universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Menschliche und automatische Evaluation von Übersetzungen von Fachtexten in Google Translate

verfasst von / submitted by

Rosita Lo Presti

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

Master of Arts (MA)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 060 331 342
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Übersetzen Deutsch Englisch
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Gerhard Budin

Danksagung

Mein Dank gilt in erster Linie meinem Betreuer, Herrn Univ.-Prof. Dr. Gerhard Budin, der mich bei der Themenfindung auf den richtigen Pfad geleitet hat und bei Unklarheiten mit seinem fachlichen Wissen jederzeit zur Verfügung stand.

Das Gelingen dieser Masterarbeit wäre ohne die wertvolle Teilnahme der Fachexperten an meiner Studie nicht möglich gewesen. Daher möchte ich mich an dieser Stelle bei Herrn Doktor Giancarlo Buccheri und Frau Doktor Grazia Dugo bedanken, die ihr medizinisches Fachwissen zur Verfügung gestellt und einen großen Beitrag zu der Interdisziplinarität dieser Arbeit geleistet haben. Bedanken möchte ich mich ebenfalls bei den Übersetzerinnen – und lieben Freundinnen – Rebecca Dattaro und Chiara Mezzasalma für die sorgfältige Durchführung des Nachbearbeitungsauftrags und die wertvollen Feedbacks. Ein besonderer Dank gilt meinem Bruder, Carlo Lo Presti, sowohl für die technische Unterstützung und die Realisierung der Webschnittstelle als auch für die Leidenschaft, mit der er an meinem Projekt teilgenommen hat und für seine Inspiration. Ich bedanke mich außerdem bei meiner lieben Freundin Roberta Sgroi, die mit Geduld alle meinen Fragen aus dem Bereich der theoretischen Informatik beantwortet hat und die mich Rat und Tat unterstützt hat.

Ich danke meinem Freund Mobarez Hazrat, der mir stets zur Seite stand und mich in schwierigen Momenten mit seiner positiven Denkweise stets ermutigt hat. Schließlich möchte ich mich ganz herzlich bei meinen lieben Eltern bedanken, die dieses Auslandsstudium an der Universität Wien ermöglicht haben, stets an mich geglaubt haben und mich während meiner gesamten Studienzeit unterstützt haben. Grazie.

Inhaltsverzeichnis

1. Einleitung

„Every time I fire a linguist,
our system performance improves.“[1]
Fred Jelinek

Seit ihren ersten Schritten innerhalb der Wissenschaftsgemeinde in den fünfziger Jahren hat die maschinelle Übersetzung (MÜ) absolute Höhe– und Tiefpunkte erlebt. Zeitweilig glaubten Forscher[2] und Forschungsförderer stark an die Simulierbarkeit des Humanübersetzens. Zu anderen Zeitpunkten wurde die MÜ hingegen als ein fehlgeschlagenes Projekt gesehen. Schrittweise hat die MÜ aber ihre Rolle nicht nur in der Wissenschaftsgemeinde, sondern auch in der Gesellschaft und im Alltag von Millionen von Menschen gefunden. Diese sind zweifelsohne die Blütezeiten der maschinellen Übersetzung.

Die MÜ wird dabei nicht nur für kommerzielle Zwecke und als Unterstützung für die Arbeit der Übersetzer eingesetzt, sondern auch Privatbenutzern greifen immer mehr auf diese Art der Übersetzung zurück. Eines der am meisten verwendeten MÜ-Systeme – Google Translate – zählte schon im Jahr 2012 200 Millionen Benutzer und eine Milliarde erstellter Übersetzungen täglich[3].

Diese Zahlen bestätigen zweifelsohne den Erfolg der maschinellen Übersetzung und sind ein Zeichen dafür, dass sich die MÜ, insbesondere seitdem die statistischen MÜ-Systeme das wissenschaftliche Paradigma darstellen, sehr verbessert hat. Obwohl bis heute nicht denkbar ist, dass die MÜ irgendwann an die Qualität der Humanübersetzung herankommt, ist die Qualität von professionell übersetzten Texten aber der goldene Standard und gilt als Maßstab für die Bewertung der maschinellen Übersetzung.

Wie können aber die Verbesserungen eines MÜ-Systems bemessen werden? Was zeichnet die Qualität maschineller Übersetzungen aus? Diese Fragen haben sich die Forscher seit den ersten Entwicklungen in der maschinellen Übersetzung gestellt und aus diesen Fragen hat sich ein Parallelbereich zur maschinellen Übersetzung entwickelt, nämlich die Evaluation maschineller Übersetzungen. Zahlreiche Parameter, Kriterien und Bewertungsskalen wurden über die Jahre entwickelt, um die Qualität einer Übersetzung objektiven Kategorien zuzuordnen. Objektivität und Übersetzung ist aber kein harmonisches Wortpaar, denn die Aspekte einer Übersetzung sind vielseitig und lassen sich nur schwer

[1] Anlässlich des Workshop on Evaluation of NLP Systems, Wayne PA, December 1988 (Jelinek 2004).
[2] Im Sinne einer geschlechtsneutralen Formulierung beziehen sich die in dieser Arbeit verwendeten männlichen Formen sowohl auf Frauen als auch auf Männer.
[3] Daten veröffentlicht von Estelle/Khare anlässlich des Treffens von Google-Entwicklern Google I/O 2013.

klassifizieren. Eine objektive Klassifikation erfordert dann auch objektive Evaluatoren. Die obersten Richter, welche die Qualität der maschinellen Übersetzung beurteilen können, sind Endbenutzer, Linguisten oder die Softwareentwickler selbst: auf jeden Fall aber Menschen. Den menschlichen Evaluatoren fehlt aber die Objektivität, die für die Bewertung und die darauffolgende Verbesserung des Systems notwendig ist. Darüber hinaus ist das Einbeziehen von Evaluatoren in große Evaluationsprojekte aus praktischen Gründen nicht immer vorteilhaft: Die menschliche Evaluation ist kostspielig und zeitintensiv.

Aus diesen Gründen wurden automatische Methoden entwickelt, um die Qualität der maschinellen Übersetzung schnell und objektiv beurteilen zu können. Diese Metriken weisen den MÜ-Systemen bessere oder schlechtere Werte zu. Die Benutzung von automatischen Systemen ist aber umstritten und das Thema der Zuverlässigkeit dieser Metriken ist sehr debattiert. Was sagen diese rein numerischen Werte über die tatsächliche Qualität einer Übersetzung aus? Ist es möglich, eine Korrelation zwischen objektiven numerischen Werten und subjektiven menschlichen Evalautionen herzustellen? Das Ziel der vorliegenden Masterarbeit besteht darin, anhand eines kritischen Vergleichs und einer Fallstudie festzustellen, inwieweit die Werte der automatischen Metriken mit menschlichen Evaluationen korrelieren und welche Informationen hinter den rein numerischen Werten stecken.

2. Evaluation maschineller Übersetzung

Die Evaluation maschineller Übersetzung (MÜ) ist ein Forschungsbereich, der sich parallel zur MÜ entwickelt hat und die Fortschritte in den MÜ-Systemen geprägt hat. Von den Ergebnissen der Evaluationen hängen nämlich die Kaufentscheidungen der Endbenutzer des Systems oder der Kapitalgeber, die die Entwicklung des Systems finanzieren, ab. Die Gründe der Durchführung einer Evaluation sind sehr unterschiedlich: Man könnte beispielweise bestimmen wollen, ob eine Software sich gut für einen bestimmten Zweck oder Fachgebiet eignet oder man könnte die Ersparnis berechnen wollen, die durch die Benutzung eines bestimmten Systems erzielt werden könnte. Die Evaluation erfolgt auch während der Anfangsphase der Entwicklung der Software oder nach Änderungen bzw. Verbesserungen am Sourcecode. Die Teilnehmer an der Evaluation sind dementsprechend auch sehr unterschiedlich: vom Endbenutzer und Post-Editor bis hin zum Softwareentwickler oder Forschungsförderer. Der Output könnte auch noch zu anderen Zwecken verarbeitet werden (*downstream tasks*), wie beispielweise zur Textklassifikation und zur Informationsrückgewinnung **(**vgl. White et al. 2000). Die Evaluation des Systems ist ein komplexes Verfahren, das eine genaue Planung erfordert, die sämtlich der erwähnten Variablen berücksichtig. Folgende Fragen bieten eine klare Darstellung der Kernelemente jeder Evaluation:

(a) In welcher Phase der Entwicklung der Software erfolgt die Evaluation?
(b) Wer sind die Teilnehmer?
(c) Was ist der Zweck der Evaluation?
(d) Was ist der Zweck des Outputs?
(e) Was ist der Input bzw. welche Merkmale haben die Texte, die maschinell übersetzt werden sollen?

Durch die Zusammensetzung dieser Fragen kann man daher folgende zwei Hauptfragen beantworten: Was zeichnet die Qualität eines MÜ-Systems aus? Mit welchen Methoden kann ein System bewertet werden?

Die Bestimmung des theoretischen Rahmens für eine übergreifende Bezeichnung der Qualität und die Bestimmung der Qualitätsparameter der maschinellen Übersetzung ist umso schwieriger, zumal die Kriterien zwei Forschungsbereiche zusammenführen bzw. integrieren sollen: die Qualitätsanalyse einer Software und die Qualitätsanalyse einer Übersetzung. Im Laufe des Kapitels werden beide Seiten der Evaluation erklärt. Im nächsten Kapitel wird

versucht, eine Korrelation zwischen den translationswissenschaftlichen Grundlagen und der maschinellen Übersetzung zu erstellen.

2.1 Skopostheorie und deren Umsetzung auf die MÜ

Um die Qualität einer Übersetzung, im Sinne einer Humanübersetzung, zu bestimmen, hat der junge Forschungsbereich der Translationswissenschaft im Laufe der letzten sechzig Jahren unterschiedliche Ansätze vorgeschlagen – kommunikative, funktionale, polysystemische, linguistische Ansätze usw. Da das Ziel dieser Arbeit auf die MÜ begrenzt ist, soll hier auf eine Ausführung der translationswissenschaftlichen Ansätze verzichtet werden. Es wird lediglich ein Ansatz erwähnt, der auch auf die MÜ anwendbar ist.

Der translationswissenschaftliche Ansatz, der sich am besten für die maschinelle Übersetzung einigt, ist die Skopostheorie von Reiß/Vermeer (vgl. Trujillo 1999:3). Die oben genannten Fragen, die zur Evaluation einer maschinellen Übersetzung dienen, entsprechen nämlich den Faktoren, die nach Reiß und Vermeer (1991) die Qualität einer Übersetzung bestimmen. „Dominante aller Translation ist deren Zweck “ (Reiß/Vermeer 1991:96), also des Skopos, d.h. die Zielvorgabe/das Ziel einer Translation (vgl. Prunç 2001:163).

Bei der Auswahl des Systems oder bei der Entwicklung bestimmter Merkmale der Software wird darauf geachtet, was der Zweck der MÜ ist. Natürlich sind die praktischen Anwendungen der Skopostheorie für Humanübersetzer ausgearbeitet worden und können nicht 1:1 für die maschinelle Übersetzung übernommen werden. Die Zwecke der Humanübersetzung können nicht vom MÜ-System ausgeführt werden, denn diese setzten das Humanverstehen voraus. Solch ein Vergleich wäre auch wenig zielführend, denn laut dem aktuellen Stand der Dinge ist die *Fully Automated Machine Translation* (FAMT) nicht mit der Humanübersetzung vergleichbar. Humanübersetzung und FAMT sind zwei unterschiedliche Übersetzungsprodukte mit unterschiedlichen Zwecken, Anforderungen und Teilnehmer an der kommunikativen Handlung.

Die Skopostheorie ist aber auch auf die maschinelle Übersetzung umsetzbar, indem der Zweck für die MÜ – genau wie die Humanübersetzung –ausschlaggebend ist. Ist beispielweise der Zweck eines MÜ-Outputs das Verstehen der Kernaussagen (*gisting*), dann ist ein korrekter Inhalt für den Zweck unabdingbar, auch im Falle, in dem der Output nicht besonders flüssig lesbar ist. Wird der Output für das Post-Editing verwendet, dann können auch Rechtschreibfehler, Wortstellung und grammatikalische Fehler die Qualität des Outputs beeinträchtigen.

In der Humanübersetzung ist der Skopos vom Auftraggeber bestimmt oder durch vom Kontext ableitbare Informationen. In der MÜ hat der Skopos zwei Facetten: Skopos des Outputs und Skopos der Evaluation. Der wichtige Unterschied zwischen Skopos bei HÜ und

bei MÜ besteht darin, dass ein Humanübersetzer in der Regel die Übersetzung je nach Skopos adaptieren kann. Ein MÜ-System ist hingegen in der Anpassung an einen Skopos sehr viel weniger flexibel. Daraus folgt, dass der Begriff des Skopos vielmehr innerhalb einer Evaluation verwendet werden kann, d.h. eine a posteriori Einschätzung des (nicht) erfolgreich erzielten Zwecks der MÜ. Der Skopos der Evaluation prägt dagegen die Merkmale der Evaluation selbst.

In den Kapiteln 5.1 bis 5.2 werden die Begriffe der Translationswissenschaft anhand eines praktischen Beispiels (die maschinelle Übersetzung von Fachtexten) auf die Analyse der MÜ umgesetzt, und es wird ersichtlich, dass Begriffe wie inhaltliche Invarianz, Äquivalenz, Informativität, Verständlichkeit, auf die MÜ umsetzbar sind.

Die Qualität der MÜ und daher auch deren Evaluation, wie schon früher erwähnt wurde, bedeutet nicht nur die Bewertung des Outputs. Jedes MÜ-System ist im Prinzip eine Software, das die Qualitätskriterien der Software erfüllen soll. Die Evaluationsmethoden der Informatik und der Sprachwissenschaft ergänzen sich gegenseitig und die tiefere Ausführung der Evaluationsmethode erfordert einen holistischen Ansatz. Im Laufe des Kapitels werden die für eine Evaluation notwendigen Faktoren (vgl. S. 3) anhand theoretischer Ansätze erklärt.

2.2 Evaluationsmethoden

Die Evaluationsmethoden der maschinellen Übersetzung sind vielseitig. Es werden daher in den nächsten Seiten die wichtigsten Anhaltspunkte jeder Evaluation vorgestellt. Die praktischen Ansätze der Evaluation teilen sich in intrinsische und extrinsische, *glass box* und *back box*, *test suite* und *test corpus*, auf. Eine weitere Klassifikation ist jene in automatische und menschliche Evaluationsmethoden, die als Ausgangspunkt dieser Arbeit gedient haben. Diese zwei Hauptkategorien sind in diesem Kapitel nicht enthalten, weil ihnen zwei eigene Kapitel (Kapitel 3 und 4) gewidmet wurden.

2.2.1 Intrinsische und extrinsische Methoden

Die erste Unterscheidung ist jene zwischen intrinsische und extrinsische Evaluationsmethoden.[4] Die **intrinsischen Evaluationsmethoden** fokussieren auf die Qualität des Outputs. Der MÜ-Output kann allein bewertet werden oder durch einen Vergleich mit anderen Outputs. Bei den intrinsischen Evaluationsmethoden können die Qualität und die

[4] Die hier genannten Methoden werden nur als Beispiel erwähnt, um die Begriffe extrinsisch und intrinsisch erklären zu können. Sie werden im Kapitel 3 ausführlicher erklärt.

Informativität des Outputs analysiert werden (vgl. Dorr et. al 2011:745). Zu den gängigsten menschlichen Methoden[5] zählen:

(a) *quality assessment:* die Bewertung von unterschiedlichen Merkmalen wie *fluency, intelligibility, adequacy /accurancy*. Diese Bewertungen können durch Evaluatoren, die Ausgangs-und Zielsprache beherrschen, durchgeführt werden oder durch einsprachige Evaluatoren, die den Output mit einer oder mehreren Referenzübersetzungen vergleichen.

(b) *translation ranking,* bei welchem die Evaluatoren gebeten werden, die Outputs von gut bis schlecht zu reihen.

(c) Fehleranalyse, in der nach Fehlern im Output gesucht wird, die dann klassifiziert und gezählt werden, um eine Häufigkeitsverteilung zu erzielen (vgl. Kit/Wong 2015).

Die **extrinsischen Methoden** sind die sogenannten *task-based* Methoden und evaluieren die Verwendbarkeit (*usability*) eines Outputs für eine spezifische Aufgabe bzw. Zweck (vgl. Dorr et al. 2011). Beispiele für diese Methoden sind:

(a) Informationsrückgewinnung: Die *usability* (Benutzbarkeit) eines Outputs ist direkt proportional zur Anzahl der korrekten Schlüssel-Informationen, die von den Evaluatoren (potentielle Endbenutzer) in einem MÜ-Output festgestellt werden.

(b) Leseverstehen des Outputs: Unter *usability* versteht man hier ein höheren Verständlichkeitsgrad des Texts, der durch die Beantwortung einigen Fragen getestet wird.

(c) Post-Editing: Der Post-Editing-Aufwand und die Post-Editing-Zeit sollen in der Regel bei besseren Outputs niedriger sein (vgl. Kit/Wong 2015).

2.2.2 *Glass box* und *black box*

Eine Evaluation – ungeachtet der Evaluationsteilnehmer – kann als *glass* box oder *black box* durchgeführt werden. Dabei handelt es sich um zwei Verfahren, die in der Informatik benutzt werden, um eine Software auf zwei Ebenen zu testen.

Bei ***black box* Evaluationen** wird das MÜ-System nämlich als eine „schwarze Box" gesehen und die vom System durchgeführten Operationen werden nur hinsichtlich des Input-Output-Verhaltens betrachtet – die systeminternen Operationen werden hier nicht berücksichtigt. Diese Evaluationen sollten nicht vom Softwareentwickler durchgeführt werden und können (aber nicht unbedingt) die tatsächlichen Endbenutzer mit einbeziehen. Wenn die Endbenutzer nicht teilnehmen, werden Merkmale wie Funktionalität und Wiederherstellung oder der Umfang der bearbeiteten Daten getestet. Wenn die Endbenutzer teilnehmen, kann die Evaluation in Form eines Labortestes oder einer Feldforschung

[5] Auf die dazu entsprechenden automatischen Methoden wird im Kapitel 4 näher eingegangen.

erfolgen, in der die Endbenutzer bei der Benutzung der Software an ihren normalen Arbeitsplätzen beobachtet werden (vgl. Trujillo 1999:256).

