datenerhebung am 15. März 2026 abgeschlossen.

4.5 Datenanalyse

Die erhobenen Daten wurden mit SoSci Survey als Microsoft-Excel-Datei exportiert. Dabei wurde der Datensatz so aufbereitet, dass Variablen- und Wertelabels sowie numerische Codierungen eindeutig abgebildet waren. Die Datenaufbereitung und -analyse erfolgte in Google Sheets und wurde anschließend kontrolliert und auf Plausibilität überprüft. Zur besseren Orientierung wurden Fragekürzel den jeweiligen Fragen sowie numerische Antwortcodes ihren Bedeutungen zugeordnet. Zudem wurden die Daten neben einer allgemeinen Übersicht zusätzlich nach Expert:innenstatus in separate Tabellenblätter aufgeteilt, um mögliche Auswertungsfehler zu vermeiden.

In Anlehnung an gängige statistische Verfahren (vgl. Bortz & Döring 2006) wurden zunächst deskriptive Analysen durchgeführt, im Rahmen derer Häufigkeiten, Prozentwerte, Mittelwerte sowie Standardabweichungen berechnet wurden. Für die Forschungsfragen 1 bis 3 wurde mit inferenzstatistischen Verfahren gearbeitet. Dabei kamen je nach Zielsetzung sowohl abhängige t-Tests (Vergleich innerhalb derselben Stichprobe) als auch unabhängige t-Tests (Vergleich zwischen Gruppen) zur Anwendung. Als Signifikanzniveau wurde $\alpha = 0{,}05$ festgelegt. Ergebnisse mit $p < 0{,}05$ wurden als statistisch signifikant interpretiert (vgl. Bortz & Döring 2006). Zur Einordnung der praktischen Relevanz wurde zusätzlich Cohen's d berechnet (vgl. Cohen 2013).

Für die Analyse der sprachlichen und stilistischen Eigenschaften wurden auf Basis der Likert-Skalen gewichtete Mittelwerte und Standardabweichungen berechnet. Diese Gewichtung war notwendig, da aufgrund der Einteilung der Teilnehmenden in vier Sets die Verteilung

nicht immer ausgewogen war. Diese Berechnungen dienten dazu, Unterschiede in der Wahrnehmung sprachlicher und stilistischer Eigenschaften sichtbar zu machen. Darüber hinaus wurden Gruppenvergleiche im Hinblick auf Expert:innenstatus, Übersetzungsart, verwendetes maschinelles Übersetzungstool und Kreativitätsniveau angestellt. Wie in Kapitel 4.2.3 festgehalten, unterschieden sich die Textausschnitte in ihrer Komplexität und wurden dahingehend für die Analyse in zwei Gruppen zusammengefasst. Textausschnitte 1–4 wiesen niedrigere kreative und stilistische Komplexität auf, Textausschnitte 5–8 hingegen höhere. Neben den Formelberechnungen in Google Sheets wurden für die t-Tests das Online-Tool GraphPad (GraphPad o.J.) und für die Effektstärken der Online-Rechner von Becker (1999) verwendet.

5 Ergebnisse

Im folgenden Kapitel werden die Ergebnisse des Fragebogens dargestellt. Nach der Beschreibung der Stichprobe, der Vorerfahrungen und der Nutzungsgewohnheiten werden die weiteren Ergebnisse nach der jeweiligen Forschungsfrage gegliedert.

5.1 Stichprobe und Vorerfahrungen

Insgesamt nahmen 93 Personen an der Befragung teil, wovon 73 Teilnehmende den Fragebogen vollständig abschlossen. Eine Person wurde zudem ausgeschlossen, weil sie mit B1–B2 nicht über die vorausgesetzten deutschen Sprachkenntnisse verfügte. Die Gesamtteilnehmer:innenzahl belief sich somit auf 72, wovon alle im Altersbereich von 18–65 Jahren lagen. Mehr als die Hälfte aller Teilnehmenden zählte zur Altersgruppe von 31–45 Jahren. 88,89 % (n = 64) davon gaben Deutsch als Erstsprache an, während eine Person ein sprachliches Niveau von C1 und sieben weitere von C2 angaben.

Nach den in Kapitel 4.1 beschriebenen Kriterien wurden zunächst 41 der Teilnehmenden automatisiert als Sprachexpert:innen sowie 32 als Lai:innen klassifiziert. Anschließend wurden die Expert:innendaten hinsichtlich des relevanten Studienabschlusses, der Berufserfahrung und der relevanten Disziplin in drei Gruppen eingeteilt: Expert:innenstatus durch Studienabschluss, Expert:innenstatus durch relevante Berufserfahrung sowie Expert:innenstatus durch Studienabschluss und relevante Berufserfahrung. Zur Überprüfung des Expert:innenstatus wurden die Daten manuell analysiert und überprüft, ob tatsächlich ein Expert:innenstatus vorlag. Hierfür waren insbesondere die Angaben im Feld Sonstiges zum relevanten Studienabschluss und der relevanten Berufserfahrung ausschlaggebend. In Einzelfällen wurde daraufhin eine Reklassifizierung von Sprachexpert:in zu Lai:in durchgeführt. Dies betraf vor allem Teilnehmende mit Studienabschlüssen, die weder der Translations- noch der Sprachwissenschaft zugeordnet werden konnten, sowie unspezifische Angaben.

Auch bei der relevanten Berufserfahrung wurde eine Reklassifizierung eines:einer Teilnehmer:in vorgenommen, der:die Literaturforschung als Beruf angab. Da diese Tätigkeit primär wissenschaftlich-analytische Textarbeit umfasst, jedoch nicht zwingend professionelles Schreiben für Zielgruppen und keine angewandte Textproduktion miteinschließt, wurde auch diese:r Teilnehmer:in als Lai:in eingestuft.

Nach der Reklassifizierung belief sich die Aufteilung auf 36 Sprachexpert:innen und 36 Lai:innen, was einer gleichmäßigen Gewichtung von jeweils 50 % gleichkommt. Die durch-

schnittliche bereinigte Ausfülldauer betrug ca. 17 Minuten, wobei Sprachexpert:innen geringfügig länger (~17,8) als Lai:innen (~16,5) brauchten.

Unter allen Teilnehmenden gaben 94,44 % (n = 68) an, bereits maschinelle Übersetzungstools oder Large Language Models für Übersetzungszwecke verwendet zu haben, wobei der überwiegende Teil wöchentlich oder monatlich darauf zurückgreift (jeweils 27,94 %; n = 19). 36,76 % (n = 25) gaben an, maschinelle Übersetzungstools oder Large Language Models täglich oder mehrmals wöchentlich zu nutzen. Während in der allgemeinen Nutzung kein Unterschied zwischen den beiden Gruppen festgestellt werden konnte, zeigt sich, dass Sprachexpert:innen maschinelle Übersetzungstools bzw. Large Language Models häufiger einsetzen als Lai:innen. So verwenden 44,12 % (n = 15) aller Sprachexpert:innen, jedoch nur 29,41 % (n = 10) aller Lai:innen täglich oder mehrmals pro Woche diese Tools.

Als Hauptanwendungsbereiche wurden von Teilnehmenden beider Gruppen private und berufliche Zwecke angegeben. Sowohl Sprachexpert:innen als auch Lai:innen nutzen maschinelle Outputs gleichermaßen zum allgemeinen Verständnis eines Originaltextes (jeweils 67,65 %; *n* = 23), während Sprachexpert:innen mit 64,71 % (*n* = 22) sie beträchtlich häufiger als Post-Editing-Basis heranziehen als Lai:innen (32,35 %; *n* = 11). Sprachexpert:innen (58,82 %; *n* = 20) tendieren auch klar häufiger als Lai:innen (23,53 %; *n* = 8) dazu, den Output zur Ideenfindung oder für Qualitätsprüfung (41,18 %; *n* = 14 vs. 29,41 %; *n* = 10) zu verwenden. Als Gründe der Verwendung wurden vor allem die Zeitersparnis (86,76 %; *n* = 59) und der leichte Zugang (61,76 %; *n* = 42) angegeben. Unterschiede zwischen den Gruppen gab es vor allem in den Kategorien Geldersparnis und Ausreichende Qualität. 20,59 % (*n* = 7) der Lai:innen, jedoch nur 5,88 % (*n* = 2) aller Sprachexpert:innen gaben an, die Outputs zu verwenden, um Geld zu sparen. Ebenso gaben 76,47 % (*n* = 26) aller Lai:innen an, dass die maschinell erstellten Übersetzungen gut genug für ihre Zwecke seien, während nur 41,18 % (*n* = 14) der Sprachexpert:innen dieser Aussage zustimmten. Unter den wenigen Teilnehmenden, die angaben, keine maschinellen Übersetzungstools bzw. Large Language Models zu nutzen, waren die Hauptgründe Ablehnung oder Gewohnheit.

Es zeigte sich eine klare Dominanz von drei Tools, die von den Teilnehmenden für Übersetzungszwecke verwendet werden. Am häufigsten wurde ChatGPT (80,88 %; *n* = 55) angegeben, dicht gefolgt von Google Translate (77,94 %; *n* = 53) und DeepL (75,00 %; *n* = 51). Abgesehen von Gemini (20,59 %; *n* = 14) blieben alle anderen Tools deutlich unter einem Wert von 10 % und zählten damit jeweils weniger als 7 Nutzer:innen (vgl. Abbildung 5). In der Kategorie Sonstige wurden unter anderem auch noch LLMs wie Claude, Copilot oder Perplexity sowie Übersetzungstools wie Pons Textübersetzer angeführt. Bei der Toolwahl konnten

nur leichte Unterschiede zwischen Sprachexpert:innen und Lai:innen festgestellt werden. Während Sprachexpert:innen DeepL und ChatGPT (jeweils 85,29 %; n = 29) als meistgenutzte Tools angaben, greifen Lai:innen tendenziell eher zu Google Translate (88,24 %; n = 30) vor ChatGPT (76,47 %; n = 26) und DeepL (64,71 %; n = 22).

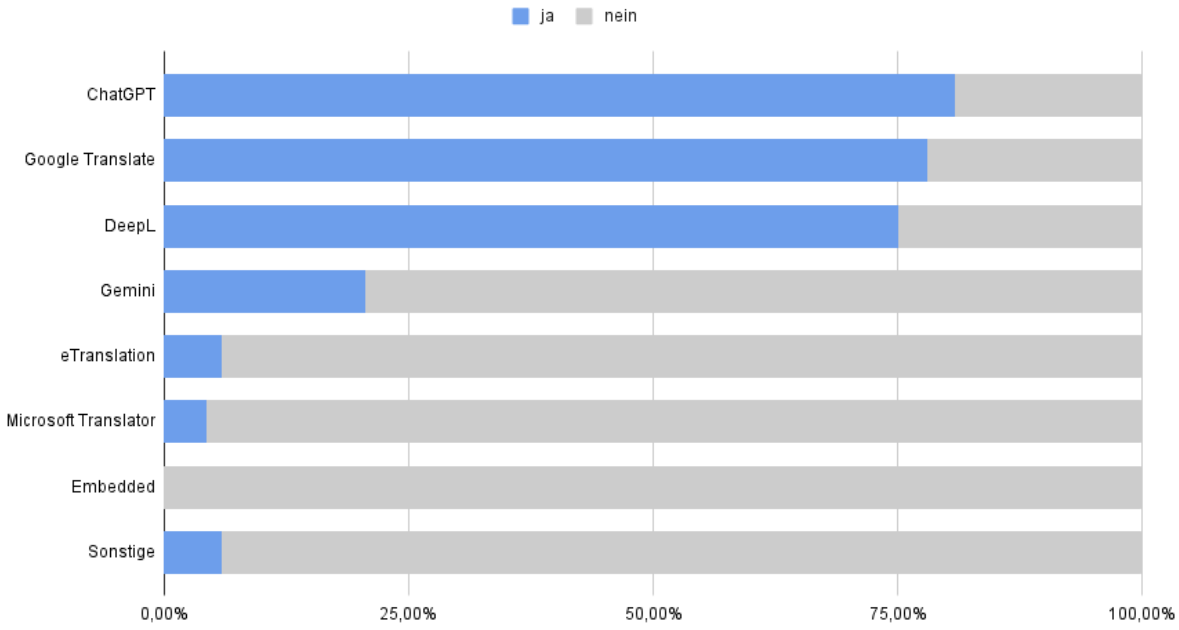

Abbildung 5: Nutzung von maschinellen Übersetzungstools und Large Language Models für Übersetzungszwecke

Hinsichtlich der Selbsteinschätzung von Teilnehmenden, ob menschliche und maschinelle Übersetzungen korrekt erkannt werden können, gaben beide Gruppen mit überwiegender Mehrheit an, dass sie sich unsicher seien. Sprachexpert:innen erzielten dabei mit 50,00 % (n = 18) einen etwas geringeren Wert als Lai:innen (52,78 %; n = 19). Mit 30,56 % (n = 11) war sich ein größerer Teil der Expert:innengruppe sicher, die maschinellen Übersetzungen erkennen zu können, während sich dies nur 19,44 % (n = 7) der Lai:innen zutrauten.

5.2 Unterscheidung zwischen maschinellen und Humanübersetzungen

Die erste Forschungsfrage bezieht sich auf die Fähigkeit von Sprachexpert:innen und Lai:innen zu erkennen, ob eine Übersetzung maschinell oder von Menschen angefertigt wurde. Zur Untersuchung dieser Forschungsfrage wurde ein abhängiger t-Test durchgeführt. Dabei wurde ver-

glichen, wie oft Humanübersetzungen bzw. maschinelle Übersetzungen korrekt als solche erkannt wurden. Da davon ausgegangen wurde, dass es Unterschiede zwischen der Erkennung von DeepL und ChatGPT gibt, wurden zwei zusätzliche abhängige t-Tests durchgeführt. Darüber hinaus wurde die Effektstärke und damit die Größe des Unterschieds mittels Cohen's d berechnet. Die detaillierten Ergebnisse dieser Vergleiche sind in Tabelle 2 zusammengefasst. Zur übersichtlichen Darstellung werden darin folgende statistische Standardabkürzungen verwendet: Mittelwert (*M*), Standardabweichung (*SD*), Stichprobengröße (*n*), t-Wert (*t*) mit den zugehörigen Freiheitsgraden (*df*), Signifikanzwert (*p*) sowie die Effektstärke (*d*).

Tabelle 2: Ergebnisse des abhängigen t-Tests zum Vergleich von Human- und maschinellen Übersetzungen

Vergleich	**Human *M* (*SD*)**	**Maschine *M* (*SD*)**	***n***	***t*(*df*)**	***p***	***d***
Human vs. Maschinell	2,15 (0,93)	2,68 (1,06)	72	3,98 (71)	< 0,001	0,47
Human vs. DeepL	2,33 (0,93)	2,82 (1,07)	39	2,47 (38)	0,018	0,40
Human vs. ChatGPT	1,94 (0,90)	2,48 (1,06)	33	3,03 (32)	0,005	0,53

Im Durchschnitt gaben die Proband:innen 4,83 von 8 möglichen richtigen Antworten bei einer Standardabweichung von 1,64. Dies entspricht einer durchschnittlichen Erkennungsrate von 60,38 % (348 von 576 bewerteten Textausschnitten) und liegt deskriptiv über dem Zufallsniveau von 50 %. Der Unterschied zwischen Human- und maschinellen Übersetzungen war statistisch signifikant ($p < 0{,}001$). Auch die Vergleiche zwischen Human- und DeepL-Übersetzungen ($p = 0{,}018$) sowie zwischen Human- und ChatGPT-Übersetzungen ($p = 0{,}005$) zeigten einen signifikanten Unterschied. Die Effektgrößen lagen dabei im Bereich kleiner bis mittlerer Effekte ($d = 0{,}40–0{,}53$) (vgl. Tabelle 2).

Abbildung 6 zeigt die Anteile korrekter und falscher Einschätzungen von Human- und maschinellen Übersetzungen. Insgesamt wurden 67,01 % (193 von 288 bewerteten Textausschnitten) aller maschinellen Übersetzungen korrekterweise als solche eingeschätzt, während nur 53,82 % (155 von 288) aller Humanübersetzungen korrekt als menschlich klassifiziert wurden. Im Umkehrschluss bedeutet dies, dass menschliche Übersetzungen in 46,18 % der Fälle

(133 von 288) fälschlicherweise für maschinell gehalten wurden. Maschinelle Übersetzungen wurden hingegen seltener fälschlich als menschlich eingestuft (32,99 %; 95 von 288).

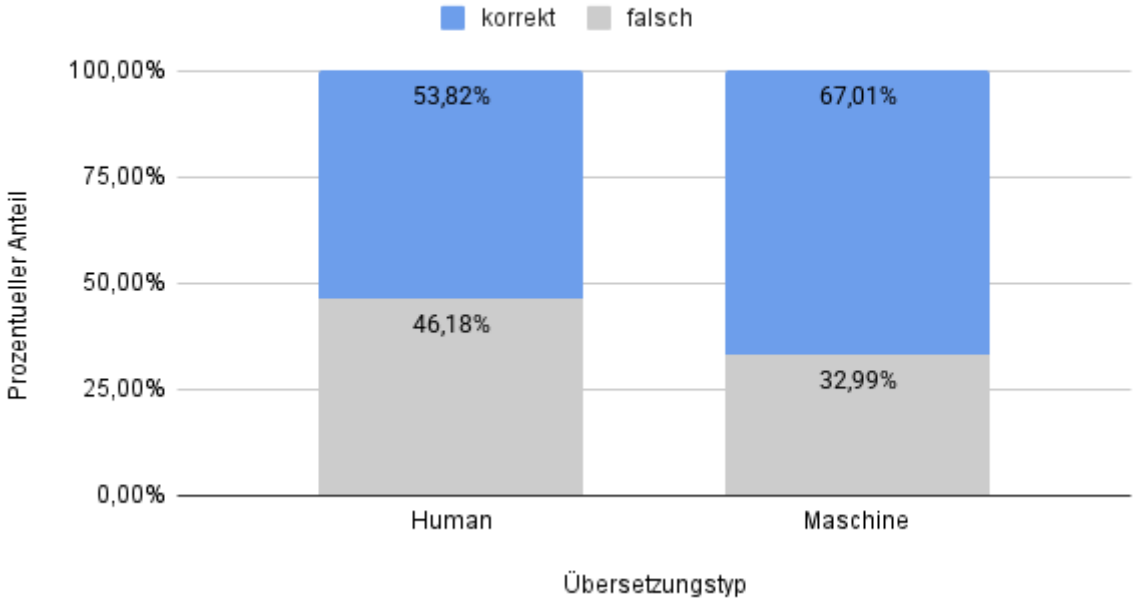

Abbildung 6: Korrekte und falsche Klassifikation von Übersetzungen (%)

Die korrekten Erkennungsraten variierten deutlich zwischen den einzelnen Textbeispielen. Sie reichten von 48,61 % (35 von 72) beim am schwersten zuzuordnenden Text bis hin zu 72,22 % (52 von 72) beim am deutlichsten erkennbaren Text, was einer beträchtlichen Differenz von nahezu 25 Prozentpunkten entspricht. Die höchsten korrekten Erkennungsraten zeigten sich bei den Textbeispielen 2, 7 und 8, bei denen rund 70 % (durchschnittlich 50 von 72 Teilnehmenden) die Übersetzungen korrekt als human oder maschinell einschätzten. Die Textausschnitte 1, 6 und 4 wurden hingegen mit rund 50 % (36 von 72) am seltensten korrekt zugeordnet. Besonders auffällig ist hierbei Textausschnitt 4, da er das einzige Beispiel darstellt, bei dem die Fehlklassifikationen (51,39 %) die korrekten Zuordnungen sogar leicht übertrafen. Die verbleibenden Textausschnitte 3 und 5 positionierten sich mit Raten von rund 58 % bzw. 64 % im Mittelfeld (vgl. Abbildung 7).

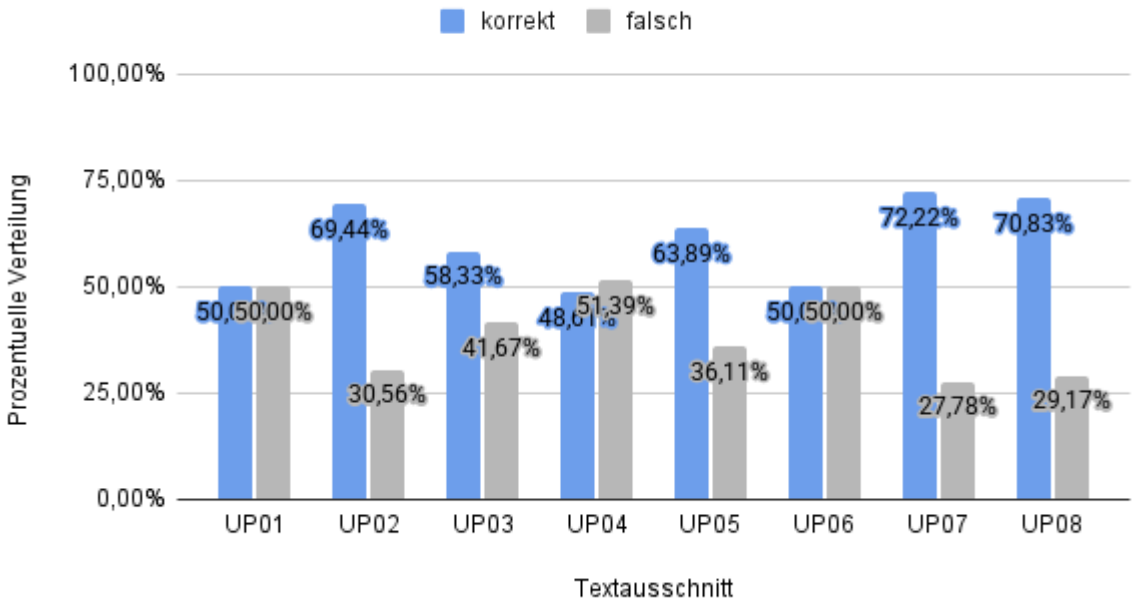

Abbildung 7: Klassifikationsverteilung nach Textausschnitt (UP) (%)

Zur weiteren Interpretation wurden die Texte in zwei Gruppen zusammengefasst (Textausschnitte 1–4, Textausschnitte 5–8) sowie mit ihrer Kennung im Fragebogen beschrieben (UP01–UP08). Textbeispiele 1–4 wurden aufgrund ihrer eher informativen und konventionellen Strukturen der ersten Textgruppe zugeordnet, während Textbeispiele 5–8 aufgrund einer höheren Dichte an kreativen und stilistischen Mitteln zur zweiten Textgruppe mit höherer Komplexität zusammengefasst wurden.

