universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Evaluation von Übersetzungen wirtschaftlicher Fachtexte in DeepL"

verfasst von / submitted by

Džemila Ribić

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

Master of Arts (MA)

Wien, 2020 / Vienna 2020

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

Danksagung

Zuallererst möchte ich vor allem meiner Betreuerin, Frau Ass.-Prof. Mag. Dr. Dagmar Gromann, BSc, danken, die mich bei der Themenfindung und der gesamten Ausarbeitung meiner Masterarbeit mit ihrem außerordentlichen fachlichen und wissenschaftlichen Wissen immerfort angeleitet und motiviert hat.

Ohne meine beiden fleißigen und verlässlichen Annotatorinnen Romana und Antonia wäre die Studie nicht möglich gewesen. Auch ihnen gebührt ein großer Dank.

Meiner wunderbaren Korrekturleserin Anna möchte ich an dieser Stelle auch ganz herzlich danken.

Danken möchte ich auch meinen lieben Eltern und meiner Schwester, die mich immer und ganz besonders im Studium ermutigt und unterstützt haben.

Zusammenfassung

Maschinelle Übersetzungssysteme sind in der heutigen schnelllebigen Welt als Hilfsmittel für das stetig wachsende Übersetzungsvolumen unabdingbar. Der neueste Ansatz dafür ist die neuronale maschinelle Übersetzung (NMÜ).

In dieser Arbeit werden wirtschaftliche Fachtexte unterschiedlicher Komplexität genutzt, die vorher in die Schwierigkeitsstufen *leicht, mittel* und *schwer* abgestuft wurden. Nach der Abstufung werden die Texte mit dem auf neuronalen Netzten basierenden online-Übersetzungsprogramm *DeepL* übersetzt. Anschließend wird eine Fehleranalyse durchgeführt. Hierfür wird eine menschliche Evaluationsmethode verwendet. Die Autorin der Arbeit und zwei weitere Annotatorinnen analysieren die übersetzten Textproben mithilfe der *SAE-J2450-Qualitätsmetrik*. Nach Abschluss der Analysearbeit werden die Werte mit *Cohen's Kappa nach Fleiss* ausgewertet. Das Ergebnis liefert einen Übereinstimmungswert.

Ziel dieser Arbeit ist es festzustellen, ob es Qualitätsunterschiede in Übersetzungen im Sprachenpaar Englisch-Deutsch von Wirtschaftsfachtexten verschiedener Abstufungen in der Fachsprachlichkeit, die mit dem vollautomatisierten Übersetzungsprogramm DeepL übersetzt wurden, gibt. Ein weiteres Ziel ist es auszumachen, ob in DeepL-Übersetzungen von Texten mit einer niedrigen Anzahl fachsprachlicher Merkmale ein höherer Qualitätsgrad als in Texten mit einer hohen Anzahl an selbigen Merkmalen gefunden werden kann.

Die Studie ergab, dass die Qualität der maschinellen Übersetzung immer noch fehlerbehaftet ist. Die Ergebnisse der Studie zeigten, dass anspruchsvollere und weniger anspruchsvolle Texte einen ähnlichen Qualitätsgrad in den Übersetzungen aufwiesen.

Abstract

Machine translation systems are indispensable in today's fast-paced world as a tool for the ever-increasing volume of translations. The latest approach to this is neural machine translation (NMT).

In this study, economic texts of varying complexity are used, which were previously graded into the difficulty levels *easy, medium* and *difficult*. After said grading, the texts are translated with the online translation program *DeepL*, which is based on neural networks. After translation, an error analysis is then carried out with a human evaluation method. The author of the thesis and two other evaluators analyze the translated text samples using the *SAE-J2450 Translation Quality Metric*. Once the analysis work is complete, the values are evaluated using *Cohen's Kappa according to Fleiss*. The result provides a match value.

The aim of this study is to determine whether there are quality differences in translations in the language pair English-German of economic texts of different gradations in technical terminology that were translated with the fully automated translation program DeepL. Another objective consists of finding out whether a higher degree of quality can be found in DeepL translations of texts with a low number of technical language features than in texts with a high number of said features.

The study finds that the quality of machine translation is still subject to errors. The results of the study showed that more demanding and less demanding texts showed a similar level of quality in the translations.

Inhaltsverzeichnis

1 Einleitung

In den letzten Jahrzehnten hat sich die Welt, in der wir leben, drastisch verändert. Einer der Hauptfaktoren dafür ist die Globalisierung, die laut Schäfer (2002: 17) in einer „internationale[n] Vernetzung des Wirtschaftsgeschehens sowie nahezu aller anderen Lebensbereiche“ besteht. Alles dreht sich um die schnelle Beschaffung und Verarbeitung von Informationen, wobei mit der Verarbeitung von Informationen auch die Übersetzung gemeint sein kann (vgl. Schäfer 2019: 17). Die globale Sprachdienstleistungsbranche hat 2018 eine Summe von 46,52 Milliarden US-Dollar generiert. Fast die Hälfte des Umsatzes wurde dabei allein in Europa erzielt. Im Jahr 2021 sollen es global sogar 56 Milliarden US-Dollar werden (vgl. Mazareanu 2019).

Ein weltweites Wachstum wird somit vorhergesagt und ein Grund dafür ist das Volumen, das übersetzt werden soll. Allein in der Europäischen Kommission werden jährlich ungefähr 1,3 Millionen Seiten übersetzt (vgl. Schäfer 2002: 17).

In unserer schnelllebigen Welt benötigt man Hilfsmittel, um bei der hohen Geschwindigkeit und dem stetig wachsenden Übersetzungsvolumen mithalten zu können. Als ein solches Hilfsmittel gilt die maschinelle Übersetzung (MÜ). Dafür existiert neben einigen Ansätzen (regel- und statistikbasierte, aber auch hybride) der durchaus neue Ansatz der neuronalen maschinellen Übersetzung (NMÜ) (Burchardt & Porsiel 2017: 11).

Pauschal kann nicht einfach behauptet werden, dass ein Ansatz besser sei als der andere. Faktoren wie vor allem die Qualität des Ausgangstextes spielen beim Übersetzungsergebnis eine bedeutende Rolle. Über die MÜ gibt es seit ihrer Einführung verschiedene Meinungen: Entweder es wird behauptet, dass MÜ für die Translation (damit ist das Übersetzen selbst und dessen AkteurInnen gemeint) nicht brauchbar ist, oder es wird zurzeit wieder die Meinung laut, dass die MÜ-Qualität in naher Zukunft sogar der Humanübersetzung entsprechen soll (vgl. Burchardt & Porsiel 2017: 11-12).

Für die Qualitätsmessung von MÜ gibt es automatische und menschliche Evaluationsmethoden, wobei letztere laut Lo Presti (2016: 1) als „goldene[r] Standard und […] als Maßstab für die Bewertung der maschinellen Übersetzung“ gelten. Auch wenn menschliche Evaluationsmethoden immer noch zu hohe Subjektivität und zu hohen Zeitaufwand aufweisen, wird heutzutage weiterhin auf sie zurückgegriffen, da sie trotz Kritik nach wie vor am verlässlichsten sind. Um den Hauptmängeln der menschlichen Evaluationsmethoden entgegenzuwirken, wurden automatische Evaluationsmethoden entwickelt.

Die Qualität kann aber auch vom Thema und dem Grad der Fachlichkeit eines Textes abhängig sein. In dieser Arbeit wird die Annahme getroffen, dass je spezifischer oder fachlicher Texte sind, desto geringer ist die Qualität der maschinellen Übersetzung. Ziel dieser Arbeit ist es, zu überprüfen, ob diese Annahme bestätigt oder verworfen werden kann.

1.1 Aufbau und Struktur dieser Arbeit

Diese Masterarbeit besteht aus elf Kapiteln. Das erste Kapitel fängt mit einer allgemeinen Einleitung an und die zu untersuchende Annahme wird beschrieben. Im zweiten Kapitel wird das Verhältnis zwischen Fachsprache und Gemeinsprache veranschaulicht. Weiters werden die Merkmale von Fachtexten und die Grundlagen zur Wirtschaftslinguistik erläutert. Im dritten Kapitel wird ein Einblick in die Geschichte der maschinellen Übersetzung gegeben. Der regelbasierte und statistikbasierte Ansatz, sowie maschinelle Übersetzung, die auf neuronalen Netzen basiert, werden im vierten Kapitel vorgestellt. Im fünften und sechsten Kapitel wird eine Einführung in die beliebtesten und gängigsten menschlichen und automatischen Evaluationsmethoden geboten. Im siebten Kapitel gibt es einen Überblick zum aktuellen Stand der Forschung. Das achte Kapitel dient zur Veranschaulichung der angewandten Methodik in der vorliegenden Arbeit. Eine detaillierte Beschreibung der Ergebnisse des praktischen Teils dieser Arbeit erfolgt im neunten Kapitel. Die Diskussion mit dem darauffolgenden Fazit bilden die letzten zwei Kapitel. Darin werden die Ergebnisse ausführlich interpretiert und diskutiert. Weiters werden Schwierigkeiten und Limitationen, die während dieser Studie auftraten, aufgezeigt, was durch eine Schlussfolgerung ergänzt wird.

2 Grundlagen

Im folgenden Kapitel wird ein Überblick über das komplexe Konzept von Fachsprache und Nicht-Fachsprache, auch Gemeinsprache genannt, gegeben. Weiters werden die Merkmale von Fachtexten genannt und die Wirtschaftslinguistik wird in Kürze umrissen.

2.1 Fachsprache vs. Nicht-Fachsprache

Spricht man von Fachsprache oder der fachsprachlichen Kommunikation, geht man davon aus, dass es auch eine Nicht-Fachsprache bzw. eine gemeinsprachliche Kommunikation, die auch Gemeinsprache genannt werden kann, gibt. Häufig wurde in der Vergangenheit versucht, *Fachsprache* und *Gemeinsprache* voneinander zu trennen (siehe z. B. Von Hahn 1981: 40-41; Adamzik 2004: 317). Diese Versuche waren jedoch nicht erfolgreich, da künstlich eine Grenze zwischen diesen beiden Begriffen gesetzt wurde. Mit der Zeit wurde festgestellt, dass eine Trennung von Fachsprache und Gemeinsprache nicht allein durch die Fachlichkeit von Denotaten erfolgen kann (vgl. Göpferich 2007: 413).

Vielmehr ist eine Abstufung der Fachsprachlichkeit notwendig, damit eine Abwendung von dem starren Zuordnen von Merkmalen in die Fachsprache oder in die Gemeinsprache erfolgt (vgl. Kalverkämper 1996: 135). Kalverkämper (1996: 135) hat eine Skala der „Graduellen Stufung von Fach(sprach)lichkeit“ entwickelt, in der er aufzeigt, dass die Qualität von Texten variieren kann: Es gibt Texte, die fachtextlicher sind bzw. mehr fachsprachliche Merkmale aufweisen als andere. Eine solche intuitive Zuordnung erfolgt nämlich auf allen Sprachebenen (wie z.B. Phoneme, Termini, Makrostrukturen, Kohärenzen und Textsortenmerkmale).

Göpferich (vgl. 2007: 413) hat Kalverkämpers Gedanken fortgesetzt und folgendes Modell der komplementären Spektren entwickelt:

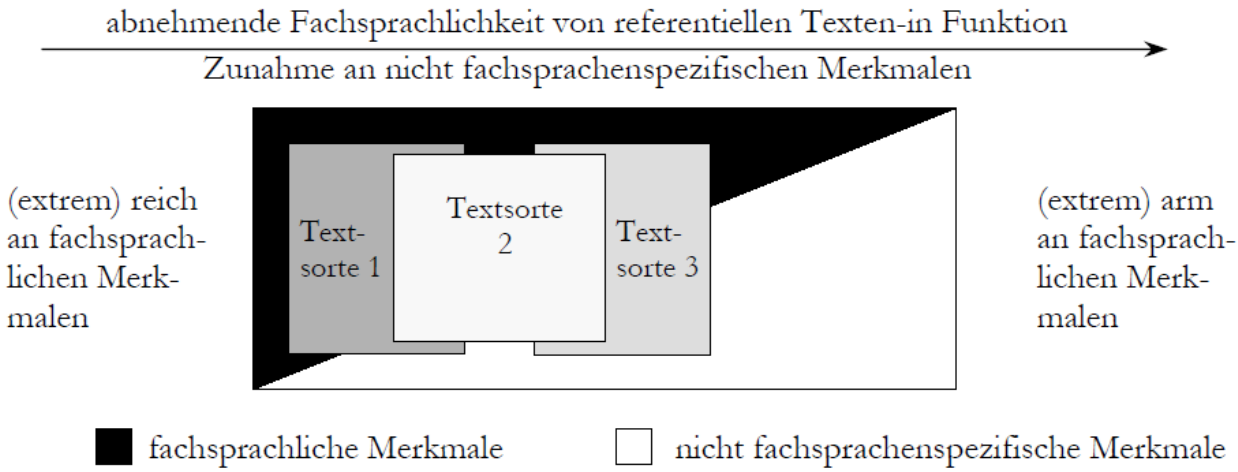

Abb. 1: Modell der komplementären Spektren (Göpferich 2007: 413)

Göpferich (vgl. 2007: 413-414) verdeutlicht in ihrem Modell der komplementären Spektren „die Fachsprachlichkeit als Eigenschaft von Text-in-Funktion“. Es gibt demnach keine Texte mehr, die nur gemeinsprachlich sind, sondern nur noch Texte, die mehr oder weniger fachsprachlich sind. Anhand des Modells zeigt Göpferich, dass Texte nicht nur fachsprachliche Merkmale, sondern auch gemeinsprachliche Merkmale enthalten können. Somit ist ersichtlich, dass man in Texten, die laut dem Modell „(extrem) reich an fachsprachlichen Merkmalen“ sind – also in sehr fachlichen Texten – auch Merkmale antreffen kann, die eher untypisch sind.

2.2 Merkmale von Fachtexten

Fachsprachliche Kommunikation wird oft mit den Wörtern *Kürze, Prägnanz* und *Sprachökonomie* (siehe z.B. Roelke 1999: 29-30; Von Hahn 1983: 117) in Verbindung gebracht. Werden diese Begriffe jedoch genauer betrachtet, erkennt man, dass es sich hierbei um Pauschalisierungen der fachsprachlichen Kommunikation handelt, welche in Theorie und Praxis wenig aussagen und nutzen. Wie erreicht man nämlich eine Kürze oder Sprachökonomie in der fachsprachlichen Kommunikation (vgl. Göpferich 2007: 412-414)? Um diese Frage zu beantworten, müssen die Funktionen von Fachtexten mit einem hohen Grad an Fachlichkeit betrachtet werden. Fachtexte kennzeichnen sich durch einen Stil, der exakt, objektiv, und ökonomisch ist (vgl. Baumann 1998: 374) und beinhalten auf allen sprachlichen Ebenen bestimmte sich immer wiederholende Strukturen (vgl. Niederhaus 2011: 211).

Komplexe Informationen werden in Fachtexten häufig durch zusätzliche Mittel wie Abbildungen, Tabellen oder Diagramme übermittelt (vgl. Buhlmann & Fearns 2000: 64).

Solche Hilfsmittel ermöglichen es, etwas zu beschreiben, ohne Wörter dafür aufwenden zu müssen, da dies wegen hoher Informationskomplexität zu zeitaufwendig sein kann (vgl. Göpferich 2007: 418).

In Texten mit einem höheren Grad an Fachlichkeit können tendenziell längere Sätze gefunden werden als in Texten mit einem niedrigeren Grad an Fachlichkeit. Dies ist ein Merkmal für Fachtexte, da davon auszugehen ist, dass längere Sätze schwieriger zu verstehen sind als kürzere (vgl. Niederhaus 2011: 211).

Zusätzlich zu den langen Sätzen gilt auch die Präzision als eine Eigenschaft von Fachtexten. Je präziser etwas geschrieben ist, desto exakter und prägnanter ist der Text und kann nicht mehr so einfach von der Allgemeinheit verstanden werden (vgl. Arntz et al. 2014: 23). Erreicht werden kann ein präziserer Schreibstil unter anderem durch die Verwendung von Attributen (z.B. der unterzeichnete Vertrag statt: der Vertrag, der unterzeichnet wurde). Durch solche Konstruktionen werden Nebenätze eingespart. Es wird angenommen, dass Texte schwerer zu verarbeiten sind, wenn sie über eine hohe semantische Dichte verfügen (vgl. Niederhaus 2011: 211).

Damit Objektivität und Allgemeingültigkeit erreicht werden können, wird in Fachtexten in einem anonymen Stil geschrieben. Um einen Sachverhalt zu anonymisieren, wird er oft abstrahiert (vgl. Oksaar 1998: 397), was dazu führt, dass die Formulierungen nicht mehr so einfach zu verarbeiten sind und von LaiInnen nicht so leicht verstanden werden können.

Nicht nur in der Sprache generell, sondern vor allem in den Fachsprachen, entstehen unaufhörlich neue Termini (vgl. Hoffmann 1976: 258). Es ergeben sich ständig neue Phänomene, Produkte oder auch Techniken, die einer Benennung oder eines Namens bedürfen, damit sie nicht umschrieben werden müssen, da auch dies zu viel Zeit beanspruchen würde. Diese Termini werden entweder beispielsweise von ForscherInnen eines bestimmten Gebiets oder von den involvierten EntwicklerInnen erfunden.

ÜbersetzerInnen spielen in diesem Prozess eine wichtige Rolle, da durch sie die Termini auch in andere Sprachen übertragen werden können (vgl. Göpferich 2007: 414). Es wird angenommen, dass Fachwörter deshalb so schwierig sind, weil man sie häufig nicht einfach vom Kontext deduzieren kann, sondern sie sich aneignen muss. Kommen in einem Text also viele Termini bzw. Wortbildungen vor, dürfte sich dies auf die Verständlichkeit des Textes auswirken (vgl. Niederhaus 2011: 212).

An dieser Stelle soll erneut auf die Annahme aufmerksam gemacht werden, dass die Häufigkeit der vorkommenden fachsprachlichen Merkmale in Texten bestimmt, ob ein Text mehr oder weniger fachlich ist (vgl. Hoffmann 1976: 201). Durch diese Herangehensweise

kann ein Vergleich von Texten bezüglich ihres Grades an Fachsprachlichkeit vorgenommen werden (vgl. Niederhaus 2011: 213).

2.3 Wirtschaftslinguistik

Nachdem im vorherigen Abschnitt ein allgemeiner Überblick über die Fachsprachenforschung und die Merkmale von Fachtexten gegeben wurde, soll im Folgenden näher auf die Wirtschaftslinguistik eingegangen werden. Die Intention der Autorin ist es, darauf hinzuweisen, wie weitreichend und umfassend die Wirtschaftslinguistik ist. Bei der Überlegung, Wirtschaftsfachtexte für die Evaluation zu verwenden, kam die Frage auf, welche Texte genau dafür verwendet werden könnten.

Die Wirtschaftslinguistik hat ihren Ursprung im frühen 20. Jahrhundert an mehreren Handelshochschulen in den Niederlanden, der Schweiz, in Deutschland und in der damaligen Tschechoslowakei. Man konzentrierte sich dabei nicht nur auf die linguistischen Aspekte wie Lexikologie, sondern auch auf Verknüpfung von Fachgebiet und Sprache, das heißt auch kulturpolitische, soziologische und sprachpsychologische Aspekte wurden neben anderen einbezogen. Ab den 1930ern wuchs die Bedeutung der Prager Schule, welche eine andere Richtung der Wirtschaftslinguistik ins Feld führte: Es wurden insbesondere semantische Aspekte durchleuchtet (vgl. Schäfer 2002: 111). Schäfer (2002: 111) zeigt des Weiteren auf, dass „auf syntaktischer Ebene […] die Differenzierung der Wirtschaftsfachsprache nach einzelnen wirtschaftssprachlichen Textsorten wie Geschäftsbrief, Börsenbericht etc. in starkem Maße in Betracht gezogen“ wurde.

Darüber hinaus ist der Begriff *Wirtschaftsfachsprache* umfassend, da auch der Wirtschaftsbereich selbst äußerst ungleichartig ist. Die Wirtschaft beinhaltet eine Vielzahl an Bereichen, wie Wirtschaftswissenschaften, Wirtschaftspolitik, Börse, aber auch Handel und Industrie. Oft spricht man in der Forschung beim Wort *Wirtschaftssprache* nur von der Sprache in der Wirtschaftspresse oder der Börse, dabei bezeichnet *Wirtschaftssprache* vielmehr eine ungleichartige Zusammensetzung von verschiedensten Fachsprachen (vgl. Schäfer 2002: 111-112).

Seilheimer (2017: 16) schlägt folgende Definition für den Begriff *Wirtschaftslinguistik* vor:

> Wirtschaftslinguistik ist derjenige Teilbereich der Angewandten Linguistik, der Sprache in einem wirtschaftlichen Kontext mit den Mitteln der Linguistik interdisziplinär und/oder interkulturell sowie anwendungs- und lösungsorientiert untersucht.

Diese Definition umfasst sehr viele Aspekte der Wirtschaftslinguistik, die im Laufe der Geschichte entstanden sind. Des Weiteren bietet Seilheimer (2017: 16) erstmals eine Definition für den weitreichenden Begriff der Wirtschaftslinguistik.

3 Geschichte der maschinellen Übersetzung

3.1 Erste Ansätze der maschinellen Übersetzung

Bereits im 17. Jahrhundert kam die Idee auf, mechanische Apparate als Hilfsmittel für sprachbedingte Hindernisse zu nutzen. Damals wurde Latein gerade von Französisch als Lingua Franca abgelöst und man war der Meinung, dass man in keiner natürlichen Sprache Gedanken und Ideen so gut ausdrücken konnte wie mit der lateinischen (vgl. Ramlow 2009: 54).