Das *black box*-Verfahren zielt daher darauf ab, das objektive Verhalten des Systems bei einem vorbestimmten Evaluationsset zu testen. Ein fairer Vergleich zweier oder mehrerer MÜ-Systeme durch diese Methoden kann nur dann erfolgen, wenn entweder zwei Systeme für die Bearbeitung von Daten ausgelegt werden, die dieselben Merkmale des Evaluationssets haben oder, wenn das nicht der Fall ist, wird darauf abgezielt, die Stabilität der Anwendung bei Datentypen, die andere Merkmale aufweisen, zu testen (vgl. Kapitel 2.2.3) – beispielweise mit unterschiedlichen Syntax-, Gender-, Stil-Merkmalen (vgl. Dorr et. al 2011:745). Die Methode, die vorsieht, dasselbe Testset für die Bewertung der Performance mehrerer MÜ-Systeme oder die Performance eines Systems nach etwaigen Änderungen zu testen, hat sich für die Evaluation der maschinellen Übersetzung als von unschätzbarem Wert erwiesen und stellt noch einen sehr aktiven Forschungsbereich dar, insbesondere im Hinblick auf die automatischen Evaluationsmethoden (vgl. Dorr et. al 2011:745). Es ist genau das Verfahren, auf das die aktuellen Evaluation-Kampagnen basieren (vgl. Kapitel 4.2).

Die ***glass box*** oder ***white box*** **Evaluation** misst die Qualität des Systems, indem sie sich auf interne Merkmale der Anwendung konzentriert. Hier werden die einzelnen Komponenten des Systems und deren gegenseitenger Einfluss getestet. Diese Methode ist besonders für Entwickler und Forscher relevant, die die Module identifizieren wollen, die einen Mangel aufweisen. Die Analyse kann statisch (die Anwendung wird kontrolliert, ohne sie auszuführen) oder dynamisch sein (vgl. Trujillo 1999:256).
Diese Methode wird insbesondere bei regelbasierten Systemen verwendet, zumal sie oftmals dazu dient, die Abdeckung der linguistischen Phänomene eines Systems und die Regeln zur Behandlung dieser Phänomenen zu testen. Dabei können die einzelnen linguistischen Komponenten erst durch eine *black box* Evaluation geprüft werden.

2.2.3 *Test suite* und *test corpus*

Eine weitere Klassifikation der Evaluationsmethoden kann durch das unterschiedliche Zusammensetzen der Testsets erfolgen. Unter Testset versteht man das Set von Texten, die für die Evaluation benutzt werden. Man unterscheidet hier *test suite* und *test corpus*.

Bei *test suite* enthält das Evaluationsset Texte, in denen ein spezifisches linguistisches Phänomen oder ein Übersetzungsproblem vorkommt. *Test suite* Evaluationen sind insbesondere für regelbasierte MÜ-Systeme relevant. Dabei kommt oftmals zum Tragen, dass das im System eingebettete linguistische Wissen fehlerhaft oder lückenhaft ist oder die grammatikalischen Regeln, die ein Phänomen behandelt, nicht richtig mit anderen Regeln funktionieren, wenn beide Phänomene gleichzeitig auftreten (vgl. Arnold 1994). Ein Beispiel

dafür ist das Behandeln der Bedeutung der Modalverben auf Englisch und die Verneinung: „*The printer can not be cleaned* (i.e. can be left uncleaned), and *The printer cannot be cleaned* (i.e. must not be cleaned) are confused.“ [Hervorhebung im Original] (Arnold 1994:167). Ein weiteres Beispiel kommt aus der von Gambäk et al. (1991) durchgeführten *test suite* Evaluation, in der unterschiedliche Transfer-Probleme getestet wurden, u.a. wird die Übersetzung des schwedischen *phrasal verb* tycka om (wörtlich "darüber denken") in unterschiedlichen Kontexten – wie Verneinung und W-Fragen - getestet (vgl. Trujillo 1999).
Test suite weisen aber drei wesentliche Probleme auf (vgl. Trujillo 1999:257):

(a) Die *test suite* Evaluationen nehmen an, dass das Verhalten eines Systems von sorgfältig konstruierten Beispielen auf reale Texte projiziert werden kann. Die Methoden testen daher keine Texte, die in der Realität vorkommen, das Verhalten wird hierbei nur angenommen.

(b) Es ist schwierig, zwei oder mehrere MÜ-Systeme durch diese Methode zu vergleichen, denn ein System kann ein sprachliches Phänomen besser „behandeln“ als andere, weist aber eine niedrige Performance bei anderen Sprachphänomenen auf. Die Entscheidung darüber, welches System besser als ein anderes ist, ist daher anhand einer *test suite* nicht möglich.

(c) Um Systeme durch *test suite* vollständig evaluieren zu können, sollte jedes Phänomen für jedes vom System unterstützte Sprachpaar getestet werden. Das ist natürlich kostspielig und zeitaufwändig.

Bei *test corpus* werden Korpora als Input für das MÜ-System verwendet. Diese Evaluation wird sowohl für kommerzielle als auch für experimentelle Systeme verwendet. Unterschiedliche Evaluationsmethoden können bei Textkorpora durchgeführt werden. Beispielweise berichten Bennet/Slocum (1988:128) über die damals laufende Evaluation des Systems METAL bei einem Korpus von ungefähr 1000 Seiten von realen Texten über einen Zeitraum von fünf Jahren. Dabei wurden zwischen 45% und 85% der Übersetzungen gefunden, die kein Post-Editing gebraucht hatten. King/Falkedal (1990) schlagen zwei Sets von Testdaten vor: Der erste Korpus enthält Datentypen, die für das System bestimmt sind; der Zweite ist ein Korpus mit nicht vorselektierten Texten. Das Ziel hierbei war, die Eignung (*suitability*) des Systems für andere Texttypen und seine Erweiterungsfähigkeit (*extendability*) zu testen. Die *corpus suite* werden heute insbesondere für automatische Evaluationsmethoden verwendet, wie beispielweise BLEU, die sich auf einem zweisprachigen Korpus stützt (vgl. Papineni et al., 2002). Scarton/Specia (2016) benutzten ein Leseverständnis-Korpus, der maschinell übersetzt wurde. Die Evaluatoren beantworten hier einige Fragen über den Text. Die Hypothese hierbei ist, dass je korrekter und ausführlicher die Fragen beantwortet werden, desto besser die Übersetzung ist (vgl. Scarton/Specia 2016).

2.3 Qualität und deren Evaluation

Im Laufe dieses Kapitels werden die wichtigsten Ansätze zur Evaluierung der maschinellen Übersetzung vorgestellt, die über die letzten siebzig Jahre entwickelt wurden und die noch heute als Grundlage der Evaluationsmethoden dienen. Im Kapitel wird ersichtlich, wie unterschiedliche Forscher der MÜ versucht haben, aus praktischen Evaluationssituationen theoretische Rahmen für die Evaluation der MÜ zu schaffen. Hierbei haben sie versucht zu definieren, wie Qualität in Bezug auf einen Text, eine Humanübersetzung und maschinelle Übersetzung zu verstehen ist, welche Rolle Evaluatoren und die Zwecke der MÜ spielen.

2.3.1 ALPAC-Bericht (1966)

Das bekannteste Ereignis in der Geschichte der MÜ ist zweifellos der Report des *Automatic Language Processing Advisory Committee* (ALPAC 1966). Obwohl die (negativen) Auswirkungen dieser Publikation bekanntlich zu einer Stillstandphase von 20 Jahren im Forschungsbereich der maschinellen Übersetzung insbesondere in den USA geführt haben, stellt der ALPAC-Report ein wichtiges Dokument für die Analyse der diachronischen Entwicklung der Evaluation dar.

Der ALPAC-Report widmet seinen Anhang X den Evaluationsmethoden der maschinellen Übersetzung. Die in diesem Anhang erklärten Methoden wurden im Rahmen des von John. B Carroll durchgeführten Experiments, das vom ALPAC finanziert wurde, umgesetzt. Schon auf den ersten Seiten des Anhangs X wird ein interessanter Aspekt hervorgehoben, und zwar wird noch nicht zwischen der Qualität der MÜ und der Qualität der Humanübersetzung unterschieden. Bei den ersten Entwicklungen in der MÜ hatte man nämlich noch stets die maschinelle Übersetzung als im Zusammenhang mit Humanübersetzung stehend verstanden. Das diskutierte Thema war die Untauglichkeit der maschinellen Übersetzung als Ersatz des Humanübersetzers. Nur sechs Jahre vor dem ALPAC-Bericht hatte nämlich Bar-Hillel das Wort *Fully Automatic High Quality Translation* (FAHQT) geprägt: „translation of the quality produced by an experienced human translator" (Bar-Hillel 1959:4). Dieser Blickwinkel auf die MÜ rechtfertigt daher die Grundlage dieser Studie, die genau geplant war, um die Voraussetzungen für ein standardisiertes Protokoll zum Messen der Qualität wissenschaftlicher Übersetzungen zu schaffen und dies, unabhängig davon, ob die Ausgangstexte maschinell oder von einem Übersetzer übersetzt worden waren (vgl. ALPAC 1966).

Es wurde aber im Bericht schon versucht, eine Standardisierung des Evaluationsprozesses zu erzielen, die das Problem der Subjektivität überwinden sollte oder zumindest abschwächen.

> „There have been other experiments on this problem […], but their methods for evaluating translations have been too laborious, too subject to arbitrariness in standards, or too lacking in reliability and/or validity to become generally accepted. The measurement procedure developed here gives promise of being amenable to refinement to the point where it will meet the requirements of relative simplicity and feasibility, fixed standards of evaluation, and high validity and reliability.“ (ALPAC 1966:67)

Die im Zitat erwähnte Vorgehensweise enthält zwei Begriffe, die noch in den heutigen Evaluationskampagnen (vgl. Kapitel 4.2) eine wichtige Rolle spielen: (a) *intelligibility* und (b) *fidelity*.

Bei der Bewertung der Verständlichkeit (*intelligibility*) hatten die Evaluatoren keinen Zugang zu den Originalsätzen. Für die Bewertung der *fidelity*, welche die Informativität einer Übersetzung in Bezug auf den Ausgangstext misst, war es hingegen nötig, den Ausgangstext vor Augen zu haben. Diese Methode ist sehr ähnlich zu der Evaluationsmethode, die noch heute verwendet wird, mit der einzigen Ausnahme, dass statt dem Originalsegment oft eine oder mehrere Referenzübersetzungen angezeigt werden (vgl. Kapitel 3.1). Das Verhältnis dieser zwei Merkmale bzw. Evaluationsmaßstäbe zueinander, hat sich allerdings auch verändert:

> „Conceptually, these characteristics are independent; that is, a translation could be highly intelligible and yet lacking in fidelity or accuracy. Conversely, a translation could be highly accurate and yet lacking in intelligibility; this would be likely to occur, however, only in cases where the original had low intelligibility.“ (ALPAC 1966:67)

Was zur Zeit des ALPAC-Berichts als Tatsache gesehen wurde, das heißt die Unabhängigkeit der *intelligibilty* von der *accurancy*, ist heute ein debattiertes Thema. Unterschiedliche Studien, die in den letzten Jahren durchgeführt wurden, haben nämlich gezeigt, dass diese zwei Merkmale sich gegenseitig beeinflussen können und aus diesem Grund wird oft eine einzige Evaluationsskala verwendet, die beide Merkmale zusammenführen sollte (vgl. Kapitel 3.1).

Dieser Report wurde für seine harte Kritik an der MÜ bekannt, die sich auf die Entwicklung der maschinellen Übersetzung in den folgenden 20 Jahren negativ ausgewirkt hat. Es kann keinen Zweifel daran geben, dass die damals untersuchten MÜ-Systeme mangelhafte und unangemessene Übersetzungen erstellt hatten, doch der Report enthielt unterschiedliche Schwächen, die zu einer unfairen Bewertung der maschinellen Übersetzung und ihres Einsatzpotentials führten. Eine wesentliche Schwachstelle in der Methodologie war beispielweise der ungünstige Vergleich zwischen den Ergebnissen der bewerteten maschinellen Übersetzungssysteme und den Outputs eines kleinen Demonstrationssystems. Einige der maschinellen Übersetzungssysteme steckten noch in der experimentellen Phase und waren nicht für einen bestimmten Zweck ausgerichtet. Auch deren Inputset wurde nicht

richtig vorbereitet (z.B. kein Update der Wörterbücher). Das kleinere Demonstrationssystem, wurde ausschließlich für ein eingeschränktes Set von Sätzen erstellt (vgl. Hutchins 1996):

> „The reader will find it instructive to compare the samples above with the results obtained on simple, or selected, text 10 years earlier (the Georgetown IBM Experiment, January 7, 1954) in that the earlier samples are more readable than the later ones." (ALPAC 1966:23)

2.3.2 Van Slype (1979)

Dreizehn Jahre nach dem ALPAC-Bericht erscheint eine weitere wichtige Publikation, welche die Evaluation der maschinellen Übersetzung dieses Mal detaillierter und strukturiert behandelt: *Critical study of methods for evaluating the quality of machine translation* (Van Slype 1979). Der Bericht ist das Ergebnis einer Studie zur Evaluation, mit der Van Slype von der Europäischen Kommission beauftragt wurde, und die eine vollständige Sammlung der Evaluationsmethoden, der Wissenschaftler, die die Methoden erstellt haben, und der Studien, in den die Methoden angewandt wurden, darstellt.
Im Gegensatz zum ALPAC-Bericht stellt Van Slype in seinem Bericht klar fest, dass Übersetzungsqualität kein absolutes Konzept ist und dass MÜ und Humanübersetzung zwei unterschiedliche Übersetzungsprodukte sind, die unterschiedliche Parameter erfordern:

> "Translation quality is not an absolute concept, and has to be assessed:
> - Relatively, applying several distinct criteria illuminating each special aspect of the quality of the translation
> - Allowing for the specific nature of MT, which is a product quite different from HT and for which a quite different market may open up." (Van Slype 1979:12)

Van Slype unterscheidet zwischen Makro-und Mikroevaluation: Zwei Ebenen, auf denen die Evaluation durchgeführt werden kann und für die unterschiedliche Qualitätsparamater und Evaluationsmethoden zu befolgen sind. Die Makroevaluation besteht darin, festzustellen, inwieweit ein System besser als ein anderes oder als letztere Versionen desselben Systems ist oder inwieweit es den Anforderungen der Anwender gerecht wird. Die Mikroevaluation geht ins Detail, und zielt darauf ab, die Verbesserungsmöglichkeiten festzustellen und eine Verbesserungsstrategie zu erstellen. Die Kriterien der Makroevaluation werden in vier Gruppen geteilt:

- Kognitive Ebene (Verständlichkeit, Inhaltstreue, Kohärenz, Brauchbarkeit, Akzeptabilität)
- Ökonomische Ebene (Lesegeschwindigkeit, Korrekturgeschwindigkeit, Übersetzungsgeschwindigkeit)

- Linguistische Ebene: Rekonstruktion der semantischen Beziehung, syntaktische und semantische Kohärenz, ‚absolute' Qualität, lexikalische Evaluation, syntaktische Evaluation, Fehleranalyse
- Operationale Ebene: automatische Spracherkennung und Überprüfung der vom Ersteller vorgelegten Ansprüche (vgl. Van Slype 1979:57-61).

Als Methoden für das Messen der ersten zwei Kriterien auf kognitiver Ebene (Verständlichkeit und Inhaltstreue) schlägt Van Slype eine Rating anhand einer Skala, ein Cloze-Test, Multiple-Choice-Fragen, ein Wissentest oder ein Noise- Test vor.
Akzeptabilität kann durch eine Umfrage unter den Endbenutzern berechnet werden. Die Verstehens-und Korrektionszeiten können auch als Faktoren berechnet werden, die in direkter Anhängigkeit zur Verständlichkeit und der Richtigkeit der Texte stehen (vgl. Van Slype 1979).
Die Kriterien und Methoden der Mikroevaluation werden vom Autor in fünf Ebenen aufgeteilt, in denen er quasi ‚medizinische Behandlung' des Systems vorschlägt (von Symptomenerkennung bis zur Überprüfung der erfolgreichen Therapie). Folgende Ebenen zählen dem Autor zufolge zu den wirksamsten:

- Grammatische symptomatische Ebene: Analyse der grammatischen Fehler
- Formale symptomatische Ebene: Klassifikation der Fehler nach der Korrektur seitens des Post-Editoren (Hinzufügen, Entfernen, Substitution oder Änderung der Stellung eines Wortes)
- Diagnostische Ebene: Analyse der Fehlerursachen (Analyse des Inputtextes, der Ausgangssprache, der Wörterbücher usw.)
- Therapeutische Ebene: Erkennen der Verbesserungen am System nach der durchgeführten Verbesserungsmaßnahmen (vgl. Van Slype 1979:14).

2.3.3 Lehrberger/Bourbeau (1988)

Einen weiteren Ansatz bieten Lehrberger/Bourbeau (1988), welche die Evaluationskriterien und Parameter nach Evaluationsteilnehmer unterteilten. Eine Evaluation kann durch die Systementwickler, Kunden oder Benutzer erfolgen. Der Systementwickler testet die Leistungsfähigkeit des Systems. Da er genau weiß, wie das linguistische Modell aufgebaut ist und wo die Schwachstellen des Systems liegen können, fokussiert er vielmehr auf die Ursache eines Fehlers als auf die Auswirkungen auf den Text (vgl. 1988:134). Die Kunden können eine Kosten-Nutzen-Evaluierung durchführen. Sie vergleichen die Kosten eines Humanübersetzers mit den Kosten der MÜ und die unterschiedlichen Kombinationen aus den beiden (MÜ plus Post-Editing vs. Humanübersetzung plus Korrekturlesen) (vgl. 1988:135).

Der Benutzer führt eine linguistische Evaluation durch. Da der Benutzer keinen Zugang zum linguistischen Modell hat, sondern nur die Effekte des Modells sehen kann, wird von den Autoren eine induktive Methode entwickelt. Die Benutzer sollen, anhand der vom System erstellen Rohübersetzung, die Fehler erkennen und sie nach linguistischen Phänomenen kategorisieren. Dadurch kann der Benutzer eine Hypothese über die Funktionsweise der grammatischen Regeln des Systems formulieren und ähnliche Beispiele erstellen. Diese ad hoc erstellten Beispiele werden dann durch das System übersetzt und können die Hypothese entweder bestätigen oder ablehnen (vgl. 1988:135f.).