Die gruppierte Betrachtung zeigt, dass Texte der Gruppe 2, also mit höherer stilistischer und kreativer Komplexität, mit 64,24 % (185 von 288) deutlich häufiger korrekt klassifiziert wurden als die weniger komplexen Texte der Gruppe 1 mit 56,60 % (163 von 288). Dieser Trend zeigt sich auch bei der gezielten Betrachtung der maschinellen Übersetzungen. Nahezu 72,37 % (110 von 152) aller maschinell generierten Textausschnitte der zweiten Gruppe wurden als solche erkannt, während dieser Wert in Gruppe 1 mit 61,03 % (83 von 136) auch hoch, jedoch deutlich geringer war.

Betrachtet man im Gegensatz dazu die Fehlklassifikationen, zeigen sich sowohl im Gesamtergebnis als auch auf Ebene der einzelnen Textbeispiele starke Schwankungen (vgl. Abbildung 8). Bei den maschinell generierten Textausschnitten, die fälschlicherweise als Humanübersetzungen eingestuft wurden, reichten die Fehlerraten bei den einzelnen Texten von 17,65 % (6 von 34, geringste Fehlquote) bei Textausschnitt 2 bis hin zu 61,76 % (21 von 34**,** höchste

Fehlquote) bei Textausschnitt 1. Bei den Humanübersetzungen, die fälschlicherweise für maschinell generiert gehalten wurden, bewegten sich die Fehlquoten der einzelnen Texte in einem etwas geringeren Bereich zwischen 35,29 % (12 von 34) bei Textausschnitt 7 und 65,79 % (25 von 38) bei Textausschnitt 4.

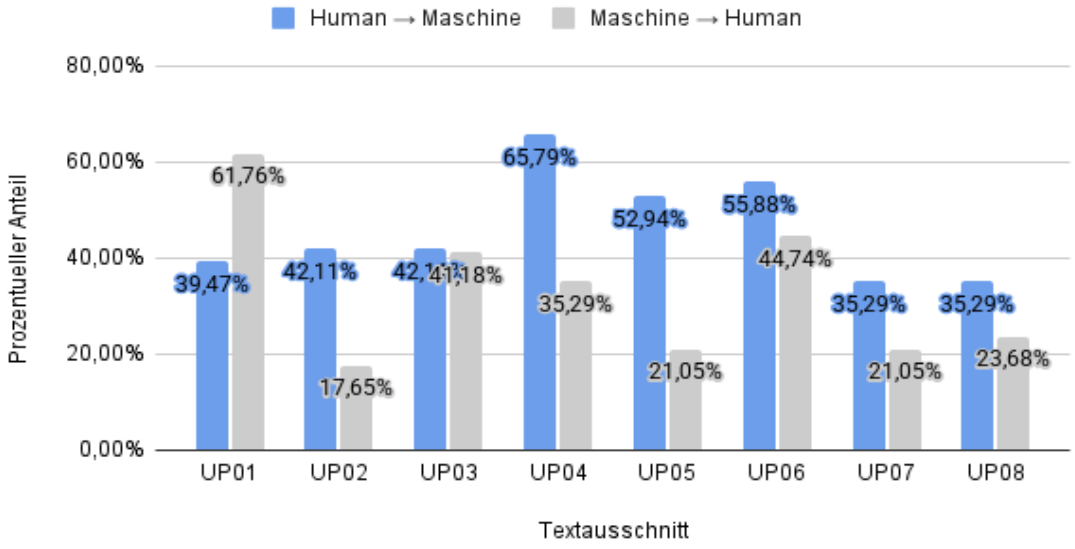

Abbildung 8: Falsche Klassifikation von Übersetzungen nach Textbeispielen (UP) (%)

Die Betrachtung der Fehlklassifikationen innerhalb der Textgruppen verdeutlicht ein durchgehendes Muster. In beiden Gruppen wurden Humanübersetzungen häufiger für maschinell gehalten als umgekehrt. In Textgruppe 1 wurden in 47,37 % der Fälle (72 von 152) Humanübersetzungen fälschlicherweise als maschinelle Übersetzungen wahrgenommen, während umgekehrt nur etwa 38,97 % (53 von 136) aller maschinellen Übersetzungen als human klassifiziert wurden. Auch in Textgruppe 2 zeigte sich mit 44,85 % (61 von 136) eine klare Tendenz zur Fehlinterpretation von humanübersetzten Texten als maschinell generierte Texte, während lediglich 27,63 % (42 von 152) aller maschinell übersetzten Textausschnitte fälschlich als human wahrgenommen wurden.

5.3 Unterschiede in der Erkennbarkeit nach verwendetem System

Die zweite Forschungsfrage bezieht sich auf den Unterschied in der Wahrnehmung zwischen den verwendeten Systemen. Zur Untersuchung dieser Forschungsfrage wurde ein unabhängiger t-Test durchgeführt. Verglichen wurde dabei die Anzahl korrekt erkannter maschineller Übersetzungen nach Tool. Pro Person konnten maximal vier korrekte Antworten erzielt werden. Die

Stichprobengrößen unterschieden sich leicht zwischen den Gruppen (DeepL: $n = 39$; ChatGPT: $n = 33$). Die Ergebnisse werden in Tabelle 3 dargestellt. Auch hier wird mit Stichprobengröße (*n*), Mittelwert (*M*) sowie Standardabweichung (*SD*) auf statistische Standardabkürzungen zurückgegriffen.

Tabelle 3: Mittelwerte und Standardabweichungen der korrekt erkannten maschinellen Übersetzungen für DeepL und ChatGPT

System	***n***	***M***	***SD***
DeepL	39	2,85	1,04
ChatGPT	33	2,48	1,06

Im Durchschnitt wurden 2,85 ($SD = 1,04$) von vier möglichen DeepL-Übersetzungen und 2,48 ($SD = 1,06$) von vier ChatGPT-Übersetzungen korrekt als maschinelle Übersetzungen erkannt (vgl. Tabelle 3). Es konnte kein statistisch signifikanter Unterschied zwischen DeepL- und ChatGPT-Übersetzungen festgestellt werden ($t(70) = 1,49$; $p = 0,14$ bei einer kleinen Effektgröße von $d = 0,35$).

Abbildung 9 zeigt die Erkennungsraten der maschinell generierten Texte je nach verwendetem System und Textausschnitt. Zwar konnte insgesamt kein signifikanter Unterschied festgestellt werden, auf Ebene der einzelnen Textbeispiele zeigten sich deskriptiv jedoch teils deutliche Abweichungen. In der ersten Textgruppe (UP01–UP04) variierten die Erkennungsraten stark. Während DeepL-Übersetzungen bei Textausschnitt 4 mit 83,33 % (15 von 18) klar häufiger erkannt wurden als ChatGPT (43,75 %; 7 von 16), verhielt es sich bei Textausschnitt 2 exakt umgekehrt. Hier war die Erkennungsrate von ChatGPT mit 93,75 % (15 von 16) sichtlich höher als jene von DeepL (72,22 %; 13 von 18).

In der zweiten Textgruppe (UP05–UP08) lagen die Erkennungsraten der beiden Tools durchgehend näher beieinander. Allerdings zeigte sich hier über alle vier Textbeispiele hinweg eine konstante Tendenz zu höheren Erkennungsraten für DeepL. Der größte Unterschied konnte bei Textausschnitt 5 festgestellt werden, bei dem 85,71 % (18 von 21) aller DeepL-Übersetzungen und 70,59 % (12 von 17) aller ChatGPT-Übersetzungen richtig zugeordnet wurden.

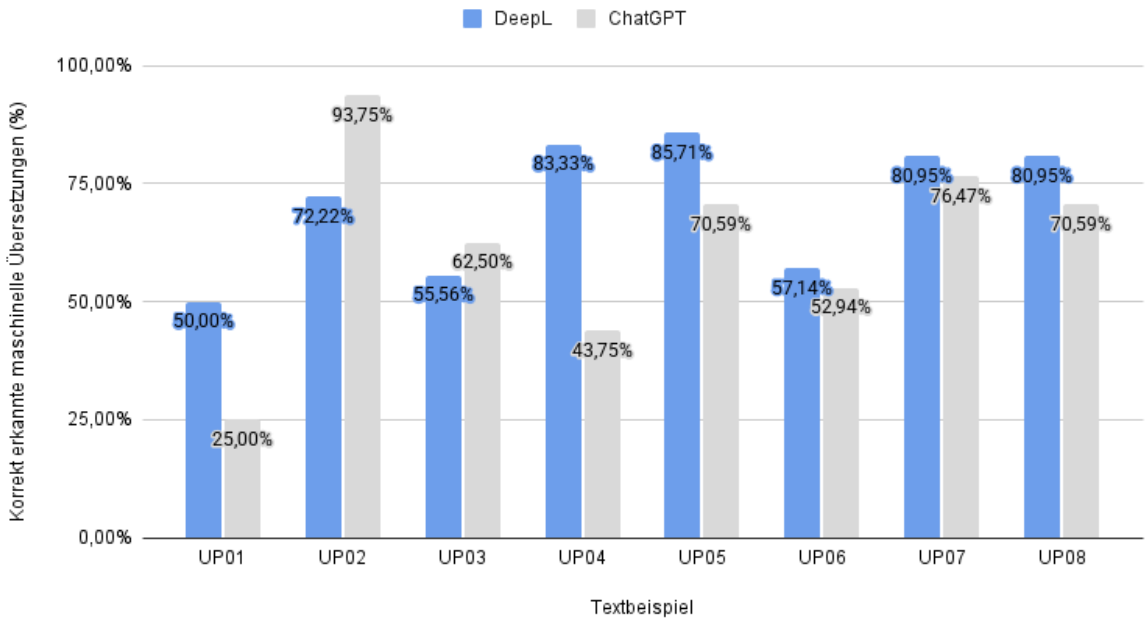

Abbildung 9: Korrekt erkannte maschinelle Übersetzungen nach Tool und Textbeispiel

5.4 Unterschiede nach Gruppen

Die dritte Forschungsfrage bezieht sich auf die Unterschiede in der Einschätzung zwischen Sprachexpert:innen und Lai:innen. Zur Untersuchung möglicher Unterschiede zwischen Lai:innen und Sprachexpert:innen wurde ein unabhängiger t-Test durchgeführt. Verglichen wurde die Gesamtzahl korrekt klassifizierter Übersetzungen pro Zielgruppe. In Tabelle 4 werden mit Stichprobengröße (*n*), Mittelwert (*M*) sowie Standardabweichung (*SD*) erneut statistische Standardabkürzungen verwendet.

Tabelle 4: Mittelwerte und Standardabweichungen der korrekt klassifizierten Übersetzungen nach Zielgruppe

Gruppe	***n***	***M***	***SD***
Lai:innen	36	4,69	1,55
Sprachexpert:innen	36	4,97	1,75

Die durchschnittliche Anzahl korrekt klassifizierter Übersetzungen lag für Sprachexpert:innen bei $M = 4{,}97$ ($SD = 1{,}75$) und bei Lai:innen bei $M = 4{,}69$ ($SD = 1{,}55$) von acht möglichen

richtigen Antworten (vgl. Tabelle 4). Es konnte kein statistisch signifikanter Unterschied zwischen beiden Gruppen festgestellt werden ($t(70) = 0{,}71$; $p = 0{,}48$) bei einer kleinen Effektgröße von $d = 0{,}17$.

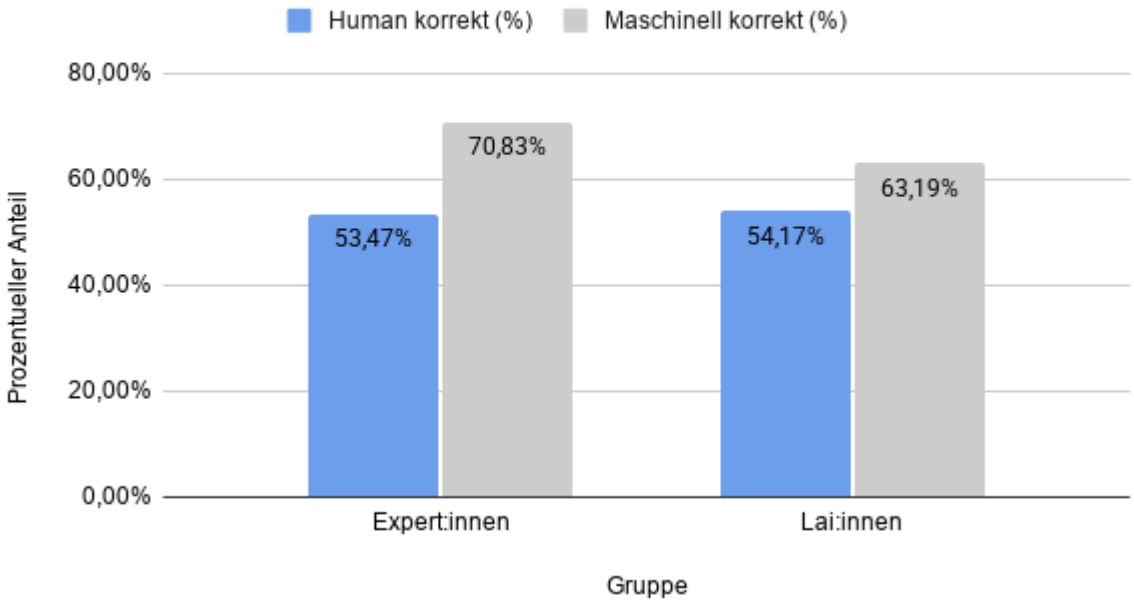

Abbildung 10: Korrekte Klassifikation durch Sprachexpert:innen und Lai:innen

Abbildung 10 zeigt die korrekten Klassifikationen von Sprachexpert:innen und Lai:innen. Beide Gruppen schätzten Humanübersetzungen nahezu gleich oft korrekt ein, wobei die Rate bei den Sprachexpert:innen bei 53,47 % (77 von 144) und bei den Lai:innen bei 54,17 % (78 von 144) lag. Bei der Einschätzung von maschinellen Übersetzungen zeigte sich hingegen, dass Sprachexpert:innen diese mit 70,83 % (102 von 144) deskriptiv häufiger korrekt erkannten als Lai:innen (63,19 %; 91 von 144).

5.5 Detailbewertung nach Kriterien

Ein weiterer zu analysierender Aspekt ist die Detailbewertung der Übersetzungen nach den ausgewählten Kriterien, wobei es besonders interessant ist, ob die Einschätzung der Übersetzung als human oder maschinell einen Einfluss auf diese Bewertung hat. Um dies zu untersuchen, wurden für jede Bewertungskategorie gewichtete Mittelwerte berechnet. Zusätzlich wurden Standardabweichungen berechnet, um die Streuung der Bewertungen innerhalb der jeweiligen Wahrnehmungskategorien klarer darzustellen. Die Bewertungskategorien waren Flüssigkeit, Natürlichkeit, Stimmigkeit der Wortwahl, Stimmigkeit des Satzbaus, die Angemessenheit von Stil und Ton im Hinblick auf den spezifischen Textkontext, und Stimmigkeit generell.

Diese Merkmale wurden jeweils auf einer fünfstufigen Skala evaluiert. Die Berechnungen basieren ausschließlich auf der Wahrnehmung der Teilnehmenden und erfolgen unabhängig davon, ob die jeweilige Einschätzung richtig oder falsch war.

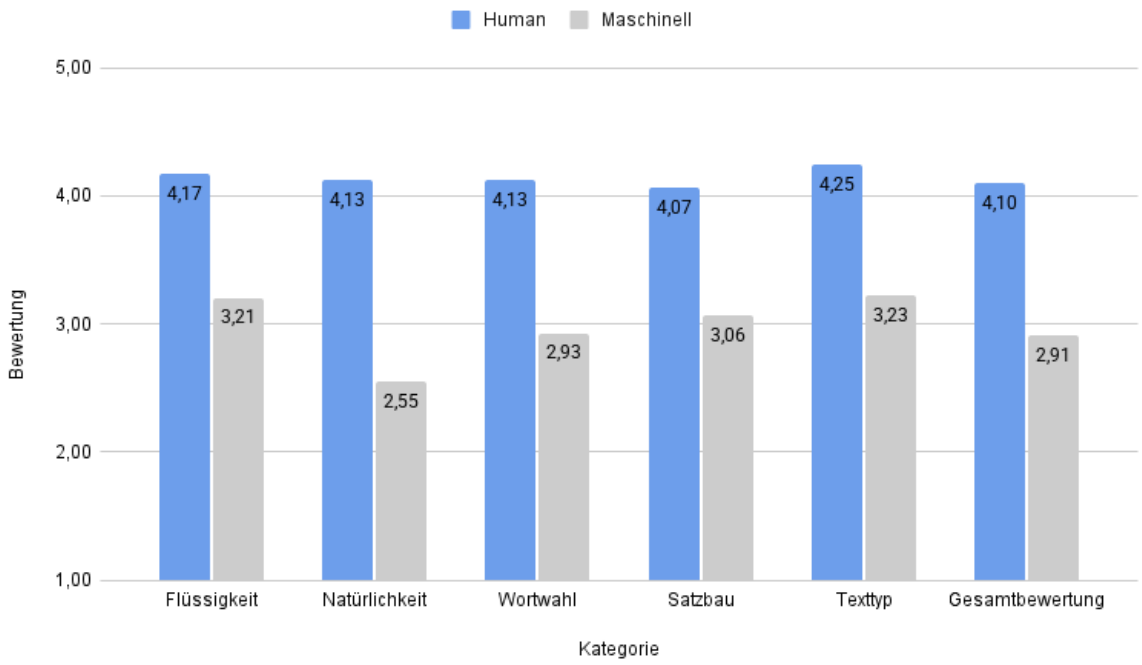

Abbildung 11: Bewertungen der Textausschnitte in Abhängigkeit von der Wahrnehmung

Die Auswertung der Bewertungen zeigte, dass Textausschnitte, die als human wahrgenommen wurden, in allen sechs Kategorien deutlich höher bewertet wurden als Texte, die maschinell wahrgenommen wurden. Besonders starke Unterschiede zeigten sich bei der Natürlichkeit (ΔM = 1,58), während auch bei allen anderen Kategorien eine Differenz von rund einem Skalenpunkt beobachtet wurde (vgl. Abbildung 11).

Die gewichteten Standardabweichungen zeigten, dass die Bewertungen bei als human wahrgenommenen Texten eine geringere Streuung aufwiesen. Die Streuung bei diesen Texten betrug zwischen 0,85 und 1,10, während sie bei als maschinell wahrgenommenen Texten zwischen 1,04 und 1,16 lag.

Auch im Vergleich zwischen Sprachexpert:innen und Lai:innen (vgl. Abbildung 12) zeigte sich, dass beide Gruppen in allen Kategorien human wahrgenommene Übersetzungen besser bewerteten als maschinell wahrgenommene Texte. Die Differenzen waren dabei jedoch klein und uneinheitlich und fielen zwischen beiden Gruppen ähnlich aus. Die größten Unterschiede zeigten sich auch im Gruppenvergleich bei der Natürlichkeit und Wortwahl sowie bei

der Gesamtbewertung. Die Grafik zeigt zudem, dass Sprachexpert:innen tendenziell sowohl human als auch maschinell wahrgenommene Übersetzungen in allen Kategorien leicht höher bewerteten als Lai:innen.

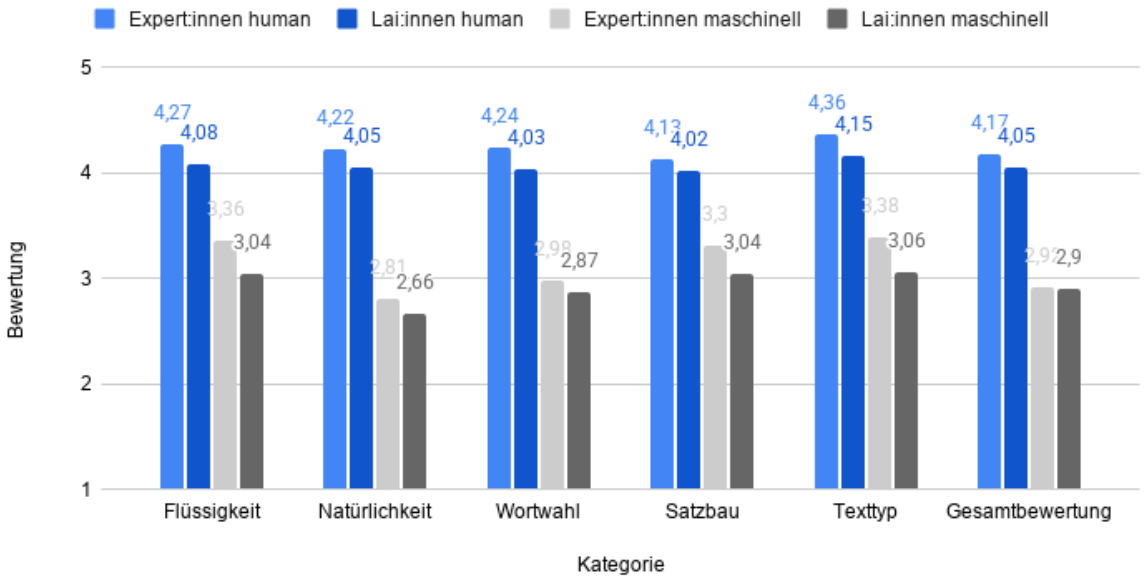

Abbildung 12: Bewertung nach Kategorie in Abhängigkeit der Wahrnehmung von Sprachexpert:innen und Lai:innen im Vergleich

Über alle Textausschnitte hinweg zeigten sich bei als humanübersetzt wahrgenommenen Textausschnitten stabil hohe Werte zwischen 4,0 und 4,3. Dies deutet darauf hin, dass der Grad kreativer bzw. stilistischer Komplexität die Bewertung human klassifizierter Texte kaum beeinflusste. Wenn Texte jedoch als maschinell wahrgenommen wurden, zeigten sich deutliche Unterschiede hinsichtlich der stilistischen und kreativen Komplexität. Bei Textgruppe 2 fiel die Differenz zwischen als human und als maschinell wahrgenommenen Texten deutlich größer aus als bei Textgruppe 1. Dieser Einbruch zeigt sich besonders prägnant in den Einzelkategorien Natürlichkeit und Wortwahl. Hier schnitten die als maschinell wahrgenommenen Texte der Gruppe 2 mit Tiefstwerten von 2,40 bzw. 2,63 Punkten signifikant schlechter ab als die übrigen Textbeispiele (vgl. Abbildung 13).