Descartes und Leibniz schlugen vor, Wörterbücher zu entwickeln, die auf universellen numerischen Codes basierten. Diese Idee wurde dann tatsächlich ausgeführt und in der Mitte des 17. Jahrhunderts machten Cave Beck (1657), Athanasius Kircher (1663) und Johann Becher (1661) ihre Wörterbücher publik (vgl. Hutchins & Somers 1992: 5; Ramlow 2009: 54). Diese Wörterbücher gingen zwar in die Richtung des maschinellen Übersetzens, doch viel wichtiger waren sie für die Erfindung von Esperanto und anderer Plansprachen sowie für die Erforschung von sprachlichen Universalien (vgl. Ramlow 2009: 54).

Ab dem 19. Jahrhundert befasste man sich mit dem Gedanken, ob es vielleicht möglich wäre, dass eine Maschine Übersetzungen anfertigte. Erst 1933 wurde dieser anfängliche Gedanke in die Tat umgesetzt, als der Franzose Georges Artsrouni und der Russe Petr Petrovich Smirnov-Troyanskii unabhängig voneinander Patente für ihre Übersetzungsmaschinen anmeldeten (vgl. Ramlow 2009: 55).

Artsrounis Maschine war genau genommen ein Speichermedium auf einem Papierstreifen, das verwendet werden konnte, um die Entsprechung eines Wortes in einer anderen Sprache zu finden. 1937 stellte Artsrouni seinen Prototyp vor. Blickt man im Kontrast dazu rückblickend auf Smirnov-Troyanskiis Maschine, erkennt man, dass seine Erfindung für die Forschung im Bereich der MÜ von größerer Bedeutung war. Troyanskii teilte den Übersetzungsprozess in drei Abschnitte ein. Im ersten Abschnitt wurde der zu übersetzende Text von einem Editor, der nur der Ausgangssprache mächtig war, analysiert und so bearbeitet, dass die Wörter im Text in ihre Grundform gesetzt wurden und ihre syntaktische Funktion angegeben wurde. Die Übersetzungsmaschine wurde dann lediglich im zweiten Abschnitt eingesetzt, um die Sequenzen der Grundformen und Funktionen in äquivalente Sequenzen in die Zielsprache zu setzen. Im dritten Absatz befasste sich ein anderer Editor, der nur die Zielsprache beherrschte, mit dem Ergebnis der Übersetzungsmaschine und bearbeitete es so lange, bis er es in die sprachlich korrekte Form brachte (vgl. Hutchins & Somers 1992: 5; Ramlow 2009: 55-56).

Smirnov-Troyanskii wollte seine Idee noch weiter ausführen und glaubte daran, dass nicht nur der zweite Abschnitt von der Maschine umgesetzt werden konnte, sondern auch die beiden anderen Abschnitte und dass somit eine Übersetzungsmaschine entwickelt werden könnte, die vollständige Sätze ohne menschliche Hilfe übersetzen würde. Außerhalb von Russland war Troyanskii nicht bekannt und als er 1939 eine verbesserte Version seiner Übersetzungsmaschine präsentierte, schenkte man seinen Bemühungen keine Beachtung. Als Folge dessen hatte er zu einem späteren Zeitpunkt auf Entwicklungen in der maschinellen Übersetzung keinen Einfluss (vgl. Ramlow 2009: 56).

3.2 Die Anfänge der maschinellen Übersetzung

Maschinelle Übersetzung ist ein junges Wissenschaftsgebiet, das sich aus der Computerwissenschaft und dem Ingenieurwesen heraus entwickelte; erst vor rund 50 Jahren wuchs der linguistische Einfluss darauf (vgl. Schwarzl 2001: 5).

Im Verlauf des Zweiten Weltkriegs wurden die ersten Computer für militärische Zwecke entwickelt und eingesetzt. In der Nachkriegszeit wurden sie auch in der Mathematik und in der Physik verwendet (vgl. Ramlow 2009: 56). Während der Gespräche zwischen Warren Weaver, dem Vizepräsidenten der Rockefeller-Stiftung, und Andrew D. Booth, einem britischen Computerpionier, kam zum ersten Mal die Idee auf, dass Computer auch für den Übersetzungsprozess eingesetzt werden könnten. Bei den Gesprächen zwischen Weaver und Booth ging es auch um die Finanzierung: Würde die Rockefeller-Stiftung die Universität von London bei der Anschaffung eines Computers finanziell unterstützen? Weavers Bedingung dazu war, dass der Computer auch für andere Bereiche – unter anderem für die Übersetzung von Texten – und nicht nur für die Mathematik und Physik eingesetzt werde (vgl. Hutchins 1986; Ramlow 2009: 56-57).

1947 entwickelten Booth und Richard H. Richens, ein Botaniker und Forscher in der Computerlinguistik, ein Übersetzungssystem, das anhand der Übersetzung von Abstracts über Pflanzengenetik in einigen Sprachenpaaren getestet wurde. Die Neuheit bei diesem Übersetzungssystem bestand in Richens Vorschlag, grammatische Informationen miteinzubeziehen, damit eine Übersetzungsmaschine entstehen würde, die mehr könnte, als bloß Wort für Wort zu übersetzen (vgl. Ramlow 2009: 57).

Entwicklungen im Bereich der maschinellen Übersetzung wurden außerhalb des Fachs kaum zur Kenntnis genommen, doch 1949, als Weaver zusammen mit Richens ein Memorandum veröffentlichte, änderte sich dies. Im Memorandum wurden Probleme

aufgelistet, die bei der maschinellen Automatisierung des Übersetzens vorlagen, doch man war sich sicher, dass diese gelöst werden könnten (vgl. Hutchins 1986; Ramlow 59). Nach der Veröffentlichung des Memorandums herrschte in den USA und in Kanada eine durchaus optimistische Stimmung. Es entstanden entsprechende Forschungszentren in ganz Nordamerika. Yehoshua Bar-Hillel wurde 1951 zum ersten Wissenschaftler für maschinelle Übersetzung am Massachusetts Institute for Technology (MIT). Ein Jahr darauf fand bereits die erste Konferenz über MÜ statt und 1954 wurde das erste MÜ-System präsentiert: 49 Sätze auf Russisch wurden ins Englische übersetzt, wobei das MÜ-System mit 250 Worteinträgen und sechs Grammatikregeln arbeitete (vgl. Hutchins & Somers 1992: 6; Schwarzl 2001: 15). Diese Veröffentlichung führte dazu, dass sich unter ExpertInnen, aber auch in der Öffentlichkeit die Meinung verbreitete, dass es sehr wahrscheinlich war, den Übersetzungsprozess bald rein maschinell ausführen zu können (vgl. Hutchins 1986; Ramlow 2009: 59.

Das Memorandum und das positive Ergebnis des MÜ-Systems für Russisch und Englisch führten dazu, dass in diesem Bereich viel Forschung betrieben wurde, die stets von großzügiger Finanzierung unterstützt wurde (vgl. Hutchins & Somers 1992: 6; Schwarzl 2001: 15). Es wurde nicht nur in den USA geforscht, sondern auch in der damaligen Sowjetunion, in Japan, in der damaligen Tschechoslowakei, in Frankreich, in Großbritannien und in Italien (vgl. Hutchins 1986; Ramlow 2009: 59).

Mit der Zeit erkannte man, dass man sich von der MÜ zu viel erwartet hatte, und es wurde deutlich, dass unzählige technische und linguistische Hindernisse fortbestanden. Bar-Hillel (1959) erklärte, dass es nicht möglich wäre, das Ziel der vollautomatisierten Übersetzung mit der Qualität einer HumanübersetzerIn zu erreichen. Weiters bemerkte er nachdrücklich, dass man die vorgesehene Qualität auch in ferner Zukunft niemals erreichen würde. Ohne das Post-Editing einer HumanübersetzerIn würde seiner Ansicht nach ein maschinell übersetzter Text nie funktionieren. Taube (1961) vertrat den gleichen Standpunkt, woraufhin immer weniger Sponsoren gefunden werden konnten, die die Forschung im Bereich der MÜ durch ihre Finanzierung unterstützten (vgl. Hutchins 1986; Ramlow 2009: 60).

3.3 Der ALPAC-Bericht

Aufgrund der immer lauter werdenden Kritik gründete die US Regierung 1964 das aus sieben Wissenschaftlern bestehende Automatic Language Processing Advisory Committee (ALPAC), um die Fortschritte im Bereich der MÜ zu begutachten (vgl. Hutchins & Somers 1992: 7; Ramlow 2009: 61).

Das Komitee führte verschiedenste Studien über die Nachfrage und Kosten der Übersetzungen, die Verfügbarkeit von menschlichen ÜbersetzerInnen und das Post-Editing durch. Zusätzlich wurden Outputs maschineller Übersetzungssysteme analysiert und bewertet (vgl. Ramlow 2009: 61).

Als Resultat der Studien wurde 1966 der ALPAC-Bericht veröffentlicht, in dem festgestellt wurde, dass Übersetzungssysteme zu jener Zeit keine ausreichende Qualität erzeugen konnten und dass es genügend ÜbersetzerInnen gab, die die Nachfrage für Übersetzungen stillen konnten (vgl. Hutchins & Somers 1992: 7; Ramlow 2009: 61).

Da im Bericht hervorgehoben wurde, dass maschinelle Übersetzungssysteme langsamer und ungenauer arbeiteten und doppelt so teuer waren wie Humanübersetzer, vertrat man die Meinung, dass weitere Investitionen in die MÜ-Forschung nicht erforderlich waren. Stattdessen wurde empfohlen, sich eher auf die Entwicklung von maschinellen Hilfsmitteln, wie elektronische Wörterbücher und Terminologiedatenbanken, für HumanübersetzerInnen zu konzentrieren (vgl. Hutchins & Somers 1992: 7; Ramlow 2009: 61).

Natürlich wurde am ALPAC-Bericht vielfach Kritik laut: Der Bericht sei engstirnig und voreingenommen. KritikerInnen akzeptierten die darin veröffentlichten Ergebnisse und Ansichten nicht, da sie der Meinung waren, dass nicht genau genug gearbeitet wurde, dass zum Teil veraltete Angaben verwendet wurden und dass Vorteile der MÜ bewusst nicht veröffentlicht wurden. Darüber hinaus wurde kritisiert, dass die die ganze Idee über die MÜ zu schnell als zwecklos beurteilt wurde und dass durch weitere Forschung bald bessere Ergebnisse erreicht werden könnten (vgl. Ramlow 2009: 62).

Nach der Veröffentlichung des ALPAC-Berichts musste man trotz dieser heftigen Kritik mit finanziellen Einbußen kämpfen: In den USA gab es statt zehn nur noch drei Forschungsteams, wobei zwei davon teilweise komplett ohne jegliche Finanzierung auskommen mussten. Zudem erlitt die maschinelle Übersetzung einen Imageverlust und außerhalb der USA, vor allem in Japan und der damaligen Sowjetunion, konnte man die verheerenden Konsequenzen des Berichts auch spüren. In Kanada, Deutschland und Frankreich setzte man im Gegensatz dazu die Forschung noch stärker fort. Vor allem Kanada hatte aufgrund der gesetzlich verankerten Amtssprachen Englisch und Französisch einen so hohen Übersetzungsbedarf, der allein durch HumanübersetzerInnen nicht gestillt werden konnte. Deshalb wurden dort einige Forschungsprojekte für die MÜ finanziell *Canadian National Research Council* unterstützt und gefördert. Europa hatte unter anderem aufgrund der Europäischen Gemeinschaft und ihrer Sprachen der Mitgliedsstaaten auch einen sehr hohen

Übersetzungsbedarf und allgemein wuchs das Interesse an der MÜ trotz des ALPAC-Berichts (vgl. Hutchins 1986; Ramlow 2009: 62-63).

Zusammenfassend ist festzustellen, dass es durch den ALPAC-Bericht nicht nur zu ökonomischen Einbußen im Bereich der Forschung zur maschinellen Übersetzung kam, sondern dass es in diesem Bereich auch motivationstechnische Verluste gab, da die MÜ plötzlich für viele als hoffnungslos galt, obwohl das bisher Erreichte in der MÜ-Forschung tatsächlich real war (vgl. Arnold et al. 1994: 14).

3.4 Die späten 1970er und die 1980er Jahre

Gegen Ende der 1970er Jahre gab es erstmals Andeutungen eines Neustarts der Forschung im Bereich der maschinellen Übersetzung. Zum ersten Mal konnte man Erfolge in Japan verzeichnen, doch dies lag nicht daran, dass es im linguistischen Bereich mehr Fortschritte gab, oder dass Computer leistungsfähiger geworden sind, sondern daran, dass man den Anwendungsbereich der MÜ verändert hatte: Sie sollte mehr in spezifischen und eingeschränkten Bereichen und Sachgebieten angewendet werden (vgl. Schäfer 2002: 23).

Maschinelle Übersetzungssysteme wurden somit viel realitätsnäher verwendet. Das Ziel war nicht mehr, Übersetzungen zu schaffen, die menschlichen Übersetzungen gleichkamen, sondern Übersetzungen, die den Inhalt eines Ausgangstextes verständlich wiedergaben (vgl. Ramlow 2009: 63).

Es wurden verschiedene Übersetzungssysteme entwickelt, wie das für die US Airforce konzipierte direkte *Systemlogos*, mit dem man Flugzeughandbücher vom Englischen ins Vietnamesische übersetzen ließ. *SYSTRAN*, ein weiteres Übersetzungssystem, wurde für zahlreiche Sprachenpaare eingesetzt (vgl. Hutchins 1986; Ramlow 2009: 64).

Der interlingua-basierte Ansatz (siehe Kapitel 4.1.2.1) war für ein System bei CETA (Centre d'Études pour la Traduction Automatique) für russisch-französische Übersetzungen und bei einem anderen System für deutsch-englische Übersetzungen an der University of Texas im Einsatz, wenngleich dieser Ansatz nicht besonders erfolgreich war (vgl. Hutchins 1986; Ramlow 2009: 64).

Beim Projekt *EURATOM* wurden transferbasierte Systeme eingesetzt, um Patente und deren Abstracts vom Russischen ins Englische und umgekehrt zu übersetzen. Nach langer Zeit begann sich wieder eine Akzeptanz für die Forschung im Bereich der MÜ zu entwickeln, was unter anderem auf elektronische Wörterbücher und Terminologiedatenbanken zurückzuführen war, die als Hilfsmittel verwendet werden konnten. Die Kommission der Europäischen

Gemeinschaft fing ab 1975 an, *SYSTRAN* für Übersetzungen zu verwenden (vgl. Hutchins 1986; Ramlow 2009: 64).

Bedeutend für die 1980er Jahre war das Projekt *EUROTRA*. Initiatorin war wieder die Europäische Gemeinschaft, die damit Übersetzungen in und aus allen Sprachen der Mitgliedsstaaten erstellen können wollte (vgl. Hutchins 1986). *EUROTRA* war womöglich das größte und mit Sicherheit das ehrgeizigste Forschungs- und Entwicklungsprojekt in der Computerlinguistik (vgl. Arnold et al 1994: 16). Obwohl das System in der Industrie auf fast kein Interesse stieß und kein Prototyp entworfen wurde, war es trotzdem bedeutend für die Forschung – vor allem in Belgien, Dänemark, Deutschland, Großbritannien und den Niederlanden – und hat einen bemerkenswerten Beitrag zur MÜ-Forschung geleistet (vgl. Hutchins 1986; Ramlow 2009: 65).

Während der 1980er Jahre entstanden noch weitere Entwicklungen, wie das MÜ-System *SPANAM* für spanisch-englische Übersetzungen der Panamerikanischen Gesundheitsorganisation, oder das *METAL*-System der Universität Texas in Austin (vgl. Arnold et al 1994: 16).

4 Verschiedene Ansätze und Methoden

Seit den 1990er Jahren sind maschinelle Übersetzungssysteme mit steigender Tendenz kommerziell im Einsatz: Vor allem große Computerfirmen, die ihre Produkte auch auf internationalen Märkten verkaufen wollten, haben MÜ für sich entdeckt und sie dabei immer mehr in ihre Produktionsabläufe eingebunden. Für solche Firmen birgt die MÜ einen großen Vorteil gegenüber der Konkurrenz, denn mit Schnelligkeit kann man sich von den anderen abheben. Übersetzt werden zum Beispiel nicht nur Software, sondern auch dazugehörige Produktbeschreibungen, Betriebsanleitungen etc. (vgl. Ramlow 2009: 69).

Klar ist, dass professionelle ÜbersetzerInnen, Unternehmen und Institutionen neben MÜ-Systemen zusätzlich maschinelle Hilfsmittel benötigen, damit sie den Übersetzungsprozess erleichtern und beschleunigen können. Zu diesen Hilfsmitteln zählen unter anderem elektronische Wörterbücher, Glossare und Terminologiedatenbanken (vgl. Hutchins 1995; Ramlow 2009: 69).

Der Gebrauch von maschinellen Übersetzungssystemen spielt nicht nur in der professionellen Welt eine Rolle (vgl. (vgl. Hutchins 1995; Ramlow 2009: 69). Seit Mitte der 1990er Jahre hat das Internet einen großen Einfluss auf die Entwicklung im Bereich der maschinellen Übersetzung. Online-Übersetzungssysteme stellen die Übersetzung von Webseiten und E-Mails zur Verfügung. MÜ-EntwicklerInnen bieten außerdem Online-Übersetzungen von On-Demand-Übersetzungen an. Die Übersetzungsqualität von diesen Online-Übersetzungssystemen ist meist schlecht, jedoch decken diese Dienste zweifellos einen erheblichen Bedarf an sofort verfügbaren Rohübersetzungen (vgl. Hutchins 2006: 17-18). Eine der größten Herausforderungen der Online-Systeme ist vor allem die Übersetzung von Abkürzungen, Wortspielen, Witzen, etc., die vor allem in E-Mails und Chatrooms vorkommen. Des Weiteren ist die im Internet verwendete Sprache oft umgangssprachlich, inkohärent und ungrammatikalisch. Die MÜ liefert bei wissenschaftlichen und technischen Texten viel bessere Ergebnisse, da sie ursprünglich auch dafür entwickelt wurden (vgl. Hutchins 2006: 17-18).

Über die Jahre hinweg haben sich bei der maschinellen Übersetzung die verschiedensten Ansätze entwickelt. Im Folgenden soll gezeigt werden, wie weit fortgeschritten die MÜ heutzutage ist, wie sich die einzelnen Ansätze voneinander unterscheiden und welche Stärken und Schwächen sie haben.

4.1 Regelbasierter Ansatz

Der regelbasierte Ansatz gehört in der MÜ-Forschung zu den klassischen Ansätzen. Abhängig vom Sprachenpaar und ob es für eine bestimmte Fachsprache verwendet wird, ob die dementsprechende Fachterminologie eingesetzt wurde oder ob es als allgemeinsprachliches System verwendet wird, kann die Qualität der Übersetzungen stark variieren, bzw. unterschiedlich ausfallen. Der regelbasierte Ansatz stützt sich auf die Analyse, den Transfer und die Synthese von sprachlichen Regeln (vgl. Stein 2009: 8).

Generell lassen sich zwei regelbasierte Ansätze in der MÜ-Forschung unterscheiden. Der erste davon ist der *direkte Ansatz* und der zweite der *indirekte Ansatz*. Der indirekte Ansatz beinhaltet wiederum zwei eigene Methoden: die *Interlingua-Methode* und die *Transfer-Methode* (vgl. Schäfer 2002: 26-27).

Mithilfe des sogenannten Vauquois-Dreiecks (Abb. 2), das nach einem der MÜ-Pioniere, Bernard Vauquois, benannt wurde, können die drei Grade der Komplexität des direkten Ansatzes und des indirekten Ansatzes mit seinen Methoden veranschaulicht werden (vgl. Schäfer 2002: 27-28):

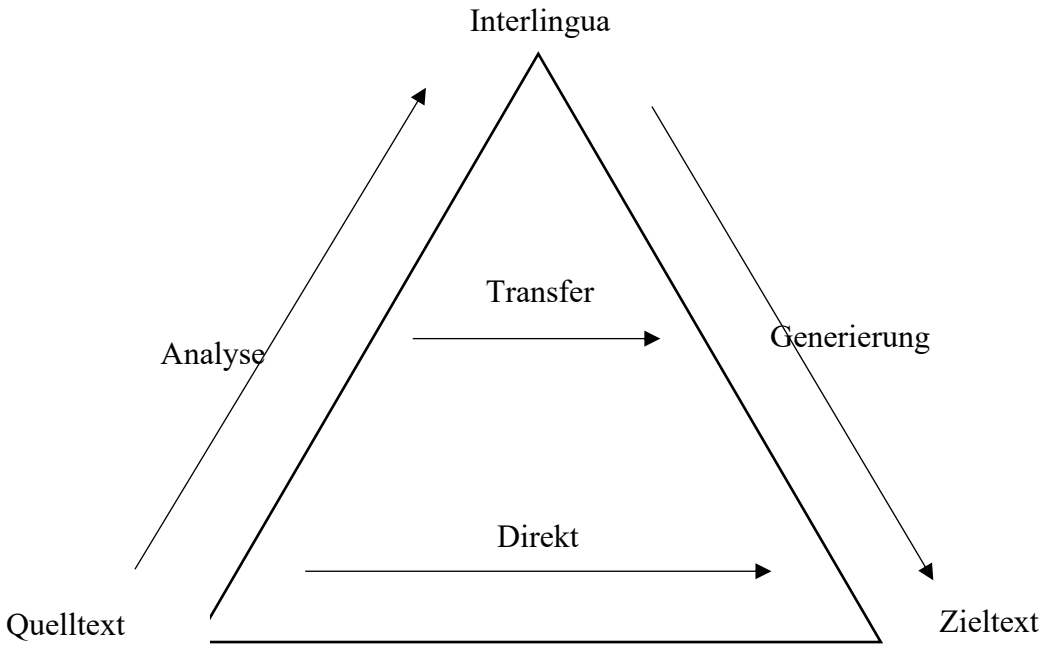

Abb. 2: Das Vasquois-Dreieck nach Schäfer (2002: 28)

Laut Stein (2009: 8) unterscheiden sich diese Methoden durch ihren Grad an Komplexität: Die direkte Methode weist den niedrigsten Grad an Komplexität auf und die Interlingua-Methode zeichnet sich durch den höchsten Grad an Komplexität aus.