Ein weiterer Faktor, den Lehrberger/Bourbeau hervorheben, ist, dass die Bedürfnisse der Benutzer für die Evaluation eine wichtige Rolle spielen. Diese teilen sich in drei Hauptgruppen:

(a) Merkmale der zu übersetzenden Texte: linguistische Merkmale (formaler oder informaler Stil, einfacher oder verworrener Satzbau, usw.) und Domain (wird ein spezifischer Wortschatz verwendet? Ist der Text für Fachleute oder für ein allgemeines Publikum bestimmt?) (vgl. 1988:140f.).
(b) Gemischter Grad an Automation des Übersetzungsprozesses: FAMT[6], HAMT[7], MAHT[8]. Natürlich führt das zu einer unterschiedlichen Bewertung der Kosten und der Qualität, welche die Übersetzung erreichen soll.
(c) Qualität, die für den Benutzer akzeptabel ist. Diese hängt vom voraussichtlichen Leser (Verteilung auf ein weites Publikum, Fachleute oder Laien) und von der Anwendungsart („*reading for full information content; - Scanning for particular information; - Instructions for carrying out specific tasks*“) ab (1988:143). Die Parameter für die Bemessung sind Inhaltstreue, Verständlichkeit und Stil (vgl. 1988:143).

2.3.4 Hutchins/Somers (1992)

Die drei oben erwähnten Kriterien werden auch von Hutchins/Somers (1992) als Parameter für die Evaluation eines MÜ-Outputs verwendet. Sie weisen jedoch auch auf die Einschränkungen dieser Parameter hin. Die zwei Autoren werfen eine Frage auf, die bis heute ungelöst bleibt: Wer sind die Evaluatoren? Welche Vor-und Nachteile hat die Miteinbeziehung bzw. Ausschließung von potentiellen Anwendern in der Evaluation?

> „A major deficiency is that many evaluations are undertaken by people with little or no expertise in MT techniques, unable to judge what is possible and what is unrealistic, unable to estimate the potential rather than current performance. On the other hand, `evaluations' made by MT researchers are often

[6] *Full automatic machine translation.*
[7] *Human aided machine translation.*
[8] *Machine aided human translation.*

> minimal and misleading: the demonstration of a system with a carefully selected set of sentences or sentence types is not the basis for claims about a Large-scale system.“ (Hutchins/Somers 1992:161)

Diese Frage ist noch immer aktuell. Auch in der im Rahmen dieser Masterarbeit durchgeführten Studie konnte dieses Problem festgestellt werden und durch eine gezielte Abklärung bzw. Einschulung der Evaluatoren teilweise gelöst (vgl. Kapitel 5.5) werden. Die Vorstellung, die potentielle Anwender (in diesem Fall die Evaluatoren) über die maschinelle Übersetzung haben, bleibt immer noch mit den Ansprüchen an die Humanübersetzung verbunden und beeinträchtigt damit die Evaluation der MÜ. Wie Hutchins/Somers (1992) behaupten, gilt es, die Wahrnehmung der MÜ in der Öffentlichkeit zu ändern, um die Tauglichkeit der Evaluatoren für die Bewertung der MÜ zu verbessern: „In view of the misconceptions and misunderstandings concerning nearly all aspects of MT, one role of evaluation must be to introduce realism in public discussions of what MT systems can and cannot do and what they may be able to do in the future.“ (Hutchins/Somers 1992:161)

Hutchins/Somers gehen dann näher auf die Beschreibung der drei Kriterien (Inhaltstreue, Verständlichkeit, und Stil), deren Evaluationsmethoden und die Einschränkungen ein. Zwei von den von den Autoren benutzten Kriterien – Inhaltstreue (*fidelity* oder *accuracy*) und Verständlichkeit (*intelligibility* oder *clarity*) wurden schon auf den letzten Seiten dargelegt. Die Definitionen von Hutchins/Somers weichen nicht von denen der vorgenannten Autoren ab. Unter Stil versteht sich, inwieweit die Übersetzung eine für einen bestimmten Kontext und Zweck geeignete Sprache verwendet (vgl. Hutchin/Somers 1992:163).

Im Hinblick auf die Inhaltstreue nennen die Autoren die vom ALPAC verwendeten Methoden – d.h. das Lesen der MÜ und des Ausgangstexts, um die Informativität zu bestimmen – und bezeichnen sie als extrem subjektiv. Als weitere möglichen Methoden erwähnen sie einerseits die *back-translation* (das Zurückübersetzen des Textes aus der Zielsprache in die Ausgangssprache), wobei die Mängel am System die Fehler verdoppelt könnten, und andererseits die Evaluation der praktischen Performance, beispielweise eine praktische Ausführung eines maschinell übersetzten Anleitungstextes.

Für die Bewertung der Verständlichkeit schlagen Hutchins/Somers die Flesch-Skala, den Cloze-Test, und den Leseverständnistest vor. Diese Verfahren seien laut den Autoren durch einen höheren Objektivitätsgrad gekennzeichnet. Dagegen scheint die Methode des Rankings von Übersetzungen von perfekt verständlich (*perfectly intelligible*) bis zu hoffnungslos unverständlich (*hopelessly unintelligible*) viel subjektiver. Die Subjektivität ist dadurch verschärft, dass in der Regel einzelne Sätze ohne den dazu gehörenden Kontext evaluiert werden (vgl. Hutchins/Somers 1992:164).

Die Beurteilung des Stils ist auch durch Subjektivität gekennzeichnet. Doch bleibt sie ein wesentliches Kriterium, weil die Unterscheidung der Fehler auch davon abhängt. Beispielweise werden Artikelwörter und Kopulas in der journalistischen Sprache

weggelassen. Das wäre bei anderen Textsorten als Fehler zu erachten (vgl. Hutchins/Somers 1992:164).

Eine weitere von den Autoren vorgeschlagene Evaluationsmethode ist eine Fehlerkategorisierung, aus welcher der erforderliche Korrekturaufwand beim Post-Editing hervorgeht. Es werden die Korrektionsmaßnahmen gezählt, die während des Post-Editings vorgenommen werden. Wie auch in den Kapiteln 3 und 6 dargelegt wird, ist diese Methode aufgrund von zwei Faktoren nicht objektiv genug: Einerseits kann die Bezeichnung von Fehler von Post-Editor zu Post-Editor abweichen, und anderseits gibt es unterschiedliche Niveaus von Akzeptabilität einer MÜ, die von externen Bedingungen abhängen (vgl. Hutchins/Somers 1992:164). Darüber hinaus ist der Korrektur-Aufwand am System bei einigen Fehlern größer als bei anderen. Einige lexikalische Fehler können beispielweise durch einfache Änderungen an den Wörterbüchern korrigiert werden, anderen beanspruchen Änderungen, die auch Implikationen auf der Ebene der einzelnen Übersetzungsmodule haben können[9]. Diese Analyse ist sicherlich für die Entwickler interessant, aber auch für potentielle Kunden, die gerne wissen möchten, ob und wie die Qualität des Systems verbessert werden kann (vgl. Hutchins/Somers 1992:164f.).

Darüber hinaus klassifizieren die zwei Autoren die Evaluationskriterien einerseits nach den Teilnehmern am Evaluationsprozesses (Entwickler, Forscher, potenzielle Anwender, Übersetzer und Rezipienten) - genau wie die anderen in diesem Kapitel genannten Autoren - und andererseits nach den Entwicklungsphasen der Software. Die genauen Phasen und deren Evaluation werden im Kapitel 2.4 betrachtet.

2.3.5 Arnold et. al (1994)

Einen anderen Blickwinkel auf die Evaluation der MÜ bieten Arnold et. al. (1994), welche die MÜ-Systeme auch im Hinblick auf ihre Umsetzung in einer Organisation bzw. einer Firma sehen, die eine großes Volumen an Texten übersetzen soll. Die Autoren versetzten sich in die Lage solcher Organisationen und analysieren Schritt für Schritt alle Faktoren, die eine Firma bei der Kaufentscheidung eines MÜ-Systems berücksichtigen würde.

Diese Faktoren sind beispielweise die organisatorischen Änderungen, d.h. die Änderungen, die sowohl die organisationsinternen Prozesse als auch die Rolle und Tätigkeiten des Personals beeinflussen werden, die technische Implementierung des MÜ-Systems in der Firma, der Status des Systemverkäufers im Hinblick auf so eine große Investition, die Geschwindigkeit des Systems[10], und die Qualität, der relevanteste Faktor in

[9] Diese Fehler und deren Korrekturmaßnahmen beziehen sich ausschließlich auf nicht statistisch basierte-MÜ-Systeme, die zur Zeit der Publikation die Regel darstellten.
[10] Für die früheren MÜ-Systeme konnte es sehr unterschiedlich sein.

diesem Kapitel und auch laut Arnold et. al eine entscheidende Bedingung für den Erfolg eines Systems, zumal eine niedrige Qualität zu höheren Post-Editing-Kosten führt.

Als Methode für die Beurteilung der Qualität eines MÜ-Outputs nennen Arnold et al. Verständlichkeit und Inhaltstreue. Die Methode, die traditionell seit dem ALPAC-Bericht verwendet wurde, das heißt die Bewertung auf einer 9-stelligen Skala, sei den Autoren zufolge zu detailliert für die MÜ-Evaluation und führe zur Dispersion der Ergebnisse. Sie schlagen eine 4-stellige Skala vor. Die Inhaltstreue (*accuracy* oder *fidelity*) kann auch mit einer 4-stelligen Skala bewertet werden (vgl. 1994:161f.).

Ein wichtiges Problem, das die Autoren hervorheben, liegt aber in der Interpretation der Ergebnisse. Es ist nämlich nicht möglich zu verstehen, was Verständlichkeit-und Inhaltstreue-Werte in Bezug auf die Kosten bedeuten – unabhängig davon, ob der Zweck des Outputs das Verstehen der Kernaussagen des Textes (*gisting*) oder eine hochqualitative Übersetzung ist (vgl. 1994:163).

> „To see this, consider the sort of quality profile you might get as a result of evaluation [...], which indicates that most sentences received a score of 3 or 4, hence of middling intelligibility. Does that mean that you can use the system to successfully gist agricultural reports? One cannot say.“ (Arnold 1994:163)

Auch für die Einschätzung der Post-Editing-Zeit, und daher auch deren Kosten, ist diese Bewertungsmethode wenig zielführend. Es kann natürlich vermutet werden, dass wenig verständliche Sätze tendenziell zu einem höhen Post-Editing-Aufwand führen, aber der Zusammenhang ist nicht klar einschätzbar.

Die Werte der Verständlichkeit und Inhaltstreue können auch nicht deutlich aussagen, ob ein System besser als ein anderes ist. Das folgende von Autoren vorgestellte Beispiel zeigt, dass eine vergleichende Evaluation sich anhand dieser Werte auch als sehr schwierig erweist:

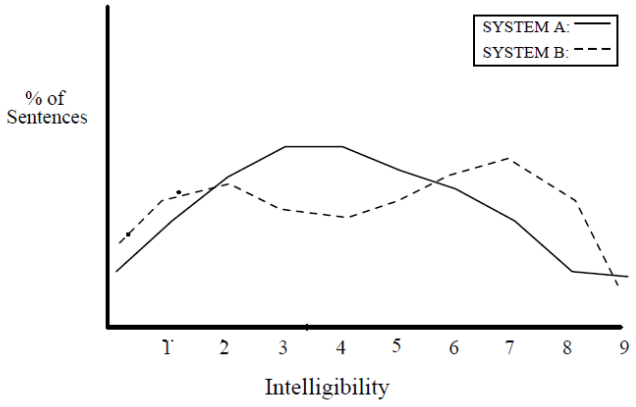

Figure 9.2 Which Performance Curve is Better?

Abb. 1: Performance-Kurven. (Arnold et al. 1994:164)

> „For system A the majority of sentences are neither very good nor bad (rating 3 or 4). System B, by comparison, tends to do either quite well (scores of 7 are common) or quite badly (scores 1, and 2 are frequent). Which system will be better in practice? It is not possible to say.“ (Arnold 1994:164)

Eine weitere von den Autoren vorgeschlagene Methode ist die Fehleranalyse mit einem Gewichtfaktor (*weighting factor)* (vgl. Arnold et. al 1994:164). Fehler, die wichtiger sind, haben ein größeres Gewicht, und umgekehrt. Jeder Fehler wird mit dem entsprechendem Gewicht multipliziert. Der Wert des Segments ergibt sich aus der Summe der gewichteten Fehler (vgl. Arnold et. al 1994:164).

2.3.6 White (2003)

Ein interessanter Ansatz stellt jener von White (2003) dar. Der Autor zieht alle Evaluationsmethoden in Zweifel, weil sie extrem subjektiv sind. Das erste Hindernis für jede Evaluation findet White in der Beurteilung jeder Übersetzung, sei es maschinelle oder von einem Übersetzer gefertigt. Die beste Übersetzung ist jene, die den Text exakt und richtig wiedergibt. Über was „richtig“ ist, gibt es aber eine von White definierte „great latitude of disagreement“ (2003:214). Aus diesem Grund sollte man vom Konzept einer möglichen „Grundwahrheit“ Abstand nehmen. „Therefore we must somehow accommodate some highly subjective judgments about which translation might be better than which other translation.” (2003:214)

Der Autor findet die Lösung in der „linguistischen Intuition“, einer menschlichen Fähigkeit, die Kommunikation zwischen Konversationsteilnehmer mit unterschiedlichen Sprachen trotz der Hindernisse der Sprache und Kultur ermöglicht. Die linguistische Intuition ist sicherlich subjektiv, aber eine Analyse der Funktionsweise der Intuition kann dabei helfen zu verstehen, welchen intuitiven Prozessen jede Evaluationsmethode unterliegt.

> „Let us imagine that we have the responsibility for determining whether a particular “into English” MT system actually can translate. Our assumptions going in are the intuitive ones about being able to determine felicity, and the idea that we should be able to make a general claim about the ability of a system to translate the infinite expressions of language, based on a infinite test set.“ (2003:214)

Der Autor erarbeitet daher drei Methoden, die er als „imaginary methods“ bezeichnet: „output only“, „input and output“, „input, two outputs“ (vgl. 2003:214-219).

Durch die reine Analyse des Outputs ist es möglich zu bewerten, inwieweit der Satz in der Zielsprache flüssig ist – in diesem Fall auf Englisch. White geht Schritt für Schritt im Prozess vor, der von der linguistischen Intuition zu den menschlichen Metriken geht: „We glance at a few of the output expressions, and realize that we are not going to find much that

is simply good English, so we devise a way to measure 'how good' an expression is." (2003:214f.)

Dabei ergeben sich folgende drei Optionen:

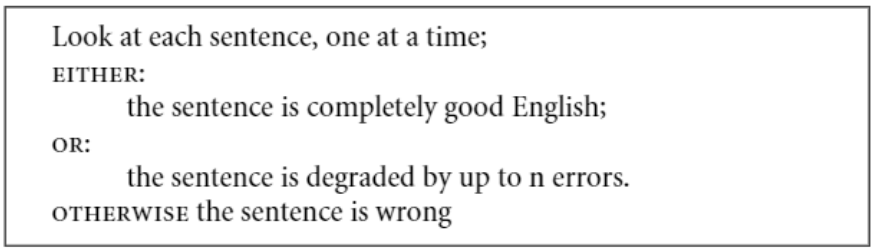
Look at each sentence, one at a time;
EITHER:
the sentence is completely good English;
OR:
the sentence is degraded by up to n errors.
OTHERWISE the sentence is wrong

Abb. 2: Methode 1, „Output only" (White 2003:215)

Diese Metrik ermöglicht es, die Vorteile der linguistischen Intuition zu nutzen. Der Nachteil hierbei ist aber, dass der Evaluator nicht weiß, woher dieser Satz stammt, und es daher nicht möglich ist, durch die reine Evaluation des Zielsatzes und deren Fehler eine Verbesserung des MÜ-Systems zu erzielen. Dafür wird in der Regel der Zielsatz mit dem entsprechenden Originalsatz verglichen. Das ist die zweite Methode: „Input and output" (vgl. 2003:216).

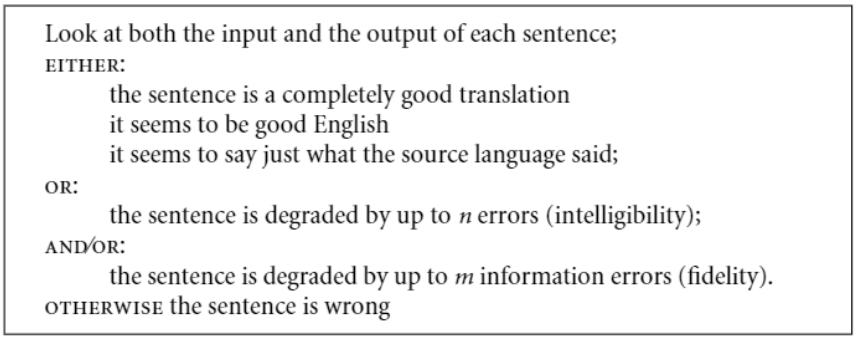
Look at both the input and the output of each sentence;
EITHER:
the sentence is a completely good translation
it seems to be good English
it seems to say just what the source language said;
OR:
the sentence is degraded by up to *n* errors (intelligibility);
AND/OR:
the sentence is degraded by up to *m* information errors (fidelity).
OTHERWISE the sentence is wrong

Abb. 3: Methode 2, „Input and output" (White 2003:217)

Diese zweite Methode ermöglicht es, sowohl die Verständlichkeit als auch die Inhaltstreue eines MÜ-Outputs zu bewerten. Wie White aber klarstellt, ist es trotzdem immer noch nicht möglich, eine generelle Behauptung über das System aufzustellen, zumal die bewerteten Sätze und deren Qualität wegen der Unbestimmtheit der Sprache den Sätzen bei wirklicher Verwendung des Systems nicht entsprechen können, wenn nicht nur zufälligerweise: „It may "accidentally" get right the sentences we have in our sample, which tells us nothing about how "extensible" the system is to the general (infinite) case." (2003:217)

Die dritte Methode sieht die Anwendung von einem Input und zwei Outputs vor. Diese Methode kann verwendet werden, wenn das Ziel der Evaluation darin besteht, die Verbesserungsmaßnahmen am System zu testen. Die Metrik wird in diesem Fall folgendermaßen aussehen:

Look at the input sentence, along with the "before" output sentence and the "after" output sentence;
EITHER:
both translations are perfect in fidelity;
AND/OR:
perfect (and possibly different!) in intelligibility;
AND/OR:
one differs from the other by *n* fidelity errors and/or *m* intelligibility errors;
OR:
one is wrong and one is not.
OTHERWISE both translations are wrong

Abb. 4: Methode 3, „Input, two outputs" (White 2003:218)

Der Vorteil dieser Methode ist natürlich, dass dadurch bestimmt werden kann, ob das System tatsächlich verbessert wurde oder nicht. Zugleich liefert die Methode eine ungefähre Vorstellung über die Erweiterungsfähigkeit des Systems, denn nach den Verbesserungsmaßnahmen sollten mehrere (oder weniger) linguistische Phänomene korrekt übersetztet werden. Ein großer Nachteil wird aber bei dieser dritten Methode klar deutlich, der sich auch bei den anderen Methoden gezeigt hatte: die ‚menschlichen Faktoren' („human factors") (vgl. 2003:218).