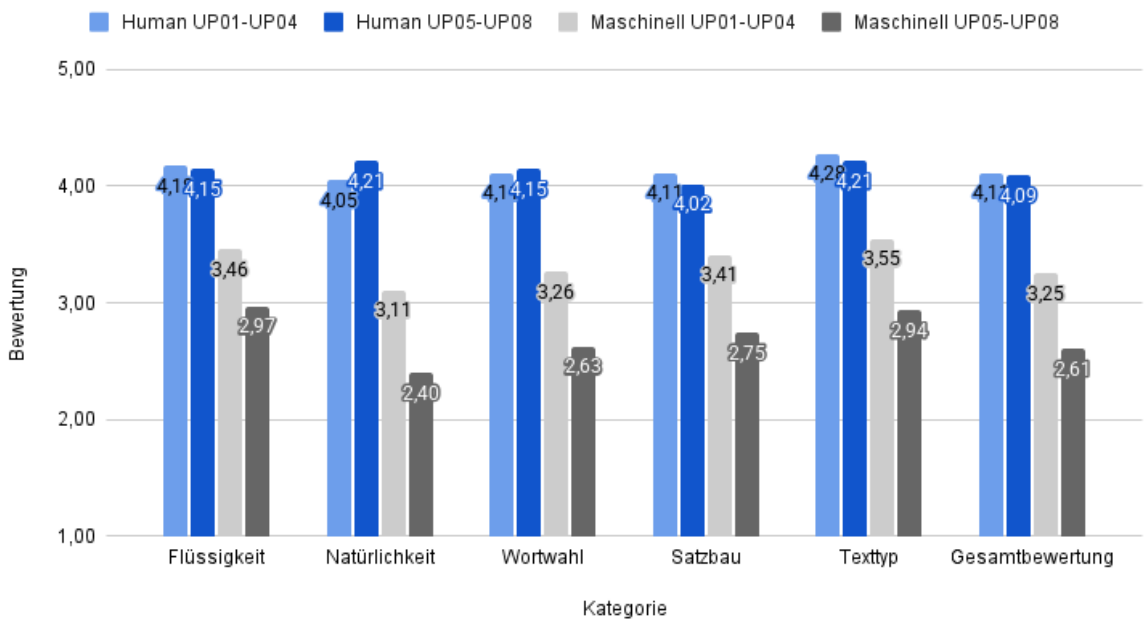

Abbildung 13: Kategoriebewertung nach Wahrnehmung, Zielgruppe und Kreativitätsniveau

5.6 Unterschiede in der Bewertung bei korrekt bzw. falsch eingeschätzten Texten

In Bezug auf die Detailbewertung ist es auch interessant zu analysieren, welche sprachlichen und stilistischen Unterschiede bei Texten erkannt wurden, die korrekt oder falsch als maschinelle bzw. von Menschen angefertigte Übersetzungen klassifiziert wurden. Zur Untersuchung dieser Frage wurden die Bewertungen nach korrekter bzw. falscher Einschätzung gruppiert und die gewichteten Mittelwerte sowie Standardabweichungen berechnet. Der tatsächliche Ursprung der Übersetzung (human oder maschinell) wurde dabei nicht berücksichtigt.

Die Ergebnisse zeigten, dass es in Abhängigkeit von der Korrektheit der Einschätzung nur geringfügige Unterschiede in der Bewertung gab. Die Differenzen der gewichteten Mittelwerte fielen in den meisten Kategorien zwischen $\Delta M = 0{,}03$ und $\Delta M = 0{,}10$. Eine Ausnahme bildete die Gesamtbewertung mit einer Differenz von $\Delta M = 0{,}26$. Die Streuungen lagen je nach Kategorie zwischen 0,10 und 0,21 (vgl. Abbildung 14).

Auffällig ist, dass Texte über nahezu alle Kategorien hinweg systematisch besser bewertet wurden, wenn sie falsch eingeschätzt wurden. Dies deutet darauf hin, dass eine höhere wahrgenommene Qualität nicht konsequent mit einer höheren Wahrscheinlichkeit einer korrekten Einschätzung einhergeht.

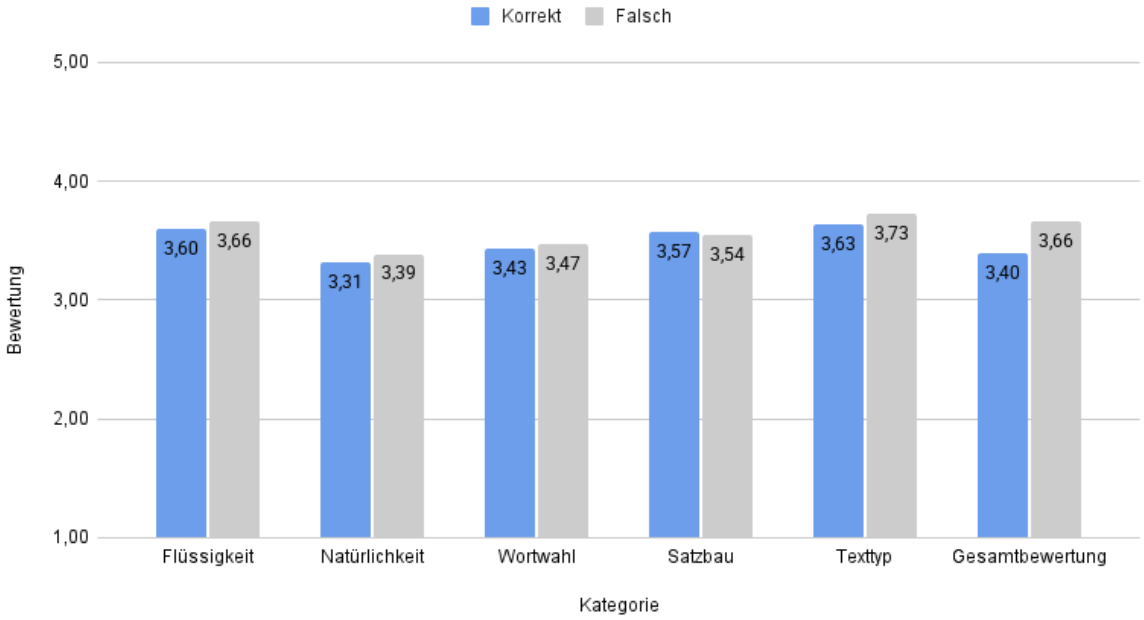

Abbildung 14: Bewertung sprachlicher und stilistischer Eigenschaften in Abhängigkeit von der Korrektheit der Einschätzung

Unter Berücksichtigung des Kreativitätsniveaus wird ersichtlich, dass stilistisch einfachere Texte unabhängig davon, ob sie korrekt oder falsch eingeschätzt wurden, tendenziell besser bewertet wurden als stilistisch komplexere Textausschnitte (vgl. Abbildung 15). Zwischen Sprachexpert:innen und Lai:innen zeigten sich lediglich geringe deskriptive Unterschiede, weshalb diese nicht gesondert dargestellt werden.

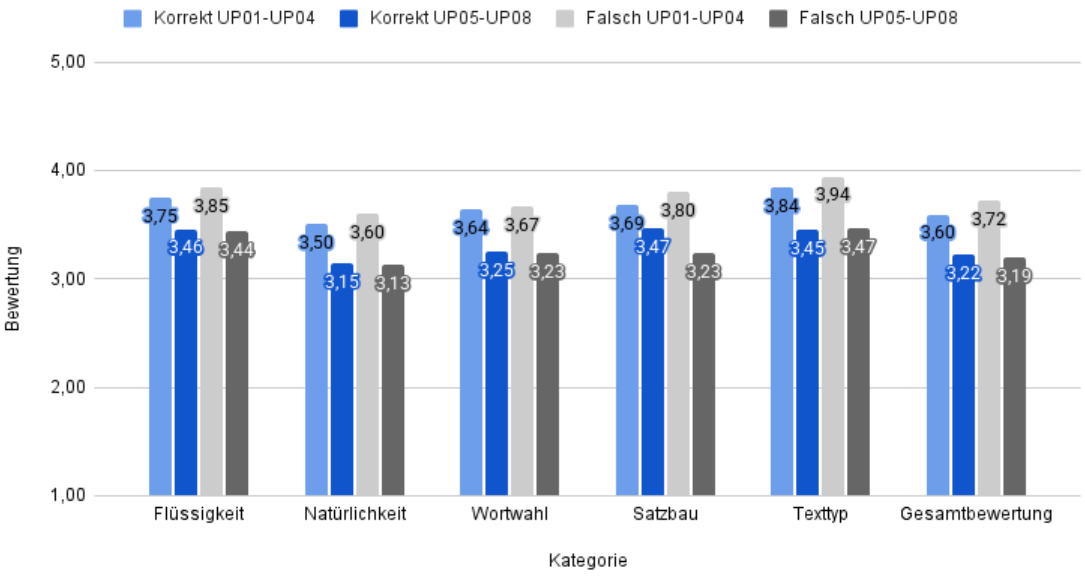

Abbildung 15: Bewertung sprachlicher und stilistischer Eigenschaften in Abhängigkeit von der Korrektheit der Einschätzung und dem Kreativitätsniveau

5.7 Unterschiede in der Bewertung zwischen Human- und maschinellen Übersetzungen

Nach der Analyse der subjektiven Wahrnehmungseffekte wurde abschließend untersucht, wie Human- bzw. maschinelle Übersetzungen tatsächlich bewertet wurden. Hierfür wurden im ersten Schritt alle Bewertungen der maschinell generierten Texte beider Systeme und der Humanübersetzungen gegenübergestellt.

Die Auswertung der gewichteten Mittelwerte über die Gesamtstichprobe hinweg zeigte, dass Humanübersetzungen ausnahmslos in allen sechs Bewertungskategorien durchgehend besser bewertet wurden als die maschinell erstellten Texte. Der deutlichste Unterschied zeigte sich in den Bewertungen der Natürlichkeit ($\Delta M = 0{,}57$), dicht gefolgt von der Angemessenheit von Stil und Ton zum Texttyp ($\Delta M = 0{,}54$) sowie bei der Flüssigkeit ($\Delta M = 0{,}48$) (vgl. Abbildung 16).

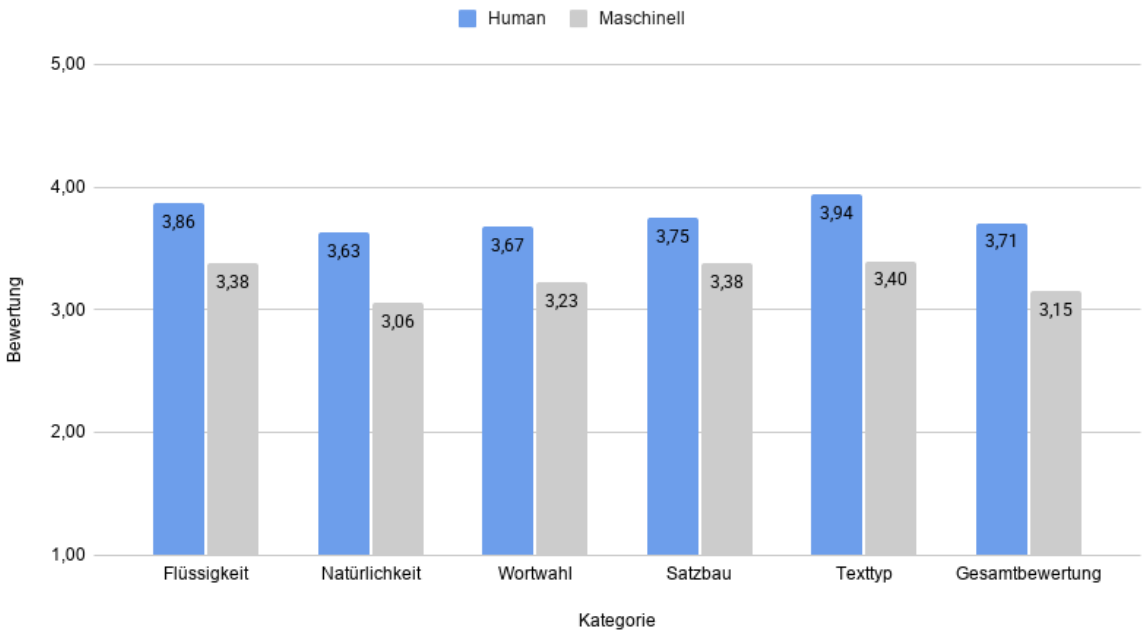

Abbildung 16: Bewertung nach Übersetzungsart (Human vs. Maschinell)

Unter Berücksichtigung der Zielgruppen konnte der übergeordnete qualitative Trend der besseren Bewertung von Humanübersetzungen sowohl bei Sprachexpert:innen als auch bei Lai:innen festgestellt werden, wobei Sprachexpert:innen über alle Kategorien hinweg sowohl Human- als auch maschinelle Übersetzungen geringfügig besser bewertete.

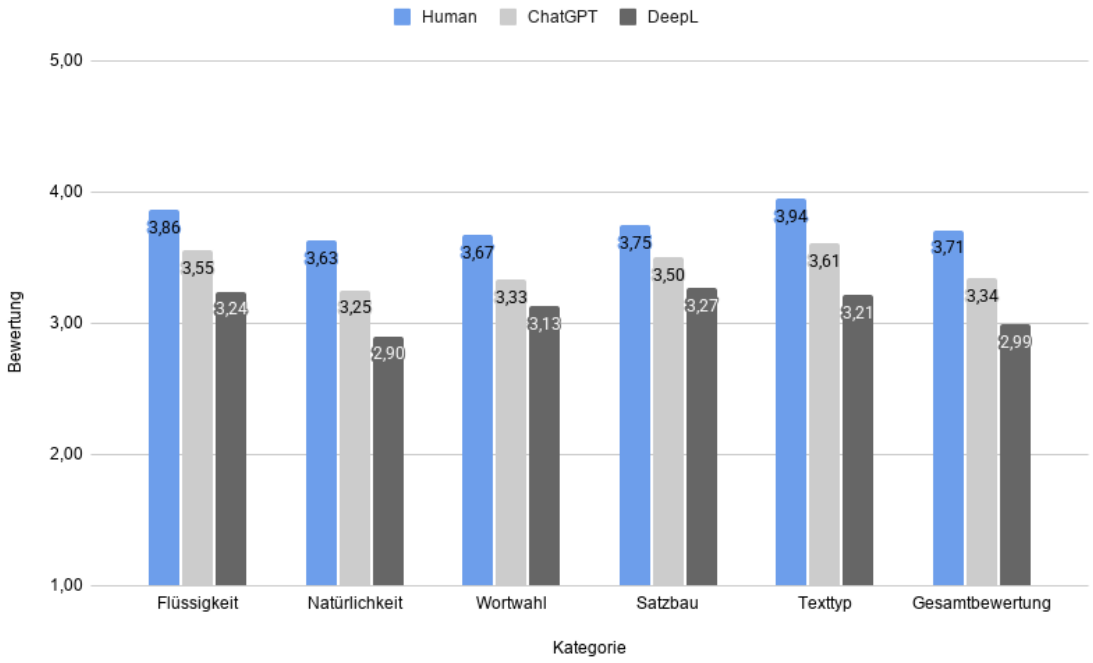

Abbildung 17: Bewertung nach Übersetzungsart (Human vs. ChatGPT und DeepL)

Im zweiten Schritt erfolgte eine detaillierte Aufschlüsselung nach dem für die Übersetzung verwendeten System, wobei sich zeigte, dass ChatGPT in allen Kategorien bessere Ergebnisse erzielte als DeepL, jedoch in keiner Kategorie die Humanübersetzungen übertreffen konnte (vgl. Abbildung 17). Bei der Stimmigkeit des Satzbaus lag ChatGPT beispielsweise mit M = 3,50 (SD = 0,98) nahezu mittig zwischen der Humanübersetzung (M = 3,75; SD = 1,00) und DeepL (M = 3,27; SD = 0,99).

Zusammenfassend lässt sich somit festhalten, dass ungeachtet der Wahrnehmung eines Textes als Human- oder maschinelle Übersetzung auch in der objektiven Detailbewertung die Humanübersetzungen am besten abschnitten, während ChatGPT im direkten Systemvergleich deutlich bessere Bewertungen als DeepL erzielen konnte.

6 Diskussion

Im folgenden Kapitel werden die in Kapitel 5 dargelegten Ergebnisse diskutiert, kritisch hinterfragt und in den Forschungskontext gesetzt.

Hinsichtlich der Fähigkeit zur Unterscheidung zwischen humanen und maschinellen Übersetzungen beträgt die Gesamtquote korrekt klassifizierter Textausschnitte 60,38 % und liegt damit nur moderat über dem Zufallsniveau. Dies deutet darauf hin, dass die Teilnehmenden nur begrenzt in der Lage waren, zwischen humanen und maschinellen Übersetzungen zu unterscheiden. Diese Erkenntnisse stehen im Einklang mit bisher durchgeführten Studien, bei denen sich die richtigen Erkennungsraten insbesondere bei appellativen, kreativen Textsorten nur knapp über dem Zufallsniveau befanden (vgl. Doru et al. 2025; Gehrmann et al. 2019).

Im Rahmen der vorliegenden Turing-Test-ähnlichen Untersuchung ist vor allem relevant, in welchem Ausmaß maschinell generierte Texte von Teilnehmenden auch korrekt als solche klassifiziert wurden. Insgesamt wurden 32,99 % der von DeepL und ChatGPT generierten Übersetzungen als humanübersetzt wahrgenommen. Dieser Wert liegt über der von Turing (1950) vorgeschlagenen Schwelle von 30 % und deutet darauf hin, dass maschinelle Übersetzungen in einem relevanten Anteil der Fälle als menschlich übersetzt wahrgenommen wurden und somit funktionale Ununterscheidbarkeit erzielen konnten. Gleichzeitig zeigt die weitaus höhere Erkennungsrate, dass diese Ununterscheidbarkeit nicht konsistent gegeben ist und nicht mit den Ergebnissen jüngster Studien zu echten Turing-Tests vergleichbar ist. Bei Jones und Bergen (2024a, 2024b, 2025) lag die Quote falsch klassifizierter KI-geführter Konversationen mit bis zu 54 % beträchtlich höher. Die Ergebnisse weisen daher nicht auf eine vollständige, wie von Hassan et al. (2018) vorgeschlagene Human Parity hin. Vielmehr bestätigen sie Läubli et al. (2018), wonach maschinelle Übersetzungssysteme bei der Erstellung von in sich geschlossenen, kohärenten Übersetzungen auf Grenzen stoßen.

Bei Betrachtung der einzelnen Textausschnitte kann festgehalten werden, dass sich die Erkennungsraten je nach Textbeispiel sehr stark unterscheiden. Textausschnitte in Gruppe 2 weisen deutlich höhere korrekte Klassifikationsraten auf als Texte der ersten Gruppe, was darauf schließen lässt, dass kreative Elemente starken Einfluss auf die Wahrnehmung von Teilnehmenden nehmen. Insbesondere die Textausschnitte von Innocent und Casper weisen eine sehr hohe Trefferrate auf, was darauf hindeutet, dass Wortneuschöpfungen, idiomatische Phrasen, Sprichwörter und Wortspiele maschinelle Übersetzungssysteme bzw. Large Language Models vor besondere Herausforderungen stellen. Ebenso hoch war die Erkennungsrate beim Textausschnitt von Atomic, der sich vor allem durch seine stark emotionale Bildsprache und seine

zusammenhängende Rahmenhandlung auszeichnet. Die Ergebnisse deuten darauf hin, dass die maschinellen Systeme auch diese textuellen Besonderheiten nicht adäquat übertragen konnten. Es lässt sich daher beobachten, dass kreative Texte nach wie vor eine große Herausforderung für die maschinelle Übersetzung darstellen, was mit den Ergebnissen anderer Studien zur Wahrnehmung von maschinell generierten kreativen Übersetzungen vergleichbar ist (vgl. Guerberof-Arenas & Toral 2022, 2024; Toral & Way 2018).

Besonders auffällig ist eine klare asymmetrische Fehlklassifikation. Humanübersetzte Texte wurden viel häufiger falsch als maschinell klassifiziert als umgekehrt. Dies deutet darauf hin, dass Teilnehmende eher dazu tendieren, humanübersetzte Texte als maschinell einzuschätzen, was zum einen an einer Wahrnehmungsverzerrung, zum anderen an der Qualität der Übersetzungen selbst liegen könnte. Vor allem zwischen den einzelnen Textausschnitten zeigen sich hier relevante Unterschiede. Humanübersetzte Texte der Gruppe 2, demnach kreativere Texte bzw. Texte mit höherer stilistischer Komplexität, wurden deutlich häufiger als maschinell generiert eingestuft als umgekehrt. Dies deutet darauf hin, dass insbesondere bei kreativen Texten sprachliche Ungereimtheiten stärker auffallen und von Teilnehmenden als Hinweis auf eine maschinell generierte Übersetzung wahrgenommen werden, unabhängig davon, ob es sich tatsächlich um eine maschinelle Übersetzung handelt oder nicht.

Die Ergebnisse stützen damit die getroffenen Grundannahmen, dass weniger komplexe maschinelle Übersetzungen häufiger als human wahrgenommen wurden, während Texte mit höherer kreativer bzw. stilistischer Komplexität häufiger korrekt als maschinell erkannt wurden.

Auffällig ist zudem, dass sich die Ergebnisse grundlegend von vergleichbaren Studien unterscheiden. So zeigen sowohl Chein et al. (2024) als auch Rathi et al. (2025), dass KI-generierte Texte häufiger als menschlich wahrgenommen werden als umgekehrt. Während Rathi et al. Chatprotokolle im Rahmen von Displaced und Inverted Turing-Tests untersuchten, stützten Chein et al. ihre Untersuchungen auf Social-Media- bzw. Blogartikel, welche hinsichtlich ihrer Textform und der Rezeption vor allem relevant für den Vergleich mit den vorliegenden Untersuchungen sind. Die unterschiedlichen Ergebnisse scheinen daher überraschend. Als mögliche Erklärungsansätze können neben der Textsorte auch die Länge der untersuchten Texte eine Rolle spielen, die im Rahmen der vorliegenden Untersuchungen mit durchschnittlich 108 Wörtern mehr als doppelt so lang wie jene bei Chein et al. (2024) waren. Zum einen bergen längere Texte mehr Fehlerpotenzial, zum anderen laufen sie Gefahr, das oben thematisierte Kohärenzniveau nicht zu erreichen.

Ebenso könnte die explizite Designierung der Textausschnitte als Übersetzungen zu Wahrnehmungsverschiebungen führen. Vor dem Hintergrund der in Kapitel 2.6 beschriebenen langjährigen qualitativen Vorbehalte gegenüber der maschinellen Übersetzung ist davon auszugehen, dass Teilnehmende sprachliche Auffälligkeiten mit dem maschinellen Ursprung eines Textes in Verbindung bringen. Diese Annahme einer translationsspezifischen Skepsis wird zusätzlich gestützt, wenn die Ergebnisse der vorliegenden Arbeit mit jenen ähnlicher Untersuchungen verglichen werden, wie beispielsweise den Turing-Tests von Jones & Bergen (2025), Rathi et al. (2025) und Chein et al. (2024).