4.1.1 Direkter Ansatz

Historisch gesehen ist der direkte Ansatz nicht nur der erste, sondern sozusagen auch der einfachste Ansatz in der maschinellen Übersetzung. Wie schon vom Namen abzuleiten ist, wird bei diesem Ansatz der Ausgangstext direkt in die Zielsprache übersetzt, ohne dass etwaige Zwischenschritte unternommen werden. Es findet keine syntaktische oder semantische Analyse statt, sondern nur eine morphologische, die aus Flexionsformen und Wortstämmen resultiert. Bei diesem Ansatz handelt es sich um ein reines Wortübersetzungssystem – die Mängel sind somit ohne ausführliche Erklärung naheliegend (vgl. Schäfer 2002: 26).

Um es vereinfachter auszudrücken, kann man den direkten Ansatz mit einer Person vergleichen, die nur sehr rudimentäre Grammatikkenntnisse in der Zielsprache besitzt und ein sehr schlechtes zweisprachiges Wörterbuch verwendet – es kommt häufig zu Fehlübersetzungen auf lexikalischer und syntaktischer Ebene. Die Mängel des direkten Ansatzes führten zur Entwicklung des indirekten Ansatzes (vgl. Hutchins & Somers 1992: 72 - 23).

4.1.2 Indirekter Ansatz

4.1.2.1 Interlingua-Methode

Bei der Interlingua-Methode wird der Ausgangstext zuerst analysiert und dann in einer sprachunabhängigen Form dargestellt, von der ausgehend dann der Zieltext erstellt wird (vgl. Hutchins & Somers 1992: 73; Ramlow 2009: 77).

Die Interlingua-Methode ist für mehrsprachige Systeme am attraktivsten. Zielsprachen haben hier keinen Einfluss auf den Analyseprozess – das Ziel der Analyse ist eben eine interlinguale bzw. sprachunabhängige Form. Diese Methode hat einen großen Vorteil: Es ist einfach, neue Sprachenpaare hinzuzufügen. Damit das System um neue Sprachen ergänzt werden kann, müssen lediglich zwei neue Elemente hinzugefügt werden: eines für die Analyse und eines für die Synthese der betreffenden Sprachen. Gibt es also beispielsweise schon zwei Analyse-Elemente und zwei Synthese-Elemente für die Sprachen Deutsch und Englisch, dann

liegen bereits zwei Sprachenpaare vor (Übersetzungen vom Deutschen ins Englische und vom Englischen ins Deutsche).

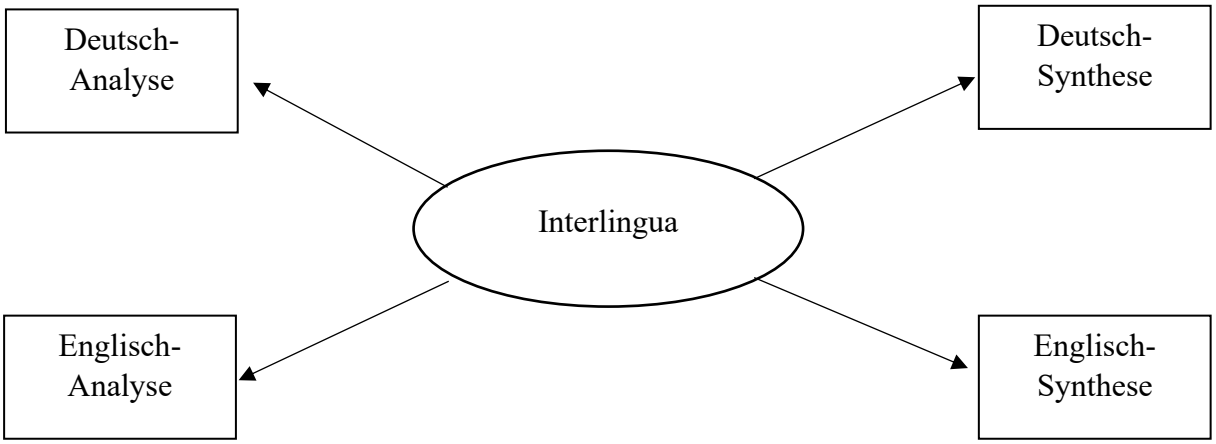

Abb. 3: Interlingua-Model mit zwei Sprachpaaren nach vgl. Hutchins & Somers (1992: 74)

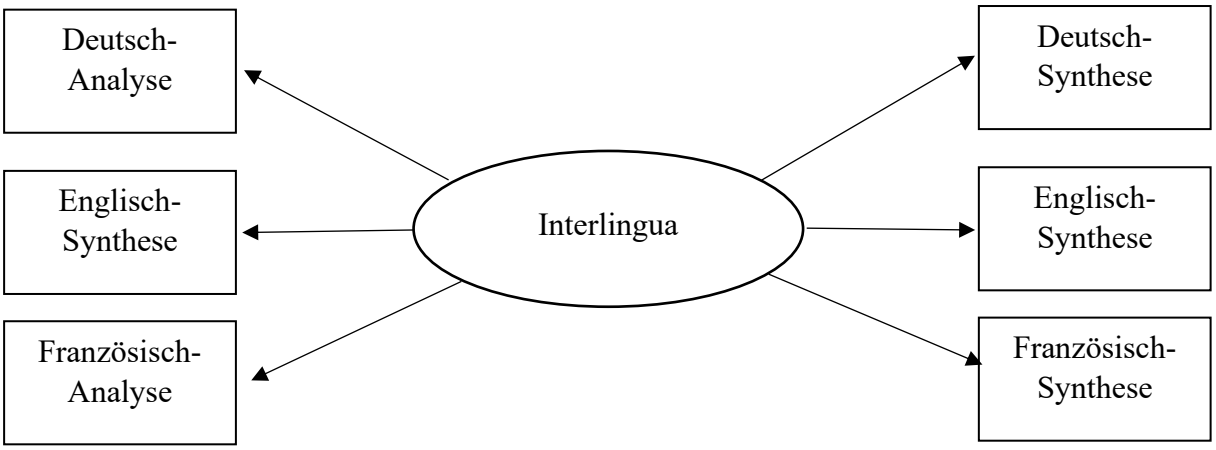

Abb. 4: Interlingua-Model mit sechs Sprachpaaren nach vgl. Hutchins & Somers (1992: 74)

Fügt man dann ein französisches Analyse-Element und ein französisches Synthese-Element hinzu, gibt es sogar sechs Sprachenpaare; es kann jetzt auch vom Französischen ins Deutsche und Englische und umgekehrt übersetzt werden (vgl. Hutchins & Somers 1992: 74; Ramlow 2009: 77).

Obwohl das Hinzufügen von Sprachenpaaren durchaus einfach erscheint (vgl. Hutchins & Somers 1992: 75) beschreibt Stein (2009: 8) sehr treffend, dass „die […] Interlingua-Übersetzung […] ein bis heute utopisches Ideal [ist], das auf der Annahme beruht, es gäbe eine universelle und völlig sprachunabhängige Art der Kodierung von sprachlichen Informationen."

Eine Universalsprache, bzw. -kodierung müsste dafür entdeckt werden, doch bis heute ist das leider noch nicht der Fall (vgl. Stein 2009: 8).

4.1.2.2 Transfer-Methode

Zum direkten Ansatz gehört auch die sogenannte Transfer-Methode, die entwickelt wurde, da sich die Interlingua-Methode ohne eine universelle Sprache als unbrauchbar erwies. Der Übersetzungsprozess bei der Transfer-Methode lässt sich in drei Phasen einteilen. Zu allererst wird der Text in der Ausgangssprache analysiert und in einer ausgangssprachlichen Repräsentation dargestellt. Im Anschluss daran wird die ausgangssprachliche Repräsentation in eine zielsprachliche Repräsentation transferiert. Die letztere Repräsentation ist dann die Basis für die Erstellung des Textes in die Zielsprache (vgl. Schäfer 2002: 27).

Ein großer Nachteil dieser Methode ist, dass die maschinelle Übersetzung nicht leichthändig um weitere Sprachen ergänzt werden kann (vgl. Schäfer 2002: 27). Zwar ist es möglich, mit diesem System in zwei verschiedene Sprachen zu übersetzen, da es sich an zweisprachigen Wörterbüchern bedient und die sogenannten Transferregeln dafür programmiert werden können. Diese Transferregeln sind jedoch häufig nicht brauchbar für eine Übersetzung in die umgekehrte Richtung und es können unter anderem lexikalische Fehler im Transfer auftreten. Deshalb wird die Transfer-Methode in der Praxis meistens für Übersetzungen in nur eine Richtung angewendet (vgl. Ramlow 2009: 77).

Als großen Vorteil des Transfer-Systems kann man hingegen die hohe Genauigkeit bestimmter sprachlicher Strukturen verzeichnen (vgl. Schäfer 2002: 27).

4.2 Statistikbasierter Ansatz

1988 wurde der statistikbasierte Ansatz von IBM-Forscher Peter Brown auf der Second TMI Conference der Carnegie-Mellon-Universität vorgestellt. Übersetzungen sollen bei diesem Ansatz mit umfangreichen Parallelkorpora – sogenannte *Translation Memorys* – erstellt werden. Zusätzlich werden Übersetzungsentscheidungen mithilfe von errechneten Wahrscheinlichkeiten getroffen (vgl. Stein 2009: 9). Es müssen genug große Textkorpora vorhanden sein, aus denen Daten extrahiert werden. Aus jenen Daten wird dann berechnet, wie wahrscheinlich es ist, dass ein Wort auf ein anderes folgt. Der zu übersetzende Ausgangstext wird in Wort- oder Satzketten segmentiert und die daraus entstandenen Segmente werden anschließend mithilfe eines zweisprachigen Textkorpus verglichen. Anschließend wird die höchste Wahrscheinlichkeit berechnet, in welcher Abfolge sich Segmente im Zieltext

aneinanderreihen und der Zieltext wird mithilfe der höchsten Wahrscheinlichkeit der Wortfolge generiert (vgl. Kurz 2014: 415-416).

Es ist äußerst unwahrscheinlich, dass man einen vollständigen Satz 1:1 im verwendeten Parallelkorpus wiederfindet. Dies ist auch der Grund, weshalb man den Satz in der statistikbasierten MÜ noch einmal in kleinere Einheiten unterteilt und diese dann genauer beleuchtet. Somit liegen auch beim statistikbasierten Ansatz zwei verschiedene Typen vor: die *wortbasierte statistische MÜ* und die *phrasenbasierte statistische MÜ* (vgl. Stein 2009: 11).

4.2.1 Wortbasierte statistische maschinelle Übersetzung

Beim ursprünglichen Typ des statistischen Ansatzes, der wortbasierten statistischen maschinellen Übersetzung, wird eine Analyse der Korpora auf der Wortebene durchgeführt. Ein Wort aus der Ausgangssprache muss somit äquivalent zu einem Wort in der Zielsprache sein. Öfters kommt es auch vor, dass ein Wort aus der Ausgangssprache in der Zielsprache mit mehreren Wörtern übersetzt werden kann. Ein Beispiel dafür ist das englische Wort *slap* und das spanische Äquivalent *dar una bofetada*. Die Übersetzung kann von der englischen Sprache in die spanische folgen, umgekehrt ist dies jedoch nicht möglich, da es für jedes Wort in der Ausgangssprache ein zielsprachliches Äquivalent geben muss. Beim Versuch einer Übersetzung vom Spanischen ins Englische würde *give a slap* herauskommen – keine einwandfreie Übersetzung. Zusammenhängende Wörter können nicht gemeinsam übersetzt werden, was ein weiteres Problem darstellt. Vor allem problematisch ist das bei mehrteiligen Verben, da es bei der Trennung dieser Verben zu einer starken Bedeutungsveränderung kommen kann (*Ich* reiste *schon nach vierzehn Tagen wieder* ab). Das gleiche Problem kommt auch bei Sprachenpaaren vor, die eine ganz unterschiedliche Syntax haben, was zum Beispiel bei der Stellung des finiten Verbes erkennbar wird (vgl. Stein 2009: 11-12).

4.2.2 Phrasenbasierte statistische maschinelle Übersetzung

Um die oben genannten Probleme lösen zu können, wurden im Laufe der Zeit neue Ansätze entwickelt. Heutzutage arbeiten bewährte MÜ-Systeme auf der Phrasenebene. Hier sind Phrasen jedoch nicht als ganze Sätze gemeint, sondern als Wortgruppen. Die phrasenbasierte statistische maschinelle Übersetzung ermöglicht die Übersetzung von mehreren Wörtern in beide Sprachrichtungen. Zudem können sprachliche Ambiguitäten durch den erweiterten Kontext entfallen, was ein weiterer Vorteil ist. Ein wortbasiertes System könnte nicht erkennen, wie das englische Wort *pretty* zu übersetzen wäre, wenn es in den Kombinationen *pretty girl*

oder *pretty much* vorkäme, doch ein phrasenbasiertes System schon. Von System zu System kann die Größe von Wortgruppen verschieden ausfallen. Je nachdem, welches System benutzt wird und wie groß die Wortgruppen sind, kann die ungleiche Syntax in Ausgangs- und Zielsprache überbrückt werden (vgl. Stein 2009: 12).

4.3 Maschinelle Übersetzung basierend auf neuronalen Netzen

Schwerpunkt dieser Arbeit ist die Analyse der Ergebnisse der neuronalen Übersetzung von Fachtexten. An dieser Stelle wird deshalb die MÜ basierend auf neuronalen Netzen vorgestellt.

Die neuronale maschinelle Übersetzung (Neural Machine Translation – NMÜ) war 2016 ein viel diskutiertes Thema, da sich jenes als das Jahr erwies, in dem die NMÜ den Durchbruch schaffte und zu einer viel bekannteren Technologie wurde, die auch außerhalb der Forschungsgemeinschaft an Bedeutung gewann. Einer der Auslöser für die Welle der Anerkennung für NMÜ war die Ankündigung des Unternehmens Facebook, Anfang 2016 vom strukturbasierten Ansatz zum neuronalen Ansatz zu wechseln. Ein weiterer Auslöser bestand in einem Google im September 2016 veröffentlichten Paper, in dem das Unternehmen neue NMÜ-Systeme ankündigte und behauptete, dass sich Humanübersetzungen und Übersetzungen, die mit Google-Systemen erstellt wurden, kaum unterschieden und dass die Übersetzungsqualität ihrer Systeme alle anderen zu jener Zeit gängigen Systeme übertraf (vgl. Vashee 2017: 44).

Der Grund, warum die NMÜ einen derartigen Überschwang auslöste, liegt auf der Hand: Der bis dato entwickelte statistikbasierte Ansatz konnte einige Jahre lang zwar weiterentwickelt werden, doch irgendwann war das Maximum erreicht, und erlangte man doch noch Verbesserungen, waren diese zu klein. Deshalb konzentrierten sich ForscherInnen auf neue Ansätze des maschinellen Lernens (große Datenmengen werden genutzt, um Computermodelle zu trainieren, bestimmte Aufgaben auszuführen [vgl. Donovan 2018]), indem sie erfolgreich neuronale netzwerkbasierte Ansätze verwendeten. Die Methode *Deep Learning* ist dabei sehr wichtig – es revolutioniert Sprachtechnologien, da es eine effektive Möglichkeit zur Verfügung stellt, um Systeme zu trainieren und damit signifikante Verbesserungen zu erzielen. Der Hauptvorteil von Deep Learning ist, dass es automatisch Merkmale aus Rohdaten lernt, ohne dass diese explizit extrahiert oder dargestellt werden müssen (vgl. Costa-jussà et al. 2016: 1).

Seit der Verwendung des Internets wächst die globale Datenmenge von Jahr zu Jahr exponentiell. Dieser Zuwachs führt dazu, dass Unternehmen unglaublich große Mengen an

Daten verarbeiten müssen, damit sie sich diese auch effektiv zu Nutzen machen können. In der Verarbeitung der Daten steckt zwar eine große Herausforderung, doch durch Übersetzungs- und Lokalisierungstechnologien stehen Unternehmen auch Möglichkeiten zur Verfügung, die sie sonst nicht hätten. Die enorme Datenmenge kann durch die vorhergenannten Technologien für die mehrsprachige Kommunikation verwendet werden, wodurch verhältnismäßig schnell auf globalen Märkten agiert werden kann (vgl. Lionbridge Marketing 2019).

Der Paradigmenwechsel innerhalb der Forschung im Bereich der maschinellen Übersetzung hat ausgezeichnete Ergebnisse hervorgebracht, indem es den Stand der Technik für einige Sprachenpaare verbessert hat. Diese neue Forschungsrichtung bereitete zudem nicht nur den Weg für neue Forschungsperspektiven, sondern skizziert auch neue MÜ-Herausforderungen, zum Beispiel in Bezug auf große Vokabulare, multimodale Übersetzungen und hohe Kosten für die stets steigende benötigte Rechenleistung (vgl. Costa-jussà et al. 2016: 2).

4.3.1 Wie funktioniert die NMÜ?

Um überhaupt verstehen zu können, wie die NMÜ funktioniert, muss zuerst erklärt werden, was ein rekurrentes neuronales Netzwerk (RNN) ist: RNN ist ein neuronales Netzwerk (Guo 2017: 5), das, ähnlich wie die menschliche Kognition, Vorhersagen machen kann und ein Gedächtnis hat (vgl. Haselhuhn 2018: 4). Es besteht aus einer Input-Ebene (x), einer versteckten Ebene (h) und einer Output-Ebene (y):

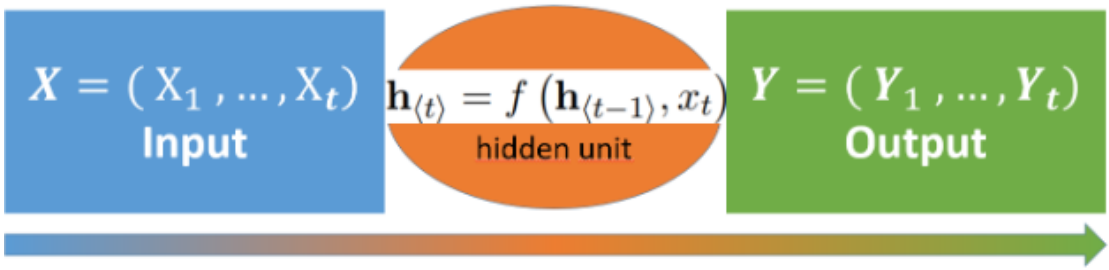

Abb. 5: Schematische Darstellung eines RNNs (Guo 2017: 5)

Die Grundstruktur eines RNNs ist in Abbildung 5 dargestellt: $x = (x_1, \ldots, x_t)$ ist die Inputebene, $h_{<t>} = f(h_{<t-1>}, x_t)$ ist die versteckte Ebene und $y = (y_1, \ldots, y_t)$ stellt die Output-Ebene dar (vgl. Guo 2017: 5). Die Information aus der vorherigen versteckten Ebene ($h_{<t-1>}$) dient als Input für die nächste versteckte Ebene ($h_{<t>}$), weshalb man hier oft von einer Art Gedächtnis vorheriger Zustände spricht. In der Output-Ebene werden Wahrscheinlichkeiten für das nächste Wort in

der Folge der Übersetzung ausgegeben, welche zu Anfang unzufriedenstellend sind, aber durch viele Wiederholungen des Trainingsvorgangs graduell verbessert werden. Am Ende ist das System in der Lage, das nächste Wort mit einer meist hohen Präzision vorherzusagen, auch bei noch nicht gesehenen Sätzen, da es Muster in den jeweiligen Sprachen lernt.

Die meisten neuronalen MÜ-Architekturen basieren auf dem sogenannten *Encoder-Decoder*, womit eine Eingabesequenz in der Ausgangssprache in einen niedrigdimensionalen Raum projiziert wird, aus dem die Ausgabefolge in der Zielsprache generiert wird (vgl. Costa-jussà et al. 2016: 3). Vaswani et al. (2017: 1) entwickelten eine neue, vereinfachte Netzwerkarchitektur, den Transformer, der ausschließlich auf Attention-Mechanismen basiert und gänzlich ohne Rekurrenzen und Convolutions auskommt.

Es gibt viele verschiedene Möglichkeiten, um den Encoder zu konzipieren. Ein erster Ansatz war die Verwendung eines einfachen neuronalen Netzwerks zur Codierung der Eingabesequenz. Das Komprimieren der Sequenz in einen Vektor mit fester Größe scheint jedoch zu eingeschränkt gewesen zu sein, um Informationen aus der Ausgangssprache beizubehalten. Aus diesem Grund wurden neue Systeme mit bidirektionalem RNN entwickelt. Quellsequenzen werden dabei in Annotationen codiert, indem die zwei erhaltenen Repräsentationen mit einer Vorwärts- bzw. Rückwärts-RNN verkettet werden. In diesem Fall enthält jeder Annotationsvektor Informationen aus der gesamten Quellsequenz, konzentriert sich jedoch auf ein bestimmtes Wort. Ein Attention-Mechanismus, der von einem vorwärtsgerichteten neuronalen Netzwerk implementiert wird, wird dann verwendet, um bestimmte Teile der Eingabe zu überwachen und eine Ausrichtung zwischen Eingabe- und Ausgabesequenz zu erzeugen (vgl. Costa-jussà et al. 2016: 3).