> „As we noted earlier, we are taking advantage of the fact that our linguistic intuitions allow for a great deal of agreement among different times, places, and between different speakers of the same language. But there are local, almost microscopic aspects that can lead us to inconsistent judgments and inaccurate conclusions about the results." (2003:218)

Diese drei Aspekte teilt White in Geschichte, Testen und Reifung ein. Unter **Geschichte** versteht White die Ereignisse der Welt (auch die scheinbar nicht linguistisch-korrelierten Ereignisse, wie beispielweise einen Börsencrash), welche die Beurteilung eines Outputs beeinflussen können. Darüber hinaus ist das **Testen** selbst ein Verfahren, das durch einen menschlichen Faktor kenngezeichnet ist. Die Evaluatoren können nämlich auf denselben Satz anders reagieren, wenn ihnen der Satz zweimal zum Testen vorgelegt wird. Das erschwert den Vergleich zwischen zwei Übersetzungen desselben Originalsatzes. Außerdem beeinflussen sich die nacheinander stehenden Sätzen gegenseitig: Nach der Bewertung einer besonders schlechten Übersetzung tendiert der Evaluator dazu, den nächsten Satz besser zu bewerten. Unter **Reifung** versteht White den unterschiedlichen Ansatz im Hinblick auf die Bewertung innerhalb eines Evaluationsprozess. Die Evaluation kann durch emotionelle und physische Faktoren des Evalautors beeinflusst werden (Müdigkeit, Hunger, mangelnde Konzentration). Das bedeutet, dass ein Satz in zwei verschiedenen Phasen der Evaluation unterschiedlich beurteilt werden kann.

2.4 Qualität des MÜ-Systems

Wie schon am Anfang dieses Kapitels erklärt wurde, erfordert die Diskussion über die Qualität der maschinellen Übersetzung einen holistischen Ansatz. Um ein System in all seinen Aspekten zu bewerten, soll man sowohl die Qualität des Outputs – also im Sinne von Übersetzung – als auch die Qualität des Systems als Software mit einbeziehen. Die Qualitätsparameter, die für die beiden Aspekte gelten, sind aber nicht klar trennbar. Sie stellen dabei zwei Aspekte desselben Sachverhalts dar, die sich oft überschneiden. Die Autoren, die sich mit der Qualität der maschinellen Überersetzung beschäftigt haben, haben oft auch Merkmale der Softwarequalität in ihre Parameter inkludiert, wie beispielweise die Eignung (*suitability*) des Systems für andere Texttypen und seine Erweiterungsfähigkeit (*extendability*) (vgl. King/Falkedal 1990). In diesem Kapitel wird dargelegt, welche Qualitätsparameter für eine Software gelten und wie sie sich auf maschinelle Übersetzungssysteme in den unterschiedlichen Phasen der Softwareentwicklung umsetzen lassen.

2.4.1 Entwicklungsphasen und Evaluation

Bevor auf die Erklärung der Parameter und Qualitätsstandards näher eingegangen wird, soll ein Überblick über die Phasen der Softwarenentwicklung gegeben werden. Je nach Phase ändern sich nämlich Qualitätskriterien, Evaluationsmethoden und Evaluationsbeteiligte.

Das Lebenszyklus einer Software (vgl. Abbildung 5) gliedert sich in folgende Phasen: Recherche, Recherche-Prototype, funktionsfähige Prototypen, Endprodukt (vgl. Hirschman/Mani 2003). In der Recherche sind die wichtigsten Akteure die Entwickler und die Forschungsförderer, weil die Technologie sich noch in einem fragilen Zustand befindet und die einzigen Evaluatoren die Entwickler selbst sind, die Tests auf Komponentenebene durchführen. Diese Evaluation wird auch **Prototype-Evaluation** genannt (vgl. Hutchins/Somers 1992).

In der zweiten Phase wird die Fähigkeit des Systems getestet, mit anderen Komponenten zu interagieren und innerhalb eines größeren Systems zu arbeiten. In dieser Phase wird einer **Embedded-Komponenten-Evaluation** (Hirschman/Mani 2003) oder wie es Hutchins/Somers (1992) definieren, eine **Entwicklungs-Evaluation** („development evalaution“), durchgeführt. Hier ist es schon möglich, Tests durchzuführen, die ‚reale‘ Benutzer einbeziehen. Diese Phase ist deshalb wichtig, weil der Entwickler durch die Evaluation überprüfen kann, ob das System die vorausgesetzten Funktionen korrekt durchführen kann und ob Anpassungen und Verbesserungen ohne radikale Änderung des

Programms durchgeführt werden können, bevor das System den Endbenutzern angeboten wird.

Die dritte Phase der Software ist der funktionsfähige Prototyp. Hier wird das System von potentiellen Käufern getestet. In der **operationellen Evaluation** werden auch Fallstudien durchgeführt, die Kosten-Nutzen der potentiellen Käufer berechnen, die Kompatibilität mit dem System des Käufers überprüfen usw.

Die Endphase ist das finale Produkt. In dieser Phase findet eine **Rezipientenevaluation** („recipient evaluation“, vgl. Hutchin/Somers 1992) statt. Hier handelt es sich um eine vergleichende Evaluation: Qualität, Geschwindigkeit und Kosten unterschiedlicher MÜ-Systeme oder der Unterschied zwischen MÜ und HÜ für die Zwecke der Firma bzw. Organisation werden verglichen.

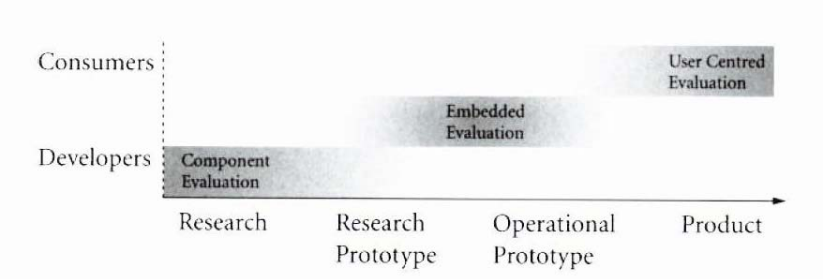

Abb. 5: Lebenszyklus einer Software (Hirschman/Inderjeet 2003:415)

2.4.2 Qualitätsstandards-und Frameworks

Die MÜ-Systeme sind im Prinzip Software und daher sollen sie die Standards und Qualitätsparameter, die für die Software gelten, erfüllen können. Die standardisierten Qualitätskriterien einer Anwendung wurden in der ISO/IEC 9126 und ISO/IEC 14598-1:1999 erfasst. Die ISO-Norm 9126 sieht die Qualität als „the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs” (ISO/IEC, 2003a). Die Qualität resultiert nach der ISO aus sechs Hauptkategorien: Funktionalität (*functionality*), Zuverlässigkeit (*reliability*), Benutzbarkeit (*usability*), Effizienz (*efficiency*), Wartbarkeit (*maintainability*), Übertragbarkeit (*portability*).

Quality characteristic	Quality sub-characteristics
Functionality	Suitability, Accuracy, Interoperability, Security, Functionality compliance
Reliability	Maturity, Fault tolerance, Recoverability, Reliability compliance
Usability	Understandability, Learnability, Operability, Attractiveness, Usability compliance
Efficiency	Time behavior, Resource utilization, Efficiency compliance
Maintainability	Analysability, Changeability, Stability, Testability, Maintainability compliance
Portability	Adaptability, Installability, Co-existence, Replaceability, Portability compliance

Abb. 6: Qualitätsmodell nach der ISO/IEC 9126 im Überblick (Estrella et al. 2009:3)

Die **Funktionalität** ist die Fähigkeit des (Software)Produktes, die Bedürfnisse und die festgelegten Anforderungen bzw. Eigenschaften zu erfüllen. Unter **Zuverlässigkeit** versteht man die Fähigkeit der Software eine bestimmte Leistung unter bestimmen Bedingungen aufrechtzuhalten. Eine Software soll zudem auch benutzbar sein (**Benutzbarkeit**), sie soll von Benutzern unter bestimmen Bedingungen gelernt, verstanden und verwendet werden können. Die **Effizienz** ist das Verhältnis zwischen Leistungsniveau der Software und dem Umfang von Ressourcen (bzw. Betriebsmitteln) unter festgelegten Bedingungen. Die **Wartbarkeit** ist die Fähigkeit der Software, geändert bzw. verbessert zu werden. Der letzte Parameter ist die **Übertragbarkeit:** Die Fähigkeit der Software in einer anderen Umgebung verwendet zu werden.

Während die ISO-Norm 9126 die Bestandteile eines allgemeinen Qualitätsmodells beschreibt, gibt die ISO-Norm 14598 Richtlinien und Beispiele vor, die bei der Planung und Durchführung der Evaluation helfen sollen. Die zwei Normen ergänzen sich gegenseitig, denn die Festlegung eines Qualitätsmodells ist Teil des Evaluationsverfahrens und das Verfahren hängt wiederum von den an der Evaluation beteiligten Akteure (Evaluatoren, Entwickler, Erwerber usw.) ab (vgl. Estrella et al. 2009:4).

Bei der Benutzung eines MÜ-Systems durch einen Endbenutzer ist die Qualität des Outputs für die Beurteilung des ganzen Systems ausschlaggebend und aus diesem Grund ist der Output in der Regel der Gegenstand der Evaluation. Wie aber die ISO klarstellt, variieren die Ansprüche an ein System stark – wie wir im Laufe dieses Kapitels gesehen haben – und hängen von Benutzer, Zweck (*task-based evaluation*), und im Endeffekt vom Kontext der Benutzung ab. Der Kontext besteht aus den Merkmalen, die von ISO als „festgelegte Bedingungen“ bezeichnet werden. Diese letzte Methode ist die sogenannte Kontext-basierte-Evaluation (vgl. Estrella et. al 2009).

Die ISO-Standards sind in unterschiedlichen Kontext-basierten Frameworks umgesetzt worden, welche die von ISO festgelegten Qualitätsparameter einer Software auf

die MÜ-Systeme anwenden. Eine der ersten Initiativen für die Erstellung eines Frameworks zur Evaluierung der MÜ, die nicht nur den Output betrachtet, ist ein Bericht von JEIDA (*Japan Electronic Industries Development Association*). In dem Bericht wurde die Qualität sowohl aus der Perspektive der Benutzer als auch als der Perspektive der Entwickler betrachtet (Nomura, 1992). Dabei wurden zwei Kriterien vorgeschlagen: Evaluatoren (Benutzer oder Entwickler) sollten einen Fragebogen über die aktuelle Arbeitssituation und einen weiteren Fragenbogen mit ihren Bedürfnissen beantworten. Danach wurden Radardiagrammen mit den Ergebnissen der beiden Fragenbogen erstellt und am Ende konnte sich der Evaluator – anhand der sich überlappenden Charts – für das System entscheiden, das seine Anforderungen am besten erfüllen konnte (vgl. Estrella et. al 2009:2).

Eine weitere Umsetzung der ISO-Normen in einem Evaluationsframework wurde im Rahmen des EU-Projektes EAGLES (Expert Advisory Group on Language Engineering Standards) von der EAGLES Evaluation Working Group durchgeführt. Hier wurde die Evaluation aus einer Benutzerperspektive geplant. Es wurde ein „consumer report paradigm" erstellt (vgl. EAGLES Evaluation Working Group, 1996), in dem Benutzer bestimmen konnten, welche Merkmale des Softwareproduktes für ihre Benutzergruppe relevant waren. Der EAGLE Framework wurde auch im Rahmen anderer Projekte verwendet und die Validität dieses Evaluationsdesigns wurde daher getestet und bestätigt (vgl. Rocca et al. 1994; TEMAA, 1996; Canelli et al. 2000).

Der EAGLE-Framework entwickelte sich im ISLE EU-Projekt im Rahmen dessen der FEMTI-Framework (vgl. Hovy et al. 2002) erstellt wurde, das die oben genannten ISO-Normen anwendet und die schon vorhandenen Methoden erweitert. Nach der Erstellung im Jahr 2002 wurde FEMTI weiter gepflegt und daraus hat sich ein interaktives Tool[11] entwickelt, das dabei hilft, ein Evaluationsdesign zu erstellen (vgl. Estrella et al. 2005).

Im FEMTI wurde das von den ISO/IEC-Normen festgelegte Modell an die maschinelle Übersetzung angepasst. Dabei wurde die Hülle bewahrt und mit anderen Qualitätsmerkmalen, nämlich den Kosten, und MÜ-spezifischen untergeordneten Qualitätsmerkmalen ergänzt. In der Abbildung 7 wird ersichtlich, wie das Merkmal der Funktionalität unterteilt wurde und auf die MÜ adaptiert wurde. Für jeden Qualitätsparameter werden dann auch die dazugehörenden Metriken angezeigt (in dem hier vorhandenen Beispiel werden die Metriken für die Lesbarkeit und Grammatik/Syntax angezeigt.

[11] Das FEMTI-Tool ist online verfügbar unter http://www.issco.unige.ch:8080/cocoon/femti/st-home.html, Stand 14.08.2016.

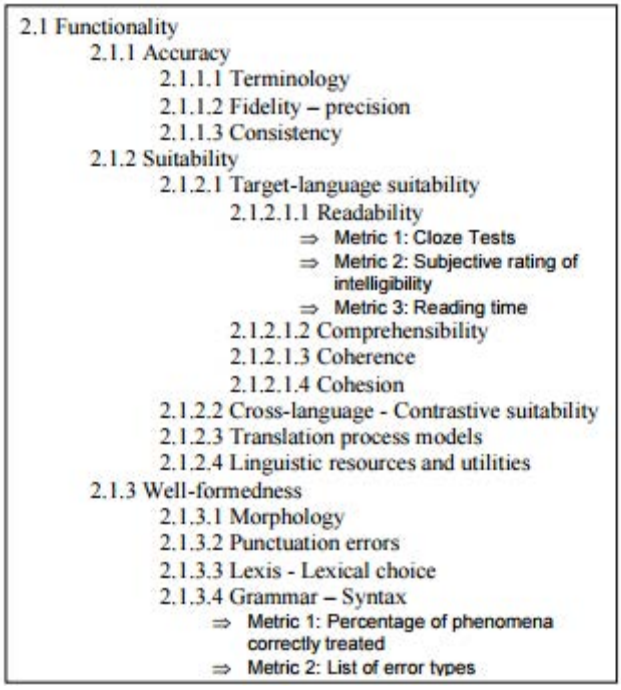

Abb. 7: Funktionalität im FEMTI (Estrella et al. 2009:4)

FEMTI ermöglicht es daher, Kontext-basierte Qualitätsmodelle zu erstellen. Das ist aber eine komplexe Aufgabe, die ein Wissen über die Funktionsweise und Qualitätsmerkmale des FEMTI voraussetzt. Dank dem interaktiven Tool hat das FEMTI seine Benutzerfreundlichkeit verbessert. Das Framework fokussiert nämlich auf eine praktische Anwendung seitens jeden Benutzer, die eine Evaluation planen möchten. Um dieses Ziel zu erreichen, wurde FEMTI im Rahmen eines Experiments getestet, in dem Experten befragt wurden, welche Methoden sie in welchem Kontext verwenden würden. Aus den Ergebnissen der Befragung wurden dann induktiv Regeln über den Zusammenhang zwischen von Kontext abhängigen Merkmalen und Qualitätsparametern abgeleitet (vgl. Estrella et al. 2008:933).

3. Menschliche Evaluationsmethoden

Die menschlichen bzw. manuellen Evaluationsmethoden stellen den goldenen Standard der Evaluation dar und werden trotz der Einführung automatischer Evaluationsmethoden immer noch verwendet. Es gibt zwei Hauptgründe, weshalb diese Methoden unentbehrlich sind. Erstens sind Übersetzungen an Humanbenutzer gerichtet und daher können nur Humanbenutzer bestimmen, welche Qualitätsparameter für einen bestimmten Zweck relevant sind und ob sie erfüllt werden. Zweitens können nur Menschen die Richtigkeit der Bedeutung einer Übersetzung beurteilen – auch im Sinne ihres praktischen Bezugs auf reale Situationen: Wenn beispielweise der Satz „there are new land-mines buried under the road to Baghdad [mit] there are no land-mines buried under the road to Baghdad“ (Dorr et al. 2011:742f.) übersetzt wird, ist es nur für einen Humanevaluator möglich, die Wichtigkeit dieser Fehler zu erkennen und die Übersetzung dementsprechend zu bewerten (vgl. 2011:742f.).

Die härteste Kritik an den menschlichen Evaluationsmethoden ist aber, dass sie zu subjektiv sind: Fehlende Inter-Annotator-Übereinstimmung (die Übereinstimmung der Antworten bzw. Bewertungen desselben Evaluators innerhalb einer Evaluation) und Intra-Annotator-Übereinstimmung (Übereinstimmung der Antworten bzw. Bewertungen unterschiedlicher Evaluatoren) sind die Stichworte der Initiatoren und Verfechter der automatischen Evaluationsmethoden. Jede menschliche Evaluation (nicht nur im Bereich der maschinellen Übersetzung) weist zweifelsohne diese beiden Nachteile zweifelsohne auf. Die Tatsache, dass die manuellen Methoden noch den goldenen Standard darstellen und in den Evaluationskampagnen als Maßstab für die Evaluation und die Meta-Evaluation (Evaluation der automatischen Methoden) verwendet werden, zeigt allerdings, dass sie bis heute den höchsten Grad an Zuverlässigkeit aufweisen. Natürlich spielen die Kosten der Durchführung einer menschlichen Evaluation auch bei der zunehmenden Verwendung der automatischen Evaluationsmetriken eine wichtige Rolle.