Insgesamt zeigt sich nur eine begrenzte positive Tendenz in Bezug auf die Annahme, dass die Unterscheidung zwischen Human- und maschineller Übersetzung korrekt möglich ist. Zwar lag die Gesamtquote korrekter Klassifikationen über dem Zufallsniveau, die Ergebnisse deuten dennoch nur auf eine begrenzte Fähigkeit zur Unterscheidung zwischen humanen und maschinellen Übersetzungen hin.

Bezüglich der Rolle des für die Übersetzungen verwendeten Systems zeigen die Ergebnisse auf aggregierter Ebene keinen signifikanten Unterschied zwischen den von DeepL und ChatGPT generierten Ausgaben. Eine Analyse über die einzelnen Textausschnitte hinweg weist jedoch auf Unterschiede in Abhängigkeit vom Kreativitätsniveau hin. Es lässt sich daraus schließen, dass das verwendete Übersetzungstool insgesamt keinen entscheidenden Einfluss auf die Wahrnehmung der Textausschnitte hat. Stattdessen scheinen die stilistischen und kreativen Merkmale der jeweiligen Textausschnitte eine deutlich größere Rolle zu spielen.

Während sich bei einfacheren Textausschnitten mit stärkeren informativen Eigenschaften und begrenzten kreativen Stilmitteln (Gruppe 1) keine klaren Muster zur Überlegenheit eines Tools feststellen lassen, zeigen die Ergebnisse bei kreativ anspruchsvolleren Texten eine leichte Tendenz zugunsten von ChatGPT. Dies bedeutet, dass ChatGPT in Gruppe 2 minimal häufiger als humanübersetzt wahrgenommen wurde als DeepL. Dieses Ergebnis lässt sich damit erklären, dass DeepL häufig zwar akkuratere Übersetzungen generiert, stilistisch aber häufig näher am Ausgangstext bleibt, während ChatGPT-Outputs als flüssiger wahrgenommen werden und sich näher an den Strukturen von Humanübersetzungen orientieren (vgl. Kamaluddin et al. 2024; Sizov et al. 2024). Dass ChatGPT bei kreativen Texten besser abschnitt, deckt sich zudem mit der objektiven Detailbewertung, bei der sich das LLM zwar in allen Kategorien hinter den durchwegs am besten bewerteten Humanübersetzungen, jedoch klar vor den von DeepL generierten Übersetzungen positionierte. Diese Ergebnisse sind insbesondere für kreativere Textsorten mit starken stilistischen Mitteln oder Storytelling-Elementen relevant, wie es in den vorliegenden Textausschnitten überwiegend der Fall ist. Besonders deutlich wird dies

am Beispiel von Atomic, bei dem ChatGPT vor allem hinsichtlich der verwendeten Terminologie schlechter bewertet wurde und dadurch als maschinell erkannt wurde. Konkret können hier die falsche Übersetzung der verschiedenen Schneekonzepte oder auch der stark konnotierte Begriff „Lines“ angeführt werden. Dies bestätigt die Annahme, dass bereits kleinste Unterschiede in Wortwahl oder Stil beträchtliche Auswirkungen auf die Wahrnehmung der Übersetzungsqualität haben können (vgl. Hiebl & Gromann 2024).

In Verbindung mit den Ergebnissen spricht die vorliegende Untersuchung dafür, dass nicht das verwendete maschinelle System, sondern vielmehr die textuellen Eigenschaften einen größeren Einfluss auf die Wahrnehmung ausüben.

Im Hinblick auf den Expert:innenstatus weisen die Ergebnisse darauf hin, dass Sprachexpert:innen nicht besser zwischen humanen und maschinellen Übersetzungen unterscheiden können. Zwar erzielten Sprachexpert:innen einen leicht höheren Mittelwert bei der korrekten Zuordnung, dieser war jedoch statistisch nicht signifikant. Auch hinsichtlich der einzelnen Textausschnitte können keine deutlichen Trends zur besseren Erkennbarkeit durch eine der beiden Gruppen festgestellt werden.

Diese Ergebnisse sind überraschend, da durch die intensive Auseinandersetzung mit Texten, sei es im Studium oder im beruflichen Umfeld, ein geschärftes Sprachgefühl und damit eine höhere Treffsicherheit bei der Erkennung von Texten maschinellen bzw. humanen Ursprungs zu erwarten war. Dies kann zum einen mit der immer besser werdenden Qualität von maschinellen Übersetzungssystemen und LLMs erklärt werden (vgl. Kamaluddin et al. 2024; Schneider 2017), zum anderen damit, dass die Wahrnehmung nicht zwangsläufig mit explizitem Fachwissen verknüpft ist. Auch Lai:innen konsumieren im Alltag eine Vielzahl unterschiedlicher Textsorten und verfügen somit über implizite Bewertungsmuster. Dies lässt darauf schließen, dass allgemeine Bewertungsstrategien bei beiden Zielgruppen ähnlich sind. Darüber hinaus ließ sich kein gruppenspezifischer Unterschied hinsichtlich der Nutzung im beruflichen Umfeld feststellen, was darauf hindeutet, dass auch Lai:innen maschinelle Übersetzungstools bzw. Large Language Models für Übersetzungszwecke in ihrem beruflichen Alltag verwenden. Auch diese regelmäßige Nutzung könnte im Zusammenhang mit dem ausgeglichenen Ergebnis zwischen den Zielgruppen stehen.

Beide Zielgruppen erkennen maschinell generierte Übersetzungen häufiger als humanübersetzte Texte, wenngleich ein geringer Unterschied in der Verteilung beobachtet wurde. Sprachexpert:innen klassifizierten maschinelle Übersetzungen häufiger korrekt als solche, was auf eine höhere Sensibilität für typische Merkmale maschineller Übersetzungen sowie mehr Expertise im Umgang mit maschinellen Übersetzungstools zurückgeführt werden könnte. Es

kann daher angenommen werden, dass klassische Fehler von maschinellen Übersetzungssystemen von der Expert:innengruppe eher wahrgenommen werden. Während andere Studien hinsichtlich der Expert:innenrolle zu dem Ergebnis kamen, dass sich professionelle Sprachdienstleister:innen länger und genauer mit den zu bewertenden Textausschnitten auseinandersetzen (vgl. Kaspere et al. 2023), konnte dies im Rahmen dieser Untersuchungen nicht statistisch signifikant bestätigt werden.

Ein weiterer zentraler Diskussionspunkt ist die wahrgenommene Qualität von Texten in Abhängigkeit vom angenommenen Ursprung. Die Ergebnisse zeigen einen klaren Zusammenhang zwischen der wahrgenommenen Herkunft eines Textes und dessen Bewertung. Über alle Bewertungskategorien hinweg wurde beobachtet, dass Übersetzungen deutlich schlechter bewertet wurden, wenn sie von Teilnehmenden als maschinell generiert wahrgenommen wurden, unabhängig davon, ob dies tatsächlich auch der Fall war. Während in vorangegangenen Studien schlechtere Bewertungen festgestellt wurden, sobald Teilnehmende wussten, dass es sich um maschinell generierte Übersetzungen bzw. Untertitel handelte (vgl. Asscher & Glikson 2023; Qiu 2025), zeigt die vorliegende Untersuchung, dass bereits die Annahme eines maschinellen Ursprungs ausreicht, um einen Text schlechter zu bewerten. Sie verdeutlicht zudem, dass die wahrgenommene Übersetzungsqualität nicht zwangsläufig mit der tatsächlichen Übersetzungsqualität übereinstimmen muss (vgl. Agrawal 2024; Kocmi et al. 2021), da dieselben Textausschnitte weitaus besser bewertet wurden, wenn sie als Humanübersetzung eingeordnet wurden. Dies geht auch mit den tatsächlichen Bewertungen nach Übersetzungsart einher. Humanübersetzungen schnitten in der objektiven Detailbewertung in allen Kategorien am besten ab, was darauf hindeutet, dass maschinelle Übersetzungen Humanübersetzungen nach wie vor qualitativ unterlegen sind. Diese kombinierten Erkenntnisse legen nahe, dass auch trotz des technologischen Fortschritts die in Kapitel 2.6 beschriebenen qualitativen Vorbehalte gegenüber maschinellen Übersetzungssystemen nach wie vor vorhanden sind und Einfluss auf die Bewertung von Übersetzungen nehmen.

Anhand der Betrachtung der einzelnen Textausschnitte wird ersichtlich, dass sich die Wahrnehmungsverzerrung verstärkt, je stilistisch komplexer die Textausschnitte werden. Dies lässt sich dadurch erklären, dass höhere Erwartungen hinsichtlich Stil, Natürlichkeit und Flüssigkeit gestellt werden, wodurch wahrgenommene Abweichungen stärker ins Gewicht fallen.

Die Ergebnisse sind über beide Zielgruppen hinweg konsistent und deuten somit auf einen vom Expert:innenstatus unabhängigen Verzerrungseffekt hin. Diese Erkenntnis zeigt, dass der Expert:innenstatus nicht nur hinsichtlich der allgemeinen Wahrnehmung, sondern auch

für die Voreinstellung wenig von Bedeutung ist. Vor dem Hintergrund der zunehmenden Wahrnehmung von KI-generierten Texten als Bedrohung für textbasierte Dienstleistungsberufe (vgl. ELIS 2024; Li 2023; Moorkens 2018) wäre zu erwarten gewesen, dass diese Voreinstellung gegenüber maschinell generierten Texten in der Expert:innengruppe noch größer ausfallen würde als in der Lai:innengruppe. Wenngleich die Unterschiede zwischen den beiden Zielgruppen sehr gering sind, bewerteten Sprachexpert:innen über alle Kategorien hinweg maschinelle Übersetzungen besser, was der Erwartung klar widerspricht. Dies ist insbesondere auffällig, da die erhobenen Daten zeigen, dass Lai:innen mit 77 % maschinelle Übersetzungsoutputs häufiger als ausreichend für ihre Zwecke einschätzen. In der Gruppe der Sprachexpert:innen gaben hingegen lediglich 41 % an, dass maschinelle Übersetzungen allgemein ihren Anforderungen entsprechen. Ähnliche Ergebnisse beschreiben auch Kaspere et al. (2023). Die Daten zeigen zudem, dass die maschinelle Übersetzung für Lai:innen viel eher ein Mittel zur Kostenersparnis ist als für Sprachexpert:innen, was mit den Erkenntnissen aus dem Bericht der europäischen Sprachindustrie übereinstimmt (vgl. ELIS 2024). Beide Erkenntnisse legen nahe, dass Sprachexpert:innen die Qualität maschineller Übersetzungen kritischer sehen als Lai:innen.

Zuletzt ist der Zusammenhang zwischen der Bewertung und der korrekten Einschätzung von Texten zu diskutieren, unabhängig von deren tatsächlichem Ursprung. Die Ergebnisse zeigen, dass über alle Kategorien kaum Unterschiede zwischen den Bewertungen von korrekt oder falsch eingeschätzten Texten bestehen. Daraus lässt sich schließen, dass bei keiner der Bewertungskategorien ein systematischer Unterschied zwischen korrekt und falsch gegeben ist.

Auffällig ist, dass falsch zugeordnete Texte in fast allen Kategorien konsistent besser bewertet wurden als korrekt identifizierte Textausschnitte. Dies weist darauf hin, dass eine hohe subjektive Qualität nicht in einem systematischen Zusammenhang mit der Korrektheit der Einschätzung steht.

Die Ergebnisse zeigen auch, dass kreativere Textausschnitte tendenziell schlechter bewertet wurden als stilistisch weniger komplexe Textausschnitte, was sowohl bei korrekten als auch falschen Bewertungen gleichermaßen beobachtet werden konnte. Eine mögliche Erklärung hierfür ist die allgemeine Schwierigkeit, Texte mit einem hohen Maß an kreativen Stilmitteln adäquat zu übersetzen. Sowohl maschinelle Übersetzungssysteme als auch Humanübersetzer:innen stehen hier vor höheren Anforderungen an Stil, Ausdruck und Flüssigkeit, weshalb ein allgemeiner schlechterer Bewertungstrend darauf zurückzuführen sein könnte.

Während die Wahrnehmung eines Textes dessen Bewertung maßgeblich beeinflusst, zeigt sich kein entsprechender Zusammenhang zwischen der Bewertung eines Textes und der

Korrektheit der Einschätzung. Die Bewertung erlaubt somit keine verlässlichen Rückschlüsse darauf, ob ein Text korrekt oder falsch zugeordnet wurde.

Auch im Hinblick auf den Expert:innenstatus lassen sich keine relevanten Unterschiede feststellen. Es kann somit festgehalten werden, dass Expertise nicht nur für die korrekte Zuordnung und Wahrnehmung, sondern auch für die Bewertung im Kontext der Korrektheit keine zentrale Rolle zu spielen scheint.

Bei der Interpretation der Ergebnisse sind jedoch einige methodische und inhaltliche Einschränkungen zu berücksichtigen. Die zwei bei der vorliegenden Arbeit für die Übersetzungsgenerierung verwendeten Tools stellen zwar sehr bekannte und gängige Systeme dar, es lassen sich dennoch auf Grundlage der Ergebnisse keine Rückschlüsse auf alle maschinellen Übersetzungssysteme ziehen. Zudem beschränkt sich die Untersuchung auf das Sprachenpaar Englisch–Deutsch, wodurch die Übertragbarkeit auf andere Sprachkombinationen eingeschränkt ist. Zwar ließ die untersuchte Zielgruppengröße aussagekräftige Analysen hinsichtlich der Wahrnehmung der maschinellen Übersetzung im Kontext der Marketingkommunikation zu, dennoch sind nicht alle Altersgruppen und Konsument:innen der Texte in den Ergebnissen vertreten. Es sollte zudem beachtet werden, dass im Rahmen dieser Arbeit nur ein sehr spezifischer Teil der Marketingkommunikation untersucht wurde und daher keine allgemeingültigen Aussagen für das gesamte Textsortenspektrum getroffen werden können.

7 Fazit

Diese Arbeit untersuchte die Wahrnehmung maschineller Übersetzungen von Marketing-Texten in einem Turing-Test-ähnlichen Setting. Untersucht wurde, ob von DeepL und ChatGPT generierte Übersetzungen von Teilnehmenden erkannt werden und ob es diesbezüglich Unterschiede in der Wahrnehmung zwischen Sprachexpert:innen und Lai:innen gibt. Ziel dieser Arbeit war es, Erkenntnisse über die aktuell wahrgenommene Qualität maschineller Übersetzung von nicht-literarischen kreativen Textsorten zu gewinnen, um damit eine Grundlage für weiterführende Forschungsaktivitäten in den Bereichen maschinelle Übersetzung und Marketingkommunikation zu schaffen.

Um stichhaltige Ergebnisse zu sammeln, wurden jeweils gleich große Gruppen an Sprachexpert:innen in der Textarbeit und Lai:innen Textausschnitte von Marketing-Website-Texten vorgelegt, die eingeschätzt und bewertet werden mussten. Es wurde festgestellt, dass Humanübersetzungen weitaus häufiger als maschinelle Übersetzungen wahrgenommen wurden als umgekehrt, sich jedoch etwa ein Drittel aller Teilnehmenden von den verwendeten maschinellen Übersetzungstools täuschen ließ, was bei einem klassischen Turing-Test als Hinweis auf teilweise funktionale Ununterscheidbarkeit interpretiert werden könnte.

Das verwendete Übersetzungstool hatte insgesamt keinen maßgeblichen Einfluss auf die Ergebnisse. Es ließ sich in der Gesamtbetrachtung keine statistisch signifikante Tendenz erkennen, wenngleich ChatGPT bei Texten mit stärker ausgeprägten kreativen Stilmitteln höhere Täuschungsraten verzeichnete als DeepL. Überraschenderweise wurde zudem festgestellt, dass der Expert:innenstatus nicht zu einem relevanten Vorteil führt, was die Annahme eines Zusammenhangs der erhöhten sprachlichen Sensibilität mit der Wahrnehmung von maschinellen Übersetzungen widerlegt.

Über beide Zielgruppen hinweg ließ sich eine sehr stark ausgeprägte Voreinstellung hinsichtlich maschineller Übersetzung feststellen, wodurch als maschinell wahrgenommene Texte unabhängig von ihrem tatsächlichen Ursprung konsistent klar schlechter bewertet wurden als menschlich klassifizierte Textausschnitte. Auch hier wurde die Relevanz des Kreativitätsniveaus deutlich, da Texte mit erhöhter stilistischer Komplexität eine weitaus höhere Verzerrung aufwiesen als weniger komplexe Texte. Im Gegensatz dazu wurde auch die Bewertung von tatsächlich korrekt und falsch zugeordneten Textausschnitten auf mögliche Ursachen, wie etwa Flüssigkeit oder Satzbau, untersucht, wobei jedoch keine der Kategorien hervorstach.

Die weitaus höher liegende korrekte Erkennungsrate von maschinellen Übersetzungen sowie die durchgehend bessere Bewertung von Humanübersetzungen zeigt, dass KI-Übersetzungssysteme wie DeepL bzw. Large Language Models wie ChatGPT zum jetzigen Zeitpunkt menschliche Marketingkommunikation in Übersetzungen noch nicht ausreichend abbilden können. Dies bedeutet, dass rohe maschinell erstellte Übersetzungen der beiden untersuchten Tools nicht der Qualität der menschlichen Übersetzung entsprechen. Insbesondere im Marketingbereich ist diese Erkenntnis relevant, da die nachgewiesene Verzerrung bei der Wahrnehmung eines Textes als maschinelle Übersetzung in Kombination mit hohen Erkennungsraten Einfluss auf die Reputation und die Kund:innenbindung einer Marke nehmen kann.

Weitere Forschungsarbeit sollte sich insbesondere auf die Abdeckung der gesamten Zielkund:innenschaft und daher auch potenzielle Nutzer:innen unter 18 Jahren bzw. über 64 Jahren konzentrieren. Aus einer Marketingperspektive wäre zudem weitere Forschung relevant, wie sich die Wahrnehmung als maschinelle Übersetzung bzw. die daraus folgende schlechtere Bewertung dieser auf das Kaufverhalten oder die Markenwahrnehmung auswirkt. Darüber hinaus könnte die vorliegende Arbeit eine gute Grundlage für die Erweiterung der Untersuchungen auf aktive Push-Marketingkommunikation wie E-Mails oder Newsletter bieten.

8 Literaturverzeichnis

Agrawal, Sweta; Farinhas, António; Rei, Ricardo & Martins, André F. T. (2024). Can Automatic Metrics Assess High-Quality Translations? In: Al-Onaizan, Yaser; Bansal, Mohit & Chen, Yun-Nung (Hg.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami: Association for Computational Linguistics, 14491–14502.

Apter, Emily (2009). *The Translation Zone. A New Comparative Literature.* Princeton: Princeton University Press.

Aslan, Erdinç (2023). Machine Translation: Perception of Translation and Interpreting Students in Turkey. *Current Trends in Translation Teaching and Learning E* 10 (1), 185–216.

Asscher, Omri & Glikson, Ella (2023). Human evaluations of machine translation in an ethically charged situation. *New Media and Society* 25 (5*)*, 1087–1107.

Atomic (2026). *Maverick & Maven.* https://www.atomic.com/en-gb/collection-maverick-maven (Stand: 19.01.2026).

Automatic Language Processing Advisory Committee (ALPAC) (1966). *Language and Machines: Computers in Translation and Linguistics.* Washington, D.C.: National Academies Press.

Bahdanau, Dzmitry; Cho, Kyunghun & Bengio, Yoshua (2014). Neural machine translation by jointly learning to align and translate. *CoRR abs/*1409.0473.

Banerjee, Satanjeev & Lavie, Alon (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Goldstein, Jade; Lavie, Alon; Lin, Chin-Yew & Voss, Clare (Hg.) *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.* Ann Arbor: Association for Computational Linguistics, 65–72.

Bapna, Ankur; Caswell, Isaac; Kreutzer, Julia; Firat, Orhan; van Esch, Daan; Siddhant, Aditya; Niu, Mengmeng; Baljekar, Pallavi; Garcia, Xavier; Macherey, Wolfgang; Breiner, Theresa; Axelrod, Vera; Riesa, Jason; Cao, Yuan; Chen, Mia Xu; Macherey, Klaus; Krikun, Maxim; Wang, Pidong; Gutkin, Alexander; Shah, Apurva; Huang, Yanping; Chen, Zhifeng; Wu, Yonghui & Hughes, Macduff (2022). Building Machine Translation Systems for the Next Thousand Languages. *CoRR abs/*2205.03983.

Becker, Lee A. (1999). *Effect Size Calculators*. https://lbecker.uccs.edu/ (Stand: 01.04.2026).

Bender, Emily M. & Koller, Alexander (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In: Jurafsky, Dan; Chai, Joyce; Schluter, Natalie

& Tetreault, Joel (Hg.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.

Bennett, Winfield S. & Slocum, Jonathan (1985). The LRC Machine Translation System. *Computational Linguistics* 11 (2–3), 111–121.

Błaszkowska, Hanka (2021). Translation und Transkreation: Zu Differenzen der Begriffssystematik zwischen Theorie und Praxis. In: Adamczak-Krysztofowicz, Sylwia; Szczepaniak-Kozak, Anna & Rybszleger, Paweł (Hg.) *Angewandte Linguistik – Neue Herausforderungen und Konzepte*. Göttingen: V&R Unipress, 343–354.

Block, Ned (1981). Psychologism and behaviorism. *The Philosophical Review* 90 (1), 5–43.