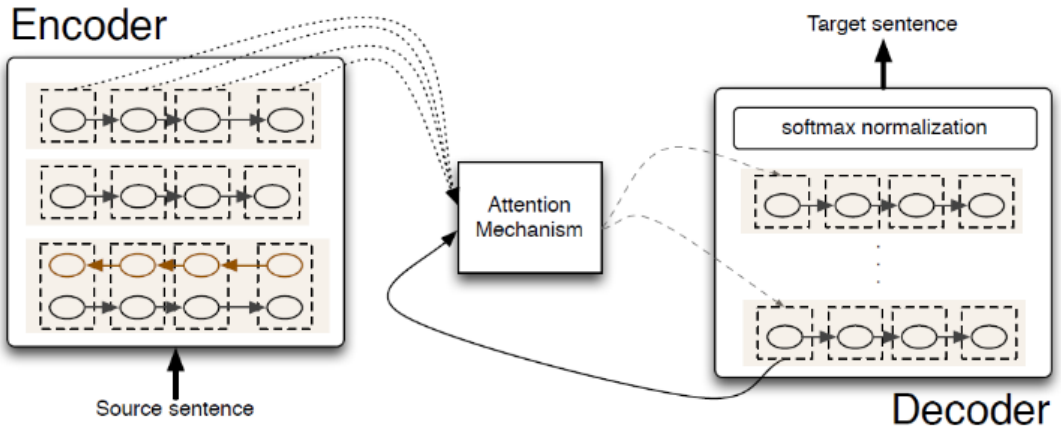

Abb. 6: Standardarchitektur von neuronalen maschinellen Übersetzungssystemen (Costa-jussà et al. 2016: 4).

4.3.2 Nachteile und Kritik an der NMÜ

Auch wenn die NMÜ unzählige Vorteile mit sich bringt, sollen an dieser Stelle auch die Nachteile und die Kritik an diesem Ansatz zusammengefasst dargestellt werden, da der neueste Übersetzungsansatz von Perfektion bzw. Humantranslationsäquivalenz noch weit entfernt ist.

NMÜ-Systeme zu trainieren ist ein sehr teurer Prozess, sowohl in Bezug auf die Rechenressourcen (spezielle Prozesstechnologien sind erforderlich) als auch in Bezug auf die Zeit (ein System wird wochenlang trainiert). Des Weiteren sind riesige Datenmengen erforderlich, um ein NMÜ-System trainieren zu können, vor allem die Menge an parallelen Daten, die bei überwachten Ansätzen erforderlich sind. Außerdem besitzen NMÜ-Systeme einen begrenzten Wortschatz, weshalb die Anzahl an unbekannten Wörtern beim Output sehr hoch ist. Darüber hinaus lassen sich NMÜ-Systeme in Bezug auf Terminologie und Stil nicht einfach an die spezifischen Anforderungen der NutzerInnen anpassen. NMÜ-Systeme sind in ihren theoretischen Grundlagen einfacher zusammengesetzt, aber es herrscht ein hohes Maß an Intransparenz darüber, wie und warum sie funktionieren. Dies macht es schwierig zu bestimmen, wie eine etwaige Fehlerbehebung vonstattengehen soll (vgl. Vashee 2017: 55).

5 Menschliche Evaluationsmethoden

Auch wenn automatische Evaluationsmethoden mittlerweile existieren, werden heute immer noch mehrheitlich menschliche Evaluationsmethoden verwendet. Dafür gibt es zwei wichtige Gründe. Zum einen sind Übersetzungen für Menschen gedacht; nur sie können also entscheiden, welche Qualitätsparameter wichtig sind. Des Weiteren ist es demnach nur für Menschen möglich, die Korrektheit einer Übersetzung festzustellen (vgl. Lo Presti 2016: 25).

Menschliche Evaluationsmethoden sind jedoch zu subjektiv und zu zeitaufwendig. Heutzutage wird immer noch auf sie zurückgegriffen, da sie trotz der Kritik nach wie vor am verlässlichsten sind. Heutige menschliche Evaluationsmethoden sind vielfältig und die Qualität der Übersetzungen kann durch direkte oder indirekte Methoden gemessen werden (vgl. Lo Presti 2016: 25). Im Folgenden wird ein kurzer Überblick über die existierenden und am häufigsten verwendeten Methoden gegeben, da man sich in dieser Arbeit der Fehleranalyse bedient.

5.1 *Adequacy* und *fluency*

Nach Callison-Burch et al. (2007: 141) ist *adequacy* und *fluency* die am häufigsten angewandte Methode. Bei dieser Methode gibt es zwei Fünf-Punkte-Skalen: eine für *adequacy* (Übereinstimmung bzw. Adäquatheit) und eine für *fluency* (Flüssigkeit). Mit der Skala für *adequacy* wird bewertet, wie viel der Bedeutung des Ausgangstextes auch im Zieltext vorkommt:

5 = alles
4 = das Meiste
3 = viel
2 = wenig
1 = keine

Die zweite Fünf-Punkte-Skala für *fluency* zeigt an, wie fließend die Übersetzung ist. Ist die Zielsprache Englisch, sieht die Skala folgendermaßen aus:

5 = makelloses Englisch
4 = gutes Englisch
3 = nicht-muttersprachliches Englisch
2 = kein fließendes English
1 = unverständlich

Es wurden zwei separate Skalen für *adequacy* und *fluency* entwickelt, da die Annahme besteht, dass eine Übersetzung zwar nicht fließend sein kann, jedoch trotzdem alle Informationen aus dem Ausgangstext beinhalten kann. Ein weiteres Problem bei dieser Methode ist, dass es keinen klaren Leitfaden gibt. EvaluatorInnen erhalten keine Anweisungen zur Quantifizierung von Bedeutung oder dazu, wie viele oder welche Grammatikfehler die verschiedenen Ebenen der fluency trennen (vgl. Callison-Burch et al. 2007: 141).

5.2 Ranking

Das Ranking wird normalerweise in Forschungskontexten und zur vergleichenden Bewertung der Outputs von verschiedenen MÜ-Systemen verwendet, die aus demselben Ausgangstext stammen. EvaluatorInnen werden gebeten, die angegebenen Sätze in der Zielsprache anhand gegebener Kriterien oder allgemeiner Konzepte, z. B. *fluency*, in eine Reihenfolge zu bringen. Der Satz in der Ausgangssprache wird bereitgestellt und zwei oder mehr MÜ-Outputs werden anonym aufgelistet, normalerweise in einer zufälligen Reihenfolge. Die EvaluatorInnen wählen dann entweder die ihrer Meinung nach beste Übersetzung aus oder ordnen die Optionen für die Zielsprache nach bestimmten Kriterien von der besten bis zur schlechtesten Übersetzung. Dieser Ansatz ist sehr beliebt, da er den Vorteil hat, einfach zu interpretierende Ergebnisse zu liefern und gleichzeitig relativ schnell und effizient zu sein (vgl. Castilho et al. 2018: 21).

5.3 Fehleranalyse

Mit den vorherigen Methoden wird bewertet, wie gut ein maschinelles Übersetzungssystem übersetzt hat. Bei der Fehleranalyse geht es im Gegensatz dazu darum zu messen, wie schlecht übersetzt worden ist. Mithilfe dieser Methode möchte man versuchen, die Stärken und Schwächen eines Systems herauszufinden. Diese Methode gilt als objektiver als jene Methoden, die Bewertungsskalen verwenden; zusätzlich ist die Inter-Rater-Übereinstimmung, also der Grad der Übereinstimmung der Evaluation zweier oder mehrerer EvaluatorInnen, höher (vgl. Lo Presti 2016: 31).

In den letzten Jahren wurden bereits einige Fehlerkategorisierungen (wie die SAE-J2450-Qualitätsmetrik, siehe Kapitel 5.3.1) entwickelt. Eine der aktuellsten ist die Fehlerkategorisierung von Vilar et al. (2006), in der die Kategorien hierarchisch aufgebaut sind:

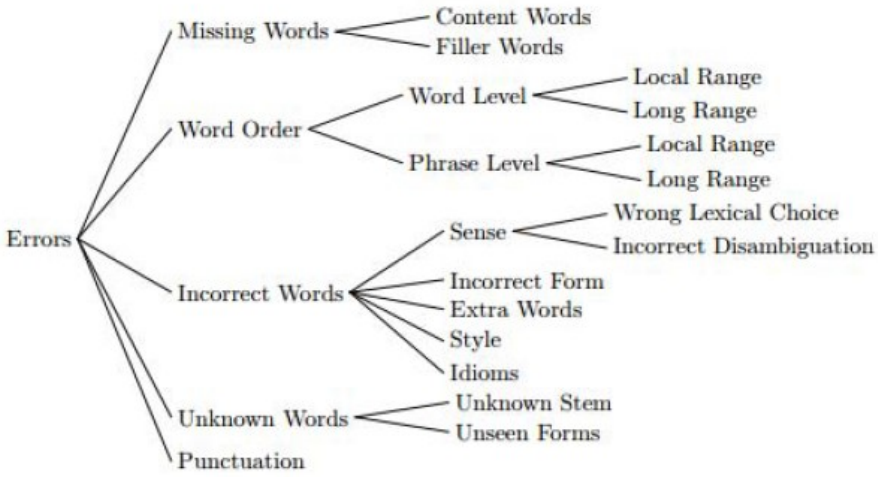

Abb. 7: Fehlerkategorisierung nach Vilar et. al (2006: 699).

Als Schwäche dieser Fehleranalyse gilt, dass auch sie immer noch zu subjektiv ist. Nicht alle EvaluatorInnen interpretieren Fehler auf die gleiche Art und Weise oder weisen die gleiche Toleranz bei Fehlern auf. Des Weiteren kann wie bei der SAE-J2450-Qualitätsmetrik auch bei anderen Fehlerkategorisierungen ein Fehler zu mehreren Kategorien gehören. Dann fällt die Entscheidung schwer, besonders, wenn keine eindeutigen Regeln gegeben sind. Hinzu kommt, dass Fehler in (Fach-)Sprachen und Textsorten oft anders eingeteilt werden: Zum Beispiel werden in journalistischen Texten Artikel häufig ausgelassen, während dies bei anderen Textsorten als grober Fehler betrachtet wird (vgl. Lo Presti 2016: 32).

5.3.1 SAE-J2450-Qualitätsmetrik

Die SAE-J2450-Qualitätsmetrik wurde 1997 von der Organisation *Society of Automotive Engineers (SAE),* einer globalen Vereinigung von IngenieurInnen sowie technischen ExpertInnen in den Bereichen Luft- und Raumfahrt, Automobilindustrie und Nutzfahrzeugindustrie entwickelt. Das Ziel bestand darin, einen einheitlichen Standard festzulegen, anhand dessen die Qualität der Übersetzung von Kfz-Serviceinformationen objektiv gemessen werden kann. Besagter Standard sollte unabhängig von der Ausgangssprache, der Zielsprache und der Art und Weise, in der die Übersetzung durchgeführt wird (menschliche oder maschinelle Übersetzung) sein. Die SAE-J2450-Qualitätsmetrik wurde von General Motors, Ford und Chrysler gemeinsam entwickelt und im August 2001 als

empfohlene Vorgehensweise veröffentlicht. Sie hat sich in der Automobilindustrie als äußerst erfolgreich erwiesen und ist weltweit anerkannt (vgl. Hui 2017: 57).

Aufgrund der starken Kritik europäischer Automobilfirmen und ihrer Übersetzungsteams am ersten Entwurf wurde jedoch Ende 2001 ein europäisches SAE-2450-Komitee gegründet, das für die Verbesserung des Entwurfs verantwortlich war. Mitglieder dieses Ausschusses waren die SprachmitarbeiterInnen aus Unternehmen wie Volvo Trucks, Daimler-Chrysler, Audi, Volkswagen sowie aus anderen europäischen Übersetzungsbüros. J2450 wurde nach dem Verhandlungsprozess zwischen Europa und den USA im Jahr 2005 endgültig zur SAE-Norm. Die SAE hat im Mai 2008 ein ergänzendes Schulungsdokument herausgegeben, das „sowohl Kunden als auch Übersetzungsanbietern einige Vorschläge zur Verfügung stellt“ (vgl. Hui 2017: 57-58).

5.3.2 Fehlerarten und Schweregrade

Der Ansatz der SAE-J2450-Qualitätsmetrik (2001) für die Beurteilung von Übersetzungsqualität ist durchaus simpel. Er basiert auf sieben Arten von Fehlern mit den folgenden Schweregraden:

1. Falsche Benennung (5/2)
2. Falsche Bedeutung (5/2)
3. Auslassung (4/2)
4. Strukturfehler (4/2)
5. Rechtschreibfehler (3/1)
6. Interpunktionsfehler (2/1)
7. Sonstige Fehler/Diverses (3/1)

Fehler in jeder Kategorie können entweder als schwerwiegend oder geringfügig eingestuft werden, wobei jedem Fehler und Schweregrad – weniger schwerwiegend (minor) oder schwerwiegend (major/serious) – eine numerische Bewertung zugeordnet wird. Die zusammengesetzte Punktzahl, die über die Übersetzungsqualität eines Textes entscheidet, ist die gewichtete Summe der Fehler, die durch die Anzahl der Wörter im Ausgangstext normiert wird (vgl. Hui 2017: 58). Diese Einschätzungen sind jedoch schwierig und müssen trotz allem subjektiv getroffen werden, da sie von jeweiligen EvaluatorInnen abhängen. Es bestehen keine festen Kriterien, nach denen sich die PrüferInnen orientieren können; es gibt nur eine

Anweisung, einen Fehler auch als schwerwiegend einzustufen, wenn die Entscheidung für einen Fehler nicht leichtfällt (vgl. Ottmann 2017: 39). Nachstehend ist eine detaillierte Beschreibung der Fehlerkategorien und der Schweregrade.

Laut der SAE-Metrik kann sich eine Benennung bzw. ein Begriff auf ein einzelnes Wort, eine Mehrwortphrase, die als einzelner oder lexikalischer Bestandteil verwendet wird, eine Abkürzung, ein Akronym, eine Zahl oder eine Ziffer oder einen Eigennamen beziehen, sei es ein Handelsname, ein Markenname, eine eingetragene Marke, ein Ortsname oder ein persönlicher Name. Eine falsche Benennung bzw. ein falscher Begriff kann ein Begriff in einer beliebigen Zielsprache sein, der: 1) gegen ein Glossar einer KundIn verstößt; 2) in klarem Widerspruch zu den gegenwärtigen Standardübersetzungen des Begriffs der Ausgangssprache im Automobilbereich steht; 3) in der Übersetzung nicht konsistent mit anderen Übersetzungen des Ausgangssprachenbegriffs im selben Dokument verwendet wird, außer der Kontext kann die Verwendung eines anderen Zielsprachenbegriffs rechtfertigen beispielsweise aufgrund der Mehrdeutigkeit des Ausgangssprachenbegriffs; 4) ein Konzept in der Zielsprache bezeichnet, das sich offensichtlich und signifikant von dem Konzept unterscheidet, das durch den Begriff der Ausgangssprache bezeichnet wird. Ein schwerwiegender falscher Begriff bekommt eine Gewichtsbewertung von 5, während ein geringfügiger falscher Ausdruck eine Gewichtsbewertung von 2 ergibt (vgl. Hui 2017: 58-59).

Bei der Fehlerkategorie „falsche Bedeutung“ handelt es sich um inhaltliche Fehler, die in Texten vorkommen können. Der Zieltext beinhaltet eine eindeutige Abweichung der Bedeutung im Ausgangstext (vgl. Dalla-Zuanna 2010: 23). Die Gewichtung für einen schwerwiegenden inhaltlichen Fehler beträgt 5, die für einen geringfügigen inhaltlichen Fehler 2.

Wird etwas aus der Ausgangssprache in der Zielsprache ausgelassen, so stellt dies in der SAE-Qualitätsmetrik einen Auslassungsfehler dar. Folgende Auslassungen können zutreffen: ein Satz, ein Wort, ein Segment oder ein ganzer Absatz, auch als Bestandteil einer Grafik (vgl. Dalla-Zuanna 2019: 24). Gleichzeitig ist hervorzuheben, dass das Auslassen nicht bedeutet, dass die Wörter der Aus- und Zielsprache 1:1 vorliegen sollten. Eine schwerwiegende Auslassung wird mit einer 4 und eine geringfügige wird mit einer 2 bewertet.

Ein Strukturfehler kann Folgendes sein: Ein ansonsten korrektes Zielsprachenwort (oder ein Ausdruck) wurde in einer falschen morphologischen Form ausgedrückt, beispielsweise Groß- oder Kleinschreibung, Geschlecht, Anzahl, Zeitform, Präfix, Suffix oder sonstige Beugungen (vgl. Hui 2017: 59). Mit dieser Kategorie werden Semantik-, Grammatik- und Syntaxfehler abgedeckt (vgl. Dalla-Zuanna 2019: 24). Die Gewichtung für einen schwerwiegenden Strukturfehler beträgt 4 und für einen geringfügigen Fehler 2.

Ein Rechtschreibfehler liegt vor, wenn ein Begriff in der Zielsprache 1) die in einem Kundenglossar angegebene Rechtschreibung verletzt; 2) gegen die anerkannten Normen für die Rechtschreibung in der Zielsprache verstößt; 3) in einem falschen oder ungeeigneten Schriftsystem für die Zielsprache geschrieben ist. Wenn ein Rechtschreibfehler als schwerwiegend eingestuft wird, erhält er eine Bewertung von 3; wenn ein solcher Fehler als geringfügiger Fehler betrachtet wird, wird er mit 1 bewertet (vgl. Hui 2017: 59-60).

Ein Interpunktionsfehler besteht dann, wenn die Zeichensetzung in einem Zielsprachentext gegen die Interpunktionsregeln der betreffenden Sprache verstößt. Die Gewichtung für einen schwerwiegenden Interpunktionsfehler beträgt 2 und die für einen geringfügigen Fehler 1. Ein schwerwiegender Interpunktionsfehler liegt vor, wenn in der Übersetzung ein Dezimalpunkt oder ein Komma fehlt (z. B. ‚116' anstelle von ‚11.6' oder ‚11, 6'. Das Fehlen des Dezimalpunktes oder des Kommas wird hier immer als Interpunktionsfehler angesehen und nicht als Auslassung (vgl. Hui 2017: 60).

Ein sonstiger Fehler ist ein sprachlicher Fehler, der sich auf den Zielsprachentext bezieht, jedoch keiner Fehlerkategorie eindeutig zugeordnet werden kann. In der SAE-Qualitätsmetrik wird betont, dass Stilprobleme nicht in diese Kategorie aufgenommen werden sollten, da Stilfehler in der Qualitätsmetrik ignoriert werden. Ein schwerwiegender sonstiger Fehler wird mit 3 bewertet und ein kleinerer Fehler mit 1 (vgl. Hui 2017: 60). An dieser Stelle soll jedoch angemerkt werden, dass während der Evaluierung der Übersetzungen wirtschaftlicher Fachtexte für diese Arbeit der Stil trotzdem in der Kategorie „sonstige Fehler/Diverses" angemerkt worden ist, da Stilfehler oft den Hauptmangel bei maschinell übersetzten Texten darstellen.

Aus der Beschreibung der einzelnen Fehlerkategorien lässt sich erkennen, dass zum Beispiel ein inhaltlicher Fehler mit einer höheren Gewichtung (5) angesetzt ist und somit auch schwerer wiegt als ein Interpunktionsfehler mit Maximalgewichtung von 3 (vgl. Ottmann 2017: 39).

5.4 Leseverständnistest

Das Verstehen von Texten ist ausschlaggebend, um an Informationen aus einem Text zu erlangen. Aus diesem Grund ist der Grad der Verständlichkeit eines mittels MÜ übersetzten Textes ein entscheidender Faktor für die Qualitätsmessung eines MÜ-Systems. Es wurden daher Methoden entwickelt, deren Grundlage auf bewährten Tests für das Leseverständnis liegen. EvaluatorInnen lesen zuerst MÜ-Texte und beantworten Fragen dazu. Anschließend

lesen sie zusätzlich die Humanübersetzung und bekommen dieselben Fragen noch einmal gestellt. Je kleiner der Unterschied zwischen den beiden Übersetzungen ist, desto besser wurde der Text also vom MÜ-System übersetzt (vgl. Lo Presti 2016: 33).

5.5 Post-Editing

Post-Editing wird bei weitem am häufigsten als Aufgabe im Zusammenhang mit MÜ in Verbindung gebracht (vgl. Allen 2003: 297). Es wird als "term used for the correction of machine translation output by human linguists/editors" definiert (Allen 2003: 297). Das Post-Editing war schon seit den Anfängen der maschinellen Übersetzung unabdinglich. Noch heute spielt die Nachbearbeitung eine große Rolle, da maschinelle Übersetzungssysteme bis heute noch nicht das Niveau von Humanübersetzungen erreicht haben (vgl. Allen 2003: 297; Lo Presti 2016: 34).

Am Post-Editing kann auch die Qualität von MÜ-Texten bestimmt werden, da hier die folgende Idee zugrunde liegt: Je besser ein Text von einem MÜ-System übersetzt worden ist, desto weniger muss ausgebessert werden. Die Qualität hat somit unmittelbaren Einfluss auf den Preis der Übersetzung, die durch das Post-Editing finalisiert wird (vgl. Lo Presti 2016: 34).