Die menschlichen Evaluationsmethoden, die heute in der wissenschaftlichen Gemeinschaft der maschinellen Übersetzung verwendet werden, sind verschieden. Die Qualität einer Übersetzung wird entweder durch direkte oder indirekte Methoden, wie einen Leseverständnistest oder andere sog. *downstream task*s, die den maschinell übersetzten Output anwenden, gemessen (vgl. Dorr et al. 2011:747). In diesem Kapitel werden die gängigsten manuellen bzw. menschlichen Methoden vorgestellt. Besonders ausführlich werden die Methoden dargelegt, die in der im Rahmen dieser Masterarbeit durchgeführten Studie verwendet wurden.

3.1 *Adequacy* und *fluency*

Zwei der heute gängigsten Qualitätsparameter sind *fluency* (Flüssigkeit) und *adequacy* (Inhaltstreue). Sie wurden vom LDC (Lingustic Data Consortium) im Jahr 2005 eingeführt. Die zwei Kriterien werden von Evaluatoren auf einer numerischen Skala bewertet (vgl. Denkowski/Lavie 2010:2). Die Bewertung dieser zwei Parameter wurde aber schon in ähnlichen Formen seit dem ALPAC-Bericht verwendet. Obwohl sich die Termini, die diese zwei Parameter bezeichnen, im Laufe der Jahre verändert haben und von den unterschiedlichen Autoren anders definiert wurden, sind die Konzepte inhaltlich unverändert geblieben. Wie im letzten Kapitel erwähnt wurde, spricht beispielweise Van Slype (1979) von *intelligibility* und *fidelity*. *Intelligibility* oder Verständlichkeit umfasst dem Autor zufolge unterschiedliche Begriffe: „*intelligibility, clarity, comprehensibility, legibility*" (1979:61). Für Halliday/Briss (1977) sind *comprehensibility* und *intelligibility* Synonyme, zwei Blickwinkel derselben Analyse.

> „Comprehensibility relates to the degree of perfection with which a complete translation can be understood (whereas the intelligibility is based on the general clarity of a translation, whether this is considered in its entirety or by segments out of context)." (Van Slype 1979:63)

Während man sich in den vergangenen Jahren auf den Begriff der Verständlichkeit eines Textes fokussiert hat, hat heutzutage der Begriff der Flüssigkeit bzw. *fluency* an Wichtigkeit gewonnen und wird als Standardparameter bei den Evaluationen verwendet. Es ist aber nicht einfach, den Begriff der *fluency* von den anderen oben genannten Merkmalen – insbesondere von der Verständlichkeit – abzugrenzen. Anhand der Flüssigkeit sollte gemessen werden können, ob ein Satz in der Zielsprache flüssig lesbar ist. Allerdings ist das Verstehen des Satzes für die Beurteilung der Flüssigkeit unabdingbar. Ein Satz kann aus unterschiedlichen Gründen, die nicht von der Flüssigkeit abhängen, unverständlich sein:

> „The main problem encountered with subjective human judgments of fluency is that understandability or comprehensibility of the translation enters into the picture. This is a problem if the judges are not familiar with the subject matter, and is a particularly acute problem if the source-language material is disfluent, as is often the case with spoken language, or unstructured texts such as web data." (Dorr et al. 2011:754)

Fidelity (Inhaltstreue – heute *adequacy*) bezeichnet Van Slype als die „subjective evaluation of the measure in which the information contained in the sentence of the original text reappears without distortion in the translation" (Van Slype 1979:72). Die Inhaltstreue kann von einem zweisprachigen Evaluator durch einen Vergleich mit dem Originalsatz bewertet werden. Anderenfalls kann die Bewertung auch durch einen einsprachigen Evaluator erfolgen, dem eine oder mehrere Humanübersetzungen vorgelegt werden. Wie schon Van

Slype (1979) bemerkte, ist eine detaillierte Analyse einer mangelnden Inhaltstreue schwierig durchzuführen, weil jeder Satz nicht nur eine oder mehrere einzelne Informationen enthält, sondern eine Serie von komplexen Nachrichten, deren jeweilige Wichtigkeit nicht leicht bewertet werden kann.

Die zwei Parameter werden auf einer numerischen Skala bewertet. Die Skala kann 5-stellig bis 9-stellig sein. Carroll (1966) verwendete für die Evaluation von MÜ für die ALPAC-Gruppe eine 9-stellige Skala. Die Skala der *fidelity*, die von Van Slype (1979) benutzt wurde, misst den Informativitätsgrad von "extrem informativ" bis "keine Informativität" (Van Slype 1979:72ff.), während die *intelligibilty* einen perfekt klaren Satz einerseits – „Perfectly clear and intelligible. Reads like ordinary text has no stylistic infelicities" – und einen unverständlichen Satz andererseits – "Hopelessly unintelligible" (Van Slype 1979:63f.) als zwei Extreme aufweist.

Die heute am meisten benutzte Skala ist die von LDC vorgeschlagene 5-stellige Skala sowohl für *adequacy* (5: All 4: Most 3: Much 2: Little 1: None) als auch für fluency (5: Flawless 4: Good 3: Non-native 2: Disfluent 1: Incomprehensible) (vgl. Denkowski/Lavie 2010:2).

Die zwei Parameter (*adequacy* und *fluency*) werden separat bewertet, aber einigen Studien, wie beispielweise die im Rahmen des 2007 ACL Workshop on Statistical Machine Translation durchgeführte Studie, haben gezeigt, dass *fluency* und *adequacy* stark korrelieren (Callison-Burch et al. 2007). Annotatoren haben Schwierigkeiten, die Bedeutung eines nicht flüssig lesbaren Satzes zu verstehen, und die *adequacy* wird konsequent auch schlecht bewertet. Umgekehrt ist eine Übersetzung, welche die ganze oder fast ganze Bedeutung eines Originalsatzes wiedergibt, in der Regel auch flüssig, denn auch kleine morphologische Fehler können die Bedeutung stark beeinflussen. Die Trennung der zwei Skalen führt auch dazu, dass sie nur schwer wieder zusammengeführt werden können. Die Zusammenführung der zwei Parameter ist für die Benutzung der Ergebnisse für andere Zwecke relevant, wie zum Beispiel die Meta-Evaluation oder das Tuning der automatischen Metriken (vgl. Denkowski/Lavie 2010:3). Aus diesem Grund wird beispielweise in der NIST Open Machine Translation Evaluation (NIST 2009) nur die Inhaltstreue bewertet, allerdings mit einer präziseren 7-stelligen Skala, welche es ermöglicht, genaue Unterscheidungen zwischen den Intervallen zu treffen.

Einigen Studien haben gezeigt, dass die Zuverlässigkeit dieser Methoden durch die mangelnde Intra-und Inter-Annotator-Übereinstimmung in Frage gestellt werden kann (vgl. Turian 2003). Das Problem der fehlenden Inter-Annotator Übereinstimmung kann durch den Einsatz mehrere Evaluatoren bewältigt werden. Die Daten der Bewertungen aller Evaluatoren können nach Kategorien eingeteilt werden und diese können dann verglichen werden, um zu überprüfen, ob die Evaluatoren damit einverstanden sind, welche Übersetzungen gut oder schlecht in Bezug auf die Inhaltstreue und die Flüssigkeit sind. Eine ähnliche Daten-Bearbeitung wurde im Rahmen dieser Studie durchgeführt (vgl. Kapitel 6.3). Wenn die Werte

numerisch sind, können sie durch eine z-Transformation normalisiert werden (vgl. Dorr 2011:755).

Die Annotator-Übereinstimmung kann durch den Cohens-Kappa berechnet werden (vgl. Denkowski/Lavie 2010:3):

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

in dem P(A) der Übereinstimmungswert der Annotatoren und P(E) die zufällig erwartete Übereinstimmung ist. Ein Beispiel für die Anwendung des K-Koeffizienten sind die Ergebnisse der WMT im Jahr 2007. Die Inter-Annotator-Übereinstimmung hatte ein K von 0,47 bei der Inhaltstreue und 0,54 bei der Flüssigkeit des Textes. Diese niedrigen Werte zeigen sehr deutlich, dass die Annotatoren diese Parameter unterschiedlich interpretiert haben. Wie Denkowski/Lavie (2010) bemerken, kann beispielweise eine einzige Negation die Bedeutung eines sehr langen Satzes komplett verändern. Wenn ein solcher Fehler in der Übersetzung vorhanden ist, sollte dann der Evaluator nur wegen diesen Fehlern eine schlechte Bewertung geben? Welcher Teil eines Satzes ist wichtiger in einem Satz, der 40-50 Wörter enthält? (vgl. Denkowski/Lavie 2010:3).

Den Evaluatoren ist es auch nicht möglich, durch die ganze Evaluation eine Übereinstimmung der Bewertungen zu erzielen (Intra-Annotator Übereinstimmung). Einige Gründe dafür wurden schon von White (vgl. Kapitel 2.3.6) vorgebracht. Die Bewertungen von schlechten und guten nacheinander stehenden Sätzen können sich gegenseitig beeinflussen. Bei sehr langen Sätzen könnte sich ändern, welche Satzteile bzw. Fehler der Annotator für wichtiger hält und bei kurzen Sätzen könnte es schwierig sein zu bestimmen, wo beispielweise die Grenze zwischen einem Wert von 3 und 4 ist.

Inter-Annotator Agreement			
Judgment Task	$P(A)$	$P(E)$	K
Adequacy	0.38	0.20	0.23
Fluency	0.40	0.20	0.25
Ranking	0.58	0.33	0.37
Intra-Annotator Agreement			
Judgment Task	$P(A)$	$P(E)$	K
Adequacy	0.57	0.20	0.47
Fluency	0.63	0.20	0.54
Ranking	0.75	0.33	0.62

Abb. 8: Inter-und Intra-Annotatoren Übereinstimmung bei WMT07 (Denkowski/Lavie 2010:3)

Wie schon oben erwähnt wurde, kann man durch eine statische Standardisierung die Werte normalisieren, um sie für andere Zwecke zu verwenden. Die mangelnde Übereinstimmung auf Segmentebene ist aber stark präsent, und wirft einen Schatten auf die gesamten Prozesse, auf welche die normalisierten Werte zurückgreifen (z.B. das Tuning der automatischen Metriken). Das folgende Histogramm aus dem WMT 2006 soll diese mangelnde Übereinstimmung veranschaulichen:

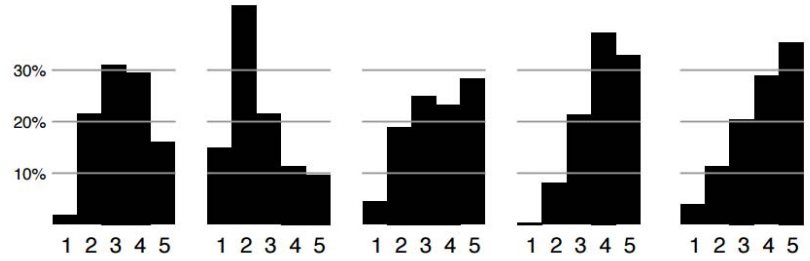

Abb. 9: Bewertung von *adequacy* auf einer 5-stelligen Skala durch 5 Evaluatoren (Koehn 2010:220)

3.2 Ranking

Die Methode des Rankings wurde erst im Jahre 2007 als Evaluationsmethode im WMT eingeführt, mit dem Ziel, eine Lösung für die sich aus den Methoden der Bewertung von *adequacy* und *fluency* ergebenden Problemen zu finden (vgl. Denkowski/Lavie 2010:3). Beim Ranking werden die MÜ-Systeme satzweise nicht im Hinblick auf ihre Qualität bewertet, sondern sie werden direkt von sehr gut bis schlecht gereiht. Im Vergleich zur Qualität-Bewertung ist das Ranking eine sehr direkte Methode, bei der die Evaluatoren folgende Frage beantworten: „Welche Übersetzung bevorzugen Sie?“ (vgl. Dorr 2011:758).

Die maschinell übersetzten Sätze der jeweiligen MÜ-Systeme können paarweise verglichen werden (dabei wird aus einem Übersetzungspaar die beste Übersetzung ausgewählt), oder sie können alle zusammen verglichen und gereiht werden. Die Qualität, die hierbei bewertet wird, ist allgemeiner Natur und nicht vordefiniert, denn es wäre für die Evaluatoren unmöglich, einzelne Qualitätsmerkmale zu isolieren – wie beispielweise semantische Inhaltstreue – und die Reihung der Sätze von einem bestimmten Merkmal abhängig zu machen (vgl. Dorr 2011:758).

Der große Vorteil des Rankings ist, dass es eine direkte Einschätzung des besten MÜ-Systems ermöglicht. Anders als bei den *fluency* und *adequacy*-Skalen, in denen Sätze, die sich nur wegen einzelnen Wörtern unterscheiden, denselben *adequacy*-Kategorien zugeordnet werden – d.h. ohne endgültige Entscheidung darüber, welcher Satz bzw. welches System das Beste ist – können hier die Systeme präziser und eindeutiger beurteilt werden. Dieser Aspekt ist insbesondere in den letzten Jahren immer wichtiger geworden, denn die meisten MÜ-

Systeme werden ständig verbessert und erstellen oft sehr ähnliche Übersetzungen. In solchen Fällen ist es schwierig anhand von den reinen Bewertungsskalen bestimmen zu können, welches System eine bessere Performance erzielt hat (vgl. Denkowski/Lavie 2010:4).

Diese Methode hat aber auch einige Nachteile. Auch hier ist eine mangelnde Annotatoren-Übereinstimmung vorhanden und in vielen Fällen berichten die Evaluatoren, die an den WMT-Kampagnen teilgenommen haben, dass das Ranking bei einigen Satzsorten schwierig ist (vgl. 2010:4). Insbesondere längere Sätze bereiten Schwierigkeit, weil es den Evaluatoren nicht möglich ist, viele lange Sätze gleichzeitig vor Augen zu halten. Das führt dazu, dass die Evaluatoren die Beurteilung unbewusst nicht auf Satzebene durchführen, sondern sich auf einige Salzelemente bzw. Phrasen fokussieren. Darüber hinaus ist die Art und Weise, wie die Evalautoren die Aufgabe der Beurteilung längerer Sätze bewältigen bzw. wie sie den Satz zur Evaluation unterteilen, sehr unterschiedlich – und die Folgen sind natürlich eine mangelnde bis fehlende Inter-Annotator-Übereinstimmung (vgl. 2010:4).

Auch im Falle einer Annotator-Übereinstimmung stellt diese Methode einen Nachteil dar: die fehlende Informativität. Denkowski/Lavie (2010) stellen als Beispiel den Fall vor, in dem bei drei kurzen Output-Sätzen (aus den MÜ-Systemen) dasselbe Wort falsch übersetzt wurde. Die Fehlertypologie ist aber durch die drei Sätze bzw. Systeme unterschiedlich. System Nr. 1 übersetzt das Wort falsch, System Nr. 2 lässt das Wort weg, System Nr. 3 lässt das Wort in der Ausgangssprache. Welche Übersetzung ist in diesem Fall die Beste bzw. wie werden diese drei Sätze gereiht? Es wäre unmöglich, zu verlangen, dass die Evaluatoren dieselbe Reihung durchführen. Sollten sie es trotzdem machen, oder die drei Sätze derselben Reihe zuordnen[12], bleibt das Problem der Informativität bestehen (vgl. 2010:4). Die unterschiedliche Bewertung von Fehlern, d.h. die Entscheidung, welche Fehler wichtiger als anderen sind, macht die Evaluation vor allem bei längeren Sätzen schwierig und weniger zuverlässig, denn solche Fehler, die schwierig zu vergleichen sind, multiplizieren sich natürlich in längeren Sätzen.

Das Kombinieren der Evaluationen von vielen Annotatoren stellt beim Ranking ein weiteres Hindernis dar. In den Bewertungsskalen kann man aus kollidierenden Evaluationen den Mittelwert bilden, sodass die theoretischen „wahren“ bzw. richtigen Werte annähernd berechnet werden können. Hingegen führen kollidierende Evaluationen beim Ranking dazu, dass sie ungültig werden (vgl. 2010:4).[13]

[12] Ausgleiche sind beim Ranking erlaubt.

[13] Dieses intrinsische Problem ist besonders für das Tuning der automatischen Evaluationsmethoden relevant (vgl. dazu Denkowski/Lavie 2010:4f.).

3.3 Fehleranalyse

Die Methoden, die bisher vorgestellt wurden, messen wie „gut“ eine Übersetzung ist. Die Fehleranalyse misst dagegen die Kehrseite der Qualität und beantwortet die Frage, wie schlecht die Übersetzung ist. Dadurch wird versucht, die Leistungsfähigkeit sowie die Einschränkungen eines Systems bei denen Fällen zu überprüfen, in denen das System am wahrscheinlichsten zu den besten oder zu den schlechten Ergebnissen führen soll. Die Fehleranalyse wird als objektivere Methode als die Qualitätsbewertung mittels Bewertungsskalen gesehen und zeigt eine bessere Inter-Annotatoren-Übereinstimmung (vgl. Schwarzl 2001). Darüber hinaus ist die Methode sowohl für die Entwickler als auch für die Benutzer (bzw. die potentiellen Käufer) informativer.

In den letzten sechzig Jahren wurden unterschiedliche Fehlerkategorisierungen entwickelt. Es wird hier eine der aktuellsten Kategorisierungen beispielhaft gezeigt (s. Abbildung 10). Die von Vilar et al. (2006) erstellte Klassifikation basiert auf der Arbeit von Llitjós et al. (2005). Die Kategorien sind hierarchisch aufgebaut. Wie aus dem Beispiel ersichtlich wird, erfolgt die Unterscheidung zwischen den Fehlerkategorien nach dem Aufwand, der für die Korrektur oder das Verstehen des Textes nötig ist. Es wird beispielweise zwischen *content words* und *fillers words* unterschieden. Erstere sind Wörter, die eine Bedeutung tragen, wie Substantive oder Verben, letztere sind beispielweise Präpositionen[14]. Auch bei der Wortstellung wird eine ähnliche Unterscheidung gemacht: Die Wörter oder Phrasen können innerhalb eines kleineren oder größeren Kontextes umgestellt werden (*local range* vs. *long range*)[15]. Nach der Kategorisierung werden die Fehler gezählt und statistisch bewertet.