Bommasani, Rishi; Hudson, Drew A.; Adeli, Ehsan; Altman, Russ; Arora, Simran; von Arx, Sydney; Bernstein, Michael S.; Bohg, Jeannette; Bosselut, Antoine; Brunskill, Emma; Brynjolfsson, Erik; Buch, Shyamal; Card, Dallas; Castellón, Rodrigo; Chatterji, Niladri; Chen, Annie; Creel, Kathleen; Davis, Jared Quincy; Demszky, Dora; Donahue, Chris; Doumbouya, Moussa; Durmus, Esin; Ermon, Stefano; Etchemendy, John; Ethayarajh, Kawin; Fei-Fei, Li; Fisch, Chelsea; Germine, Trevor; Gillespie, Lauren; Goel, Karan; Goodman, Noah; Grossman, Shelby; Guha, Neel; Hashimoto, Tatsunori; Henderson, Peter; Hewitt, John; Ho, Daniel E.; Hong, Jenny; Hsu, Kyle; Huang, Jing; Icard, Thomas; Jain, Saahil; Jurafsky, Dan; Kalluri, Pratyusha; Karamcheti, Siddharth; Keeling, Geoff; Khani, Feresheh; Khattab, Omar; Koh, Pang Wei; Krass, Mark; Krishna, Ranjay; Kudipitigi, Rohith; Kumar, Ananya; Ladhak, Faisal; Lee, Mina; Lee, Tony; Leskovec, Jure; Levent, Isabelle; Li, Xiang Lisa; Li, Xuechen; Liu, Tengyu; Malik, Ali; Manning, Christopher D.; Mirchandani, Suriv; Mitchell, Eric; Munyikwa, Zanele; Nair, Suraj; Narayan, Avanika; Narayanan, Deepak; Newman, Ben; Nie, Allen; Niebles, Juan Carlos; Nilforoshan, Hamed; Nyarko, Julian; Ogut, Giray; Orr, Laurel; Papadimitriou, Isabel; Park, Joon Sung; Piech, Chris; Portalene, Eva; Potts, Christopher; Raghunathan, Aditi; Reich, Rob; Ren, Hongyu; Roohani, Yusuf; Ruiz, Camilo; Ryan, Jack; Ré, Christopher; Sadigh, Dorsa; Sagawa, Shiori; Santhanam, Keshav; Shih, Andy; Srinivasan, Krishnan; Tamkin, Alex; Taori, Rohan; Thomas, Armin W.; Tramèr, Florian; Wang, Rose E.; Wang, William; Wang, Boham; Wu, Jiajun; Wu, Yuhuai; Xie, Sang Michael; Yasunaga, Michihiro; You, Jiaxuan; Zaharia, Matei; Zhang, Michael; Zhang, Tianyi; Zhang, Xikun; Zhang, Yuhui; Zheng, Lucia; Zhou, Kaitlyn & Liang, Percy (2021). On the Opportunities and Risks of Foundation Models. *CoRR abs/*2108.07258.

Bortz, Jürgen & Döring, Nicola (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler.* 4. Aufl. Heidelberg: Springer.

Bowker, Lynne (2021). Machine Translation Literacy Instruction for Non-Translators: A Comparison of Five Delivery Formats. In: Mitkov, Ruslan; Sosoni, Vilelmini; Giguère, Julie Christine; Murgolo, Elena & Deysel, Elizabeth (Hg.) *Translation and Interpreting Technology Online.* Shoumen: INCOMA, 25–36.

Brown, Peter F.; Cocke, John; Della Pietra, Stephen A.; Della Pietra, Vincent J.; Jelinek, Fredrick; Lafferty, John D.; Mercer, Robert L. & Roossin, Paul S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics* 16 (2), 79–85.

Brown, Peter F.; Della Pietra, Stephen A.; Della Pietra, Vincent J. & Mercer, Robert L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* 19 (2), 263–311.

Brown, Tom B.; Mann, Benjamin; Ryder, Nick, Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prafulla; Neelakantan, Arvind; Shyam, Pranav; Sastry, Girish; Askell, Amanda; Agarwal, Sandhini; Herbert-Voss, Ariel; Krueger, Gretchen; Henighan, Tom; Child, Rewon; Ramesh, Aditya; Ziegler, Daniel M.; Wu, Jeffrey; Winter, Clemens; Hesse, Christopher; Chen, mark; Sigler, Eric; Litqin, Mateusz; Gray, Scott; Chess, Benjamin; Clark, Jack; Berner, Christopher; McCandlish, Sam; Radford, Alec; Sutskever, Ilya & Amodei, Dario (2020). Language Models are Few-Shot Learners. *CoRR abs/*2005.14165.

Büttner, Gesa (2021). *Dolmetschvorbereitung digital*. Berlin: Frank & Timme.

Callison-Burch, Chris; Fordyce, Cameron; Koehn, Philipp; Monz, Christof & Schroeder, Josh (2007). (Meta-) Evaluation of Machine Translation. In: Callison-Burch, Chris; Koehn, Philipp; Fordyce, Cameron Shaw & Monz, Christof (Hg.) *Proceedings of the Second Workshop on Statistical Machine Translation*. Prag: Association for Computational Linguistics, 136–158.

Carroll, John B. (1966). An Experiment in Evaluating the Quality of Translations. *Mechanical Translation and Computational Linguistics* 9 (3, 4), 55–66.

Casper (2026). *The Dog Bed.* https://casper.com/products/dog-mattresses (Stand: 27.01.2026).

Catford, John C. (1965). *A Linguistic Theory of Translation. An Essay in Applied Linguistics.* London: Oxford University Press.

Chandioux, John (1976). METEO : un système opérationnel pour la traduction automatique des bulletins météorologiques destinés au grand public. *Meta* 21 (2), 127–133.

Chatterji, Aaron; Cunningham, Tom; Deming, David; Hitzig, Zoe; Ong, Christopher; Shan, Carl & Wadman, Kevin (2025). *How People Use ChatGPT*. https://cdn.openai.com/pdf/a253471f-8260-40c6-a2cc-aa93fe9f142e/economic-research-chatgpt-usage-paper.pdf (Stand: 19.05.2026).

Chatzikoumi, Eirini (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering* 26 (1), 137–161.

Chein, J.M.; Martinez, S.A. & Barone, A.R. (2024). Human intelligence can safeguard against artificial intelligence: individual differences in the discernment of human from AI texts. *Scientific Reports* 14 (1), 25989.

Chen, Pingqing (2024). The Impact of Generative AI on the Role of Translators and Its Implications for Translation Education. *Education Insights* 1 (2), 31–40.

Cho, Kyunghyun; van Merrienboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger & Bengio, Yoshua (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *CoRR abs/*1406.1078.

Ciubotaru, Bogdan-Iulian (2025). Generative Artificial Intelligence and Large Language Models: A Systematic Review of Architectures, Applications, and Future Directions. In: *25th International Conference on Control Systems and Computer Science (CSCS)*. New York: Institute of Electrical and Electronics Engineers, 380–388.

Cohen, Jacob (2013). *Statistical Power Analysis for the Behavioral Sciences*. Burlington: Elsevier Science.

Colby, Kenneth M.; Hilf, Franklin D.; Weber, Sylvia & Kraemer, Helena C. (1972). Turing-like Indistinguishability Tests for the Validation of a Computer Simulation of Paranoid Processes. *Artificial Intelligence* 3 (1), 199–221.

Crego, Josep; Kim, Jungi; Klein, Guillaume; Rebollo, Anabel; Yang, Kathy; Senellart, Jean; Akhanov, Egor; Brunelle, Patrice; Coquard, Aurélien; Deng, Yongchao; Enoue, Satoshi; Geiss, Chiyo; Johanson, Joshua; Khalsa, Ardas; Khiari, Raoum; Ko, Byeongil; Kobus, Catherine; Lorieux, Jean; Martins, Leidiana; Nguyen, Dang-Chuan; Priori, Alexandra; Riccardi, Thomas; Segal, Natalia; Servan, Christophe; Tiquet, Cyril; Wang, Bo; Yang, Jin; Zhang, Dakun; Zhou, Jing & Zoldan, Peter (2016). SYSTRAN's Pure Neural Machine Translation Systems. *CoRR abs*/1610.05540.

Daems, Joke; Macken, Lieve & Ruffo, Paola (2025). The Role of Translation Workflows in Overcoming Translation Difficulties: A Comparative Analysis of Human and Machine Translation (Post-Editing) Approaches. In: Vanroy, Bram; Lefer, Marie-Aude; Macken, Lieve; Ruffo, Paola; Guerberof-Arenas, Ana & Hansen, Damien (Hg.) *Proceedings of the Second Workshop on Creative-text Translation and Technology (CTT)*. Genf: European Association for Machine Translation, 1–13.

DeepL (2025). *DeepL Übersetzer Sprachen.* https://support.deepl.com/hc/de/articles/360019925219-DeepL-%C3%9Cbersetzer-Sprachen (Stand: 16.11.2025).

DeepL (2026). *DeepL Übersetzer*. https://www.deepl.com/de/translator (Stand: 02.02.2026).

Delorme Benites, Alice; Cotelli Kureth, Sara; Lehr, Caroline & Steele, Elizabeth (2021). Machine translation literacy: A panorama of practices at Swiss universities and implications for language teaching. In: Zoghlami, Naouel; Brudermann, Cédric; Sarré, Cedric; Grosbois, Muriel; Bradley, Linda & Thouësny, Sylvie (Hg.) *CALL and professionalisation: Short papers from EUROCALL 2021*. [o.O.]: Research-publishing.net, 80–87.

Devlin, Jacob; Chang, Ming-Wie; Lee, Kenton & Toutanova, Kristina (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR abs/* 1810.04805.

Doru, Berin; Maier, Christoph; Busse, Johanna Sophie; Lücke, Thomas; Schönhoff, Judith; Enax-Krumova, Elena; Hessler, Steffen; Berger, Maria & Tokic, Marianne (2025). Detecting Artificial Intelligence–Generated Versus Human-Written Medical Student Essays: Semirandomized Controlled Study. *JMIR Med Educ* 11 (1), e62779.

Doyon, Jennifer B.; White, John S. & Taylor, Kathryn B. (1999). Task-Based Evaluation for Machine Translation. In: *Proceedings of MT Summit VII*. Singapore: Kent Ridge Digital Labs, 574–578.

Dugast, Loïc; Sennelart, Jean & Koehn, Philipp (2007). Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. In: Callison-Burch, Chris; Koehn, Philipp; Fordyce, Cameron Shaw & Monz, Christof (Hg.) *Proceedings of the Second Workshop on Statistical Machine Translation*. Prag: Association for Computational Linguistics, 220–223.

Elhamayed, Sanaa Abou & Nour, Mohamed (2025). Overview of deep learning and large language models in machine translation: a special perspective on the Arabic language. *Journal of Electrical Systems and Information Technology* 12 (1), Artikel 27.

ELIS (2024). *European Language Industry Survey 2024. Trends, expectations and concerns of the European language industry.* https://elis-survey.org/wp-content/uploads/2024/03/ELIS-2024-Report.pdf (Stand: 11.11.2025).

ELIS (2025). *European Language Industry Survey 2025. Trends, expectations and concerns of the European language industry.* https://elis-survey.org/wp-content/uploads/2025/03/ELIS-2025_Report.pdf (Stand: 11.11.2025).

ELIS (2026). *European Language Industry Survey 2026. Trends, expectations and concerns of the European language industry.* https://elis-survey.org/wp-content/uploads/2026/03/ELIS-2026-Report.pdf (Stand: 01.04.2026).

Elnashar, Ashraf; Fu, Quchen; Gilbert, Henry; Hays, Sam; Olea, Carlos; Sandborn, Michael; Schmidt, Douglas C.; Spencer-Smith, Jesse & White, Jules (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *CoRR abs/*2302.11382.

Europäischer Referenzrahmen (o.J.). *Gemeinsamer Europäischer Referenzrahmen für Sprachen (GER)*. https://www.europaeischer-referenzrahmen.de/ (Stand: 12.12.2025).

Farley, Belmont & Clark, Wesley (1954). Simulation of self-organizing systems by digital computer. *Transactions of the I.R.E. Professional Group on Information Theory* 4 (4), 76–84.

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

French, Robert (2000). The Turing Test: The first 50 years. *Trends in Cognitive Sciences* 4 (3), 115–121.

Gaspari, Federico (2024). The History of Translation Technologies. In: Lange, Anne; Monticelli, Daniele & Rundle, Christopher (Hg.) *The Routledge Handbook of the History of Translation Studies.* London: Routledge, 324–338.

Gaspari, Federico & Hutchins, W. John (2007). Online and Free! Ten Years of Online Machine Translation: Origins, Developments, Current Use and Future Prospects. In: Maegaard, Bente (Hg.) *Proceedings of Machine Translation Summit XI: Papers*. Kopenhagen: MTSummit. https://aclanthology.org/2007.mtsummit-papers.27.pdf (Stand: 12.11.2025).

Gehrmann, Sebastian; Strobelt, Hendrik & Rush, Alexander (2019). GLTR: Statistical Detection and Visualization of Generated Text. In: Costa-jussà, Marta R.& Alfonseca, Enrique (Hg.) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florenz: Association for Computational Linguistics, 111–116.

Ghassemiazghandi, Mozhgan (2024). An Evaluation of ChatGPT's Translation Accuracy Using BLEU Score. *Theory and Practice in Language Studies* 14 (4), 985–994.

Goodfellow, Ian; Bengio, Yoshua & Courville, Aaron (2016). *Deep Learning*. Cambridge/London: The MIT Press.

GraphPad (o.J.). *T test calculator.* https://www.graphpad.com/quickcalcs/ttest1/?format=c (Stand: 01.04.2026).

Gritsch, Simone (2012). Meinungen abbilden. Die Likert-Skala. *Ergopraxis* 5 (1), 16–17.

Guerberof-Arenas, Ana & Toral, Antonio (2022). Creativity in translation. Machine translation as a constraint for literary texts. *Translation Spaces* 11 (2), 184–212.

Guerberof-Arenas, Ana & Toral, Antonio (2024). To be or not to be: A translation reception study of a literary text translated into Dutch and Catalan using machine translation. *Target International Journal of Translation Studies* 36 (2), 215–244.

Guerreiro, Nuno M.; Alves, Duarte M.; Waldendorf, Jonas; Haddow, Barry; Birch, Alexandra; Colombo, Pierre & Martins, André F. T. (2023). Hallucinations in Large Multilingual Translation Models. *CoRR abs/*16104.

Han, Lifeng (2018). *Machine Translation Evaluation Resources and Methods: A Survey.* https://doras.dcu.ie/24493/7/20201006_Machine%20Translation%20Evaluation%20Resources%20and%20Methods_%20A%20Survey.pdf (Stand: 16.11.2025).

Hassan, Hany; Aue, Anthony; Chen, Chang; Chowdhary, Vishal; Clark, Jonathan; Federmann, Christian; Huang, Xuedong; Junczys-Dowmunt, Marcin; Lewis, William; Li, Mu; Liu, Shujie; Liu, Tie-Yan; Luo, Renqian; Menezes, Arul; Qin, Tao; Seide, Frank; Tan, Xu; Tian, Fei; Wu, Lijun; Wu, Shuangzhi; Xia, Yingce; Zhang, Dongdong; Zhang, Zhirui & Zhou, Ming (2018). Achieving Human Parity on Automatic Chinese to English News Translation. *CoRR abs/*1803.05567.

Hauenschild, Christa (2004). Maschinelle Übersetzung: Die gegenwärtige Situation. In: Kittel, Harald (Hg.) *Übersetzung: ein internationales Handbuch zur Übersetzungsforschung. 1. Teilband*. Berlin: de Gruyter, 756–766.

Hiebl, Bettina & Gromann, Dagmar (2024). Comparative Quality Assessment of Human and Machine Translation with Best-Worst Scaling. In: Scarton, Carolina; Prescott, Charlotte; Bayliss, Chris; Oakley, Chris; Wright, Joanna; Wrigley, Stuart; Song, Xingyi; Gow-Smith, Edward; Bawden, Rachel; Sánchez-Cartagena, Víctor M; Cadwell, Patrick; Lapshinova-Koltunski, Ekaterina; Cabarrão, Vera; Chatzitheodorou, Konstantinos; Nurminen, Mary; Kanojia, Diptesh & Moniz, Helena (Hg.) *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1).* Sheffield: European Association for Machine Translation (EAMT), 507–536.

Hisense (2026). *Televisions.* https://www.hisense-usa.com/televisions (Stand: 19.01.2026).

House, Juliane (1977). A Model for Assessing Translation Quality. *Meta* 22 (2), 103–109.

House, Juliane (2015). *Translation Quality Assessment. Past and Present.* 2. Auflage. New York: Routledge.

Hutchins, W. John (1995). Machine Translation: A brief History. In: Koerner, E. F. K. & Asher, R. E. (Hg.) *Concise History of the Language Sciences. From the Sumerians to the Cognitivists.* Cambridge: Cambridge University Press, 431–445.

Hutchins, W. John (2001a). Machine translation over fifty years. *Histoire, Epistemologie, Langage* 22 (1), 7–31.

Hutchins, W. John (2001b). Machine translation and human translation: in competition or in complementation? *International Journal of Translation* 13 (1), 5–20.

Hutchins, W. John (2005). Towards a Definition of Example-based Machine Translation. In: *Workshop on example-based machine translation.* Phuket: MTSummit, 63–70.

Hutchins, W. John (2014). The history of machine translation in a nutshell. https://aclanthology.org/www.mt-archive.info/10/Hutchins-2014.pdf (Stand: 02.11.2025).

Hutchins W. John & Somers, Harold L. (1992). *An introduction to machine translation*. London: Academic Press.

Innocent (2026). *Super Smoothies.* https://www.innocentdrinks.co.uk/things-we-make/super-smoothies (Stand: 27.01.2026).

Ji, Ziwei; Lee, Nayeon; Frieske, Rita; Yu, Tiezheng; Su, Dan; Xu, Yan; Ishii, Etsuko; Bang, Yejin; Chen, Delong; Dai, Wenliang; Chan, Ho Shu; Madotto, Andrea & Fung, Pascale (2024). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys* 55 (12), Artikel 248.

Johnson, Lisa (2016). Tomatoes and Google connect a Canadian farmer to new Syrian neighbour. *CBC News*, 12.04.2016. https://www.cbc.ca/news/canada/british-columbia/tomatoes-google-kelowna-curtis-stone-1.3532564 (Stand: 16.11.2025).

Johnson, Melvin; Schuster, Mike; Le, Quoc V.; Krikun, Maxim; Wu, Yonghui; Chen, Zhifeng; Thorat, Nikhil; Viégas, Fernanda; Wattenberg, Martin; Corrado, Greg; Hughes, Macduff & Dean, Jeffrey (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics* 5 (1), 339–351.

Jones, Cameron R. & Bergen, Benjamin K. (2024a). Does GPT-4 pass the Turing test? *CoRR abs/*2310.20216.

Jones, Cameron R. & Bergen, Benjamin K. (2024b). People cannot distinguish GPT-4 from a human in a Turing test. *CoRR abs/*2405.08007.

Jones, Cameron R. & Bergen, Benjamin K. (2025). Large Language Models Pass the Turing Test. *CoRR abs/*2503.23674.

Kalchbrenner, Nal & Blunson, Phil (2013). Recurrent Continuous Translation Models. In: Yarowsky, David; Baldwin, Timothy; Korhonen, Anna; Livescu, Karen & Bethard, Steven (Hg.) *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle: Association for Computational Linguistics, 1700–1709.

Kamaluddin, Mohamad Ihsan; Rasyid, Moch. Wildan Khoerul; Abqoriyyah, Fourus & Saehu, Andang (2024). Accuracy Analysis of DeepL: Breakthroughs in Machine Translation Technology. *Journal of English Education Forum* 4 (2), 122–126.

Kasperė, Ramunė; Motiejūnienė, Jurgita; Patašienė, Irena; Patašius, Martynas & Horbačiauskienė, Jolita (2023). Is machine translation a dim technology for its users? An eye tracking study. *Frontiers in Psychology* 14 (1), Artikel 1076379.

Kenny, Dorothy (2018). Sustaining Disruption? On the Transition from Statistical to Neural Machine Translation. *Revista Tradumàtica* 16 (1), 59–70.

Kenny, Dorothy (2022). Human and machine translation. In: Kenny, Dorothy (Hg.) *Machine translation for everyone. Empowering users in the age of artificial intelligence.* Berlin: Language Science Press, 23–49.

Kim, Sejoon; Sung, Mingi; Lee, Jeonghwan; Lim, Hyunkuk & Gimenez Perez, Jorge Froilan (2024). Efficient Terminology Integration for LLM-based Translation in Specialized Domains. In: Haddow, Barry; Kocmi, Tom, Koehn, Philipp & Monz, Christof (Hg.) *Proceedings of the Ninth Conference on Machine Translation*. Miami: Association for Computational Linguistics, 636–642.

Kocmi, Tom; Federmann, Christian; Grundkiewicz, Roman; Junczys-Dowmunt, Marcin; Matsushita, Hitokazu & Menezes, Arul (2021). To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In: *Proceedings of the Sixth Conference on Machine Translation (WMT)*. [o.O.]: Association for Computational Linguistics, 478–494.

Koehn, Phillip (2010). *Statistical machine translation.* Cambridge: Cambridge University Press.

Koehn, Phillip (2020). *Neural machine translation.* Cambridge: Cambridge University Press.

Koller, Werner (1979). *Einführung in die Übersetzungswissenschaft*. Heidelberg: Quelle & Meyer.

Läubli, Sam; Sennrich, Rico & Volk, Martin (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. *CoRR abs/*1808.07048.

Lennon, Brian (2014). Machine Translation: A Tale of Two Cultures. In: Bermann, Sandra & Porter, Catherin (Hg.) *A companion to translation studies.* Somerset: John Wiley & Sons, Incorporated, 135–146.

Li, Fengqi (2023). A Survey of Translation Learners' Uses and Perceptions of Neural Machine Translation. *Theory and Practice in Language Studies* 13 (11)*,* 3039–3048.

Likert, Rensis (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology* 140 (1), 1–55.

Lommel, Arle; Uszkoreit, Hans & Burchardt, Aljoscha (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista Tradumàtica* 12 (1), 455–462.

Luo, Junliang; Li, Tianyu; Wu, Di; Jenkin, Michael; Liu, Steve & Dudek, Gregory (2024). Hallucination Detection and Hallucination Mitigation: An Investigation. *CoRR abs/*2401.08358.

Luo, Yingfeng; Zheng, Tong; Mu, Yonghu; Li, Bei; Zhang Qinghong; Gao, Yongqi; Xu, Ziqiang; Feng, Peinan; Liu, Xiaoqian; Xiao, Tong & Zhu, Jingbo (2025). Beyond Decoder-only: Large Language Models Can be Good Encoders for Machine Translation. In: Che, Wanxiang; Nabende, Joyce; Shutova, Ekaterina & Pilehvar, Mohammad Taher (Hg.) *Findings of the Association for Computational Linguistics: ACL 2025.* Vienna: Association for Computational Linguistics, 9399–9431.

Lush (2026a). *Crush*. https://www.lush.com/au/en/p/crush-gift?queryId=e0a92f3b69335200da534768b2b8abc5 (Stand: 19.01.2026).

Lush (2026b). *Valentine's Day Gift Ideas*. https://www.lush.com/au/en/c/valentines-day-gifts (Stand: 19.01.2026).

Lutz, Marlene; Sen, Indira; Ahnert, Georg; Rogers, Elisa & Strohmaier, Markus (2025). The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models. *CoRR abs/*2507.16076.