Die Methode der Nachbearbeitung ist selbstredend subjektiv. Post-EditorInnen haben jeweils ihr eigenes Tempo, mit dem sie Texte überarbeiten. Außerdem kennen sie sich in unterschiedlichen Themen bzw. Fachbereichen individuell besser oder schlechter aus (vgl. Lo Presti 2016: 34).

6 Automatische Evaluationsmethoden

Da bei den menschlichen Evaluationsmethoden den Hauptmangel der Subjektivität vorliegt, und zusätzlich viel Zeit und Geld investiert werden müssen, wurden automatische Evaluationsmethoden entwickelt. Mit der Zeit benötigte man in der Wissenschaft Methoden, die schneller, einheitlicher und kostengünstiger waren. Die heute gängigsten Methoden sind die Folgenden: *BLEU, METEOR, WER, TER, HTER* und *TERp* (vgl. Lo Presti 2016: 37).

BLEU (Bilingual Evaluation Understudy) ist die beliebteste automatische Evaluationsmetrik. Hier werden N-Gramme (Sequenzen von aufeinanderfolgenden Wörtern) verschiedener Längen gezählt, die im Output der MÜ, aber auch in verwendeten Referenzübersetzungen vorzufinden sind. Bei BLEU besteht die Neuerung darin, dass nicht nur eine Referenzübersetzung, sondern mehrere verwendet werden können (vgl. Lo Presti 2016: 39; Koehn 2010: 226).

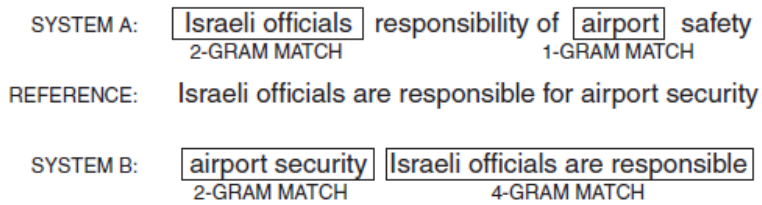

Abb. 8: BLEU basiert auf N-Gramm-Übereinstimmungen mit Referenzübersetzungen (vgl. Koehn 2010: 26)

In der vorliegenden Arbeit wird jedoch mit der Fehleranalyse eine menschliche Evaluationsmethode verwendet, weshalb an dieser Stelle nicht weiter auf automatische Evaluationsmethoden eingegangen wird. Snover et. al (2009) und Argarwal & Lavie (2008) bieten einen guten Überblick über die anderen hier erwähnten Methoden.

7 Aktueller Stand der Forschung

Seit ihren ersten Schritten hat sich die MÜ deutlich weiterentwickelt, wie in den obigen Kapiteln gezeigt wurde. Unter den unterschiedlichen Konzepten bzw. Übersetzungsmethodensticht die NMÜ heraus (Sennrich et al. 2016). Studien zeigen sogar, dass sie anderen maschinellen Übersetzungsmethoden qualitativ überlegen ist (Bentivogli et al. 2016; Volkart et al. 2018). De-facto ist die NMÜ mittlerweile der Standard bei maschineller Übersetzung (vgl. Läubli et al. 2018: 4791).

Aktuelle Vertreter von MÜ mit neuronalen Netzwerken sind dabei beispielsweise *Google Translate* und *DeepL*. Dabei ist die Entwicklung dieser Programme mittlerweile soweit vorangeschritten, dass die Meinung vertreten wird, die Qualität der MÜ werde bald an die der humanen Übersetzung heranreichen (vgl. Burchardt & Porsiel 2017: 11-12). Auch die Qualität der verschiedenen Programme untereinander unterscheidet sich. Macketanz et al. (2018) zeigen in ihrer Arbeit beispielsweise auf, dass DeepL bessere Ergebnisse als Google Translate liefert.

Aber wie gut ist die MÜ wirklich? Zu dieser Frage lassen sich einige Publikationen in der Literatur finden, die die Qualität der MÜ mit der Qualität der humanen Übersetzung vergleichen (Ahrenberg 2017; Hutchins 2001; Läubli et al. 2018). Bisher lässt sich keine Überlegenheit der MÜ gegenüber Humanübersetzungen feststellen. Des Weiteren gewähren mehrere Studien Einblicke in die Evaluierung der Qualität der MÜ (Grundvall 2019; Shafia 2018; Wiesmann 2019; etc.). So weist Grundvalls (2019) Arbeit Parallelen zum Studiendesign der vorliegenden Masterarbeit auf.

Die Texte, die in den meisten Studien übersetzt wurden, um die Qualität der MÜ zu evaluieren, entstammten hauptsächlich aus dem rechtlichen oder technischen Bereich. Studien, die MÜ mit Wirtschaftsfachtexten evaluieren, wurden von der Autorin während ihrer Recherche ebenso wenig gefunden als auch Arbeiten, die unterschiedliche Textstufen einer Textgattung untersuchen.

Für die Evaluierung maschinell übersetzter Texte gibt es, wie bereits diskutiert, zahlreiche Methoden. Es werden automatische, aber auch immer noch menschliche Evaluationsmethoden angewandt. Ahrenberg (2017) verwendete für seine Evaluation die gängigste automatische Evaluationsmethode – den BLEU-Score. Grundvall (2019) führte ihre Studie mit dem Qualitätsmodel von House (1977) durch.

Diese zwei Evaluationsmethoden werden mitunter sogar kombiniert, beispielsweise von Koehn & Monz (2006). Hier wurden Texte aus den Sprachen Spanisch, Französisch und Deutsch ins Englische und umgekehrt übersetzt und anschließend mit der beliebtesten

automatischen Evaluationsmethode, dem BLEU-Score, und der am häufigsten verwendeten menschlichen Evaluationsmethode *adequacy* und *fluency* evaluiert.

8 Methodik

Zu Beginn wird an dieser Stelle kurz der Ablauf dieser Studie beschrieben. Der erste Schritt bestand in der Recherche und Selektion von neun Wirtschaftsfachtexten aus verschiedenen Medien. Diese wurden anschließend in drei Stufen eingeteilt (3, 2, 1) und gekürzt, worauf die Übersetzung mit dem DeepL-Übersetzungsprogramm (DeepL 2019) vom Englischen ins Deutsche erfolgte. Danach wurden die Übersetzungen mithilfe der SAE-J2450-Qualitätsmetrik von drei Personen annotiert (eine davon war die Autorin der vorliegenden Masterarbeit), die Ergebnisse wurden ausgewertet und die Übereinstimmung der Annotatorinnen wurde mit Cohen's Kappa nach Fleiss verglichen. Abbildung 9 bietet eine kurze grafische Zusammenfassung.

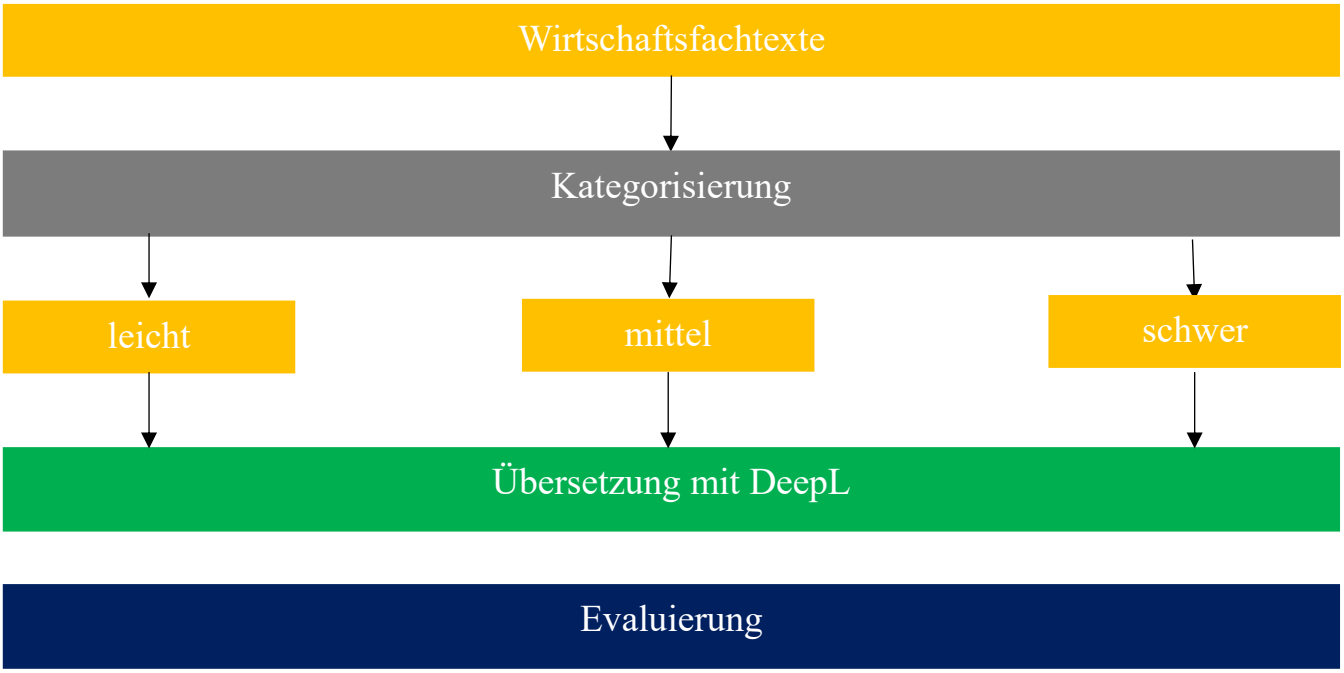

Abb. 9: Schema des methodischen Ablaufs

8.1 Auswahl der Texte

Die Entscheidung, für die vorliegende Arbeit Wirtschaftsfachtexte zu verwenden, fußte auf mehreren Gründen. Ein Grund ist das große Interesse der Autorin an dem komplexen Themenbereich der Wirtschaft, ein anderer die Verfügbarkeit von Wirtschaftsfachtexten für verschiedene Grade der Fachsprachlichkeit besonders umfassend. Des Weiteren ist die Verfügbarkeit von AnnotatorInnen für diesen Bereich auch leicht gegeben. Zuletzt ist der Wirtschaftsbereich ein Bereich, der unabhängig von individuellem Interesse und Wissensstand

alle Menschen tangiert, weshalb eine Vielzahl an Medien zur Verfügung stehen, die Wirtschaftsthemen abdecken und die Suche nach Wirtschaftstexten erleichtern.

Die Anzahl der Texte wurde aufgrund des Kapazitätsrahmens der Masterarbeit auf neun Stück limitiert. Alle Texte sind zwischen den Jahren 2013 und 2019 erschienen und wurden auf maximal ca. 150 Wörter gekürzt, wobei manche wesentlich kürzer sind, wie beispielsweise Text 9 mit 66 Wörtern. Die Kürzung wurde vorgenommen, da die Texte nur als Stichproben dienen sollten und andernfalls den Rahmen der Masterarbeit gesprengt hätten. Die gekürzten Versionen der Texte sind dem Anhang zu entnehmen. Die ausgewählten Themen wurden aufgrund von Aktualität, Interesse, Verfügbarkeit und Verständlichkeit erwählt.

Tabelle 1: Überblick der ausgewählten Texte

Nr.	**Thema**	**Name**	**Paper/Magazin/ Zeitung/Zeitschrift**	**Stufe**
1	Bitcoin	Bitcoin as Money?	Paper - Current Policy Perspectives	3
2	Bitcoin	A Complete Beginner's Guide to Bitcoin in 2018	Artikel - Forbes Magazine	2
3	Bitcoin	Geek's Gold: What is Bitcoin, why has the price collapsed and how can you buy the cryptocurrency?	Artikel - The Sun	1
4	Financial Crisis 2008	The Bankruptcy of Lehman Brothers: Causes of Failure & Recommendations Going Forward	Artikel - Social Science Research	3
5	Financial Crisis 2008	Crash course; The Origins of the Financial Crisis	Artikel - The Economist	2
6	Financial Crisis 2008	Why did Lehman Brothers collapse and what caused the global financial crisis in 2008?	Artikel - The Sun	1
7	US-China Trade War	The US-China Trade War: Cause-Effect Analysis	Artikel - The EUrASEANs: journal on global socio-economic dynamics	3
6	US-China Trade War	Everything you need to know about US-China Trade War	Artikel - Forbes India	2
9	US-China Trade War	Trading Blows: Donald Trump's trade war with China explained	Artikel - The Sun	1

Es sollten Themen sein, die in der Gesellschaft breit diskutiert werden und demnach möglichst viele Menschen betreffen. Die Texte wurden in drei Stufen eingeteilt: Stufe 3 (schwer), Stufe 2 (mittel) und Stufe 1 (leicht). Die Stichproben sollten in allen drei Stufen einen ähnlichen Inhalt wiedergeben. Jedes Thema enthält somit drei Texte in jeder Schwierigkeitsstufe.

Für die Einteilung der Texte in den Schwierigkeitsstufen wurden die Medien, in denen sie erschienen sind, als Orientierung genutzt. Angenommen wurde, dass in wissenschaftlichen Fachzeitschriften wie *The EUrASEANs* (EUrASEANs 2020) Texte vorkommen, die für Fachleute aus einem bestimmten Fachbereich – in diesem Fall Wirtschaft – geschrieben wurden. Zeitschriften oder Magazine wie *The Forbes Magazine* (Forbes 2020) oder *The Economist* (Economist 2020) sind für ein an einem Fachbereich sehr Publikum konzipiert, das also ein gewisses Fachwissen besitzt. Das Medium *The Sun* (Sun 2020) hingegen ist eine täglich erscheinende britische Boulevardzeitung, die von einem breiten Publikum gelesen und verstanden werden kann, weshalb daraus die Texte für Stufe 1 entnommen wurden.

Aus Tabelle 1 wird zusammengefasst ersichtlich, welche Themen und welche Texte aus welchem Medium ausgewählt und in welche Stufe eingeteilt wurden.

8.2 Auswahl des Übersetzungstools: DeepL

Nach Auswahl, Einteilung und Kürzung der Texte wurden sie mithilfe des DeepL-Übersetzungsprogramms übersetzt. Das Programm wurde dabei direkt auf der Website von DeepL (DeepL 2019) benutzt. Das DeepL-Übersetzungsprogramm wurde 2017 vorgestellt und stammt vom Technologieunternehmen DeepL GmbH mit dem Sitz in Köln, Deutschland (vgl. DeepL 2020)

Die Auswahl für die Übersetzungen dieser Arbeit fiel auf DeepL, da das Unternehmen mit seiner Pressemitteilung vom August 2017 das Interesse der Autorin an der Qualität des maschinellen Outputs von DeepL geweckt hatte (DeepL 2017). Mithilfe von Blindtests wurde laut Eigenangaben der Plattform festgestellt, dass die Ergebnisse von DeepL durch professionelle ÜbersetzerInnen als „drei Mal häufiger als besser“ (DeepL 2017) bewertet wurden als bei anderen öffentlichen maschinellen Übersetzungssystemen auf dem Markt. Dies sollen auch automatisierte Tests wie der BLEU-Score nachgewiesen haben (vgl. DeepL 2017). Die Autorin dieser Arbeit war daher an einer weiteren Evaluierung des Translats von DeepL nach terminologischen Standards interessiert und wollte herausfinden, ob diese positiven

Ergebnisse auch auf die ausgewählten Texte verschiedener Fachsprachlichkeit im wirtschaftlichen Themenbereich zutrafen.

8.3 Annotationsvorgang

Nach der Übersetzung wurden die Texte zur besseren Möglichkeit der Fehleranalyse in sinnvolle kleine Bedeutungseinheiten aufgeteilt (siehe Abbildung 10). Manchmal sind es Einheiten von einzelnen Wörtern, manchmal von mehreren. Es wurde bei der Einteilung besonderes darauf geachtet, Adjektive nicht vom zu beschreibenden Nomen zu trennen, Pronomen nie allein stehen zu lassen oder eine Zahl vom darauffolgenden Prozent zu trennen. Das alles diente dazu, den Annotationsvorgang zu erleichtern, da es einfacher ist, zusammenhängende Einheiten zu annotieren, statt einzelne Wörter.

Nr.	Segment
139	**US-China Handelskrieg**
140	**Aktuelle Situation**
141	Bis zum 10. Mai 2019
142	zahlte
143	China
144	25 Prozent Zölle
145	nur
146	auf Waren
147	im Wert von
148	50 Milliarden Dollar
149	und
150	weniger als
151	10 Prozent
152	auf Waren
153	im Wert von
154	200 Milliarden Dollar.

Abb. 10: Aufteilung der übersetzen Texte in Bedeutungseinheiten

Der Evaluierungsprozess wurde von drei Studentinnen durchgeführt, die für die Evaluierung der Übersetzungen im Rahmen der Masterarbeit zugezogen wurden. Annotatorin 1 (gleichzeitig die Autorin der vorliegenden Masterarbeit) ist Masterstudentin der Translation mit dem

Schwerpunkt *Fachübersetzen und Sprachindustrie*, Erstsprache Deutsch und Zweitsprache Englisch und weist beträchtliche Erfahrung mit Korrekturlesen auf. Auf Annotatorin 2 treffen alle genannten Charakteristika von Annotatorin 1 auf. Annotatorin 3 hingegen ist Masterstudentin der Translation mit dem Schwerpunkt *Konferenzdolmetschen* und der Wirtschaftsphilosophie. Zwar liegt der Schwerpunkt bei der letztgenannten Annotatorin nicht auf dem Übersetzen, doch hat sie mit Abstand die meiste Erfahrung im Korrekturlesen, da sie neben dem Studium freiberuflich professionell wissenschaftliche Arbeiten korrekturliest.

Alle drei Annotatorinnen haben eine Excel-Datei mit den vorsegmentierten Texten und den Fehlerkategorien mit den jeweiligen Gewichtungen der SAE-J2450-Qualitätsmetrik (im nachfolgenden Kapitel 8.4 genauer beschrieben) erhalten. Des Weiteren haben die Annotatorinnen 2 und 3 eine PDF-Datei mit genauen Instruktionen zur Annotation (siehe Anhang) bekommen. Alle Annotatorinnen haben die Evaluierung innerhalb von zwei Wochen durchgeführt.

8.4 Evaluierung nach SAE-J2450-Qualitätsmetrik

Die SAE-J2450-Qualitätsmetrik wurde als Vorlage in eine Exceldatei eingebaut (siehe Tabelle 2). Nach der Segmentierung der Texte wurden die Textsegmente und die einzelnen Kategorien mit ihren Gewichtungen in die Vorlage eingefügt. Des Weiteren wurde auch eine Spalte für Kommentare eingefügt, damit die Annotatorinnen vor allem in der Kategorie 7 (Diverse Fehler) die Möglichkeit hatten, anzugeben, welchen Fehler sie nicht in die restlichen sechs Kategorien einordnen konnten. Um die Ergebnisse statistisch einfacher errechnen zu können, sollten die Annotatorinnen in der Fehlerspalte entweder 1 für „ja“ oder 0 für „nein“ angeben. Jede Zelle sollte in der Fehlerspalte ausgefüllt werden. Bei den einzelnen Fehlerkategorien und ihren Gewichtungen war jedoch nur eine 1 zu setzen, falls ein Fehler vorhanden war. Eine genauere Erklärung zum Annotationsvorgang ist in den *Instruktionen zur Annotation* (siehe Anhang) ersichtlich.

Die Zuordnung in die Fehlerkategorien ist, wie bereits diskutiert, nicht immer eindeutig, weshalb sich die Annotatorinnen an folgenden Regeln orientieren sollten:

- Sollte es dazu kommen, dass mehrere Fehlerkategorien zutreffen könnten, ist die erste Kategorie auf der Liste auszuwählen.
- Fällt die Entscheidung bei der Gewichtung schwer, ob ein Fehler als schwerwiegend oder geringfügig ist, ist er immer als schwerwiegend einzustufen (vgl. Dalla-Zuanna 2019: 24).

Die SAE-J2450-Qualitätsmetrik wurde für die vorliegende Arbeit ausgewählt, da sie im Vergleich zu anderen Evaluierungsmethoden (z.b. Fehlerkategorisierung nach Vilar et. al [2006]) einfacher im Rahmen einer Masterarbeit durchzuführen sowie leicht verständlich ist und schnell mit der Evaluierung begonnen werden kann. Es ist verhältnismäßig einfach, das Prinzip dieser Qualitätsmetrik und deren Regeln schnell anderen AnnotatorInnen zu erklären. Die Durchführung lässt sich zügig durchführen und die Daten sind einfach statistisch zu berechnen. Aus diesen Daten soll zudem im besten Fall ersichtlich werden, wie gut die Qualitätsmetrik generell anwendbar ist als menschliche Evaluation für potenziell alle Fachbereiche.

Tabelle 2: Evaluierungsbeispiel mit der SAE-J2450-Qualitätsmetrik

Segment	F	K1		K2		K3		K4		K5		K6		K7		Kom.
		5	2	5	2	4	2	4	2	3	1	2	1	3	1	
Um																
ein Bitcoin zu verwenden																
reicht																

F = Fehler
K1 = Falsche Benennung
K2 = Falsche Bedeutung
K3 = Auslassung
K4 = Strukturfehler
K5 = Rechtschreibfehler
K6 = Interpunktionsfehler
K7 = Diverses
Kom = Kommentar

8.5 Auswertung der Evaluierung mit Cohen's Kappa nach Fleiss

Nach Abschluss der Analysearbeit erfolgte die Auswertung der Evaluierung mit Cohen's Kappa nach Fleiss. Die Evaluierungen der Annotatorinnen wurden mit der Evaluierung der Autorin der Masterarbeit miteinander verglichen und ein Übereinstimmungsgrad wurde bestimmt.