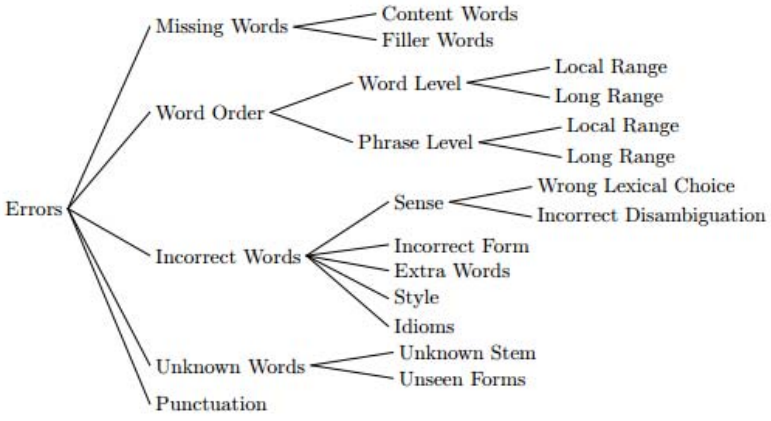

Abb. 10: Fehlerklassifizierung nach Vilar et al. (2006:699)

[14] Die Zuordnung ist aber nicht eindeutig. Eine (falsche) Präposition könnte beispielweise die Bedeutung des Satzes stark beeinflussen.

[15] „The difference between local and long range is difficoult to define in absolute terms, but it tries to express the difference between having to reorder the words only in a local context (within the same syntactic chunk) or having to move words in another chunk.“ (Vilar et al. 2006:698)

Wie auch Vilar et al. (2006) angeben, ist die Fehlerkategorisierung keinesfalls eindeutig (2006:698). Das, was als Fehler interpretiert wird, weicht von Evaluator zu Evaluator ab und hängt von der Toleranz der Evaluatoren für die möglichen Fehler ab. Ein Fehler kann zudem gleichzeitig unterschiedlichen Kategorien angehören bzw. kann schwierig zu kategorisieren sein. Darüber hinaus können die Fehler je nach MÜ-Systemfähigkeiten, (Fach)sprachen und Textsorten anders klassifiziert werden (vgl. Lehrberger/Bourbeau 1988). Beispielweise werden, wie schon im Kapitel 2 dargelegt wurde, Artikelwörter und Kopulas in der journalistischen Sprache weggelassen. Das wäre bei anderen Textsorten als Fehler zu erachten (vgl. Hutchins/Somers 1992:164).

3.4 Informationsgewinnung

Das ultimative Ziel des Endbenutzers, der einen Text maschinell übersetzt, besteht darin, u.a. Informationen zu suchen. Die entsprechende Evaluationsmethode ist daher das Messen der Anzahl von Informationen, die ein Evaluator aus einem maschinell übersetzten Text gewinnen kann. Eine Studie, die von Taylor/White (1998) durchgeführt wurde, zeigt (unter den unterschiedlichen Evaluationsaufgaben) die Informationsgewinnung (*extraction task*), und erklärt, wie sie als Evaluationsmethode eingesetzt werden kann. Sie erkennen drei Ebenen der Informationsgewinnung:

> „(1) one level will test the reporting of individual entities (names, places, organizations) found in a text, (2) a second level will test the reporting of relationships (person-to-place, person-to-organization. etc.) in a text and (3) the final level will test the recognition of a particular event type and the reporting of standard, pertinent information related to the event (e.g., for a bombing event, the location, date, time, type of device, persons injured or killed, buildings damaged or destroyed, group claiming responsibility).“ (Taylor/White 1998:369).

Diese Methode wurde in einer semi-automatischen Metrik von Lo/Wu (2011) – MEANT – umgewandelt, die darauf zielt, die Benutzbarkeit eines Outputs durch die Anzahl der erfolgreich gewonnenen Informationen zu messen. Die Informationen, die gewonnen werden sollen, sind aber nicht einzelne Daten. Die semantischen Rollen der Wörter werden in Verbindung mit dem *frame* gesehen: „(a) all core semantic roles should be checked, and (b) not only should we evaluate the presence of semantic role fillers in isolation, but also their relations to the frames' predicates.“ (2011:221) In der folgenden Tabelle (Abbildung 11) werden die von Lo/Wu (2011) bearbeiteten semantischen Rollen gezeigt. Aufgabe der Evaluatoren ist es, den Inhalt der semantischen Rolle zu vergleichen und feststellen, ob die in der MÜ gefundenen Rollen korrekt, teilweise korrekt oder falsch sind (vgl. Kit/Wong 2015:224).

Semantic role	*Event*	*Semantic role*	*Event*
Agent	who	Location	where
Action	did	Purpose	why
Experiencer	what	Manner	how
Patient	whom	Degree or extent	how
Temporal	when	Other adverbial arguments	how

Source: Lo and Wu (2011: 225)

Abb. 11: Korrespondenz zwischen semantischen Rollen und Event-Information (Lo/Wu 2011:225 in Kit/Wong 2015:224)

3.5 Leseverständnistest

Um Informationen aus einem Text zu gewinnen, ist das Verstehen des Textes eine wesentliche Voraussetzung. Der Grad der Verständlichkeit eines MÜ-Textes ist daher auch ein ausschlaggebender Faktor für das Messen der Qualität eines Systems. Um dieser Parameter zu bewerten, wurden Methoden entwickelt, die auf den gängigsten Tests für das Leseverständnis basieren. Evaluatoren lesen die MÜ-Texte und beantworten die Fragen zu dem Text. Danach lesen sie die Humanübersetzung und beantworten dieselben Fragen. Je geringer der Unterschied zwischen den zwei Performances ist, desto besser wurde der Text vom System übersetzt.

Tomita (1992) und Tomita et al. (1993) haben beispielweise die Texte der TOEFL-Sprachtest verwendet. Die Texte wurden sowohl maschinell als auch manuell ins Japanische übersetzt. An der Studie hat eine Gruppe japanischer Studierender teilgenommen, die die Übersetzung gelesen und die Fragen beantwortet haben. Das Textverständnis der Studierenden wurde durch die TOEFL-Werte beurteilt (vgl. Kit/Wong 2015:224).

Somers/Wild (2000) schlagen einen Cloze-Test vor. Diese Methode wurde in den 60er und 70er Jahren schon oft verwendet. In den nächsten 20 Jahren hat man sich dann von dieser Methode abgewandt (vgl. 2000:1). Die Cloze-Prozedur besteht im Herausnehmen einiger Wörter aus einem Text (ca. ein gelöschtes Wort pro 5 Wörter) (vgl. 2000:1f.). Danach werden die Studienteilnehmer befragt, die Lücke im Text zu ergänzen. Das Wort „cloze“ stammt vom Terminus „clozure“ (Gestaltschließung) und wurde zum ersten Mal von Taylor (1953) geprägt: „gestalt psychology applies to the human tendency to complete a familiar but not-quite finished pattern—to "see" a broken circle as a whole one, for example, by mentally closing up the gaps.” (Taylor 1953:415)

Somers/Wild (2000) haben zwei Experimente durchgeführt, in den drei MÜ-Texte und HÜ-Texte unterschiedlicher Texttypen verwendet wurden, mit dem Ziel, zu überprüfen, ob diese Prozedur als indirekte Ranking-Methode der MÜ-Systeme verwendet werden kann. In der von den zwei Autoren durchgeführten Studie wurde die Standard-Cloze-Prozedur erweitert bzw. leicht verändert (Prozentsatz an abgedeckten Wörtern, Einführung

unterschiedlicher Grade von falschen/richtigen Antworten, Berücksichtigung von Synonymen) (vgl. 2000:1).

3.6 Post-Editing

Das Post-Editing ist der bei weitem am meisten verbreitete Verwendungszweck der maschinellen Übersetzung. Das Post-Editing wird als „correction of machine translation output by human linguists/editors“ bezeichnet[16] (Allen 2003:297). Schon seit der ersten Entwicklungen im Bereich der maschinellen Übersetzung ist das Nachbearbeiten der maschinell übersetzten Texte (oder Post-Editing) für die praktische Anwendung des MÜ-Outputs unabdingbar. Das Post-Editing ist dabei für die maschinelle Übersetzung von großer Wichtigkeit, zumal die MÜ bis heute noch nicht an das Niveau der Humanübersetzung herangekommen ist. Die sogenannte Fully Automatic High Quality Machine Translation – FAHQMT, die von Bar-Hillel im Jahr 1960 als ein Traum, der sich nicht in absehbarer Zeit verwirklichen wird, bezeichnet wurde, bleibt noch heute eine Utopie. Das Post-Editing stellt daher nicht nur die einzige Möglichkeit dar, um das MÜ-Output auf eine veröffentlichbare Qualität zu bringen, sondern es hat sich zu einem eigenständigen Bereich der Übersetzung entwickelt (vgl. Allen 2003).

Da das Post-Editing noch ein relativ junger Sektor der Übersetzungsindustrie ist, sind die Rolle und die Aufgaben der Post-Editoren nicht vordefiniert. Diese variieren je nach Anforderungen der Kunden oder der Firma bzw. des Übersetzungsbüros, in denen das Post-Editing eingesetzt wird (vgl. Allen 2003:298f.). Die Arten von Post-Editing und die Richtlinien, die im Laufe der Jahre entwickelt wurden, werden hier nicht beleuchtet, weil sich dieses Kapitel hauptsächlich dem Post-Editing als Evaluationsmethode widmet.

Das Nachbearbeiten von MÜ-Texten ist gleichzeitig ein Zweck der MÜ und ein Maßstab für die Bemessung von deren Qualität. Die Hauptidee diese Methode ist folgende: Je besser ein System den Text übersetzt hat, desto weniger muss das Output nachbearbeitet werden. Die Qualität des Outputs hat einen direkten Einfluss auf die Preise der nachbearbeiteten Übersetzung.

Die zwei Parameter für die Bewertung der Qualität sind daher die Zeit, welche die Post-Editoren für das Nachbearbeiten des Segments aufgewendet haben, und der Post-Editing-Aufwand, der in der Regel als der Prozentsatz der Änderungen am Output-Segment verstanden wird. Für die Aufzeichnung dieser Parameter werden bestimmte Softwares eingesetzt, die nicht nur im MÜ-System integriert sind und das Post-Editing ermöglichen,

[16] „An increasing number of professional translators at the Commission do take raw machine output as a first draft, which they then polish up to produce a finished translation. It is perhaps more appropriate to refer to this activity as the "correcting" rather than the ‘post-editing’ of MT.“ (Senez 1998:4)

sondern auch einen Bericht der durchgeführten Aktivitäten (Zeit, Art und Anzahl der Änderungen usw.) erstellen[17].

Diese Methode ist zweifelsohne sehr subjektiv. Die Zeit, die für das Nachbearbeiten aufgewendet wird, hängt von vielen Faktoren ab: Jeder Post-Editor hat seine eigene Geschwindigkeit bzw. seinen eigenen Rhythmus, der Post-Editor kann gut oder wenig gut mit einem bestimmen Thema bzw. einer Fachsprache vertraut sein. Darüber hinaus spielt die Vertraulichkeit des Post-Editors mit dem Post-Editing-Ablauf und den Qualitätskriterien bzw. die Erfahrung mit Post-Editing auch eine Rolle. Der Prozentsatz des veränderten Textes kann auch von den Vorgaben des Post-Editings selbst abhängen, wie die Übersetzung verwendet wird, oder von der persönlichen Neigung des Post-Editors, viele (oder wenige) Korrekturen vorzunehmen (vgl. „red pen syndrome", Allen 2003:305).

> „The first case is that of over-correcting whereby the post-editor spends too much time on the post-editing process (also referred to as "over-engineering", Godden, 1999), or secondly that of undercorrecting whereby the post-editor does not su¹ciently review the document and lets significant errors appear in the resulting final text." (Allen 2003:305)

Die Post-Editing Zeit ist sicherlich direkt mit dem Post-Editing-Aufwand verbunden. Was aber unter ‚Aufwand' verstanden wird und wie dieser gemessen werden kann, ist nicht einheitlich. Zudem ist seine Bemessung komplizierter als eine reine Berechnung des Prozentsatzes des korrigierten Segments. So auch Allen (2003) dazu:

> „This human effort can be measured as the cognitive effort exerted to identify the corrections (especially since post-editing is a different task from translating or revising), as well as the manual effort to make the corrections on paper and/or on-line." (Allen 2003:298)

Der Aufwand, der von Allen als manueller Aufwand bezeichnet wurde, ist die praktische Verwicklung eines kognitiven Aufwands. Unterschiedliche Fehler oder unbekannte Wörter, die Kenntnis des Themas oder der Ausgangssprache können je nach Evaluator einen größeren kognitiven Aufwand bedeuten. Die Korrekturmaßnahmen werden demzufolge unterschiedlich viel Zeit in Anspruch nehmen. Einige Anhaltspunkte, die bei der Feststellung der Gründe eines höheren oder niedrigeren kognitiven Aufwands helfen, können von den Post-Editoren durch eine Selbst-Analyse des Post-Editing-Verfahrens angegeben werden[18]. Eine subjektive Analyse so eines subjektiven Verfahrens kann aber selbstverständlich nicht präzise und vollständig sein.

Eine erfolgreiche neue und holistische Methode, die neueste Technologien, Neurologie, Linguistik, Übersetzung, kognitive Psychologie, Softwareentwicklung und

[17] Ein Bespiel dafür ist die im Rahmen dieser Studie verwendete webbasierte Software MateCat (vgl. Kapitel 5.4.1).
[18] Diese Methode wird in der hier durchgeführten Studie verwendet (vgl. Kapitel 5.4).

Computerlinguistik vereint, ist das **Eye-Tracking** oder Blickerfassung (vgl. O'Brien et al. 2014).

Das Eye-Tracking kann eingesetzt werden, um das Verfahren der Evaluation zu verbessern. Es wird auch im Rahmen anderer Evaluationsmethoden verwendet: Doherty et al. (2010) haben das Eye-Tracking angewandt, um die Verständlichkeit eines französischen MÜ-Outputs zu bewerten. Dabei haben sie die Augenbewegungen von französischen Muttersprachlern beim Lesen des Textes aufgezeichnet und anschließend die Daten mit HTER[19]-Werten verglichen. Stymne et al. (2012) haben diese Methode im Bereich der Fehleranalyse eingesetzt und herausgefunden, dass lang dauernde Blicke und eine hohe Anzahl an Fixationen mit einer höheren Fehleranzahl im Output korrelieren (vgl. Guzmán et al. 2015).

Vieira (2014) wendet das Eye-Tracking auf das Post-Editing mit dem Ziel an, das Verständnis in Bezug auf das Post-Editing zu verbessern. Er erforscht die Prädiktoren für den kognitiven Aufwand beim PE (Ausgangstext, MÜ-Output und Merkmale der Teilnehmer), die Rolle der Kenntnisse von Ausgangs-und Zielsprache und den potentiellen Zusammenhang zwischen WMC (*working memory capacity*) und der Kenntnis der Ausgangssprache.

[19] Human-targeted Translation Edit Rate (vgl. Kapitel 5.3).

4. Automatische Evaluationsmethoden

Im letzen Kapitel wurden neben den Vorteilen der menschlichen Evaluation auch die Schwächen dargelegt, die alle manuellen Methoden kennzeichnen, unabhängig von den Evaluationsteilnehmern, Zwecken und Ausführungsmethoden. Der Hauptmangel ist – wie schon im Laufe der letzten Kapitel ausführlich erklärt wurde – die Subjektivität der Evaluation. Darüber hinaus ist die manuelle Evaluation zeit-und kostenaufwendig. Aus diesen Gründen eignen sich die menschlichen Evaluationsmethoden nicht für das schrittweise Testen bei der Softwareentwicklung oder für einen objektiven Vergleich zwischen unterschiedlichen Systemen bzw. unterschiedlichen Versionen desselben Systems. Die wissenschaftliche Gemeinschaft bedarf Methoden, um Evaluationsdaten einheitlich, einfach, schnell und zu geringen Kosten zu testen.

Schon Mitte der 90er Jahre hatten die Ergebnisse der ARPA-Evaluation von maschinellen Übersetzungen die Validität und die Reliabilität der menschlichen Bewertung von *fluency* und *adequacy* in Frage gestellt (vgl. Dorr 2011:774) und die ersten automatischen Metriken, die auf der Edit-Distanz basierten, wurden als Evaluationsmethoden der maschinellen Übersetzung verwendet (vgl. Frederking/Nirenburg 1994; Knight/Chander 1994; King 1996).

Die meisten automatischen Metriken basieren auf dem Vergleich zwischen MÜ-Output und Referenzübersetzungen (Humanübersetzungen). In diesem Sinne kann die automatische Evaluation als die reine Berechnung der Ähnlichkeit von Texten gesehen werden. Die Grundidee der automatischen Methoden ist, dass ein MÜ-Output umso besser ist, desto ähnlicher er der Referenzübersetzung ist. Bei der Humanübersetzung kann ein Text natürlich auch unterschiedlich übersetzt werden. Abgesehen davon, dass die MÜ nicht im Stande ist, hochqualitative Übersetzungen zu erstellen, ist eine komplette Ähnlichkeit nicht möglich. Allerdings wird bei diesen Methoden angenommen, dass gleiche Wörter oder Phrasen der MÜ auch in der Referenzübersetzung zu finden sind. Die Ähnlichkeit stellt hier den Qualitätsparameter dar.

Im Laufe des Kapitels werden einige der heute am meisten benutzten Methoden und deren Funktionsweise ausgeführt: BLEU, METEOR, WER, TER, HTER, TERp. Eine ausführlichere Beschreibung der Methoden, die in der hier durchgeführten Studie verwendet wurden – nämlich TER und HTER – erfolgt in Kapitel 5.3.