Macken, Lieve; van Hest, Ella; Tezcan, Arda; Lumingo, Michael; Maryns, Katrijn & De Wilde, July (2024). MaTIAS: Machine Translation to Inform Asylum Seekers. In: Scarton, Carolina; Prescott, Charlotte; Bayliss, Chris; Oakley, Chris; Wright, Joanna; Wrigley, Stuart; Song, Xingyi; Gow-Smith, Edward; Forcada, Mikel & Moniz, Helena (Hg.) *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*. Sheffield: European Association for Machine Translation (EAMT), 6–7.

Majji, Eswara Rao; Ampalam, Srinivasa Babu & Raj, Chingilipalli (2025). Evaluating Google Translate as an AI-Powered Translation Tool: Strengths, Limitations, and Implications for Multilingual Communication. *Journal of Information Systems Engineering and Management* 10 (1), 135–141.

Mauldin, Michael L. (1994). CHATTERBOTs, TINYMUDs, and the Turing Test Entering the Loebner Prize Competition. In: *AAAI-94 Proceedings.* Seattle: AAAI, 16–21.

Mazda (2026). *Mazda CX-80.* https://www.mazda.co.uk/cars/mazda-cx-80/ (Stand: 19.01.2026).

Melby, Alan K. (2018). *Human and machine translation quality. Definable? Achievable? Desirable?* https://www.ttt.org/wp-content/uploads/2019/01/2018-12-17-LACUS2012-melby-formatted-v9h.pdf (Stand: 29.11.2025).

Milička, Jiří; Marklová, Anna; Drobil, Ondřej & Pospíšilová, Eva (2025). Learning to detect AI texts and learning the limits. *PLoS One* 20 (10), e0333007.

Moorkens, Joss (2018). What to expect from Neural Machine Translation: a practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer* 12 (4), 375 –387.

Moorkens, Joss; Toral, Antonio; Castilho, Sheila & Way, Andy (2018). Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces* 7 (2), 240–262.

Neufeld, Eric & Finnestad, Sonje (2020). In Defense of the Turing Test. *AI & SOCIETY* 35 (2), 819–827.

Nida, Eugene Albert (1964). *Toward a science of translating: with special reference to principles and procedures involved in Bible translating.* Leiden: Brill.

Nießen, Sonja; Och, Franz Josef; Leusch, Gregor & Ney, Hermann (2000). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In: Gavrilidou, G.; Carayannis, G.; Markantonatou, S.; Piperidis, S. & Stainhauer, G. (Hg.) *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athen: European Language Resources Association (ELRA), 39–45.

Nord, Christiane (1997). *Translating as a purposeful activity: functional approaches explained.* New York: Routledge.

Oatly (2026). *Chilled Oatmilk Low Fat.* https://www.oatly.com/en-us/products/chilled-oatmilk/chilled-oatmilk-low-fat-64-oz (Stand: 19.01.2026).

OpenAI (2023). GPT-4 Technical Report. *CoRR abs/*2303.08774.

OpenAI (2026). *ChatGPT.* https://chatgpt.com/ (Stand: 02.02.2026).

Papineni, Kishore; Roukos, Salim; Ward, Todd & Zhu, Wei-Jing (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In: Pierre, Isabelle; Charniak, Eugene & Lin, Dekang (Hg.) *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia: Association for Computational Linguistics, 311–318.

Pawar, Saurav; Tonmoy, S. M. Towhidul Islam; Zaman, S. M. Mehedi; Jain, Vinija; Chadha, Aman & Das, Amitava (2024). The What, Why, and How of Context Length Extension Techniques in Large Language Models - A Detailed Survey. *CoRR abs/*2401.07872.

Perez, Sarah (2025). Meta rolls out AI-powered translations to creators globally, starting with English and Spanish. *TechCrunch*, 19.08.2025. https://techcrunch.com/2025/08/19/meta-rolls-out-ai-powered-translations-to-creators-globally-starting-with-english-and-spanish/ (Stand: 16.11.2025).

Popel, Martin; Tomkova, Marketa; Tomek, Jakub; Kaiser, Łukasz; Uszkoreit, Jakob & Bojar, Ondřej (2020). Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications* 11 (1), Artikel 4381.

Puma (2026a). *Lifestyle.* https://uk.puma.com/uk/en/collections/lifestyle (Stand: 19.01.2026).

Puma (2026b). *Select.* https://uk.puma.com/uk/en/collections/select (Stand: 19.01.2026).

Pyka, Sebastian & Furchheim, Pia (2017). *Experimentelle Marktforschung – Eine Einführung in die sozialwissenschaftliche Experimentalforschung.* (Chemnitz Economic Papers 14) (1). https://www.econstor.eu/bitstream/10419/170674/1/CEP014.pdf (Stand: 02.02.2026).

Qiu, Juerong (2025). Can Viewers Identify Raw Machine Translation in Subtitles, and What It Means for Reception? *Journal of Audiovisual Translation* 8 (1), 1–21.

Rathi, Ishika M.; Taylor, Sydney; Bergen, Benjamin & Jones, Cameron (2025). GPT-4 is Judged More Human than Humans in Displaced and Inverted Turing Tests. In: Alam, Firoj; Nakov, Preslav; Habash, Nizar; Gurevych, Iryna; Chowdhury, Shammur; Shelmanov, Artem; Wang, Yuxia; Artemova, Ekaterina; Kutlu, Mucahid; Mikros, George (Hg.) *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*. Abu Dhabi: International Conference on Computational Linguistics, 96–110.

Rei, Ricardo; Steward, Craig; Farinha, Ana C. & Lavie, Alon (2020). COMET: A Neural Framework for MT Evaluation. In: Webber, Bonnie; Cohn, Trevor; He, Yulan & Liu, Yang (Hg.) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. O. O.: Association for Computational Linguistics, 2685–2702.

Reiß, Katharina (1971). *Möglichkeiten und Grenzen der Übersetzungskritik: Kategorien und Kriterien für eine sachgerechte Beurteilung von Übersetzungen*. München: Hueber.

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

Rothkegel, Annely (2004). Geschichte der maschinellen und maschinenunterstützten Übersetzung. In: Kittel, Harald; Frank, Armin Paul; Greiner, Norbert; Hermans, Theo; Koller, Werner; Lambert, José & Paul, Fritz (Hg.) *Übersetzung - Translation - Traduction. 1. Teilband*. Berlin: De Gruyter, 748–756.

Sahoo, Pranab; Kumar Singh, Ayush; Saha, Sriparna; Jain, Viija; Mondal, Samrat & Chadha, Aman (2025). A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *CoRR abs/*2402.07927.

Saygin, Ayse Pinar; Cicekli, Ilyas & Akman, Varol (2000). Turing Test: 50 Years Later. *Minds and Machines* 10 (1), 463–518.

Schmalz, Antonia (2019). Maschinelle Übersetzung. In: Wittpahl, Volker (Hg.) *Künstliche Intelligenz*. Berlin: Springer, 194–211.

Schneider, Richard (2017). DeepL, der maschinelle Übersetzungsdienst der Macher von Linguee: Ein Quantensprung? *UEPO*, 30.08.2017. https://uepo.de/2017/08/30/deepl-der-maschinelle-uebersetzungsdienst-der-macher-von-linguee-ein-quantensprung/ (Stand: 02.11.2025).

Schwartz, Lane (2017). The history and promise of machine translation. In: Lacruz, Isabel & Jääskeläinen, Rita (Hg.) *Innovation and Expansion in Translation Process Research: Vol. XVIII.* Amsterdam/Philadelphia: John Benjamins, 161–190.

Searle, John (1980). Minds, brains, and programs. *The Behavioral and Brain Sciences* 3 (1), 417–457.

Sejnowski, Terrence J. (2023). Large Language Models and the Reverse Turing Test. *Neural Computation* 35 (1), 309–342.

Sennrich, Rico; Haddow, Barry & Birch, Alexandra (2015). Neural Machine Translation of Rare Words with Subword Units. *CoRR abs/*1508.07909.

Shah, Huma & Warwick, Kevin (2010). Testing Turing's Five Minutes, Parallel-Paired Imitation Game. *Kybernetes* 39 (3), 449–465.

Shieber, Stuart M. (1994). Lessons from a Restricted Turing Test. *Communications of the Association for Computing Machinery* 37 (6), 70–78.

Sizov, Fedor; España-Bonet, Cristina; van Genabith, Josef; Xie, Roy & Chowdhury, Koel Dutta (2024). Analysing Translation Artifacts: A Comparative Study of LLMs, NMTs, and Human Translations. In: Haddow, Barry; Kocmi, Tom; Koehn, Philipp & Monz, Christof (Hg.) *Proceedings of the Ninth Conference on Machine Translation.* Miami: Association for Computational Linguistics, 1183–1199.

Snover, Matthew; Dorr, Bonnie; Schwartz, Richard; Micciulla, Linnea & Makhoul, John (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Cambridge: Association for Machine Translation in the Americas, 223–231.

Snover, Matthew G.; Madani, Nitin; Dorr, Bonnie & Schwartz, Richard (2009). TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation* 23 (2/3), 117–127.

SoSci Survey (2026). *SoSci Survey – die Lösung für eine professionelle Onlinebefragung.* https://www.soscisurvey.de/ (Stand: 23.12.2025).

Sreelekha, S. (2017). Statistical Vs Rule Based Machine Translation; A Case Study on Indian Language Perspective. *CoRR abs/*1708.04559.

Stahlberg, Felix (2020). Neural Machine Translation: A Review. *Journal of Artificial Intelligence Research* 69 (1), 343–418.

Stiegelbauer, Laura-Rebeca; Pantea, Dumitra Madalina & Balas, Raluca (2024). Challenges of translating advertising texts. *Studii de ştiinţă şi cultură* 20 (4), 124–136.

Sutskever, Ilya; Vinyals, Oriol & Le, Quoc V. (2014). Sequence to Sequence Learning with Neural Networks. *CoRR abs/*1409.3215.

Systran (o.J.a). *50 years of MT innovation*. https://www.systransoft.com/systran/translation-technology/systran-50-years-of-mt-innovation/?utm_source=chatgpt.com (Stand: 02.11.2025).

Systran (o.J.b). *Systran.* https://www.systransoft.com/systran/ (Stand: 02.11.2025).

Systran (o.J.c). *Defense & Security*. https://www.systransoft.com/industries/defense-security/ (Stand: 08.12.2025).

Tillmann, Christoph; Vogel, Stephan; Ney, Hermann; Zubiaga, Alex & Sawaf, Hassan (1997). Accelerated DP-based Search for Statistical Translation. In: *Proceedings of Eurospeech 1997.* Rhodos: European Speech Communication Association, 2667–2670.

Toma, Peter (1976). SYSTRAN. In: Hays, David G. & Mathias, J. (Hg.) *Foreign Broadcast Information Service Seminar on Machine Translation*. Rosslyn: Association for Computational Linguistics, 40–45.

Toral, Antonio & Way, Andy (2018). What Level of Quality can Neural Machine Translation Attain on Literary Text? *CoRR abs/*1801.04962.

Torresi, Ira (2021). *Translating promotional and advertising texts*. 2. Aufl. London: Routledge.

Turing, Alan (1950). Computing Machinery and Intelligence. *Mind* 49 (1), 433–460.

Universität Wien (o.J.). *Transkulturelle Kommunikation (Bachelorstudium).* https://auf-nahmeverfahren.univie.ac.at/transkulturelle-kommunikation (Stand: 22.01.2026).

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukasz & Polosukhin, Ilia (2017). Attention is all you need. In: von Luxburg, Ulrike; Guyon, Isabelle; Bengio, Samy; Wallach, Hanna & Fergus, Rob (Hg.) *Advances in Neural Information Processing Systems 30. 31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach: Curran Associates, 5999–6009.

Wakefield, Jane (2019). The hobbyists competing to make AI human. *BBC,* 14.09.2019. https://www.bbc.com/news/technology-49578503 (Stand: 29.11.2025).

Wang, Haifeng; Wu, Hua; He, Zhongjun; Huang, Liang & Church, Kenneth Ward (2021). Progress in Machine Translation. *Engineering* 18 (1), 143–153.

Wang, Xindi; Salmani, Mahsa; Omidi, Parsa; Ren, Xiangyu; Rezagholizadeh, Mehdi & Eshaghi, Armaghan (2024). Beyond the Limits: A Survey of Techniques to Extend the Context Length in Large Language Models. In: Larson, Kate (Hg.) *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. Jeju: International Joint Conferences on Artificial Intelligence, 8299–8307.

Wang, Yuxia; Wang, Minghan; Manzoor, Muhammad, Arslan; Liu, Fei; Georgiev, Georgi; Das, Rocktim Jyoti & Nakov, Preslav (2024). Factuality of Large Language Models: A Survey. *CoRR abs/*2402.02420.

Warwick, Kevin & Shah, Huma (2016). *Turing's Imitation Game: Conversations with the Unknown.* Cambridge: Cambridge University Press.

Weaver, Warren (1949). *Translation*. https://www.mt-archive.net/50/Weaver-1949.pdf (Stand: 02.11.2025).

Weizenbaum, Joseph (1966). ELIZA – a Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM* 9 (1), 36–45.

White, John S.; O'Connell, Theresa & O'Mara, Francis (1994). The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In: *Proceedings of the First Conference of the Association for Machine Translation in the Americas.* Columbia: AMTA, 193–205.

Wu, Yonghui; Schuster, Mike; Chen, Zhifeng; Le, Quoc V.; Norouzi, Mohammad; Macherey, Wolfgang; Krikun, Maxim; Cao, Yuan; Gao, Qin; Macherey, Klaus; Klingner, Jeff; Shah, Apurva; Johnson, Melvin; Liu, Xiaobing; Kaiser, Łukasz; Gouws, Stephan; Kato, Yoshikiyo; Kudo, Taku; Kazawa, Hideto; Stevens, Keith; Kurian, George; Patil, Nishant; Wang, Wei; Young, Cliff; Smith, Jason; Riesa, Jason; Rudnick, Alex; Vinyals, Oriol;

Corrado, Greg; Hughes, Macduff & Dean, Jeffrey (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR abs/*1609.08144.

Xiao, Yimin; Zhang, Yongle; Ki, Dayeon; Bao, Calvin; Martindale, Marianna J.; Vaughn, Charlotte; Gao, Ge & Carpuat, Marine (2025). Toward Machine Translation Literacy: How Lay Users Perceive and Rely on Imperfect Translations. In: Christodoulopoulos, Christos; Chakraborty, Tanmoy; Rose, Carolyn & Peng, Violet (Hg.) *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou: Association for Computational Linguistics, 33997–34014.

Yamada, Masaru (2023). Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability. *CoRR abs/*2308.01391.

Zhang, Lechen; Ergen, Tolga; Logeswaran, Lajanugen; Lee, Moontae & Jurgens, David (2026). SPRIG: Improving Large Language Model Performance By System Prompt Optimization. *CoRR abs/*2410.14826.

Zhang, Tianyi; Kishore, Varsha; Wu, Felix; Weinberger, Kilian Q. & Artzi, Yoav (2020). BERTScore: Evaluating Text Generation with BERT. *CoRR abs/*1904.09675.

Zhao, Wayne Xin; Zhou, Kun; Li, Junyi; Tang, Tianyi; Wang, Xiaolei; Hou, Yupeng; Min, Yingqian; Zhang, Beichen; Zhang, Junjie; Dong, Zican; Du, Yifan; Yang, Chen; Chen, Yushuo; Chen, Zhipeng; Jiang, Jinhao; Ren, Ruiyang; Li, Yifan; Tang, Xinyu; Liu, Zikang; Liu, Peiyu; Nie, Jian-Yun & Wen, Ji-Rong (2026). A Survey of Large Language Models. *Frontiers of Computer Science* 12 (20), Artikel 2012627.

Zuo, Fei; Chen, Kehai; Zhang, Yu; Xue, Zhengshan & Zhang, Min (2025). InImageTrans: Multimodal LLM-based Text Image Machine Translation. In: Che, Wanxiang; Nabende, Joyce; Shutova, Ekaterina; Pilehvar & Mohammad Taher (Hg.) *Findings of the Association for Computational Linguistics: ACL 2025.* Vienna: Association for Computational Linguistics, 20256–20277.

9 Anhang

A1 Textausschnitte

Textausschnitt 1

Kürzel	UP01
Quelle	https://www.mazda.co.uk/cars/mazda-cx-80/ (Stand: 19.01.2026)
Ausgangstext	A breathtaking example of Japanese craftsmanship, the all-new Mazda CX-80 is an elegant three-row SUV offering seating for up to seven passengers. Spacious and versatile, every element is carefully crafted for everything your life has to offer. In Japan, empty space is seen as a canvas. A space you can fill with family and adventure. From city streets to country roads, the all-new Mazda CX-80 comes equipped with a range of drive modes and features that deliver a superior driving experience – no matter the weather. With a choice of two powerful engines that offer you the perfect balance of performance and efficiency, you'll love the highly responsive driving dynamics every time you hit the open road.
DeepL	Der brandneue Mazda CX-80 ist ein atemberaubendes Beispiel japanischer Handwerkskunst und ein eleganter SUV mit drei Sitzreihen, der Platz für bis zu sieben Passagiere bietet. Er ist geräumig und vielseitig, und jedes Element wurde sorgfältig für alle Aspekte Ihres Lebens entwickelt. In Japan wird leerer Raum als Leinwand betrachtet. Ein Raum, den Sie mit Familie und Abenteuern füllen können. Von Stadtstraßen bis hin zu Landstraßen – der brandneue Mazda CX-80 ist mit einer Reihe von Fahrmodi und Funktionen ausgestattet, die ein überragendes Fahrerlebnis bieten – unabhängig vom Wetter. Mit einer Auswahl von zwei leistungsstarken Motoren, die Ihnen die perfekte Balance zwischen Leistung und Effizienz bieten, werden Sie die reaktionsschnelle Fahrdynamik jedes Mal lieben, wenn Sie auf die Straße fahren.

ChatGPT	Ein atemberaubendes Beispiel japanischer Handwerkskunst: Der brandneue Mazda CX-80 ist ein eleganter SUV mit drei Sitzreihen und Platz für bis zu sieben Passagiere. Großzügig und vielseitig – jedes Detail ist sorgfältig gestaltet für alles, was das Leben bereithält. In Japan gilt leerer Raum als Leinwand. Ein Raum, den Sie mit Familie und Abenteuern füllen können. Von urbanen Straßen bis zu ländlichen Wegen ist der brandneue Mazda CX-80 mit einer Auswahl an Fahrmodi und Funktionen ausgestattet, die ein herausragendes Fahrerlebnis bieten – bei jedem Wetter. Mit der Wahl zwischen zwei leistungsstarken Motoren, die die perfekte Balance aus Performance und Effizienz bieten, werden Sie die hochpräzise, reaktionsschnelle Fahrdynamik bei jeder Fahrt auf offener Straße lieben.

Human	Ein Meisterwerk japanischer Handwerkskunst: Der neue Mazda CX-80 ist ein eleganter SUV mit drei Sitzreihen, geschaffen für bis zu sieben Passagiere. Großzügig, vielseitig und bis ins kleinste Detail durchdacht – jedes Element wurde mit höchster Präzision gestaltet und passt sich nahtlos Ihrem Lebensstil an. In Japan gilt freier Raum als ein unbeschriebenes Blatt, das erst durch Ihre Erlebnisse zum Leben erwacht. Der neue Mazda CX-80 bietet Ihnen genau diesen Raum für neue Möglichkeiten – mit ausgiebig Platz für Ihre Familie, Ihre Abenteuer und Entdeckungen. Ob pulsierende Stadt oder einsame Landstraßen: Dank intelligenter Fahrfunktionen und durchdachter Technik genießen Sie jederzeit ein dynamisches und harmonisches Fahrerlebnis – unabhängig vom Wetter. Zwei leistungsstarke Antriebsoptionen bieten Ihnen die ideale Balance aus Performance und Effizienz. Ganz gleich, wohin Ihr Weg Sie führt: Der Mazda CX-80 lässt Sie Fahrfreude in ihrer reinsten Form erleben – präzise, direkt und inspiriert von echter Kunstfertigkeit.

Textausschnitt 2

Kürzel	UP02
Quelle	https://www.atomic.com/en-gb/collection-maverick-maven (Stand: 19.01.2026)

Ausgangstext	Introducing MAVERICK and MAVEN, two new all-mountain skis inspired by the pioneers, peaks, and the endless variety of North American powder, windblown snow and all-mountain conditions. Built with simple versatile, state of the art all-mountain technology, these are the go-to skis for every intrepid skier who wants to keep exploring new lines in a new world. Tested and validated by everyday skiers, Atomic traversed North America to design and build this ski, hitting the hill on good days, deep days, and zero visibility days with some of the biggest personalities in town, sharing both secret and famous lines, and finding new ways to describe this sport we all love so much. And like always, ending each day with a beer, sharing jokes in the warm, welcoming glow of the local ski shops.
DeepL	Wir stellen vor: MAVERICK und MAVEN, zwei neue All-Mountain-Ski, inspiriert von Pionieren, Gipfeln und der unendlichen Vielfalt des nordamerikanischen Pulverschnees, des vom Wind verwehten Schnees und der All-Mountain-Bedingungen. Ausgestattet mit einfacher, vielseitiger und modernster All-Mountain-Technologie sind dies die idealen Ski für jeden unerschrockenen Skifahrer, der weiterhin neue Spuren in einer neuen Welt erkunden möchte. Atomic hat diesen Ski von alltäglichen Skifahrern testen und validieren lassen und ist dafür durch Nordamerika gereist, um ihn zu entwerfen und zu bauen. Dabei war Atomic an guten Tagen, an Tagen mit tiefem Schnee und an Tagen mit null Sicht mit einigen der größten Persönlichkeiten der Stadt auf der Piste unterwegs, hat geheime und berühmte Routen geteilt und neue Wege gefunden, diesen Sport zu beschreiben, den wir alle so sehr lieben. Und wie immer endete jeder Tag mit einem Bier und Witzen in der warmen, einladenden Atmosphäre der örtlichen Skigeschäfte.