8.5.1 Cohen's Kappa

In der vorliegenden Arbeit werden Bewertungen über den Output von DeepL unabhängig voneinander abgegeben. Es wird die Übereinstimmung der Beurteilerinnen oder auch die Interrater-Reliabilität bestimmt. Der Cohen's Kappa-Koeffizient wird am häufigsten verwendet, um die Übereinstimmung zwischen BeurteilerInnen zu bewerten (vgl. Hammann et al. 2014: 1).

Die Berechnungsformel für Cohen's Kappa lautet wie folgt:

$$k = \frac{p_o - p_e}{1 - p_e}$$

Abb. 11: Cohen's-Kappa-Formel (IDioStatics 2016)

Dabei steht die Variable p_0 für den Anteil der tatsächlich beobachteten Übereinstimmungen, während p_e für den Anteil zufälliger Übereinstimmungen verwendet wird (Hammann et al. 2014).

Da die Übersetzungen der ausgewählten Wirtschaftsfachtexte von zwei Annotatorinnen und der Autorin dieser Arbeit beurteilt werden, gibt es somit insgesamt drei Annotatorinnen. Daher wird Cohen's Kappa nach Fleiss (1971) verwendet, da das Originalmaß nach Kappa nur auf zwei AnnotatorInnen anwendbar ist und Fleiss eine Erweiterung auf mehrere AnnotatorInnen bietet.

Als Ergebnis der Kappa-Berechnung erhält man eine Zahl im Bereich von 0 bis 1, wobei auch negative Werte möglich sind. Landis und Koch (1977) schlagen eine hilfreiche Interpretation der Aussagekraft der Werte vor. Beträgt das Kappa (κ) weniger als 0, dann liegt eine schlechte bzw. keine Übereinstimmung vor. Wenig Übereinstimmung ist dann gegeben, wenn der Wert κ zwischen 0 und 0,20 liegt. Liegt ein Wert zwischen 0,21 und 0,40 vor, dann

ist eine ausreichende Übereinstimmung vorhanden. Eine mittelmäßige Übereinstimmung liegt vor, wenn κ zwischen 0,41 und 0,60 liegt. Ein Wert zwischen 0,61 und 0,80 bedeutet, dass eine beachtliche Übereinstimmung vorliegt. Liegt der Wert κ zwischen 0,81 und 1,00, dann deutet das auf eine (fast) vollkommene Übereinstimmung hin.

9 Ergebnisse

9.1 Überblick durchschnittlicher Fehleranzahl der Texte

Tabelle 3 soll einen Überblick darüber geben, wie viele Fehler die einzelnen Texte bei allen drei Annotatorinnen durchschnittlich ergeben haben. Text 1 hat in der Schwierigkeitsstufe 3 durchschnittlich eine Fehleranzahl von 10,00 bzw. 11,24 % der Gesamtwortanzahl. Bei Text 2 beträgt die durchschnittliche Fehleranzahl 16,33 (10,96 %) in der Stufe 2. Text 3 hat in der Stufe 3 eine durchschnittliche Fehleranzahl von 13,66 (12,09 %). Eine etwas höhere Fehleranzahl, sowohl absolut als auch prozentual an der Gesamtwortanzahl, hat Text 4 mit 18,33 Fehlern bzw. 17,13 % in der Stufe 3. Die Fehleranzahl in Text 5 beträgt in der Stufe 2 13,00 (11,93 %). Text 6 hat in der Stufe 1 eine durchschnittliche Fehleranzahl von 15,66 (14,64 %). Bei Text 7 ergibt die durchschnittliche Fehleranzahl in der Stufe 3 mit 18,33 (19,92 %) wieder eine etwas höhere Fehleranzahl, genau wie bei Text 4 (ebenfalls in der Stufe 3). Text 8 weist in der Stufe 2 eine Fehleranzahl von 16,00 (10,06 %) auf und mit 16,33 hat Text 9 in der Stufe 1 eine fast gleiche Fehleranzahl wie der vorherige Text in der Stufe 2, jedoch gemessen an der Gesamtwortzahl mit 24,74 % die höchste Fehlerquote.

Tabelle 3: Durchschnittliche Fehleranzahl

Text	Stufe	Durchschnittliche Fehleranzahl
Text 1	3	10,00 (11,24 %)
Text 2	2	16,33 (10,96 %)
Text 3	1	13,66 (12,09 %)
Text 4	3	18,33 (17,13 %)
Text 5	2	13,00 (11,93%)
Text 6	1	15,66 (14,64 %)
Text 7	3	18,33 (19,92 %)
Text 8	2	16,00 (10,06 %)
Text 9	1	16,33 (24,74 %)

9.2 Fehlerauswertung der Textstufen

Mithilfe der folgenden Tabelle 4 soll die Gesamtfehleranzahl der Fehler in den drei Textstufen von jeder Annotatorin veranschaulicht werden.

Tabelle 4: Gesamtfehleranzahl der Textstufen pro Annotatorinnen

Textstufe	A1*	A2*	A3*
Stufe 1 – leicht	30	30	77
Stufe 2 – mittel	34	32	70
Stufe 3 – schwer	40	33	67

** A1 = Annotatorin 1*

** A2 = Annotatorin 2*

** A3 = Annotatorin 3*

Die Gesamtanzahl in der Textstufe 1 beträgt bei der Annotatorin 1 30 Fehler, bei Annotatorin 2 ebenfalls 30 Fehler und bei Annotatorin 3 mit 77 Fehlern mehr als das Doppelte als bei Annotatorin 1 oder 2. Bei der Textstufe 2 beträgt die Gesamtfehleranzahl bei der 1. Annotatorin 34, bei der 2. Annotatorin 32 und bei der 3. Annotatorin 70. Die Textstufe 3 hat bei der Annotatorin 1 40 Fehler, bei der Annotatorin 2 33 Fehler und bei der Annotatorin 3 67 Fehler. Zusammenfassend zeigt sich, dass Annotatorin 3 wesentlich mehr Fehler identifiziert hat als die anderen beiden Annotatorinnen. Zudem konnte sich keine deutlich höhere Anzahl an Fehlern für eine bestimmte Textstelle über alle Annotatorinnen hinweg zeigen.

9.3 Auswertung der einzelnen Fehlerkategorien der Qualitätsmetrik

In Tabelle 5 werden die Gesamtfehleranzahlen in den einzelnen Fehlerkategorien pro Annotatorin mit der jeweiligen Gewichtung aufgelistet. Annotatorin 1 hat in der Fehlerkategorie *Falsche Benennung* (Beispiel: [on the public transaction ledger] wurde zu [im öffentlichen Transaktionsledger] übersetzt) 14 Fehler in der Gewichtung 5 gefunden, Annotatorin 2 hat 33 Fehler und Annotatorin 3 hat 12 Fehler gefunden. Bei der Gewichtung 2 in derselben Kategorie wurden von der 1. Annotatorin 2 Fehler, von der 2. Annotatorin 5 Fehler und von der 3. Annotatorin 10 Fehler gefunden. In der nächsten Fehlerkategorie *Falsche Bedeutung* (Beispiel hierfür: [Blaming the decline on "trade protectionism"] wurde zu [die den Rückgang auf den "Handelsprotektionismus" zurückführen sollte] übersetzt) wurde Annotatorin 1 in der Gewichtung 5 auf 5 Fehler, Annotatorin 2 auf 24 Fehler und Annotatorin 3 auf 43 Fehler aufmerksam. In der Gewichtung 2 wurden von Annotatorin 1 insgesamt 3 Fehler entdeckt, bei der Annotatorin 2 war es 1 Fehler und bei Annotatorin 3 waren es 13 Fehler.

Die Fehlerkategorie *Auslassung* (Beispiel hierzu: […by the Trump administration, China, on Monday decided to retaliate…] wurde zu […durch die Trump-Administration, China, am Montag beschlossen…] übersetzt, dabei wurde ein Verb wie. „hat", welches nötig im deutschen Kontext ist, ausgelassen) hatte bei der 1. Annotatorin in der Gewichtung 4 6 Fehler, bei der 2. Annotatorin 4 und bei der 3. Annotatorin 0 Fehler. In der Gewichtung 2 kam bei keiner der drei Annotatorinnen ein Fehler vor. Annotatorin 1 hat in der nächsten Fehlerkategorie *Strukturfehler* (Beispiel: […spread over all securitization vehicles, in which they are designed to allow a company or a bank holding assets with little liquidity…] wurde zu […erstreckte sich auch auf alle Verbriefungsträger, bei denen sie so konzipiert sind, dass sie es einem Unternehmen oder einer Bank, die Vermögenswerte mit geringer Liquidität hält…] übersetzt) in der Gewichtung 4 33 Fehler gefunden und Annotatorin 2 und 3 haben beide 13 Fehler gefunden. In der Gewichtung 2 kamen bei Annotatorin 1 7 Fehler vor, bei Annotatorin 2 4 Fehler und bei Annotatorin 3 18 Fehler.

Tabelle 5: Fehleranzahl in den einzelnen Fehlerkategorien

Fehlerkategorie	Gewichtung	A1*	A2*	A3*
1 – Falsche Benennung	5	14	33	12
	2	2	5	10
2 – Falsche Bedeutung	5	5	24	43
	2	3	1	13
3 – Auslassung	4	6	4	0
	2	0	0	2
4 – Strukturfehler	4	33	13	13
	2	7	4	18
5 – Rechtschreibfehler	3	6	2	0
	1	0	0	0
6 – Interpunktionsfehler	2	1	3	3
	1	0	1	5
7 – Diverses	3	6	16	5
	1	3	2	93

** A1 = Annotatorin 1*

** A2 = Annotatorin 2*

** A3 = Annotatorin 3*

In der Kategorie *Rechtschreibfehler* (Beispiel: [25%] wurde zu [25%] übersetzt und aufgrund des fehlenden Leerzeichens als Rechtschreibfehler tituliert) hat Annotatorin 1 in der Gewichtung 3 6 Fehler angegeben, Annotatorin 2 hat 2 Fehler angegeben und Annotatorin 3 hat 0 Fehler angegeben. In der Gewichtung 1 wurden von allen drei Annotatorinnen keine Rechtschreibfehler angegeben.

In der Kategorie *Interpunktionsfehler* (Beispiel hierzu: [US-China Trade War Current Situation] wurde zu [US-China Handelskrieg Aktuelle Situation] übersetzt und der fehlende Doppelpunkt bzw. Bindestrich wurde als Interpunktionsfehler gewertet) hat Annotatorin 1 in der Gewichtung 2 1 Fehler gefunden, während es bei Annotatorin 2 und 3 jeweils 3 Fehler waren. In der Gewichtung 1 kamen bei Annotatorin 1 keine Interpunktionsfehler vor, bei Annotatorin 2 kam 1 Fehler vor und bei Annotatorin 3 waren es 5 Fehler.

In der letzten Fehlerkategorie *Diverses* (Beispiel: [The Bitcoin protocol – the rules that make Bitcoin work – say…] wurde zu [Das Bitcoin-Protokoll - die Regeln, nach denen Bitcoin funktioniert – besagt…] übersetzt und als Personifikation evaluiert) hat Annotatorin 1 in der Gewichtung 3 6 Fehler angegeben, Annotatorin 2 hat 16 Fehler angegeben und Annotatorin 3 hat 5 Fehler angegeben. Annotatorin 1 hat in der Gewichtung 1 insgesamt 3 Fehler gefunden, Annotatorin 2 hat 2 Fehler und Annotatorin 3 hat 93 Fehler gefunden.

Insgesamt gab es unterschiedliche Schwerpunkte in den Kategorien bei den Annotatorinnen. Kategorie 4, mit insgesamt 40 Fehlern, war bei Annotatorin 1 die Kategorie mit der größten Anzahl an Fehlern, bei Annotatorin 2 war dies die Kategorie 1 mit insgesamt 38 Fehlern. Annotatorin 3 hatte die meisten Fehler in der Kategorie 7, *Diverses*, wo 98 Fehler notiert wurden. Die Fehler in der Kategorie *Diverses* unterteilten sich in Stilfehler, Personifikationsfehler, fehlendes Gendern, Uneinheitlichkeit und fehlende Übersetzung. Bei der Betrachtung der Gewichtungen zeigte sich, dass Gewichtung 5 und Gewichtung 4 wesentlich höhere Häufigkeiten aufweisen als die anderen Gewichtungen. Eine Ausnahme dazu bildet die Gewichtung 1 in der Kategorie 7 von Annotatorin 3, die in dieser Kategorie 93 Mal angemerkt wurde.

9.4 Ergebnisse des Cohen's Kappa nach Fleiss

Der Kappa-Wert für die Übereinstimmung in der Gesamtheit der Fehlerkategorien aller drei Annotatorinnen beträgt 0,0636. Werden die Gewichtungen in den einzelnen Fehlerkategorien ausgelassen, beträgt der Kappa-Wert aller drei Annotatorinnen 0,0878. Ein Kappa-Wert von 0,0969 kommt heraus, wenn man die Fehlerkategorie *Diverses* bei der Berechnung der

Übereinstimmung weglässt. Nach Landis und Koch (siehe Kapitel 7.5.1) besteht bei diesen drei Kappa-Werten (0,0636 0,0878, 0,0969) wenig Übereinstimmung zwischen den Annotatorinnen. Bei der Angabe, ob Fehler in den Texten vorhanden sind (ja/nein), wurde ein Kappa-Wert von 0,235 berechnet. Dieser Wert zeigt demnach eine ausreichende Übereinstimmung nach Landis und Koch an.

9.5 Rückmeldungen der Annotatorinnen

Nach Abschluss der Evaluation wurde ein kurzes Gespräch mit den Annotatorinnen geführt, in dem sie der Autorin dieser Masterarbeit Fragen darüber beantworteten, wie ihr Eindruck des Outputs von DeepL im Hinblick auf die Qualität war, ob sie viel recherchieren mussten, um sicherzugehen, dass DeepL die richtigen Termini in der Übersetzung verwendet, und ob die SAE-J2450-Qualitätsmetrik im aktuellen Design für eine gute Evaluierung von maschinell übersetzten Texten ausreicht.

Annotatorin 2 meinte dazu, dass die Übersetzungen von DeepL für den privaten Gebrauch auf jeden Fall genügend und hilfreich seien. Man erkenne meistens, worum es in einem Text geht. Möchte man jedoch Übersetzungen für einen offiziellen Vortrag oder ähnliches erstellen, seien die Übersetzungen von DeepL nicht verwendbar. Die Qualitätsmetrik sei ohne eine Kategorie für die Evaluierung von Stil nicht ausreichend, ansonsten reichten die anderen vorhandenen Fehlerkategorien aus. Annotatorin 2 musste während ihrer Evaluation laut eigenen Angaben mehrfach recherchieren, um die Korrektheit der Übersetzungen zu überprüfen.

Annotatorin 3 war der Ansicht, dass die Qualität bei den meisten Texten oberflächlich auf den ersten Blick nicht schlecht sei. Bei einem groben Durchgang hätte sie nicht viel ausgebessert, doch beim nächsten Durchgang, in dem sie viel genauer auf die Texte eingegangen sei und auch die Gesamttexte berücksichtigt habe, seien ihr dann am Ende doch sehr viele Fehler aufgefallen. Annotatorin 3 gibt auch an, in diesem Fall sehr perfektionistisch veranlagt zu sein. Für die kleinen Stichproben der Gesamttexte reichte das allgemeine Wissen eines wirtschaftsinteressierten Durchschnittspublikums aus. Eine aufwändige Recherche sei nicht notwendig gewesen und man könne auf die Übersetzungen von DeepL vertrauen. Würden jedoch die Gesamttexte für Textstufe 3 evaluiert werden, sei dies etwas Anderes, denn dann sei eine genauere Recherche notwendig, um sicherzugehen, dass auch die exakten Fachtermini verwendet werden.

Zusammenfassend sind also beide Annotatorinnen sind der Meinung, dass die Übersetzungen von DeepL noch immer sehr viele Fehler aufweisen und dass in der SAE-J2450-Qualitätsmetrik für die Evaluation auch der Stil berücksichtigt werden sollte.

10 Diskussion

Die Ergebnisse dieser Studie zeigen, dass weder das Thema noch der Grad der Fachlichkeit eines Textes die Qualität der maschinellen Übersetzung direkt beeinflussen. Die Annahme, dass das maschinelle Übersetzungssystem DeepL fachsprachlich anspruchsvollere Texte schlechter übersetzt als Texte mit einem geringeren Grad an Fachsprachlichkeit, konnte in dieser Studie nicht bestätigt werden. Als Folge der Evaluierung und des starken Auftretens von Fehlern in relativ kurzen Sequenzen liegt die Aussage nahe, dass maschinell erstellte Texte von DeepL nicht direkt für den professionellen Gebrauch geeignet sind und menschliches Post-Editing oder von Beginn an einer Humanübersetzung bedürfen.

Ursprünglich sollten die ausgewählten Texte genau nach den Merkmalen von Fachtexten klassifiziert und eingeteilt werden, doch dies stellte sich für die vorliegende Studie als nicht anwendbar heraus, da Texte mit einer höheren Fachsprachlichkeit auch Elemente von Gemeinsprache enthalten können, und umgekehrt. Des Weiteren wurden die Texte so gekürzt, dass sie in allen drei Textstufen inhaltlich hinreichend deckungsgleich waren. Aus diesem Grund waren auch keine Formeln und Grafiken vorhanden, die vor allem in anspruchsvolleren Texten vorkommen. Deshalb wurde letztendlich bei der Auswahl der Texte das Medium als Orientierung verwendet.

Zusammenfassend lässt sich hier kein klarer Trend erkennen, dass Texte der anspruchsvolleren Stufen 3 und 2 eine höhere Fehleranzahl haben als Texte aus der Stufe 1 mit einfacheren Ausgangstexten (siehe Tabelle 3). Die absolute Fehleranzahl betrug zwischen 10,00 und 18,33 Fehlern. Die relative Fehleranzahl, gemessen an der Gesamtwortanzahl der Texte, variierte von 10,06 % und 24,74%. Trotz zwei prozentualer Ausreißer nach oben (19,92 % und 24,74 %) blieb die relative Fehleranzahl in einem engen Bereich und es konnte keine Abhängigkeit von der Textstufe erkannt werden.

Eine mögliche Erklärung dafür ist, dass Texte in den höheren Textstufen 3 und 2 zwar eine größere Anzahl an Fachtermini beinhalten, die während der Evaluierung recherchiert werden müssen, doch Texte aus der niedrigsten Textstufe 1 beinhalten dafür beizeiten saloppe und umgangssprachliche Formulierungen, die zusätzlich bestimmte Emotionen bei den LeserInnen hervorrufen sollen. Fachtermini, die spezifisch für das Thema *Bitcoins* sind, wie etwa das englische Wort *blockchain*, werden üblicherweise im Deutschen beibehalten, doch DeepL hat es mit *Blockkette* übersetzt. Der Begriff *transaction ledger* wurde hingegen mit *Transaktionsledger* übersetzt, wobei der zweite Teil der deutschen Übersetzung nur aus dem Englischen übernommen wurde.

Als Beispiel für saloppe Formulierungen aus der Textstufe 1 kann die Übersetzung von *China's action* herangezogen werden. Übersetzt wurde dieser Teil mit *chinesische Aktion*, was im Deutschen inhaltlich und stilistisch nicht vertretbar ist. Ein amüsantes Beispiel für einen Bedeutungsfehler ist die Übersetzung von DeepL beim Wort *miners*, dass im Bitcoin-Kontext die Personen beschreibt, die Bitcoins erstellen können. Übersetzt wurde dieses Wort von DeepL mit *Bergleuten*, was für den besagten Kontext gänzlich unpassend ist.

Die Ergebnisse in der Tabelle 4 zeigen, dass Annotatorin 1 und Annotatorin 2 sehr ähnliche Gesamtfehleranzahlen der Fehler in den drei Textstufen haben, wobei Annotatorin 3 vor allem in den Textstufen 3 und 2 fast mehr als doppelt so viele Übersetzungsfehler bemängelte. Dies kann womöglich damit erklärt werden, dass Annotatorin 3 regelmäßig professionell Texte von verschiedensten AutorInnen korrekturliest und damit eine viel höhere und automatisiertere Wahrnehmung und eine präzisere Vorgehensweise für die Fehlerfindung hat. Außerdem ist sie beim Korrekturlesen nach eigenen Angaben perfektionistisch veranlagt, was dazu führt, dass sie eine geringere Toleranz gegenüber Fehlern besitzt. Des Weiteren kann die Erwartungshaltung gegenüber maschinellen Übersetzungssystemen stark variieren.

In der Tabelle 5, in der die Gesamtfehleranzahlen pro Annotatorin in den einzelnen Fehlerkategorien mit dazugehörigen Gewichtungen aufgezeigt sind, wird ersichtlich, warum der berechnete Kappa-Wert für die Übereinstimmung in allen Fehlerkategorien aller drei Annotatorinnen nur 0,0636 beträgt – es herrschte keine Übereinstimmung. Eine wahre Übereinstimmung ist nur bei der Fehlerkategorie *Rechtschreibung* in der Gewichtung 1 gegeben: Alle drei Annotatorinnen stießen auf 0 Fehler. Die von Annotatorin 3 angegebenen 93 stilistischen Fehler in der Gewichtung 1 sprengen dagegen jegliche Interpretation der kleinsten möglichen Übereinstimmung der drei Annotatorinnen, da dieser Abstand erneut klar zeigt, dass Annotatorin 3 sich mit ihren bemängelten Übersetzungsstellen deutlich von den anderen zwei Annotatorinnen abhebt.