4.1 Metriken und Stand der Technik

Die automatischen Evaluationsmethoden, die auf dem Vergleich zwischen MÜ und Referenzübersetzungen basieren, verwenden zwei Metriken: *precision* (Präzision) und *recall* (Vollständigkeit). Da diese Begriffe für das Verstehen der Funktionsweise der automatischen Methoden zentral sind, ist eine Erklärung erforderlich. Die Anwendung dieser Parameter auf die Evaluation der maschinellen Übersetzung wird von Koehn (2010) anhand folgendem Beispiel klar erklärt:

„SYSTEM 1: Israeli officials ~~responsibility~~ of airport ~~safety~~

REFERENCE: Israeli officials are responsible for airport security

SYSTEM 2: Airport security Israeli officials are responsible“[20]

Das Beispiel zeigt zwei MÜ-Outputs. Die Präzision ist das Verhältnis zwischen der Anzahl der richtigen Wörter in der MÜ (richtig im Sinne von gleich wie bei der Referenz) und der gesamten Anzahl der Wörter in der MÜ. Das System 1 hat 3 richtige Wörter von 6 (50%) und das System 2 hat 6/6 (100%) richtige Wörter ergeben. Abgesehen von der falschen Wortstellung, könnte das System 2 ein perfektes Match mit der Referenz darstellen. Nur bei System 2 fehlt ein Wort (*for*). Dafür wird die Metrik der *recall* verwendet, die die Anzahl von Wörtern, die korrekt sein *sollten*, berechnet. Es wird daher das Verhältnis zwischen der Anzahl der richtigen Wörter und der Referenzlänge berechnet.

$$precision = \frac{correct}{output-length} \qquad recall = \frac{correct}{reference-length}$$

Die zwei Metriken werden in einem Einheitsmaß kombiniert (f-Wert).[21]

Eine der ersten automatischen Evaluationsmethoden ist die **Word Edit Rate** (**WER**). Sie basiert auf der **Levenshtein-Distanz** bzw. Editierdistanz: die minimale Anzahl der Edits (Einfügen, Löschen, und Ersetzen), die notwendig sind, um zwei Segmenten zu matchen (vgl. Koehn 2010:224). Die Anzahl der Edits wird dann durch die Referenzlänge normalisiert:

$$\text{WER} = \frac{substitutions+insertions+deletions}{reference-length}$$

[20] Koehn (2010:223f.)
[21] Für eine ausführlichere Beschreibung des *f-Wertes* vgl. Koehn (2010:224).

Der Nachteil der Levenshtein-Distanz und der WER liegt darin, dass die falsche Wortstellung innerhalb des Segmentes als zwei Änderungen bzw. Operationen zählt, nämlich das Löschen des Wortes und das Wiedereinfügen des Wortes an der richtigen Stelle.

Dieser Mangel wurde von anderen Metriken beseitigt, die sich aus der WER entwickelt haben, u.a. TER. Die **Translation Edit Rate**[22] (TER) wurde von Snover et al. 2006 entwickelt. Hier wird die Wortumstellung (Umstellung eines Wortes oder Wortsequenz) als eine Änderung bzw. *shift* betrachtet. Die Translation Edit Rate hat zwei wesentliche Nachteile: Die optimale Berechnung der Edit-Distanz durch *shifts* ist ein NP-vollständiges Problem (vgl. Lopresti/Tomkins 1997), das heißt, dass es sich nicht effizient lösen lässt. Darüber hinaus werden Wörter, die sich nur leicht ändern (zum Beispiel wegen eines Rechtschreibfehlers), oder Synonyme als Fehler gesehen.

Eine Lösung dafür stellt die **TERp** dar (Snover et al. 2008; 2009). Die TERp berücksichtigt übereinstimmende Wortstämme und Synonyme und Phrasensubstitution. Anders als TER, bei der jede Änderung mit 1 bemessen wird, variieren die Kosten einer Substitution in TER, wenn die ersetzten Wörter Synonyme sind, dieselben Wortstämme haben oder Paraphrasen sind.

Eine weitere Metrik, die aus dem TER entwickelt wurde, ist die HTER (Snover et al. 2006). **HTER** (Human-targeted Translation Edit Rate) oder *human-in the-loop-evalutation* besteht aus einem Verfahren, um gezielte Referenzen (*targeted reference*s - TARG) zu erstellen. Die TARGs sollten das Segment darstellen können, das von einem Post-Editor nachbearbeitet wird. Dabei wird der Inhalt des Segments beibehalten und nur minimal erforderliche Änderungen werden vorgenommen.

BLEU – Bilingual Evaluation Understudy (Papineni et al. 2001) – wurde als eine der ersten und einflussreichsten Metriken anerkannt. Sie erfolgt durch das Zählen der N-Gramme (Sequenzen von aufeinanderfolgenden Wörtern) von unterschiedlicher Länge, die sowohl im MÜ-Output als auch in einer oder mehreren Referenzübersetzungen vorkommen. Die Innovation von BLEU besteht nämlich darin, dass mehrere Referenzübersetzungen benutzt werden können (vgl. Koehn 2010:229). Obwohl BLEU eine der am meisten benutzten Metriken ist, weist sie einige Nachteile auf. Dabei wird nicht zwischen Wörtern, die wichtiger für die Bedeutung des Satzes sind (wie ein Substantiv oder eine Negation) und Wörtern, die weniger wichtig sind, wie Bestimmungswörter bzw. Artikel und Interpunktion, unterschieden. Außerdem arbeitet BLEU auf einer lokalen Ebene und berücksichtigt nicht die (grammatikalische) Kohäsion des Textes. Ein Satz kann im Bezug auf N-Gramme gut sein, aber auf grammatikalischer Ebene durcheinandergewürfelt sein (vgl. Koehn 2010:229).

[22] „Within the GALE community, the TER error measure is referred to as Translation Error Rate , derived from the Word Error Rate (WER) metric in the automatic speech recognition community. The name is regrettable for its implication that it is the definitive MT measure. The authors make no such claim, and have adopted the name Translation Edit Rate for use in this paper and the wider community.“ (Snover et. al. 2006:2)

Schließlich sind die Werte von BLEU sogar weniger informativ als die Werte anderer Metriken:

> „The actual BLEU scores are meaningless. Nobody knows what a BLEU score of 30% means, since the actual number depends on many factors, such as the number of reference translations, the language pair, the domain, and even the tokenization scheme used to break up the output and reference into words." (Koehn 2010:229).

Während BLEU eine Präzisions-orientierte Metrik ist, entwickeln Banerjee/Lavie (2005) eine *recall*-orientierte Metrik. METEOR (Metric for Evaluation of Translation with Explicit Ordering) sieht ein Wort-Alignment von Output und Referenzübersetzung vor. Um die Wahrscheinlichkeit eines Matches zu erhören, werden drei Mapping-Kriterien eingesetzt: (1) exakte Sequenz von Buchstaben, (2) gleicher Wortstamm, (3) Synonyme (vgl. Kit/Wong 2015:228). Anders als bei BLEU werden daher Wörter u.a. gleichen Wortstamms (wie beispielweise bei *responsible* und *responsibility*) oder Synonyme (*safety*, *security*) beachtet und gematcht. Ein Nachteil von METEOR ist, dass die Formel viel komplizierter als jene von BLEU ist. Das Matching erfolgt durch ein rechenintensives Wort-Alignment und es gibt viel mehr Parameter, die einzustellen sind (vgl. Koehn 2010:228).

4.2 Meta-Evaluation

Die Anmerkung, die Koehn (2010) zur Informativität der BLEU-Werte gemacht hat (vgl. Kapitel 4.1), kann zweifelsohne auf die ganze automatische Evaluation ausgeweitet werden. Was bedeuten eigentlich diese Werte? Was sagen diese Werte über die Qualität des Systems aus? Diese Fragen bleiben bis heute offen. So Koehn dazu:

> „The state of the current debate on automatic evaluation metrics evolves around a general consensus that automatic metrics are an essential tool for system development of statistical machine translation systems, but not fully suited to computing scores that allow us to rank systems of different types against each others. Developing evaluation metrics for this purpose is still an open challenge to the research community." (Koehn 2010:232)

Wie kann dann festgestellt werden, ob die automatischen Methoden zuverlässig sind? Diese Frage versucht man im Rahmen der jährlich durchgeführten Meta-Evaluationen zu beantworten. Die größten Meta-Evaluationen finden im Rahmen der Evaluationskampagnen statt. Bei den Evaluationskampagnen werden hauptsächlich MÜ-Systeme getestet. Indirekt stellen sie aber auch einen wichtigen Prüfstand für die manuellen und automatischen Evalautionsmethoden dar, denn die aktuellsten Erfindungen und Entwicklungen im Bereich

der Evaluation werden während der Evaluationskampagnen verwendet und deren Zuverlässigkeit wird dabei bemessen.

Die erste MÜ-Evaluationskampagne, in der sowohl statistische als auch Regel-basierte Systeme getestet wurden, wurde von ARPA (Advanced Research Projects Agency) Anfang der 90er Jahre organisiert (vgl. White/O'Connel 1994). In den letzten fünfzig Jahren sind die Evaluationskampagnen wahrhafte MÜ-Wettbewerbe geworden, an denen sowohl von Institutionen entwickelte als auch kommerzielle MÜ-Systeme teilnehmen. Seit 2001 organisiert DARPA jährliche Evaluationen, die heute von den NIST Open Machine Translation Evaluation (OpenMT) koordiniert werden. Eine bedeutende Rolle haben die Workshops on Machine Translations (WMT). Dieser Workshop, der zum neunten Mal und im Jahr 2016 als Konferenz veranstaltet wird[23], sieht unterschiedliche Aufgaben (*tasks*) vor, in der Regel einen *translation task* und *evaluation task*, und eventuelle zusätzliche Aufgaben (2016 wird u.a. das automatische Post-Editing getestet). Im Rahmen der Evaluationsaufgaben werden menschliche und automatischen Evaluationsmetriken verwendet. Dank des großen Umfangs der Evaluationssets – jedes Jahr werden bis zu ca. 70 Systeme in unterschiedlichen Sprachpaaren auf einer bestimmten Domain getestet) können hier die Metriken leicht getestet und evaluiert werden.

Die einzige Methode zum Testen der Zuverlässigkeit einer Metrik ist die Korrelation zwischen der Metrik und den Bewertungen von Annotatoren, die durch die Korrelationskoeffizienten nach Pearson oder Spearman berechnet wird. Die menschlichen Bewertungen, die für diesen Zweck am meisten benutzt werden, sind die *fluency*-und *adequacy*-Skalen und das Ranking der Systeme (vgl. Kapitel 3).

Die reinen Korrelationskoeffizienten sagen aber nichts über die tatsächliche Zuverlässigkeit einer Metrik aus, zumal es zu viele Variablen gibt, welche die Ergebnisse der Metriken und/oder der menschlichen Methoden und daher die Korrelation beider beeinflussen, wie beispielweise das Volumen an Evaluationsdaten (BLEU braucht mindesten mehr als 500 Segmente, (vgl. Thurmair 2005), die Länge der Texte (die Inter-Annotator-Übereinstimmung ist bei kürzeren Texten niedrigen), die einbezogenen Sprachpaare usw. Wenn sich all diese Variablen ändern, ändert sich auch die Korrelation der Metrik mit menschlichen Bewertungen. Darüber hinaus basieren die Metriken auf Referenzübersetzungen. Je wörtlicher ein Referenztext übersetzt wird, desto bessere Werte bekommt die MÜ: „The human translations that scored poorly were generally 'freer' translations" (Culy 2003:6). Wenn sich wörtliche Übersetzungen für Fachübersetzungen eignen können, werden hingegen allgemeine Texte freier übersetzt.[24] Das führt dazu, dass

[23] ACL 2016, First Conference on Machine Translation (WMT) (http://www.statmt.org/wmt16/, Stand:14.08.2016).

[24] Für eine ausführlichere Erklärung des Zusammenhanges zwischen Textsorte, Sprache und MÜ vgl. Kapitel 5.1.

Übersetzungen von allgemeinen Texten a priori schlechtere Werte als Fachtexte erhalten könnten.
Das sind nur einige für alle Metriken geltende Gründe, welche die Zuverlässigkeit der Metriken entkräften. Jede Metrik hat dazu auch ihre eigene intrinsische Schwäche, wie im Laufe des Kapitels ersichtlich wurde.

Die Metrics for Machine Translation Challange (NIST 2010), in der die automatischen Metriken offiziell getestet werden, hat die Schwäche der automatischen Evaluation im Stand der Technik folgendermaßen zusammengefasst:

> „• They have not yet been proved able to consistently predict the usefulness, adequacy, and reliability of MT technologies.
> • They have not demonstrated that they are as meaningful in target languages other than English.
> • They need more insights into what properties of a translation should be evaluated and into how to evaluate those properties.“ (Kit/Wong 2015:231).

Obwohl der Grund des Bedarfes an automatischen Metriken die Suche nach einer objektiven und zuverlässigen Methode zur Evaluation der MÜ ist, kann festgestellt werden, dass bisher dieses Ziel nur teilweise erreicht wurde und die menschliche Evaluation – trotz ihrer Subjektivität – noch immer unersetzbar bleibt, denn „even though human evaluation of MT is itself inconsistent and not very reliable, automatic MT evaluation measures are even less reliable and are still very far from being able to replace human judgement.“ (Turian et al. 2003:8)

5. Forschungsdesign des Evaluationsprojekts

Die zentrale Fragestellung der vorliegenden Masterarbeit – so soll diese hier nochmals in Erinnerung gerufen werden – besteht darin, anhand eines kritischen Vergleichs festzustellen, inwieweit die Werte der automatischen Metriken mit menschlichen Evaluationen korrelieren und welche Informationen hinter den rein numerischen Werten stecken. Um eine Beantwortung dieser Fragen zu ermöglichen, wurde im Rahmen dieser Arbeit eine Studie durchgeführt. Der Forschungsgegenstand besteht aus vier Evaluationsmethoden: zwei Evaluationsmetriken, TER und HTER, und zwei menschliche Evaluationskriterien, der Post-Editor-Aufwand und die Bewertungen von Endbenutzern.

Für die Auswahl der Evaluationsmethoden war das FEMTI-Framework ein wertvolles Instrument (vgl. Kapitel 2.4.2). Dank einer interaktiven Auswahl von Kontext, Ziel und Teilnehmern der Evaluation wurde durch das von ISI (Information Science Institute, Univesity of California) und ISSCO (University of Geneva) hergestellte Framework der Grundstein für eine initiale Planung der Studie gelegt. Wie bereits im Kapitel 2 beschrieben wurde, sind bei der Bewertung eines MÜ-Outputs die Qualitätskriterien, Ziele der Übersetzung und Endbenutzer untrennbar miteinander verbunden.

In der hier durchgeführten Studie wurden zwei praktische (bzw. fiktive) Situationen angenommen. Die Benutzer des MT-Systems sind (a) Post-Editoren, die den Text mittels eines MÜ-Systems vorübersetzt haben, mit dem Ziel, das Output zu verbessern bzw. nachzubearbeiten und einen Text für eine reprofähige Druckvorlage liefern zu können und (b) zwei Ärzte, die den Text aus beruflichem Interesse lesen möchten.

Im Fall (a) ist daher die Qualität der Übersetzung indirekt proportional zum Post-Editing-Aufwand, der durch eine Zusammensetzung unterschiedlicher Parameter zu verstehen ist: die Zeit, die zur Verbesserung des Segments aufgebracht wurde, der Prozentsatz des bearbeiteten Segments und etwaige Schwierigkeiten, die während des Post-Editings entstanden sind, von der Outputqualität verursacht wurden sind und die zu einem kognitiven Aufwand geführt haben. Im Fall (b) ist die Qualität der Übersetzung als *fluency* und *adequacy* zu verstehen. Die Ergebnisse der zwei menschlichen Evaluationsmethoden werden anschließend als Interpretationsschlüssel der TER-und HTER-Werte verwendet. In den folgenden Absätzen werden die unterschiedlichen Teile des Forschungsdesigns ausführlich beschrieben und anhand theoretischer Ansätze begründet.

5.1 Auswahl des Ausgangstextes

Der Ausgangstext, der maschinell übersetzt und evaluiert wurde, ist ein Patent aus dem Fachgebiet der in der Rheumatologie angewandten Molekularbiologie. Der Text ist in der Datenbank Google Patent öffentlich zugänglich und auch im Anhang dieser Arbeit zu finden. Im nächsten Schritt wird auf den Aspekt der Tauglichkeit der Fachtexte bzw. der Fachsprache für die maschinelle Übersetzung näher eingegangen. Um einen vollständigen Überblick über die Umsetzung der MÜ für Fachtexte zu geben, werden diese Merkmale durch einen zweigleisigen Ansatz analysiert: Zu den sprach-und translationswissenschaftlichen Argumentationen kommen computerlinguistische Erklärungen dazu. Darauffolgend werden die sprachlichen Merkmale der medizinischen Fachsprache im Hinblick auf die maschinelle Übersetzung beschrieben.

5.1.1 Fachtexte in der maschinellen Übersetzung

Bekanntlich produzieren MÜ-Systeme bei literarischen Texten schlechte bis unbrauchbare Outputs. Die Qualität der Übersetzung literarischer Texte mittels eines MÜ-Systems wird von Arnold et al. (1994) mit einer einfachen – aber plastischen – Metapher geschildert: „The critisism that MT systems cannot translate Shakespeare is a bit like critisism of industrial robots for not being able to dance Swan Lake“ (1994:6). Die Entscheidung, einen Fachtext als Ausgangstext für die vorliegende Studie zu wählen, stützt sich daher auf die weit verbreitete These, dass Fachtexte sich besonders gut für die maschinelle Übersetzung eignen. Eine These, die in diesem Absatz anhand theoretischer Ansätze begründet wird. Insbesondere wird hier beschrieben, welche sprachlichen Merkmale der Fachtexte für die maschinelle Übersetzung besonders geeignet sind, sowie auch deren Einschränkungen.

Der Grund warum MÜ-Systeme bessere Leistungen bei Fachsprachen (bzw. Fachtexten) als bei Gemeinsprachen erzielen, liegt in den Merkmalen der Fachsprachen selbst, die von Wilss (1977:148ff.) folgendermaßen zusammengefasst wurden:

1. Der semantische Bestandteil eines Terminus ist bei Fachtermini intralingual und interlingual relativ präzis und gegeneinander abgrenzbar. Diese intralinguale und interlinguale Abgrenzbarkeit wurde von Kocourek (1972) folgendermaßen formuliert:

> „The semantic comparison of terms belonging to different languages aimed at the eventual determination of equivalence is here called equivalence study. […] Two important properties of the term come into prominence in equivalence study. Firstly, being a lexical unit, a term may be constituted by a word or by a lexicalized word-group. That is why equivalence study does not compare individual corresponding words but complex lexical units as a whole […]. Secondly, a term is a defined lexical

> unit, so that if the same form is used in other defined senses, it is considered as homonymous within terminology, and if it is used in undefined senses, it is considered as polysemic and outside terminology.“ (Kocourek 1972:190)

2. Für jede Fachsprache ist es möglich, ein zwei-oder mehrsprachiges Glossar der Termini zu erstellen, die inhaltlich und formal äquivalent sind. Das gewährleistet einen „verhältnismäßig reibungslosen, ökonomischen, kognitv intereferenzfreien Ablauf fachsprachlicher Kommunikation“ (Wilss 1977:149). Als Beispiel nennt Wilss *communicale diseas* (Engl.) und *übertragbare Krankheiten*. Wie wir im Absatz 7.2 sehen werden, ist natürlich eine eins-zu-eins-Übersetzung der Fachtermini nicht immer möglich, und stellt insbesondere in der medizinischen Fachsprache unterschiedliche Einschränkungen dar.