ChatGPT	Wir präsentieren MAVERICK und MAVEN – zwei neue All-Mountain-Ski, inspiriert von Pionieren, Gipfeln und der endlosen Vielfalt nordamerikanischen Pulverschnees, verwehten Schnees und sämtlicher All-Mountain-Bedingungen. Ausgestattet mit einfacher, vielseitiger und hochmoderner All-Mountain-Technologie sind sie die erste Wahl für alle abenteuerlustigen Skifahrer, die immer weiter neue Lines in einer neuen Welt entdecken wollen. Getestet und bestätigt von Skifahrern aus dem Alltag ist Atomic quer durch Nordamerika gereist, um diesen Ski zu entwickeln und zu bauen – auf der Piste an guten Tagen, an Tiefschneetagen und bei null Sicht, gemeinsam mit einigen der prägendsten Persönlichkeiten der Szene. Dabei wurden geheime wie legendäre Lines geteilt, neue Wege gefunden, diesen Sport zu beschreiben, den wir alle so sehr lieben. Und wie immer endete jeder Tag mit einem Bier, mit gemeinsamen Witzen im warmen, einladenden Licht der lokalen Skishops.
Human	Mit MAVERICK und MAVEN präsentiert Atomic zwei neue All-Mountain-Ski – inspiriert von Pioniergeist, markanten Gipfeln und der endlosen Vielfalt nordamerikanischer Schneelandschaften. Reduziert auf das Wesentliche, vielseitig im Charakter und ausgestattet mit modernster All-Mountain-Technologie, sind sie die verlässliche Wahl für Skier, die den Berg immer wieder neu erleben wollen. Zusammen mit echten Skifahrer:innen hat Atomic diese Modelle auf den Hängen Nordamerikas getestet und perfektioniert. Egal ob bei schlechter Sicht, im Tiefschnee oder an Sonnentagen: Einige der größten Ski-Legenden haben ihre Geheimtipps verraten, die bekanntesten Pisten neu definiert und ihre Leidenschaft für den Sport geteilt. Und wie es die Tradition verlangt, endeten diese Tage meist mit Lachen, einem kühlen Bier und vertrauter Atmosphäre – irgendwo zwischen Berg, Freunden und Geschichten.

Textausschnitt 3

Kürzel	UP03
Quelle	https://www.hisense-usa.com/televisions (Stand: 19.01.2026)
Ausgangstext	Find the perfect Hisense TV just for you. You deserve high-end performance, quality, and immersive features — and Hisense brings all of this to the table with a versatile lineup of TVs that are hard not to fall in love with. Whether you're into movies, video games, or just catching the big game, Hisense has a TV that's right for you. As one of the top TV brands worldwide—and the #1 brand in 100" TVs—Hisense is redefining big-screen entertainment across the globe. From cutting-edge Mini-LED and QLED displays to next-generation picture processing, Hisense TVs deliver premium viewing experiences at every size and price point.
DeepL	Finden Sie den perfekten Hisense-Fernseher für sich. Sie verdienen High-End-Leistung, Qualität und beeindruckende Funktionen – und Hisense bietet Ihnen all das mit einer vielseitigen Auswahl an Fernsehern, die Sie begeistern werden. Egal, ob Sie Filme oder Videospiele mögen oder einfach nur das große Spiel verfolgen möchten, Hisense hat den richtigen Fernseher für Sie. Als eine der weltweit führenden TV-Marken – und die Nummer 1 bei 100-Zoll-Fernsehern – definiert Hisense das Unterhaltungserlebnis auf Großbildschirmen weltweit neu. Von hochmodernen Mini-LED- und QLED-Displays bis hin zur Bildverarbeitung der nächsten Generation bieten Hisense-Fernseher ein erstklassiges Seherlebnis in jeder Größe und Preisklasse.

ChatGPT	Finden Sie den perfekten Hisense TV für sich. Sie verdienen erstklassige Performance, höchste Qualität und mitreißende Features – und genau das liefert Hisense mit einem vielseitigen TV-Portfolio, in das man sich kaum nicht verlieben kann. Ob Filme, Videospiele oder das große Sportereignis: Hisense hat den passenden Fernseher für Sie. Als eine der weltweit führenden TV-Marken – und die Nummer 1 bei 100-Zoll-Fernsehern – definiert Hisense großformatige Unterhaltung rund um den Globus neu. Von hochmoderner Mini-LED- und QLED-Technologie bis hin zur nächsten Generation der Bildverarbeitung bieten Hisense TVs erstklassige Seherlebnisse in jeder Größe und Preisklasse.
Human	Finde den Hisense-TV, der perfekt zu dir passt. Starke Performance, cleveres Design und innovative Technologien – Hisense bringt Qualität zum Verlieben direkt in dein Wohnzimmer. Egal ob Blockbuster-Marathon, intensive Gaming-Sessions oder Sport-Highlights zum Mitfiebern: Wir haben den richtigen Fernseher für dich. Als globaler Vorreiter und Nummer 1 bei 100-Zoll-Displays krempeln wir die Welt des Großbild-Entertainment völlig um. Von wegweisender Mini-LED- und QLED-Technologie bis zum Bildprozessor der nächsten Generation: Erlebe Premium-Qualität, die in jeden Alltag und zu jedem Budget passt.

Textausschnitt 4

Kürzel	UP04
Quellen	https://uk.puma.com/uk/en/collections/lifestyle (Stand: 19.01.2026) https://uk.puma.com/uk/en/collections/select (Stand: 19.01.2026)

Ausgangstext	**PALERMO** Out of the archive, onto the streets. The ’80s classic Palermo is here with bold, vibrant colors and a football-inspired design. **EASY RIDER** The street is your track. The city is your field. Easy Rider, a retro running shoe turned streetwear staple. **SUEDE XL** Durability, a sturdy midsole, and padded upper make the Suede XL beloved by skaters on the board and off. **SELECT** PUMA SELECT exists in the intersection of sport, style, and creativity. Through bold collaboration, it embraces the constantly evolving worlds of streetwear and fashion, elevating the brand and pushing the culture forward.

DeepL	**PALERMO** Aus dem Archiv auf die Straße. Der Klassiker Palermo aus den 80ern ist zurück – mit kräftigen, lebendigen Farben und einem vom Fußball inspirierten Design. **EASY RIDER** Die Straße ist deine Rennstrecke. Die Stadt ist dein Spielfeld. Easy Rider, ein Retro-Laufschuh, der zum Streetwear-Klassiker wurde. **SUEDE XL** Langlebigkeit, eine robuste Zwischensohle und ein gepolstertes Obermaterial machen den Suede XL bei Skatern auf dem Board und außerhalb beliebt. **SELECT** PUMA SELECT befindet sich an der Schnittstelle zwischen Sport, Stil und Kreativität. Durch mutige Kooperationen greift es die sich ständig weiterentwickelnde Welt der Streetwear und Mode auf, hebt die Marke hervor und treibt die Kultur voran.

ChatGPT	**PALERMO** Aus dem Archiv direkt auf die Straße. Der 80er-Jahre-Klassiker Palermo ist zurück – mit kräftigen, lebendigen Farben und einem fußballinspirierten Design. **EASY RIDER** Die Straße ist deine Rennstrecke. Die Stadt dein Spielfeld. Easy Rider – ein Retro-Running-Schuh, der zum festen Bestandteil der Streetwear geworden ist. **SUEDE XL** Robustheit, eine stabile Mittelsohle und ein gepolsterter Schaft machen den Suede XL zum Liebling der Skater – auf dem Board und daneben. **SELECT** PUMA SELECT bewegt sich an der Schnittstelle von Sport, Style und Kreativität. Mit starken Kollaborationen greift es die sich ständig wandelnden Welten von Streetwear und Fashion auf, hebt die Marke auf ein neues Level und treibt die Kultur voran.

Human	**PALERMO** Legende trifft auf Straßenkultur: Der Palermo bringt den Spirit der 80er Jahre zurück – laut, farbenfroh und inspiriert vom Fußball. **EASY RIDER** Mach die Stadt zu deiner Laufstrecke: Easy Rider verbindet Retro-Lauf-Vibes mit urbanem, modernem Lifestyle. **SUEDE XL** Mit robustem Design, stabiler Mittelsohle und gepolstertem Obermaterial ist der Suede XL der perfekte Begleiter für Skater – auf dem Board und abseits davon. **SELECT** PUMA SELECT steht für den kreativen Raum, der Sport und Lifestyle verbindet. Die Linie lebt von Kollaborationen und prägt die Entwicklung und den Dialog zwischen Streetwear, Fashion und Kultur.

Textausschnitt 5

Kürzel	UP05
Quellen	https://www.lush.com/au/en/c/valentines-day-gifts (Stand: 19.01.2026) https://www.lush.com/au/en/p/crush-gift?queryId=e0a92f3b69335200da534768b2b8abc5 (Stand: 19.01.2026)

Ausgangstext	Love is all you need, but there's oh so many ways to do it – our Valentine's Day gift range celebrates romance in all its forms. We have fragrant melts for mates, dusting powders for hot dates and if you don't believe in all that fate, sweet, solo baths await… We can't keep it a secret any longer … We have a crush on this gift! But with how this trio of shower and body care delights leaves skin feeling gorgeously hydrated and seductively scented, we think the feeling is mutual. Can't work up the courage to confess your love? Let this gift do the talking — if the name isn't a big enough hint, the raspberry-rose splendour inside will win them over.
DeepL	Liebe ist alles, was man braucht, aber es gibt so viele Möglichkeiten, sie zu zeigen – unsere Valentinstagsgeschenke zelebrieren Romantik in all ihren Formen. Wir haben duftende Melts für Freunde, Puder für heiße Dates und wenn Sie nicht an das Schicksal glauben, erwarten Sie süße Solo-Bäder... Wir können es nicht länger geheim halten... Wir sind verliebt in dieses Geschenk! Aber da dieses Trio aus Dusch- und Körperpflegeprodukten die Haut herrlich mit Feuchtigkeit versorgt und verführerisch duften lässt, glauben wir, dass das Gefühl auf Gegenseitigkeit beruht. Sie trauen sich nicht, Ihre Liebe zu gestehen? Lassen Sie dieses Geschenk für Sie sprechen – wenn der Name nicht schon Hinweis genug ist, wird die Himbeer-Rosen-Pracht darin Ihr Gegenüber überzeugen.

ChatGPT	Liebe ist alles, was man braucht – doch es gibt unzählige Wege, sie zu feiern. Unsere Valentinstags-Geschenkauswahl zelebriert Romantik in all ihren Facetten. Duftende Melts für Freunde, Puder für heiße Dates – und für alle, die nicht an das Schicksal glauben, warten süße Solo-Bäder ... Wir können es nicht länger geheim halten ... Wir haben uns in dieses Geschenk verliebt! Und so, wie dieses Trio aus Dusch- und Körperpflege die Haut herrlich mit Feuchtigkeit versorgt und verführerisch duften lässt, ist das Gefühl ganz sicher gegenseitig. Du traust dich nicht, deine Liebe zu gestehen? Lass dieses Geschenk für dich sprechen – und falls der Name noch nicht eindeutig genug ist, wird die himmlische Kombination aus Himbeere und Rose garantiert Herzen erobern.
Human	Liebe ist alles, alles was wir brauchen – und unsere Valentinstag- Kollektion feiert die vielen Formen des Verliebtseins: vom zärtlichen Crush bis zum selbstbewussten „Ich gönn mir was“. Schmelzende Düfte für gute Freunde, seidige Puder für aufregende Dates und für alle, die lieber alleine lieben – ein Bad, so süß wie Selbstliebe selbst... Wir können es nicht länger verheimlichen: Wir sind schockverliebt in dieses Geschenk! Und so wie dieses Trio aus Dusch- und Körperpflege die Haut rundum verwöhnt und verführerisch duften lässt, beruht das Gefühl sicher auf Gegenseitigkeit. Du traust dich nicht, deine Liebe zu gestehen? Lass dieses Geschenk für dich sprechen – wenn der Name nicht schon Wink genug ist, wird der Mix aus Himbeere und Rose jedes Herz im Sturm erobern.

Textausschnitt 6

Kürzel	UP06
Quelle	https://www.oatly.com/en-us/products/chilled-oatmilk/chilled-oatmilk-low-fat-64-oz (Stand: 19.01.2026)

Ausgangstext	Some people just prefer less fat, no matter how good the fat in question happens to be. And if you're one of those nice people, how cool and weird is it that you're reading this paragraph on this website at this moment about this amazing Low-Fat Oatmilk with all the flavor of beautiful gluten-free oats and only one tiny gram of unsaturated fat per serving? It also has heart-healthy beta-glucans* (big, scientific word for soluble fiber from oats) and no dairy, nuts, soy, gluten or GMOs, but right now we're talking about less fat. Or karma. Well, whatever we're talking about, it's the good kind.
DeepL	Manche Menschen bevorzugen einfach weniger Fett, egal wie gut das betreffende Fett auch sein mag. Und wenn Sie zu diesen netten Menschen gehören, wie cool und seltsam ist es dann, dass Sie gerade diesen Absatz auf dieser Website über diese erstaunliche fettarme Hafermilch lesen, die den ganzen Geschmack von wunderschönem glutenfreiem Hafer hat und nur ein winziges Gramm ungesättigtes Fett pro Portion enthält? Sie enthält außerdem herzgesunde Beta-Glucane* (ein großes wissenschaftliches Wort für lösliche Ballaststoffe aus Hafer) und keine Milchprodukte, Nüsse, Soja, Gluten oder GVO, aber im Moment sprechen wir über weniger Fett. Oder Karma. Nun, worüber auch immer wir sprechen, es ist etwas Gutes.
ChatGPT	Manche Menschen bevorzugen einfach weniger Fett – ganz egal, wie gut das jeweilige Fett auch sein mag. Und wenn du zu diesen netten Menschen gehörst: Wie cool und irgendwie schräg ist es eigentlich, dass du genau jetzt auf dieser Website diesen Absatz über diese großartige fettarme Hafermilch liest – mit dem vollen Geschmack feiner, glutenfreier Haferflocken und nur einem winzigen Gramm ungesättigtem Fett pro Portion? Außerdem enthält sie herzgesunde Beta-Glucane* (ein großes, wissenschaftliches Wort für lösliche Ballaststoffe aus Hafer) und kommt ganz ohne Milchprodukte, Nüsse, Soja, Gluten oder Gentechnik aus. Aber im Moment reden wir ja über weniger Fett. Oder Karma. Wie auch immer – es ist auf jeden Fall die gute Sorte.

Human	Es gibt Menschen, die wollen einfach weniger Fett. Punkt. Völlig egal, wie gesund oder unfassbar gut das betreffende Fett eigentlich ist. Wenn du so jemand bist – was für ein witziger Zufall ist es bitte, dass du genau jetzt, auf dieser Website, diesen Text über unsere großartige Low-Fat Hafermilch liest? Sie schmeckt wie die feinsten glutenfreien Haferflocken (nur flüssiger), enthält genau ein winziges Gramm ungesättigtes Fett pro Portion und trotzdem richtig viel Geschmack. Dass da noch Beta-Glucane (ein kompliziertes Wort für lösliche Ballaststoffe – gutes Wort, um damit anzugeben) drin sind und wir auf Milch, Nüsse, Soja, Gluten und Gentechnik verzichten, ist toll, aber gerade geht's hier nur um weniger Fett. Oder ums Karma. Egal, worüber wir gerade reden – es fühlt sich verdammt gut an, oder?

Textausschnitt 7

Kürzel	UP07
Quelle	https://www.innocentdrinks.co.uk/things-we-make/super-smoothies (Stand: 27.01.2026)
Ausgangstext	**Citrus defence** Fences famously don't get rid of themselves. So if you're looking to defence, you'll need to stop browsing for smoothies on the internet, put your DIY gear on and get yourself outside. But if it's defence you're after (especially the fruity sort), you can just enjoy this vitamin-packed super smoothie, bursting with mandarin, carrot, apple and ginger. It's a tasty way to give your immune system a helping hand, and it won't lead to more small talk with the neighbours. Phew.

DeepL	**Zitrus-Abwehr** Zäune verschwinden bekanntlich nicht von selbst. Wenn Sie also Ihren Zaun entfernen möchten, sollten Sie aufhören, im Internet nach Smoothies zu suchen, Ihre Heimwerkerausrüstung anlegen und nach draußen gehen. Wenn Sie jedoch auf der Suche nach Abwehrkräften sind (insbesondere der fruchtigen Art), können Sie einfach diesen vitaminreichen Super-Smoothie genießen, der voller Mandarinen, Karotten, Äpfeln und Ingwer steckt. Das ist eine leckere Möglichkeit, Ihr Immunsystem zu unterstützen, ohne dass Sie sich mit Ihren Nachbarn unterhalten müssen. Puh.
ChatGPT	**Zitrus-Abwehr** Zäune verschwinden bekanntlich nicht von selbst. Wenn du also „ent-zäunen“ willst, musst du aufhören, im Internet nach Smoothies zu stöbern, dein Heimwerker-Equipment anziehen und nach draußen gehen. Wenn du hingegen auf der Suche nach Abwehr bist – vor allem nach der fruchtigen Art –, kannst du diesen vitaminreichen Super-Smoothie einfach genießen, randvoll mit Mandarine, Karotte, Apfel und Ingwer. Eine leckere Art, deinem Immunsystem unter die Arme zu greifen – ganz ohne zusätzlichen Smalltalk mit den Nachbarn. Puh.
Human	**Zitrus-Abwehr** Abwehr ist eine Frage des Typs: Manche bauen Zäune, um sich zu schützen (was wahnsinnig viel Arbeit macht), andere trinken einfach unseren vitaminhaltigen Super-Smoothie. Falls du dich für Letzteres entscheidest: verstehen wir. Statt Säge, Schleifpapier und Schutzbrille gibt es hier die volle Ladung Mandarine, Apfel, Karotte und Ingwer. Das stärkt deine Abwehrkräfte deutlich schneller, als du „Do-it-yourself“ buchstabieren kannst. Und das Beste daran? Du musst dabei nicht so tun, als hättest du Ahnung von Gartenbau, wenn die Nachbarn über den Zaun spähen. Win-win!

Textausschnitt 8

Kürzel	UP08
Quelle	https://casper.com/products/dog-mattresses (Stand: 27.01.2026)
Ausgangstext	**Let sleeping dogs lie, on foam.** We combined pressure-relieving memory foam and durable support foam to create the best bed that loves your furry friend back. The supportive foam bolsters of the bed also create a safe space for dogs to lay their heads — literally and figuratively. **Dogs will dig it, and scratch it.** Our durable cover material is one of the strongest bonded microfibers available. It features excess material on top mimicking the sensation of pawing at loose earth. **Perfect for humans, too.** The practically impenetrable nylon cover is easy to remove and machine wash. Sewn-in pockets hide the metal zippers, so they won't scratch your floors.

DeepL	**Lass schlafende Hunde liegen, auf Schaumstoff.** Wir haben druckentlastenden Memory-Schaumstoff und strapazierfähigen Stützschaumstoff kombiniert, um das beste Bett zu schaffen, das Ihren vierbeinigen Freund liebt. Die stützenden Schaumstoffpolster des Bettes schaffen außerdem einen sicheren Platz, an dem Hunde ihren Kopf ablegen können – im wörtlichen und im übertragenen Sinne. **Hunde werden es lieben und daran kratzen.** Unser strapazierfähiges Bezugsmaterial ist eine der stärksten gebundenen Mikrofasern, die es gibt. Es verfügt über überschüssiges Material auf der Oberseite, das das Gefühl nachahmt, mit den Pfoten in loser Erde zu scharren. **Auch für Menschen perfekt geeignet.** Der praktisch undurchdringliche Nylonbezug lässt sich leicht abnehmen und in der Maschine waschen. Eingenähte Taschen verbergen die Metallreißverschlüsse, sodass sie Ihren Boden nicht zerkratzen.

ChatGPT	**Schlafende Hunde soll man ruhen lassen – auf Schaumstoff.** Wir haben druckentlastenden Memory-Schaum mit langlebigem Stützschaum kombiniert, um das beste Bett zu schaffen – eines, das deinen Vierbeiner genauso liebt wie er dich. Die stützenden Schaumstoffränder bieten Hunden außerdem einen sicheren Platz zum Ablegen des Kopfes – im wörtlichen wie im übertragenen Sinn. **Hunde werden es lieben – und daran kratzen.** Unser strapazierfähiges Bezugsmaterial gehört zu den stärksten gebondeten Mikrofasern überhaupt. Überschüssiges Material auf der Oberseite ahmt das Gefühl nach, mit den Pfoten in lockerer Erde zu scharren. **Perfekt auch für Menschen.** Der nahezu unverwüstliche Nylonbezug lässt sich leicht abnehmen und in der Maschine waschen. Eingenähte Taschen verbergen die Metallreißverschlüsse, sodass sie deinen Boden nicht zerkratzen.

Human	**Schlafende Hunde soll man nicht wecken – aber weich betten.** Wir haben druckentlastenden Memory-Foam mit stützendem Schaum kombiniert und damit ein Hundebett geschaffen, das deinen Vierbeiner genauso liebt, wie er dich – Pfote drauf. Die stabilen Schaumstoff-Ränder stützen den Kopf und schaffen einen sicheren Ort, an dem dein Hund die Welt um sich herum vergessen kann. **Da wird der Hund in der Pfanne verrückt (vor Freude).** Unser Bezug aus hochwertiger Mikrofaser ist so widerstandsfähig, dass Hunde darin nicht nur schlafen, sondern auch nach Herzenslust „buddeln" können: Das zusätzliche Material auf der Oberseite erinnert an locker aufgescharrte Erde. **Zum Wohlfühlen – auch für Zweibeiner.** Der robuste Nylonbezug ist im Handumdrehen in der Waschmaschine. Eingenähte Taschen verstecken die Metallreißverschlüsse – damit dein Boden kratzerfrei bleibt und du dich entspannt zurücklehnen kannst.