Allgemein gesagt besteht zwischen den drei Annotatorinnen keine Übereinstimmung (κ = 0,0636). Lässt man die Gewichtungen der einzelnen Fehlerkategorien aus oder lässt man nur die Fehlerkategorie *Diverses* aus, ändert sich die Übereinstimmung nicht signifikant. Nur bei der Überprüfung, ob ein Fehler bei einem Segment besteht oder nicht, wurde mit dem Kappa-Wert 0,235 zumindest eine ausreichende Übereinstimmung berechnet. Somit kann zumindest mit einer gewissen Sicherheit gesagt werden, dass Fehler identifiziert wurden.

Die fehlende Übereinstimmung zwischen den Annotatorinnen kann mit der Art der getätigten Evaluation begründet werden. Die SAE-J2450-Qualitätsmetrik sieht in der Regel keine stilistischen Fehler vor, doch die Autorin dieser Arbeit wollte bewusst auch stilistische

Mängel hinzufügen, da dies bei maschinellen Übersetzungssystem oft den Hauptmangel darstellt, was von den beiden anderen Annotatorinnen bestätigt wurde. Da sich NMÜ-Systeme in Bezug auf Terminologie und Stil nicht einfach an die spezifischen Benutzeranforderungen anpassen lassen (vgl. Vashee 2017: 55), macht eine Evaluierung ohne eine Kategorie für Stil nur wenig Sinn, da dieser dann vor allem für eine gute offizielle Endübersetzung funktionieren muss. Des Weiteren ist die SAE-J2450-Qualitätsmetrik trotz der dazu verfassten Regeln noch nicht einheitlich genug, da es weiterhin sehr viel Raum für Subjektivität gibt – eine Aussage, die durch die mangelnde Übereinstimmung dieser Studie zumindest teilweise bestätigt werden konnte.

Bei der Literaturrecherche nach Studien mit ähnlichem Design hat die Autorin vor allem Studien gefunden, in denen man sich mit diversen maschinellen Übersetzungssystemen auseinandergesetzt oder verschiedene Übersetzungssysteme miteinander verglichen hatte. In den Studien wurde festgestellt, dass die allgemeine Qualität des MÜ-Outputs teilweise mangelhaft war. Zum Beispiel hat Wiesmann (2019) in ihrer Studie ermittelt, dass maschinelle Übersetzungssysteme typische Fehler wie Terminologiefehler, Syntaxfehler, Grammatikfehler und Stilfehler machen – häufig ist der Grund dafür, dass das System den Kontext falsch interpretiert. Der MÜ-Output ist aufgrund der hohen Fehleranzahl ohne aufwendiges Post-Editing unbrauchbar. Shafia (2018) kommt zur gleichen Erkenntnis, und zwar, dass die Ergebnisse unzureichend sind.

Grundvall (2019) kommt zu dem ähnlichen Schluss, dass Termini oft nicht übersetzt bzw. neue Wörter oder Wortformen gebildet werden. Auch hier wurde festgestellt, dass viele Fehler aufgrund von Fehlinterpretationen des Programms passieren. Schreibstil und Ton werden zudem häufig nicht getroffen, da das System keine Analyse dazu machen und dann zusätzlich nicht darüber entscheiden kann, welcher Stil in der Zielsprache richtig wäre. Grundvalls (2019) Studie hat Ähnlichkeiten zum Studiendesign der vorliegenden Masterarbeit. Sie ist vom Aufbau her ähnlich, untersucht jedoch weniger Texte, aber dafür längere Textproben; außerdem verwendet Grundvall (2019) das Qualitätsmodell von House (1977).

Eine eindeutige Beantwortung der Frage, ob sich die SAE- J2450-Qualitätsmetrik generell als Metrik für die Evaluation von maschinell übersetzen Texten eignet, ist in dieser Form nicht möglich, da die vorliegende Arbeit als Limitation nur 9 Stichproben und 3 Evaluatorinnen hatte. Diese Zahlen sind selbstredend zu klein, um repräsentative und signifikante Ergebnisse liefern zu können.

11 Fazit

In dieser Arbeit wurde die Qualität von Übersetzungen wirtschaftlicher Fachtexte mit unterschiedlichen Textabstufungen evaluiert. Die Evaluierung erfolgte durch die Autorin mit zwei zusätzlichen Annotatorinnen, die anhand der SAE-J250-Qualitätsmetrik mit Berücksichtigung des Stiles durchgeführt wurde.

Dabei konnten Einblicke darüber gegeben werden, dass maschinelle Übersetzungssysteme noch sehr weit von der Qualität einer HumanübersetzerIn entfernt sind. Es wurde festgestellt, dass sich die Übersetzungen nicht proportional zum textlichen Schwierigkeitsniveau verhalten, sondern dass die Übersetzungen alle ungefähr das gleiche Fehlerniveau aufwiesen. Dabei spielte es keine Rolle, ob der Text ein Fachtext war oder aus einer Boulevardzeitung stammte.

Die Datenmenge dieser Studie ist jedoch relativ klein, weshalb es empfehlenswert ist, eine ähnliche Studie mit größerer Datenmenge (Textanzahl, AnnotatorInnen) sowie modifizierter SAE-J2450-Qualitätsmetrik durchzuführen, um die Schlussfolgerung dieser Studie zu verfestigen oder doch vollständig zu verwerfen.

Es wäre interessant zu analysieren, wie das Ergebnis mit einer höheren Anzahl und längeren Texten, sowie mehr EvaluatorInnen aussehen würde, und ob eine höhere Übereinstimmung erzielt werden könnte. Ein weiterer interessanter Aspekt wäre der Vergleich der Übersetzungsqualität unterschiedlicher Fachbereiche, wie zum Beispiel Wirtschaft und Medizin, sowie der Vergleich zwischen unterschiedlichen Textarten untereinander.

Literaturverzeichnis

Adamzik, Kirsten (2004). *Sprache: Wege zum Verstehen*. Tübingen / Basel: A. Francke Verlag.

Agarwal, Abhaya & Lavie, Alon (2008). METEOR, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output. In: *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, Ohio, USA: 115-118.

Ahrenberg, Lars (2017). Comparing machine translation and human translation: A case study. In: Temnikova, Irina; Orasan, Constantin; Corpas, Gloria & Vogel, Stephan (Hg.) *RANLP 2017 The First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT) Proceedings of the Workshop*. Linköping University, Schweden: 21-28.

Allen, Jeffrey (2003). Post-editing. In: Somers, Harold L. (Hg.) *Computers and Translation: A Translator's Guide*. Amsterdam/Philadelphia: John Benjamins.

Arnold, Doug; Balkan, Lorna; Humphreys, R. Lee, Meijer, Siety & Sadler, Louisa (1994). *Machine Translation: An Introductory Guide.* Oxford: NCC Blackwell.

Arntz, Reiner; Picht, Heribert & Schmitz, Klaus-Dirk (2014). *Einführung in die Terminologiearbeit.* Hildesheim: Georg Olms Verlag.

Azadinamin, Amirsaleh (2013). The Bankruptcy of Lehman Brothers: Causes of Failure & Recommendations Going Forward. *Social Science Research*, 4-5.

Bar-Hillel, Yehoshua (1959). *Report on the state of machine translation in the United States and Great Britain*. Jerusalem: Hebrew University.

Baumann, Klaus-Dieter (1998). Das Postulat der Exaktheit für den Fachsprachengebrauch. In: Hoffmann, Lothar; Kalverkämper , Hartwig & Wiegand, Herbert Ernst (Hg.) *Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft.* Berlin, 373-377.

Bentivogli, Luisa; Bisazza, Arianna; Cetollo, Mauro & Federico, Marcello (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Austin, Texas, USA: 257-267.

Brown, Peter F.; Stephen, John C.; Della Pietra, A.; Della, Vincent J.; Jelinek, Pietra F.; Lafferty, John D.; Mercer, Robert L. & Roossin, Paul S. (1990). A statistical approach to machine translation. *Computational Linguistics* 1990: Volume 16, Number 2, 79-85.

Burchard, Aljoscha & Porsiel, Jörg (2017). Vorwort: Was kann maschinelle Übersetzung und was nicht? In: Porsiel, Jörg (Hg.) *Maschinelle Übersetzung. Grundlagen für den professionellen Einsatz.* Berlin: BDÜ Weiterbildungs- und Fachverlagsgesellschaft mbh, 11-17.

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

Canfora, Carmen & Ottmann, Angelika (2017). Übersetzungen richtig bewerten. *Technische Kommunikation* 2017: Ausgabe 6, 35-40.

Castilho, Sheila; Doherty, Stephen; Gaspari, Federico & Moorkens, Joss (2018). Approaches to human and machine translation quality assessment. In: Moorkens, Joss; Castilho, Sheila; Gaspari, Federico & Doherty, Stephen (Hg.) *Translation Quality Assessment*. *From Principles to Practice.* Cham: Springer, 9-38.

Costa-jussà, Marta R.; Allauzen, Alexandre; Barrault, Loïc; Kyunghun, Cho & Schwenk, Holger (2016). Introduction to the Special Issue on Deep Learning Approaches for Machine Translation. *Computer Speech & Language* 2017: Volume 46, 367-373.

Dalla-Zuanna, Jean-Marc (2010). Evaluierung fremdsprachiger technischer Dokumentation im Bereich Vertrieb After Sales der Marke Volkswagen Pkw. Direkte Qualitätsmessung. *MDÜ – Fachzeitschrift für Dolmetscher und Übersetzer* 2010: Ausgabe 2, 20-25.

DeepL (2017). Pressemitteilung DeepL Übersetzer. https://www.deepl.com/press.html (Stand: 24.02.2020).

DeepL (2019). DeepL Übersetzter. https://www.deepl.com/translator (Stand 19.11.2019).

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

Donovan, Molly (2018). Maschinelle Übersetzung: Wichtige Begriffe kurz erklärt. *Lionbridge.* https://www.lionbridge.com/de/blog/translation-localization/machine-translation-in-translation/ (Stand: 17.12.2019).

Economist (2013). Crash course; The origins of the financial crisis. *The Economist*, 07.09.2013. https://pdfs.semanticscholar.org/99ed/12612243475696e54225daa3dc619fd3b1d2.pdf (Stand: 08.10.2019).

Economist (2020). https://www.economist.com/ (Stand 04.04.2020)

EUrASEANs (2020). The EUrASEANs: journal on global socio-economic dynamics. https://euraseans.com/index.php/journal (Stand: 04.04.2020).

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*: *76*(5), 378.

Forbes Magazine (2020). https://www.forbes.com/#653f96012254 (Stand: 04.04.2020).

Grundvall, Jessica (2019). *Was heißt „gut genug“ bei maschinell übersetzten Texten? Eine Studie der Qualität an zwei von DeepL übersetzten Textbeispielen.* Masterarbeit, Universität Vaasa.

Guo, Yupeng (2017). *NLP-Machine Translation. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation.* Forschungsseminar, Universität Leipzig.

Göpferich, Susanne (2007). Kürze als Prinzip fachsprachlicher Kommunikation. In: Bär, Jochen A.; Roelcke, Thorsten & Steinhauer Anja (Hg.) *Sprachliche Kürze. Konzeptuelle, strukturelle und pragmatische Aspekte.* Berlin: Walter de Gruyter, 412-433.

Hammann, Markus; Jördens, Janina & Schecker, Horst (2014). Übereinstimmung zwischen Beurteilern: Cohens Kappa (κ). In Krüger, Dirk; Parchmann, Ilka; Schecker, Horst (Hg.) *Methoden in der naturwissenschaftsdidaktischen Forschung*. Berlin: Springer.

Haselhuhn, Andreas (2018). *Rekurrente/rückgekoppelte neuronale Netzwerke.* Forschungsseminar, Universität Leipzig.

Hoffmann, Lothar (1976). *Kommunikationsmittel Fachsprache. Eine Einführung.* Berlin: Akademie-Verlag.

House, Juliane (1977). *A Model for Translation Quality Assessment*. Tübingen: Günter Narr Verlag.

Hui, Liu (2017). *A Study on Translation Quality Management of Automative Information Based on SAE J2450 Translation Quality Metric*. Dissertation, Universität Wien.

Hutchins, John W. (1986). *Machine Translation: past, present, future.* http://www.hutchinsweb.me.uk/PPF-TOC.htm (Stand: 23.11.2019).

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

Hutchins, John W. (1995). *Machine Translation: A Brief History.* In: Koerner, Ernst F. H. & Asher, E. A. (Eds.) Concise history of the language sciences: from the Sumerians to the

cognitivists. Oxford: Pergamon, 431-445. *Hutchinsweb.* http://hutchinsweb.me.uk/ConcHistoryLangSci-1995.pdf (Stand: 29.11.2019).

Hutchins, John W. (2001). Machine Translation and Human Translation: In Competition or in Complementation? *International Journal of Translation 2001*: Volume 13, Nummer 1-2, 5-20.

Hutchins, John W. (2006). Machine Translation: a concise history. *Hutchinsweb.* http://www.hutchinsweb.me.uk/CUHK-2006.pdf (Stand: 29.11.2019).

Hyatt, Edward & Lockett Jon (2018). Geek's Gold. What is Bitcoin, why has the price collapsed and how can you buy the cryptocurrency? *The Sun*, 16.10.2018. https://www.thesun.co.uk/money/3000715/bitcoin-what-is-price-gbp-usd-today-value-cryptocurrency-buy/ (Stand: 08.10.2019).

IDoStatistics (2016). Cohen's kappa free calculator. https://idostatistics.com/cohen-kappa-free-calculator/ (Stand: 31.05.2019).

Kurz, Christopher (2014). Maschinelle Übersetzung – Vom streng geheimen Forschungsprojekt zum alltäglichen Einsatz. In: Baumann, Klaus-Dieter & Kalverkämper, Hartwig (Hg.) *Theorie und Praxis des Dolmetschens und Übersetzens in fachlichen Kontexten*. Berlin: Frank & Timme GmbH, 397-426.

Kalverkämper, Hartwig (1990). Der Einfluß der Fachsprachen auf die Gemeinsprache. In: Stickel, Gerhard (Hg.) *Deutsche Gegenwartssprache. Tendenzen und Perspektiven.* Berlin: Walter de Gruyter, 88–133.

Kalverkämper, Hartwig (1996). Im Zentrum der Interessen: Fachkommunikation als Leitgröße. *HERMES-Journal of Language and Communication in Business.* (16), 117-176.

Koehn, Philipp (2010). *Statical Machine Translation*. Cambridge: Cambridge University Press.

Läubli, Samuel; Sennrich, Rico & Volk, Martin (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* Brüssel, Belgien: 4791-4796.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*: 159-174.

Lionsbridge Marketing (2019). Neuronale maschinelle Übersetzung: Künstliche Intelligenz erleichtert die mehrsprachige Kommunikation. https://www.lionbridge.com/de/blog/translation-localization/neural-machine-translation-artificial-intelligence-works-multilingual-communication/ (Stand: 17.12.2019).

Lo Presti, Rosita (2016). *Menschliche und automatische Evaluation von Übersetzungen von Fachtexten in Google Translate.* Masterarbeit, Universität Wien.

Lo, Stephanie & Wang, Christina J. (2014). *Bitcoin as Money?* Boston Federal Reserve, Current Policy Perspectives.

Marr, Bernard (2018). A Complete Beginner's Guide To Bitcoin In 2018. *Forbes Magazine*, 17.01.2018. https://www.forbes.com/sites/bernardmarr/2018/01/17/a-complete-beginners-guide-to-bitcoin-in-2018/#cc9c884418df (Stand: 08.10.2019).

Mazareanu, E. (2019). Language services industry in the U.S. - Statistics & Facts. https://www.statista.com/topics/2152/language-services-industry-in-the-us/ (Stand: 01.07.2019).

MediStat (2019). Cohen’s Kappa Koeffizient. https://www.medistat.de/glossar/analyse-von-haeufigkeiten/cohens-kappa-koeffizient/ (Stand: 23.02.2020)

Niederhaus, Constanze (2011). Die Komplexität von Fachtexten verschiedener Berufsfelder– Eine korpuslinguistische Untersuchung des Fachsprachlichkeitsgrades von Lehrbuchtexten der Berufsfelder Körperpflege und Elektrotechnik. In: Granato, Mona; Münk, Dieter & Weiß, Reinhold (Hg.) *Migration als Chance.* Bonn, 209-224.

Onyusheva, Irina V.; Thinn Nain, Cherry & Zaw, Aung Lin (2019). The US-China Trade War: Cause-Effect Analysis. *The EUrASEANs: journal on global socio-economic dynamics* 1 (14), 8-12.

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

Roelke, Thorsten (1999). *Fachsprachen*. Berlin: Schmidt.

Rogers, Jon (2019). Trading Blows: Donald Trump's trade war with China explained. *The Sun*, 27.08.2019. https://www.thesun.co.uk/news/9082139/us-china-trade-war-trump-jinping/ (Stand: 15.10.2019).

Sarda, Pranit (2019). Everything you need to know about US-China Trade War. *Forbes India*, 14.05.2019. http://www.forbesindia.com/article/special/everything-you-need-to-know-about-uschina-trade-war/53525/1 (Stand: 15.10.2019).

Schäfer, Falko (2002). *Die maschinelle Übersetzung von Wirtschaftsfachtexten. Eine Evaluierung anhand des MÜ-Systems der EU-Kommission, SYSTRAN, im Sprachenpaar Französisch-Deutsch.* Frankfurt am Main: Peter Lang GmbH.

Schwarzl, Anja (2001). *The (Im)Possibilities of Machine Translation.* Frankfurt am Main: Peter Lang GmbH.

Seilheimer, Andrea (2017). Entwicklungen und Forschungsdesiderata in der romanistischen Wirtschaftslinguistik. In: Hennemann, Anja; Lobin, Antje: Plötner, Kathleen & Schlaak,

Claudia (Hg.) *Werbesprache pluridisziplinär – Aktuelle Tendenzen in der romanistischen Werbesprachenforschung*. Berlin: Frank & Timme GmbH, 15-41.

Sennrich, Rico; Haddow Barry & Birch Alexandra (2016). Edinburgh Neural Machine Translation Systems for WMT 16. In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Deutschland: 371-376.

Shafia, Parnian (2018). *Disregarding linguistics: A critical study of Google Translate's syntactic/semantic errors in rendering multiword units in English to Persian translations.* Masterarbeit, Carleton University.

Snover, Matthew; Madnani, Nitin; Dorr, Bonnie J. & Schwartz, Richard (2009). Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athen, Griechenland: 259-268.

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

Stow, Nicola (2018). Death Bank. Why did Lehman Brothers collapse and what caused the global financial crisis in 2008? *The Sun*, 15.12.2018. https://www.thesun.co.uk/news/7265360/lehman-brothers-bank-collapse-crisis-bankruptcy/ (Stand: 08.10.2019).

Sun (2020). https://www.thesun.co.uk/ (Stand 04.04.2020).

Taube, Mortimer (1961). *Computers and common sense*. New York: Columbia University Press.

Vashee, Kirti (2017). Neural Machine Translation; A Practitioner's Viewpoint. In: Porsiel, Jörg (Hg.) *Maschinelle Übersetzung. Grundlagen für den professionellen Einsatz.* Berlin: BDÜ Weiterbildungs- und Fachverlagsgesellschaft mbH, 44-58.

Vaswani, Ashish; Shazeer, Noam, Parmar, Niki; Uszkoreit, Jakob, Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz & Polosukhin, Illia (2017). Attention is All you Need. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, CA, USA, 1-11.

Vilar, David; Xu, Jia; D'Haro, Luis Fernando & Ney, Hermann (2006). Error Analysis of Statistical Machine Translation Output. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06)*. Genoa, Italy, 697–702.

Volkart, Lise; Bouillon, Pierette & Girletti, Sabrina (2018). Statistical vs. Neural Machine Translation: A Comparison of MTH and DeepL at Swiss Post's Language Service. In: *Proceedings of the 40th Conference Translating and the Computer.* Université de Genève, Schweiz: 145-150.

Von Hahn, Walther (1981). *Fachsprachen.* Darmstadt: Wissenschaftliche Buchgesellschaft.

Von Hahn, Walther (1983). *Fachkommunikation: Entwicklung, Linguistische Konzepte, Betriebliche Beispiele.* Berlin/New York: Walter de Gruyter.

Wiesmann, Eva (2019). Machine Translation in the Field of Law: A Study of the Translation of Italian Legal Texts Into German. *Comparative Legilinguistics.* Volume 37, 117-153.

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang

Instruktionen zur Annotation

Vielen Dank für die Bereitschaft der Teilnahme an dieser Studie! Anbei erhalten Sie zuerst einen allgemeinen Überblick, eine Beschreibung zu den einzelnen Fehlerkategorien der SAE-J2450-Qualitätsmetrik und anschließend einen detaillierten Leitfaden zur Vorgehensweise bei der Evaluierung.

Allgemeiner Überblick

Die von DeepL übersetzten Texte werden mit Hilfe der SAE-J2450-Qualitätsmetrik evaluiert. Die Texte wurden in drei verschiedene Schwierigkeitsstufen eingeteilt: rot – schwer, gelb – mittel, grün – leicht. Die übersetzten Texte wurden dann in Segmente eingeteilt.
Da es schwierig ist, die Texte allein durch Segmente zu lesen, erhalten Sie als Annotatorinnen neben dem Evaluierungsbogen zusätzlich die vollständigen Texte, was zu einem besseren Überblick und Verständnis beitragen soll.