A2 Fragebogen

Fragebogendesign

Abbildung A2.1: Einleitung und Abschnitt 1

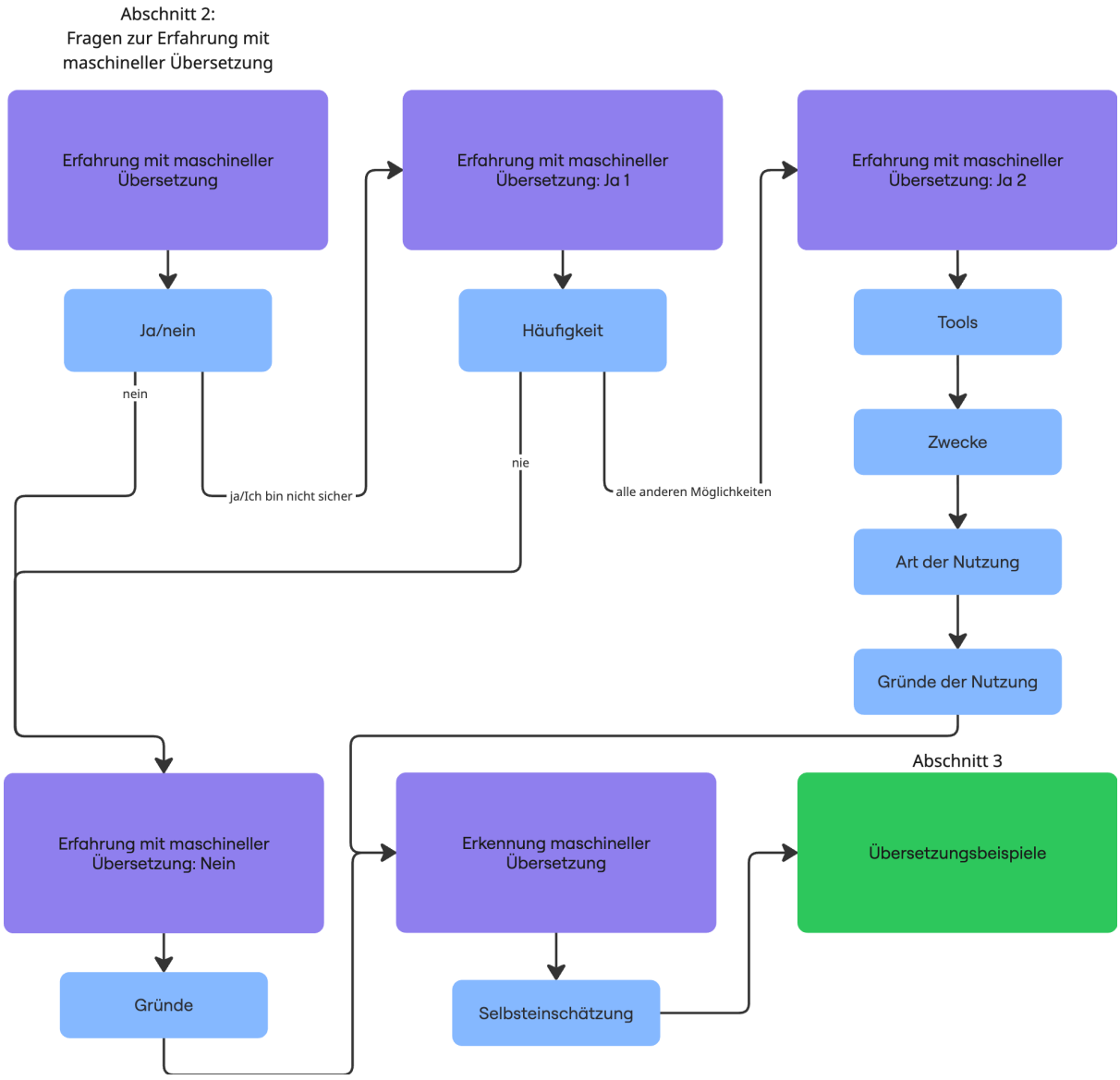

Abbildung A2.2: Abschnitt 2

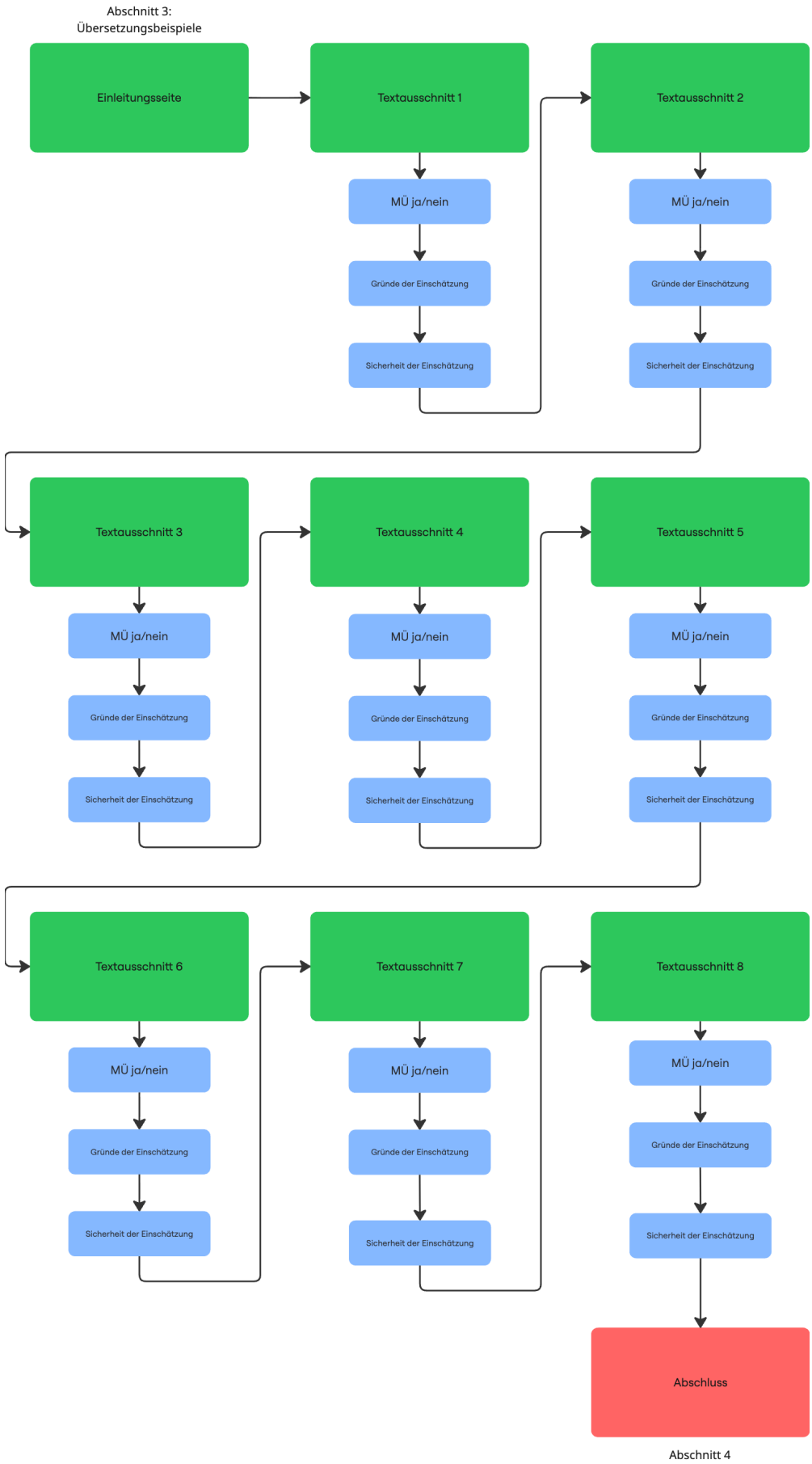

Abbildung A2.3: Abschnitt 3

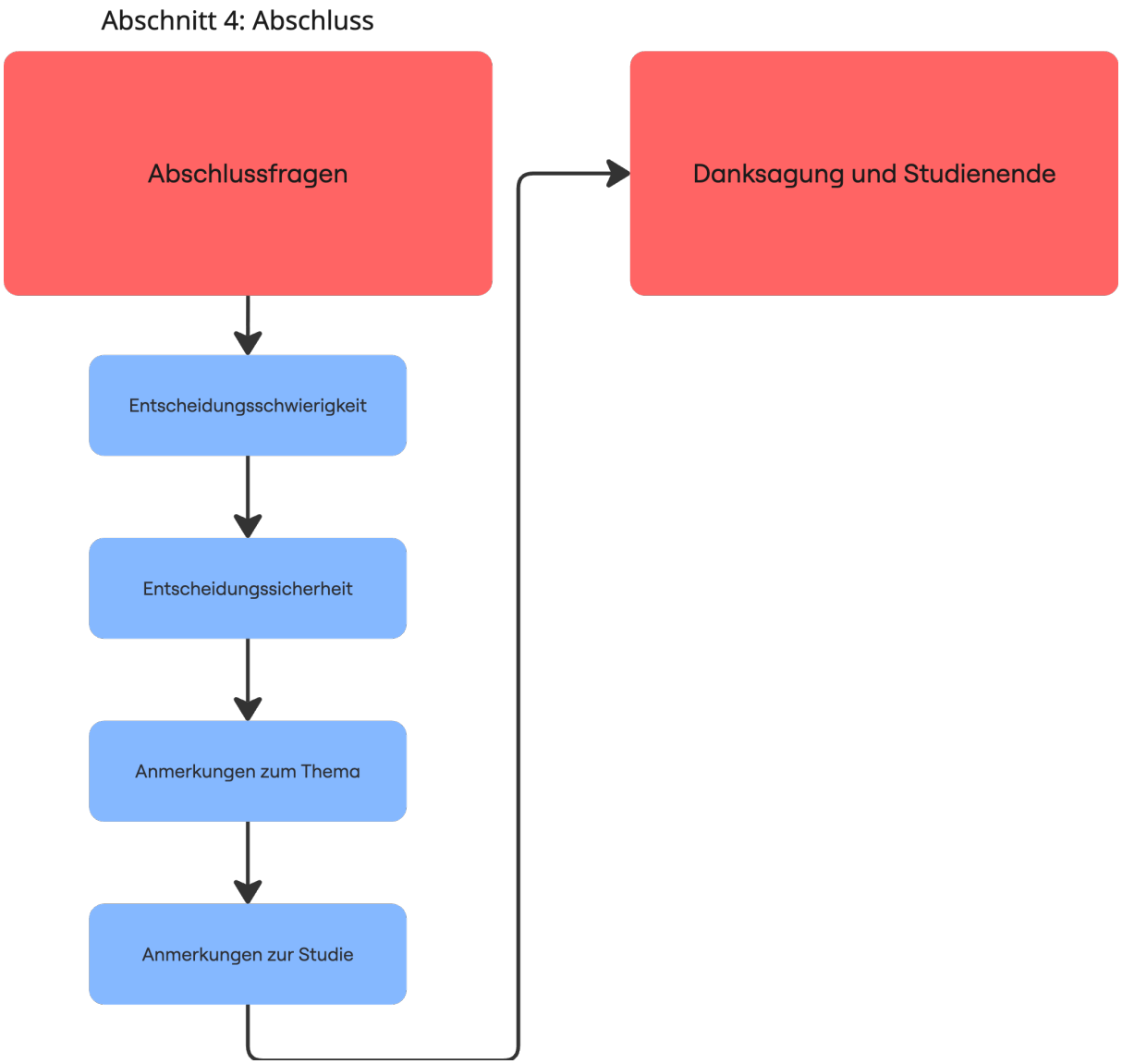

Abbildung A2.4: Abschnitt 4

Fragen

Einleitungsseite

Vielen Dank für Ihr Interesse an dieser Studie.

Im Rahmen dieser Untersuchung wird erforscht, wie unterschiedliche Übersetzungen wahrgenommen und eingeschätzt werden. Dabei werden Ihnen mehrere kurze Textausschnitte vorgelegt, zu denen Sie verschiedene Fragen beantworten.

Die Teilnahme an dieser Studie ist **freiwillig**. Sie können die Befragung jederzeit abbrechen, ohne dass Ihnen daraus Nachteile entstehen. Die Bearbeitung des Fragebogens dauert etwa **10–15 Minuten**.

Alle erhobenen Daten werden **anonym** gespeichert und ausschließlich zu wissenschaftlichen Zwecken im Rahmen einer Masterarbeit ausgewertet. Es werden keine personenbezogenen Daten erhoben, die eine Identifikation Ihrer Person ermöglichen.

Mit dem Klick auf **„Weiter“** bestätigen Sie, dass Sie die oben genannten Informationen gelesen und verstanden haben und **freiwillig an der Studie teilnehmen**.

Abschnitt 1: Soziodemografische Fragen

SM01: Wie alt sind Sie?

☐ unter 18

☐ 18–30

☐ 31–45

☐ 46–65

☐ über 65

SM02: Ist Deutsch Ihre Erstsprache?

☐ Ja → SM05

☐ Nein → SM03

SM03: Welche Sprache ist Ihre Erstsprache?

☐ Albanisch

☐ Arabisch

☐ Bosnisch/Kroatisch/Serbisch

☐ Bulgarisch

☐ Dänisch

☐ Englisch

☐ Estnisch

☐ Finnisch

☐ Französisch

☐ Griechisch

☐ Irisch/Gälisch

☐ Italienisch

☐ Katalanisch

☐ Litauisch

☐ Luxemburgisch

☐ Maltesisch

☐ Niederländisch

☐ Österreichische Gebärdensprache

☐ Persisch

☐ Polnisch

☐ Portugiesisch

☐ Romani

☐ Rumänisch

☐ Schwedisch

☐ Slowakisch

☐ Slowenisch

☐ Spanisch

☐ Tschechisch

☐ Türkisch

☐ Ungarisch

☐ Andere:

SM04: Wie schätzen Sie Ihre sprachliche Kompetenz in Deutsch ein?

☐ Anfänger/in (A1-A2) → ST01

☐ Mittelstufe (B1-B2) → ST01

☐ Fortgeschritten (C1) → SM05

☐ Ausgezeichnet/(nahezu) auf Erstsprachenniveau (C2) → SM05

SM05: Haben Sie ein abgeschlossenes Studium im Bereich der Sprach-/Übersetzungswissenschaft? *Bitte wählen Sie den höchsten erreichten Abschluss aus.*

☐ Nein → SM07

☐ Ja, Bachelor (Bakk./BA) → SM06

☐ Ja, Master bzw. einschlägiges Diplom (Mag./MA/Dipl.-Übersetzer:in) → SM06

☐ Ja, Doktorat (Dr./PhD) → SM06

SM06: In welcher Disziplin haben Sie einen Abschluss? *Bitte wählen Sie alle zutreffenden Optionen aus.*

☐ Transkulturelle Kommunikation/Translation

☐ Linguistik

☐ Deutsche Philologie/Germanistik

☐ DaF/DaZ

☐ Sonstiges:

SM07: Wie viel einschlägige Berufserfahrung haben Sie mit Textarbeit? *Dazu zählen z. B. Übersetzung, Lokalisierung, Copywriting oder Lektorat.*

☐ Keine → MT01

☐ Weniger als 1 Jahr → SM08

☐ 1 bis unter 3 Jahre → SM08

☐ 3 Jahre oder mehr → SM08

SM08: In welcher Art der Textarbeit haben Sie diese Erfahrung? *Bitte wählen Sie alle zutreffenden Optionen aus.*

☐ Übersetzung

☐ Lokalisierung

☐ Copywriting

☐ Lektorat

☐ Sonstiges:

Abschnitt 2: Fragen zur Erfahrung mit maschineller Übersetzung

MT01: Haben Sie schon einmal maschinelle Übersetzungstools oder Large Language Models für Übersetzungszwecke verwendet? *z. B. Google Translate, DeepL oder ChatGPT.*

☐ Ja → MT02

☐ Nein → MT07

☐ Ich bin mir nicht sicher → MT07

MT02: Wie häufig nutzen Sie maschinelle Übersetzungstools oder Large Language Models für Übersetzungen?

☐ Nie → MT07

☐ Jährlich → MT03

☐ Monatlich → MT03

☐ Wöchentlich → MT03

☐ Mehrmals pro Woche → MT03

☐ Täglich → MT03

MT03: Welche maschinellen Übersetzungstools oder Large Language Models haben Sie schon einmal für Übersetzungen verwendet? *Bitte wählen Sie alle zutreffenden Optionen aus.*

☐ Google Translate

☐ DeepL

☐ eTranslation

☐ Embedded

☐ Microsoft Translator

☐ ChatGPT

☐ Gemini

☐ Sonstige:

MT04: Wofür nutzen Sie maschinelle Übersetzungstools oder Large Language Models für Übersetzungen? *Bitte wählen Sie alle zutreffenden Optionen aus.*

☐ Private Zwecke

☐ Ausbildungszwecke

☐ Berufliche Zwecke

☐ Sonstiges:

MT05: Wofür nutzen Sie maschinelle Übersetzungstools oder Large Language Models im Übersetzungsprozess? *Bitte wählen Sie alle zutreffenden Optionen aus.*

☐ Rohübersetzung zum Verständnis des Originaltexts

☐ Rohübersetzung zur Nachbearbeitung

☐ Ideenfindung für den Übersetzungsprozess

☐ Nachbearbeitung einer bestehenden Übersetzung

☐ Qualitätssicherung

☐ Sonstiges:

MT06: Aus welchen Gründen nutzen Sie maschinelle Übersetzungstools oder Large Language Models für Übersetzungszwecke? *Bitte wählen Sie alle zutreffenden Optionen aus.*

☐ Sie sind leicht zugänglich

☐ Ich spare dadurch Zeit

☐ Ich spare dadurch Geld für professionelle Übersetzungsdienste

☐ Die Qualität ist gut genug für meine Zwecke

☐ Sonstiges:

MT07: Warum nutzen Sie keine maschinellen Übersetzungstools? *Bitte wählen Sie alle zutreffenden Optionen aus.*

☐ Aus Gewohnheit

☐ Ich kenne keine maschinellen Übersetzungstools

☐ Ich weiß nicht, wie ich maschinelle Übersetzungstools verwende

☐ Die Qualität ist nicht gut genug

☐ Die Vor- und Nachbereitung ist zu aufwendig

☐ Generelle Ablehnung von maschineller Übersetzung

☐ Sonstiges:

MT08: Glauben Sie erkennen zu können, ob es sich um eine maschinell angefertigte Übersetzung handelt?

☐ Ja

☐ Nein

☐ Ich bin mir nicht sicher

Mitteleinleitung

Nachfolgend werden Ihnen acht kurze Textausschnitte gezeigt.

Bitte lesen Sie jeden Text aufmerksam durch und beantworten Sie anschließend die dazugehörigen Fragen.

Abschnitt 3: Übersetzungsbeispiele

Anmerkung: jeweils pro Textausschnitt.

UP01–08: Handelt es sich bei der Übersetzung um eine maschinelle oder Humanübersetzung?

☐ Maschinelle Übersetzung

☐ Humanübersetzung

UP09, UP26–UP32: Bitte bewerten Sie den Text anhand der folgenden Fragen.

Wie flüssig ist der Text?

☐ ☐ ☐ ☐ ☐

gar nicht sehr

Wie natürlich klingt der Text?

☐ ☐ ☐ ☐ ☐

gar nicht sehr

Wie stimmig ist die Wortwahl?

☐ ☐ ☐ ☐ ☐

gar nicht sehr

Wie stimmig ist der Satzbau?

☐ ☐ ☐ ☐ ☐

gar nicht sehr

Wie gut passen Stil und Ton zum Texttyp?

☐ ☐ ☐ ☐ ☐

gar nicht sehr

Wie stimmig wirkt der Text insgesamt?

☐ ☐ ☐ ☐ ☐

gar nicht sehr

UP33–UP40: Sonstige Anmerkungen (optional)

UP18–UP25: Wie sicher sind Sie sich bei Ihrer Einschätzung?

☐ ☐ ☐ ☐ ☐

Sehr unsicher Eher unsicher Neutral Eher sicher Sehr sicher

Abschnitt 4: Abschluss

AB01: Wie schwierig war es für Sie, eine Entscheidung zu treffen?

☐	☐	☐	☐	☐
Sehr schwierig	Eher schwierig	Neutral	Eher leicht	Sehr leicht

AB02: Wie sicher waren Sie sich insgesamt bei Ihren Entscheidungen?

☐	☐	☐	☐	☐
Sehr unsicher	Eher unsicher	Neutral	Eher sicher	Sehr sicher

AB03: Haben Sie sonstige Anmerkungen zum Thema der Studie? (optional)

AB04: Haben Sie Anmerkungen zum Fragebogen und zur Studie selbst? (optional)

Abbruch bei Nichtqualifizierung (ST01)

Vielen Dank für Ihr Interesse an dieser Studie.

Für die Teilnahme an dieser Untersuchung ist ein sehr hohes Sprachverständnis erforderlich. Leider erfüllen Sie dieses Kriterium nach Ihren Angaben nicht vollständig.

Ihre Teilnahme endet daher an dieser Stelle.
Vielen Dank für Ihre Zeit und Ihr Interesse an der Studie.

Abstract (DE)

Die vorliegende Masterarbeit untersucht die Wahrnehmung maschineller Übersetzungen anhand eines Turing-Test-ähnlichen Settings. Im Fokus steht die Frage, ob Leser:innen zwischen Humanübersetzungen und maschinellen Übersetzungen kreativer Marketingtexte unterscheiden können und inwiefern die verwendeten Tools DeepL und ChatGPT sowie der Expert:innenstatus der Teilnehmenden die Wahrnehmung beeinflussen. Dazu wurden Sprachexpert:innen in der Textarbeit und Lai:innen Ausschnitte aus Marketing-Webseitentexten vorgelegt, die hinsichtlich ihres Ursprungs eingeschätzt und anhand sprachlicher sowie stilistischer Kriterien bewertet werden sollten.

Die Ergebnisse zeigen, dass rund ein Drittel der maschinell übersetzten Textausschnitte von Teilnehmenden als humanübersetzt wahrgenommen wurde, wobei kein klarer Vorteil des Expert:innenstatus festgestellt werden konnte. Darüber hinaus wurde kein signifikanter Unterschied zwischen den beiden verwendeten maschinellen Tools nachgewiesen, wenngleich ChatGPT bei kreativeren Texten höhere Täuschungsraten erzielte. Die Untersuchungsergebnisse zeigen eine systematische Verzerrung bei der Beurteilung von Übersetzungen, die als maschinell wahrgenommen wurden, da diese konsistent schlechter bewertet wurden.

Die Arbeit zeigt damit, dass rohe Outputs maschineller Übersetzungssysteme im Bereich kreativer Marketingkommunikation von Leser:innen überwiegend erkannt werden, ihre Bewertung jedoch stark vom angenommenen Ursprung beeinflusst wird. Die Ergebnisse dieser Arbeit leisten somit einen Beitrag zur Einschätzung des aktuellen Entwicklungsstands maschineller Übersetzung im kreativen Bereich und bieten Anknüpfungspunkte für weitere Untersuchungen zu den Auswirkungen maschinell wahrgenommener Übersetzungen auf Markenreputation und Kund:innenbindung.

Abstract (EN)

This master's thesis examines the perception of machine translations using a Turing Test-like setting. The focus lies on the question of whether readers can distinguish between human translations and machine translations of creative marketing texts, and to what extent the tools used (DeepL and ChatGPT) and the participants' expert status influence this perception. For this purpose, experts in text work and laypeople were presented with excerpts from marketing website texts, which they were asked to assess in terms of their origin and evaluate on the basis of linguistic and stylistic criteria.

The results show that around a third of the machine-translated text excerpts were perceived by participants as having been translated by a human, with no clear advantage being observed for participants with expert status. Furthermore, no significant difference was found between the two machine translation systems used, although ChatGPT achieved higher deception rates with more creative texts. The findings reveal a systematic bias in the evaluation of translations perceived as machine-generated, as these were consistently rated lower.

The study thus demonstrates that readers are generally able to recognise raw outputs from machine translation systems in the field of creative marketing communication, although their assessment is strongly influenced by the perceived origin of the text. The findings of this paper thus contribute to an assessment of the current state of development of machine translation for creative purposes and provide a basis for further research into the effects of machine-translated content on brand reputation and customer loyalty.