Fehlerkategorien der SAE-J2450-Qualitätsmetrik

Im Folgenden werden die Fehlerkategorien beschrieben, die in der SAE-J2450-Qualitätsmetrik beinhaltet sind (vgl. Jean-Marc Dalla-Zuanna 2010: 23-24):

1. **Falsche Benennung (WT, Wrong Term) – Gewichtung 5/2:**
   Die Benennung im Zieltext entspricht nicht der vorgegebenen Terminologie oder die im Zieltext verwendete Terminologie ist nicht einheitlich. Es kann vorkommen, dass es einen Fehler in dieser Kategorie gibt, wenn zum Beispiel etwas von DeepL nicht übersetzt worden ist, wofür es im Deutschen jedoch sehr wohl ein Wort dafür gibt.

2. **Falsche Bedeutung (WM, Wrong Meaning) – Gewichtung 5/2:**
   Hier ist auch der Inhalt gemeint. Die Bedeutung des Zieltextes weicht eindeutig von der des Ausgangstextes ab. Es kann auch sein, dass etwas vom Programm sehr schlecht übersetzt worden ist, dass man einfach nicht versteht, was ein Segment/Satz aussagen soll, also ist das auch als Fehler zu zählen.

3. **Auslassung (OM, Omission) – Gewichtung 4/2:**
   Eine zusammenhängende Textstelle (Wort, Satz, Segment, Absatz) fehlt im Zieltext.

4. **Strukturfehler (SE, Structural Error) – Gewichtung 4/2:**
   Diese Kategorie deckt Semantik-, Grammatik- und Syntaxfehler ab.

5. **Rechtschreibfehler (SP, Misspelling) – Gewichtung 3/1:**
   Die Schreibweise verstößt gegen die Regeln der offiziellen Rechtschreibung.

6. **Interpunktionsfehler (PE, Punctuation Error) – Gewichtung 2/1:**
   Die Zeichensetzung verstößt gegen die Regeln der Interpunktion.

7. **Diverses (ME, Miscellaneous Error) – Gewichtung 3/1:**
   Sonstige Fehler, die nicht eindeutig in die Hauptkategorien 1 bis 6 zugeordnet werden können (z. B. unübersetzte Textstelle). Hier ist auch anzumerken, dass die SAE-J2540-

Qualitätsmetrik keine Kategorie für „Stil“ beinhaltet, deswegen ist ein Stilfehler bitte bei Diverses zu vermerken.

Evaluierung und Gewichtung

Ist in einem Segment ein Fehler enthalten, setzen Sie bitte in die jeweilige Zelle der vorhandenen „Fehler“-Spalte eine „1“ für „ja“, gibt es in einem Segment keinen Fehler, so setzen Sie bitte in die jeweilige Zelle eine „0“ für „nein“ ein.

Ordnen Sie bitte entdeckte Fehler in eine der vorhandenen Fehlerkategorien ein. Dazu gibt es zu jeder Fehlerkategorie eine bestimmte Gewichtung, in der Sie bitte erneut für „ja“ eine „1“ oder für „nein“ eine „0“ in die angemessene Gewichtung setzen. Geben Sie die Fehler entweder als „serious“ oder „minor“ Fehler an. Schwere Fehler bzw. grobe Fehler (hohe Zahl bei Gewichtung) sind Fehler, die zu einer Irritation beim Leser führen können. Es ist eine Sinnänderung im übersetzten Text vorhanden. Leichte Fehler (niedrige Zahl bei Gewichtung) sind Fehler, die beim Leser keine bzw. nur eine leichte Irritation beim Leser hervorrufen. Leichte Fehler beinhalten jedoch im Vergleich zu schweren Fehlern keine Sinnänderung. Ist es schwer sich dafür zu entscheiden, ob es sich um einen schweren oder leichten Fehler handelt, so stufen Sie den Fehler bitte immer als „serious“ ein. Trifft ein Fehler auf mehrere Kategorien zu, so wählen Sie bitte immer die zuerst aufgeführte Kategorie aus (vgl. Jean-Marc Dalla-Zuanna 2010: 23-24).

Falls Sie einen vorhandenen Fehler nicht in eine der vorgegebenen Kategorien einteilen können, dann geben Sie den Fehler bitte bei der Kategorie “Diverses” an, indem Sie diese Kategorie mit „1“ in der entsprechenden Gewichtung markieren und im Feld “Kommentar/Anmerkungen” erklären, um welchen Fehler es sich handelt. Sollten Sie generelle Anmerkungen zu einer bestimmten Textstelle haben, können Sie auch bei anderen Fehlerkategorien das Feld “Kommentar/Anmerkungen” nutzen.

Geben Sie bitte auch Ihre Beurteilung des stilistischen GesaMÜeindrucks der zu evaluierenden Text an. Dafür gibt es ganz unten im Dokument eine eigene kleine Tabelle, in der Sie zwischen „positiv“, „mittelmäßig“ und „negativ“ auswählen können. Setzen Sie dafür in der Spalte „p/m/n“ für „ja“ wieder eine „1“ und für „nein“ eine „0“ ein.

Ausgewählte Texte und ihre Übersetzung

Thema 1

Bitcoins

Textstufe 3 – Text 1

Lo, Stephanie & Wang, Christina J. (2014). *Bitcoin as Money?* Boston Federal Reserve, Current Policy Perspectives.

By Stephanie Lo and Christina J. Wang | September 4th, 2014

To use a bitcoin, a user submits her account number (“public key”) and password (“private key”) for verification on the public transaction ledger, known as the ‘”block chain.” Individuals (“miners”) use their computing power to verify that the transaction is real by solving a computationally intensive problem (finding the “hash” of a “nonce”).[2] In return for verifying the transaction, the first individual to post the solution to the problem is rewarded with a given number of bitcoin, adding to the stock of bitcoin and hence constituting money creation. […]

Übersetzung DeepL:

Um ein Bitcoin zu verwenden, reicht ein Benutzer seine Kontonummer ("öffentlicher Schlüssel") und sein Passwort ("privater Schlüssel") zur Überprüfung im öffentlichen Transaktionsledger, der so genannten "Blockkette", ein. Einzelpersonen ("Miner") nutzen ihre Rechenleistung, um zu überprüfen, ob die Transaktion real ist, indem sie ein rechenintensives Problem lösen (das Auffinden des "Hash" eines "Nonce").2 Als Gegenleistung für die Überprüfung der Transaktion wird die erste Person, die die Lösung des Problems veröffentlicht, mit einer bestimmten Anzahl von Bitcoin belohnt, indem sie den Bitcoin-Bestand erhöht und damit Geld schafft. […]

Übersetzt mit www.DeepL.com/Translator

Textstufe 2 – Text 2

Marr, Bernard (2018). A Complete Beginner's Guide To Bitcoin In 2018. *Forbes Magazine*, 17.01.2018. https://www.forbes.com/sites/bernardmarr/2018/01/17/a-complete-beginners-guide-to-bitcoin-in-2018/#cc9c884418df (Stand: 08.10.2019).

By Bernard Marr | January 17th, 2018

How do you "mine" Bitcoins?

People—or more accurately extremely powerful, energy-intense computers—"mine" Bitcoins to make more of them. There are currently about 16 million Bitcoins in existence, and that leaves only 5 more million available to mine because Bitcoins developers capped the quantity to 21 million. Ultimately, each Bitcoin can be divided into smaller parts with the smallest fraction being one hundred millionth of a Bitcoin called a "Satoshi," after the founder Nakamoto. The mining process involves computers solving an extremely challenging mathematical problem that progressively gets harder over time. Every time a problem is solved, one block of the Bitcoin is processed and the miner gets a new Bitcoin. A user establishes a Bitcoin address to receive the Bitcoins they mine; sort of like a virtual mailbox with a string of 27-34 numbers and letters. Unlike a mailbox, the user's identity isn't attached to it. […]

Übersetzung DeepL:

Wie "mine" man Bitcoins?

Menschen - oder genauer gesagt extrem leistungsfähige, energieintensive Computer - "bauen" Bitcoins, um mehr aus ihnen zu machen. Es gibt derzeit etwa 16 Millionen Bitcoins, und das lässt nur 5 weitere Millionen für mich übrig, weil die Entwickler von Bitcoins die Menge auf 21 Millionen begrenzt haben. Letztendlich kann jede Bitcoin in kleinere Teile aufgeteilt werden, wobei der kleinste Teil ein Hundertstel eines Bitcoins ist, der nach dem Gründer Nakamoto "Satoshi" genannt wird. Der Mining-Prozess beinhaltet Computer, die ein extrem anspruchsvolles mathematisches Problem lösen, das mit der Zeit immer schwieriger wird. Jedes Mal, wenn ein Problem gelöst wird, wird ein Block des Bitcoins verarbeitet und der Miner erhält einen neuen Bitcoin. Ein Benutzer richtet eine Bitcoin-Adresse ein, um die von ihm abgebauten Bitcoins zu erhalten; eine Art virtuelles Postfach mit einer Zeichenkette von 27-34 Zahlen und Buchstaben. Im Gegensatz zu einem Postfach ist die Identität des Benutzers nicht daran gebunden. […]

Übersetzt mit www.DeepL.com/Translator

Textstufe 1 – Text 3

Hyatt, Edward & Lockett Jon (2018). Geek's Gold. What is Bitcoin, why has the price collapsed and how can you buy the cryptocurrency? *The Sun*, 16.10.2018. https://www.thesun.co.uk/money/3000715/bitcoin-what-is-price-gbp-usd-today-value-cryptocurrency-buy/ (Stand: 08.10.2019).

By Jon Lockett and Edward Hyatt | October 16th, 2018

The Bitcoin protocol – the rules that make Bitcoin work – say that only 21 million Bitcoins can ever be created by miners.

However, these coins can be divided into smaller parts with the smallest divisible amount one hundred millionth of a Bitcoin.

This is called a "Satoshi", after the founder.

To receive a Bitcoin, a user must have a Bitcoin address - a string of 27-34 letters and numbers - which acts as a kind of virtual postbox.

Since there is no register of these addresses, people can use them to protect their anonymity when making a transaction.

These addresses are in turn stored in Bitcoin wallets, which are used to manage savings. […]

Übersetzung DeepL:

Das Bitcoin-Protokoll - die Regeln, nach denen Bitcoin funktioniert - besagt, dass nur 21 Millionen Bitcoins jemals von Bergleuten geschaffen werden können.

Diese Münzen können jedoch in kleinere Teile mit der kleinsten teilbaren Menge von einem hundert Millionstel einer Bitcoin unterteilt werden.

Dies wird nach dem Gründer als "Satoshi" bezeichnet.

Um einen Bitcoin zu erhalten, muss ein Benutzer eine Bitcoin-Adresse - eine Zeichenkette von 27-34 Buchstaben und Zahlen - haben, die wie eine Art virtueller Briefkasten wirkt.

Da es kein Register dieser Adressen gibt, können Menschen diese verwenden, um ihre Anonymität bei einer Transaktion zu schützen.

Diese Adressen werden wiederum in Bitcoin-Wallets gespeichert, die zur Verwaltung von Einsparungen verwendet werden. […]

Übersetzt mit www.DeepL.com/Translator

Textstufe 3 – Text 4

Azadinamin, Amirsaleh (2013). The Bankruptcy of Lehman Brothers: Causes of Failure & Recommendations Going Forward. *Social Science Research*, 4-5.

By Amirsaleh Azadinamin | August 2013

Subsequently, these failures caused a chain reaction on the markets. The failure also spread over all securitization vehicles, in which they are designed to allow a company or a bank holding assets with little liquidity to group them together and sell them to a specialized entity that is created for the same purpose (Morin and Maux, 2011).
Securitization therefore enables an organization to dispose of assets while immediately obtaining capital in exchange, a process which represents a new means of financing for these entities. Credit then becomes liquid; however, if it is based on a poor risk, someone will eventually have to pay the piper (p. 41).

Übersetzung DeepL:

Das Scheitern erstreckte sich auch auf alle Verbriefungsträger, bei denen sie so konzipiert sind, dass sie es einem Unternehmen oder einer Bank, die Vermögenswerte mit geringer Liquidität hält, ermöglichen, diese zusammenzufassen und an eine spezialisierte Einheit zu verkaufen, die für denselben Zweck geschaffen wurde (Morin und Maux, 2011).
Die Verbriefung ermöglicht es einem Unternehmen daher, Vermögenswerte zu veräußern und gleichzeitig sofort Kapital im Austausch zu beschaffen, ein Prozess, der ein neues Finanzierungsinstrument für diese Unternehmen darstellt. Der Kredit wird dann liquide; wenn er jedoch auf einem schlechten Risiko basiert, muss irgendwann jemand den Pfeifer bezahlen (S. 41).

Übersetzt mit www.DeepL.com/Translator

Textstufe 2 – Text 5

Economist (2013). Crash course; The origins of the financial crisis. *The Economist*, 07.09.2013. https://pdfs.semanticscholar.org/99ed/12612243475696e54225daa3dc619fd3b1d2.pdf (Stand: 08.10.2019).

By Anonymous | September 7th, 2013

Low interest rates created an incentive for banks, hedge funds and other investors to hunt for riskier assets that offered higher returns. They also made it profitable for such outfits to borrow and use the extra cash to amplify their investments, on the assumption that the returns would exceed the cost of borrowing. The low volatility of the Great Moderation increased the temptation to "leverage" in this way. If short-term interest rates are low but unstable, investors will hesitate before leveraging their bets. But if rates appear stable, investors will take the risk of borrowing in the money markets to buy longer-dated, higher-yielding securities. That is indeed what happened.

Übersetzung DeepL:

Niedrige Zinssätze schufen einen Anreiz für Banken, Hedgefonds und andere Investoren, nach risikoreicheren Anlagen mit höheren Renditen zu suchen. Sie machten es auch profitabel für solche Outfits, sich Geld zu leihen und das zusätzliche Geld zu verwenden, um ihre Investitionen zu verstärken, unter der Annahme, dass die Renditen die Kosten der Kreditaufnahme übersteigen würden. Die geringe Volatilität der Großen Moderation erhöhte die Versuchung, auf diese Weise "zu hebeln". Wenn die kurzfristigen Zinssätze niedrig, aber instabil sind, werden die Anleger zögern, bevor sie ihre Wetten nutzen. Aber wenn die Zinsen stabil erscheinen, gehen die Anleger das Risiko ein, an den Geldmärkten Kredite aufzunehmen, um länger laufende, höher verzinsliche Wertpapiere zu kaufen. Das ist in der Tat geschehen.

Übersetzt mit www.DeepL.com/Translator

Textstufe 1 – Text 6

Stow, Nicola (2018). Death Bank. Why did Lehman Brothers collapse and what caused the global financial crisis in 2008? *The Sun*, 15.12.2018. https://www.thesun.co.uk/news/7265360/lehman-brothers-bank-collapse-crisis-bankruptcy/ (Stand: 08.10.2019).

By Nicola Stow | December 15th, 2018

There were a huge number of factors that led to the collapse of the global economy, but it was mostly abuse of mortgage lending without strict regulation or understanding.

These included:

- Low interest rates which encouraged mortgage lending
- Securitisation, which is when mortgages were bundled together and sold as low-risk securities
- Lax regulation allowed predatory lending in the private sector
- The Community Reinvestment Act (CRA), a 1977 US federal law designed to help low- and moderate-income Americans get mortgage loans encouraged banks to grant mortgages to higher risk families
- Mortgage guarantees. Many of the subprime (high risk) loans were bundled and sold, finally accruing to quasi-government agencies

Übersetzung DeepL:

Es gab eine Vielzahl von Faktoren, die zum Zusammenbruch der Weltwirtschaft führten, aber es war vor allem der Missbrauch von Hypothekarkrediten ohne strenge Regulierung oder Verständnis.
Dazu gehörten:

- Niedrige Zinssätze, die die Hypothekarkredite begünstigten.
- Verbriefung, d.h. wenn Hypotheken gebündelt und als risikoarme Wertpapiere verkauft wurden.
- Laxe Regulierung erlaubt Verdrängungskredite im privaten Sektor
- Der Community Reinvestment Act (CRA), ein US-Bundesgesetz von 1977, das Amerikanern mit niedrigem und mittlerem Einkommen helfen soll, Hypothekendarlehen zu erhalten, ermutigte die Banken, Hypotheken an Familien mit höherem Risiko zu vergeben.
- Hypothekengarantien. Viele der Subprime-Kredite (mit hohem Risiko) wurden gebündelt und verkauft, die schließlich an Quasi-Regierungsstellen gingen.

Übersetzt mit www.DeepL.com/Translator

US-China Trade War

Textstufe 3 – Text 7

Onyusheva, Irina V.; Thinn Nain, Cherry & Zaw, Aung Lin (2019). The US-China Trade War: Cause-Effect Analysis. *The EUrASEANs: journal on global socio-economic dynamics* 1 (14), 8-12.

By Irina V. Onyusheva, Cherry Thinn Nain & Aung Lin Zaw | 2019

Timeline of the US-China Trade War and the Current Actions

The ongoing US-China trade war started in 2018, and its timeline was briefly as follows (Trump tariffs: the US escalates trade threats to China, 2018):
January 2018: the US government announced new tariffs on the import of solar panels and washing machines;

March 2018: the US government announced that it is planning to impose 25% tariff on steel and 10% on aluminum imports; April, same year: China retaliated back by announcing tariffs' rise (from 15% up to 25%) on the US imports;

Übersetzung DeepL:

Zeitrahmen des US-amerikanischen und chinesischen Handelskrieges und der laufenden Maßnahmen
Der andauernde Handelskrieg zwischen den USA und China begann 2018, und sein Zeitrahmen war kurz wie folgt (Trump tariffs: the US escalates trade threats to China, 2018):
Januar 2018: Die US-Regierung kündigte neue Zölle auf den Import von Solarmodulen und Waschmaschinen an;

März 2018: Die US-Regierung kündigte an, dass sie plant, im April desselben Jahres 25% Zölle auf Stahl und 10% auf Aluminiumimporte zu erheben: China schlug zurück, indem es die Erhöhung der Zölle (von 15% auf 25%) für die US-Importe ankündigte;

Übersetzt mit www.DeepL.com/Translator

Textstufe 2 – Text 8

Sarda, Pranit (2019). Everything you need to know about US-China Trade War. *Forbes India*, 14.05.2019. http://www.forbesindia.com/article/special/everything-you-need-to-know-about-uschina-trade-war/53525/1 (Stand: 15.10.2019).

By Pranit Sarda | May 14th, 2019

US-China Trade War Current Situation

Until **May 10**, **2019,** China was paying 25 percent tariffs only on $50 billion worth of goods and a lower 10 percent on $200 billion worth of goods. Trade talks between the two countries to quell the trade tensions fell through on Friday without reaching an agreement, and on May 5, 2019, President Trump tweeted saying that the US will raise its tariffs on all $200 billion worth of Chinese goods to 25 percent from the previous 10 percent.

The US imported $539 billion worth of Chinese goods in 2018. As it stands, a 25 percent tariff is imposed on $250 billion worth of goods currently. Trump has also threatened to levy an additional 25% tariff on goods worth $325 billion.

China's retaliation: Responding to the increase in tariffs by the Trump administration, China, on Monday decided to retaliate by increasing tariffs on $60 billion worth of US goods, effective from June 1, 2019.

Übersetzung DeepL:

US-China Handelskrieg Aktuelle Situation

Bis zum 10. Mai 2019 zahlte China 25 Prozent Zölle nur auf Waren im Wert von 50 Milliarden Dollar und weniger als 10 Prozent auf Waren im Wert von 200 Milliarden Dollar. Die Handelsgespräche zwischen den beiden Ländern, um die Handelsspannungen zu überwinden, fielen am Freitag ohne eine Einigung, und am 5. Mai 2019 tweeted Präsident Trump, dass die USA ihre Zölle auf alle chinesischen Waren im Wert von 200 Milliarden Dollar von zuvor 10 Prozent auf 25 Prozent anheben werden.

Die USA importierten 2018 chinesische Waren im Wert von 539 Milliarden Dollar. Gegenwärtig wird ein Zoll von 25 Prozent auf Waren im Wert von 250 Milliarden Dollar erhoben. Trump hat auch gedroht, einen zusätzlichen Zoll von 25% auf Waren im Wert von 325 Milliarden Dollar zu erheben.

Chinas Vergeltung: Als Reaktion auf die Erhöhung der Zölle durch die Trump-Administration, China, am Montag beschlossen, Gegenmaßnahmen zu ergreifen, indem die Zölle für US-Waren im Wert von 60 Milliarden Dollar mit Wirkung zum 1. Juni 2019 erhöht wurden.

Übersetzt mit www.DeepL.com/Translator

Textstufe 1 – Text 9

Rogers, Jon (2019). Trading Blows: Donald Trump's trade war with China explained. *The Sun*, 27.08.2019. https://www.thesun.co.uk/news/9082139/us-china-trade-war-trump-jinping/ (Stand: 15.10.2019).

By Jon Rogers | August 27th, 2019

Blaming the decline on "trade protectionism", China's action followed Donald Trump's threat last week to slap punitive tariffs to an additional $300billion of Chinese imports.

The Ministry of Commerce announced it was suspending promised purchases of American farm products.

As a result, technology stocks bore the brunt of the selling, with Apple sliding 5.2 per cent. The tech depends on Chinese factories to assemble its iPhones.

Übersetzung DeepL:

Die chinesische Aktion, die den Rückgang auf den "Handelsprotektionismus" zurückführen sollte, folgte der Drohung von Donald Trump letzte Woche, Strafzölle auf zusätzliche 300 Milliarden Dollar an chinesischen Importen zu erheben.

Das Handelsministerium teilte mit, dass es die zugesagten Käufe von amerikanischen Agrarprodukten aussetzt.

Infolgedessen trugen Technologiewerte die Hauptlast der Verkäufe, wobei Apple um 5,2 Prozent nachgab. Die Technologie hängt von chinesischen Fabriken ab, um ihre iPhones zu montieren.

Übersetzt mit www.DeepL.com/Translator