3. Dank der Internationalität der Forschung kommt es zu einer internationalen Terminologieangleichung, die durch beispielweise lexikalische Entlehnungsvorgänge erkennbar wird (vgl. Wilss 1977:149). In der medizinischen Fachsprache gibt es außerdem weitere Formen von Internationalisierung, d.h. eine Reihe von Klassifikationen, Terminologien und Standards im Gesundheitswesen, wie u.a. die Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme (ICD-10-GM), die Internationale Klassifikation der Funktionsfähigkeit, Behinderung und Gesundheit (ICF) oder die Amtliche Nomenklatur zur Verschlüsselung von Medizinprodukten (UMDNS).

Fachtexte zeichnen sich daher durch einen höheren Objektivitätsgrad und Bestimmbarkeit aus. Die Allgemeintexte und insbesondere literarischen Texte sind dagegen von einem höheren Subjektivitätsgrad geprägt. Die Texte, „wo die pragmatische, referenzsemantische Inhaltsorientierung […] ausschlaggebend ist, wo individualistische Gesichtspunkte beim interlingualen Transfer zweitranging sind und wo konnotative Textelemente eine untergeordnete Rolle spielen oder gar nicht zum Tragen kommen“ (Wilss 1977:147), erzielen daher bessere Ergebnisse, denn Subjektivität eines Textes kann beim interlingualen Transfer nur durch den menschlichen Verstehensprozess des Übersetzers erkannt werden.

5.1.2 Das Verstehen im Übersetzungsprozess und die *Word Sense Disambiguation*

Das menschliche Übersetzungsverfahren beginnt mit dem Verstehen des Ausgangstextes, denn „eine Übersetzung [hat] dann besondere große Chancen, ihrem Original zu entsprechen, wenn der Übersetzer das Original verstanden hat“ (Figge 1989:302). Das Haupthindernis von Computerprogrammen ist, dass sie den Ausgangstext nicht verstehen können. Wilss unterscheidet in Anlehnung zu Enkvist (1987) drei Analyseebene des Verstehens:

syntaktisches, lexikalisches und *pragmatisches Verstehen.* Während die ersten zwei Eigenschaften in einem bestimmten Umfang von einem Computer erkannt und bearbeitet werden können, ist der Zugang zu den pragmatischen Probleme nicht möglich (Wilss 1994:170f.). Das pragmatische Verstehen setzt das Weltwissen voraus. Bei der Analyse eines Textes ist der Übersetzer dank eines *impliziten Wissens* (*tacit knowledge*) in der Lage, die „an der Textoberfläche nicht ausgedrückten semantischen Beziehungen zwischen den verschiedenen Bestandteilen einer Wortzusammensetzung" zu erkennen (Wilss 1994:171f.).
Diese Fähigkeit besitzt der Computer nicht, weil ihm die enorm große Anzahl von Informationen aus dem Weltwissen fehlt, die in der Regel ein Humanübersetzer hinzuzieht, um logische Schlussfolgerung ziehen zu können (vgl. Ramlow 2009:122). Dieses Weltwissen wird im Bereich der Künstlichen Intelligenz (KI) als ***common-sense reasoning***[25] bezeichnet. In Bezug auf die Computerlinguistik und Natural Language Processing (NLP) ist common-*sense reasoning* nur beschränkt simulierbar:

> „[…]computer cannot perform common-sense reasoning. There are several reasons for this, but perhaps the most serious is the fact that common-sense reasoning involves literally millions of facts about the world […]. The task of coding up the vast amount of knowledge required is daunting. In practice, most of what we understand by "common-sense reasoning" is far beyond the reach of modern computers." (Arnold 2003:122)

Die MÜ-Systeme können nicht sprachliches mit außersprachlichem Wissen korrelieren und können daher Ambiguitäten weder auf einer lexikalischen noch auf einer syntaktischen Ebene erkennen bzw. auflösen. Das Problem der Ambiguitäten bzw. der semantischen Disambiguierung war schon ganz zu Beginn der MÜ eine viel debattierte Frage. Gerade auf der Disambiguierung gründete Bar-Hillel seine harte Kritik (1960) über die maschinelle Übersetzung, die er anhand des bekannten Beispiels der „box in the pen"[26] veranschaulicht:

> „I now claim that no existing or imaginable program will enable an electronic computer to determine that the word pen in the given sentence within the given context has the second of the above meanings, whereas every reader with a sufficient knowledge of English will do this "automatically." Incidentally, we realize that the issue is not one that concerns translation proper, […] but a preliminary stage of this process, of the determination of the specific meaning in context of a word which, in isolation, is semantically ambiguous […]." (Bar-Hillel 1960:158)

Wie John Hutchins (1999) berichtet, ist die vom Logiker und Philosophen Bar-Hillel vorgebrachte Argumentation, obwohl diese stark theoretisch ist, trotzdem sehr schlagkräftig, denn „the full resolution of all ambiguities demands human-like understanding of reality;

[25] Für eine Vertiefung des Themas *common-sense reasoning* in der KI vgl. Naidenova (2010).
[26] „The linguistic context from which this sentence is taken is, say, the following: Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy. Assume, for simplicity's sake, that pen in English has only the following two meanings: (1) a certain writing utensil, (2) an enclosure where small children can play." (Bar-Hillel, 1960:158)

human quality translation is not a realistic goal for MT research, even perhaps as a 'futuristic' long-term project" (1999:20f.).

Fast 60 Jahre nach der geharnischten Aussage Bar-Hillels und der atemberaubenden Entwicklung der MÜ-Industrie, die er als ein „multimillion dollar affair" (vgl. Hutchins 1999:20) mit einem falschen und unerreichbaren Ziel beschrieb, bleibt die Disambiguierung eine ungelöste Frage und hat sich zu einem spezifischen Forschungsbereich der MÜ entwickelt: *word sens disambiguation*, d.h. das Bestimmen des richtigen *word sens* in einem gegebenen Kontext (vgl. Koehn 2010:44).

> „The definition of word sense, like everythinelse in semantics, is a difficult business. How many senses does the word *interest* have? […] A word has multiple senses if it has multiple translations into another language. For istance, the three meanings of *interest* we listed bove result in different translations nto German: *Interesse* (curiosity sense), *Anteil* (stake sense) and *Zins* (money sense)." (Koehn 2010:44)

Die Forschung im Bereich *word sense disambiguation* hat gezeigt, dass ein möglicher Weg für die Auflösung der Ambiguitäten in den benachbarten Wörtern liegt: „ […] closely neighboring words and content words[27] in a larger window, is a good indicator for word sense" (Koehn 2010:44).

Laut dem aktuellen Stand der Dinge gibt es unterschiedliche Komponenten zur Disambiguierung, die in die MÜ-Systeme integriert werden können. Sie können **wissensbasiert** oder **korpusbasiert** sein. Bei ersteren erfolgt die Disambiguierung durch die Benutzung von Informationen eines Lexikons (ein maschinell lesbares Wörterbuch oder ein Thesaurus, wie u.a. WordNet[28] und LDOCE[29]) und bei zweiteren erfolgt die Disambiguierung durch eine trainierte Informationsextraktion aus einem Korpus. Es gibt auch die **hybride Methode,** die sowohl wissensbasierte als auch korpusbasierte Modelle anwendet (vgl. ACL[30]).

Die Komponenten werden in der Regel nicht in die SMÜ integriert: „Statisical machine translation system consider local context in the form of language models […] and within phrase translation models […][31], so they typically do not require a special word sense disambiguation component" (Koehn 2010:44). Ein weiterer Trend der MÜ-Forschung ist es aber nicht zuletzt, solche Disambiguierung-Komponenten auch in der SMÜ zu verwenden (vgl. Chan et. al. 2007:34), obwohl es noch umstritten ist, ob die Performance von SMÜ durch WSD tatsächlich verbessert werden kann.

---

[27] „Content words refer to objects, actions or properties. Function words are for instance prepositions that express how the content words relate to each other." (Koehn, 2010:38)

[28] „WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. " (In: http://wordnet.princeton.edu/, Stand 14.08.2016)

[29] Longman English Dictionary Online.

[30] ACL Home Association for Computational Linguistics.

[31] Hier soll auf eine nähere Ausführung verzichtet werden; für eine umfassendere Erklärung der linguistischen Modelle vgl. Koehn 2010, Kap. 5 und 7.

Die Tatsache, dass MÜ-Systeme die Bedeutung der Wörter nicht verstehen können, spiegelt sich nicht nur auf der semantischen Ebene des Satzes wider, sondern auch in der Kohäsion des ganzen Textes. Die maschinelle Übersetzung ist nämlich auf die Satzebene beschränkt. Während ein Humanübersetzer die äußerlichen Markierungen der Textkohäsion (Konnektive, Rekurrenz, Textdeixis, usw.) erkennt und konsequent übersetzt, werden diese Markierungen vom MÜ-System nicht analysiert. Dies führt dazu, dass die Bedeutung des Texts beeinträchtig wird (vgl. Ramlow 2009:122).

Die Disambiguierung auf Text-Ebene stellt bis heute noch ein weiteres Hindernis für die Qualität des MÜ-Outputs dar. Es werden aber immer größere Fortschritte in dieser Richtung gemacht (vgl. Wong/Kit, 2012:1060). Zu den letzten Versuchen, die MÜ-Qualität in diesem Sinne zu verbessen, zählen beispielweise Murata et al. (2001), die eine Regel entwickelt haben, um die referentiellen Merkmale der japanischen Nominalphrasen zu erkennen und damit die Auflösung von Anaphern auf Japanisch durch die Generierung von Artikeln auf Englisch zu erleichtern; Hutchinson (2004) hat sich mit der Disambiguierung von Konnektiven („discourse markers“) beschäftigt, und Cartoni et. al. (2011) haben im Rahmen des COMTIS-Projektes eine Methode entwickelt, um Kohäsionsmerkmale (Konnektive, Tempus / Modus /Aspekt des Verbs) in die SMÜ-Systeme zu integrieren.

Abgesehen von den letzten Entwicklungen in Bereich WSD und den etwaigen Regeln, die entwickelt wurden – wie bei den o. g. Studien – bleibt ein Grundproblem bestehen: Eine Regel wird nie präzise genug sein, dass sie von einem Computer berechnet werden kann:

> „Precisely formulated rules are required because they must, ultimately, be interpreted in terms of the normal operations of computer hardware. Much of the difficulty of natural language processing in general, and MT in particular, arises from the difficulty offinding sufficiently precise formulations of intuitively very straightfor-ward ideas like In English, the subject usually comes before the verb" (the really problematic word here is usually, of course). Moreover, a precise formulation is not enough. There are problems for which rules can be formulated precisely, but for which solutions still cannot always be computed (any task that involves examining every member of an infinite set, for example).“ (Arnold 2003:122)

Zusammenfassend kann aus den vorangehenden Seiten herausgefiltert werden, dass je weniger ein Text verstanden (im Sinne vom menschlichen Verstehen) werden soll, um korrekt übersetzt zu werden, desto besser kann dieser Text maschinell übersetzt werden. Ein Text, der weniger oder kaum polysemische Wörter enthält und dessen Sprache genaue und wiederkehrende syntaktische Normen und Regeln folgt, kann am besten maschinell berechnet werden. MÜ-Systeme erzielen daher die beste Performance bei Texten, die in einer Fachsprache verfasst werden. Das bekannteste – und auch das erfolgreichste – Beispiel, das die Verwendung von Fachsprachen in der maschinellen Übersetzung veranschaulicht, ist das TAUM-Météo, ein MÜ-System, das zur Übersetzung von Wetterberichten vom Englischen ins Französische entwickelt wurde. Die Fachsprache ist hier durch einen bestimmten

Wortschatz und einen telegraphischen Stil gekennzeichnet. Die Verben haben nur zwei Zeitformen und die Syntax ist eingeschränkt: keine Relativsätze, keine Passivformulierungen, keine Artikel usw. (vgl. Kittredge1982:99). Der Erfolg von Météo stützt sich daher nicht nur auf das System, sondern auch auf die ‚kontrollierte' Sprache. Gerade wegen dieser kontrollierten Benutzung der Fachsprache[32] der Meteorologie, ist es durchaus möglich gewesen, hochqualitative Texte zu produzieren, was in der Geschichte der MÜ ein Einzelfall ist. Diese Ausnahme bestätigt aber die Regel der Umsetzbarkeit der MÜ für die Übersetzung von Fachtexten.

Bisher wurde in der vorliegenden Arbeit die Umsetzbarkeit der MÜ für Fachtexte nur auf der Ebene des Verstehens und der Analyse des Textes argumentiert. Aus translationswissenschaftlicher Sicht simuliert die maschinelle Übersetzung von Fachtexten aber auch bei der Phase der Neukodierung die Humanübersetzer besser als bei Sachtexten oder literarischen Texte.

Bei der (Human)Übersetzung von Texten verständigt sich der Übersetzer mit dem Sender und dem Adressat über den von Sabatini (1999) formulierten *vincolo interpretativo*: der Grad der interpretatorischen Freiheit bzw. Unfreiheit. Er unterscheidet zwischen stark, mittel und wenig interpretatorisch bindenden Texten (*molto, mediamente* und *poco vincolanti*). Die Fachtexte gehören in diesem Sinne zu den stark interpretatorisch bindenden Texten:

> „Vi sono rapporti comunicativi nei quali l'emittente avverte come imprescindibile, e talora anche dichiara il bisogno di restringere al massimo e comunque di regolare esplicitamente la libertà di interpretazione del testo da parte del destinatario: è questo chiaramente il caso delle leggi scritte ufficiali […], delle definizioni scientifiche ridotte all'essenziale […], delle istruzioni per l'uso di apparecchi o sostanze (ad es. i medicinali) […]. Tali rapporti e i testi che li rispecchiano, sono da definire "**fortemente vincolanti**." (Sabatini 1999:148)

Der Übersetzer, als Teilnehmer an der Kommunikation, muss sich daher bei der zielsprachigen Wiedergabe an diese kommunikative Übereinkunft (*patto comunicativo*) halten. Bei Fachtexten kommt die kommunikative Übereinkunft durch die Beschränkung der Interpretation des Textes zum Tragen. Die Interpretation der Fachtermini und der ganzen Fachtexte ist tendenziell eindeutig, und das lässt dem Übersetzer weniger Spielraum für die zielsprachliche Wiedergabe. In diesem Sinne halten sich MÜ-Systeme bei der Übersetzung von Fachtexten auch an diesen ethischen Pakt. Die Entscheidungen, die ein MÜ-System trifft, werden zwar auf statistischer Basis getroffen, aber bei der Übersetzung von Fachtexten hat das MÜ-System höhere (als bei allgemeinen Texte) Chancen, die Entscheidung des Übersetzers zu simulieren und damit auch die kommunikative Übereinkunft zu respektieren. Bei mittel oder wenig interpretatorisch bindenden Texten hat dagegen der Übersetzer seine

[32] In diesem Fall kann man von s*ublanguage* sprechen (vgl. dazu Kapitel 5.1.3).

eigene Erfahrung, Vorwissen, und Weltwissen mitzubringen, was für das MÜ-System nicht möglich ist.

5.1.3 MÜ und medizinische Fachsprache: die Grenzen der semantischen Eindeutigkeit

Die medizinischen Texte weisen die Merkmale jeder Fachsprache auf, die von Wilss im Kapitel 5.1 beschrieben wurden. Außerdem ist die Fachsprache der Medizin, wie aus der Definition von Gotti (1991:17ff.) hervorgeht, durch Eindeutigkeit, Transparenz, Nicht-Emotivität und Knappheit gekennzeichnet.

Wie im letzten Kapitel erklärt wurde, ist die semantische Eindeutigkeit (vgl. „monoreferenzialità“, Gotti 1991:17) die wichtigste Voraussetzung für die WSD. Die medizinische Sprache, genau wie alle anderen Fachsprachen auch, weist aber eigentümliche Ausnahmen von der Eindeutigkeit auf:

> „An important point for scientific translation is that, of all the components of language, technical terminology has the highest probability of one-to-one equivalence in translation. The correspondence is, it should be stressed, by no means complete; but once terminological equivalents are established, they cause relatively little trouble. It is not true, however, that the whole of the language of a scientific text, including its grammar and non-technical lexis, is similarly likely to yield one-to-one equivalents in translation.“ (Halliday et al. 1965:129)

Man kann verschiedene Fälle von Polysemie und Synonymie (wobei die erste aus translatorischer Sicht problematischer ist als die zweite) feststellen. In der medizinischen Fachsprache entstehen solche sprachlichen Phänomene beispielweise aus einer diachronischen Variation eines Terminus oder aus der Verwendung von Eponymen. Die diachronische Variation führt dazu, dass die zwei Termini (der alte und der neue Terminus) für eine gewisse Zeit als Synonyme verwendet werden oder nur eine leicht unterschiedliche Konnotation aufweisen. Ein Beispiel dafür ist das Wort *therapy*, das nach und nach das Wort *healing* ersetzt hat. Die Variation kann kulturell bedingt sein, wie beispielweise das italienische Wort *handicappato*, das bis zum Ende der neunziger Jahre keine schlechte Konnotation hatte: „For the moment *handicappato* remains in Italian, possibly because, being a calqued expression, it does not have the same connotation as handicapped. But “portatore di handicap” is already current“ (Taylor 1998:87). Wie Canepari (2013:114) feststellt, sind aber zehn Jahre nach der Aussage von Taylor die politisch korrekten Versionen *disabile* und *diversamente abile* die Regel.

Die durch die Eponyme entstehende Synonymie ist auch kein seltenes Phänomen in der medizinischen Sprache. Wie Gotti (1991) erklärt, wird oft eine wissenschaftliche Erfindung von verschiedenen Experten beansprucht:

> „[…] per cui si ha conseguentemente una moltitudine di eponimi per riferirsi allo stesso termine: il megacolon, ad esempio, è il *morbo di Hirschsprung* [Kursivsetzung im Original] per i danesi, il *morbo di Ruysch* per gli olandesi, e corrisponde addirittura a due eponimi (da usarsi a seconda della «scuola» di appartenenza, vale a dire il *morbo di Battini* e il *morbo di Mya*) per gli italiani." (Gotti, 1991:36f.)

Während die Synonymie nur zu leichten Problemen bei einer MÜ führen kann (etwa wie die Verwendung eines politisch nicht korrekten Worts), scheint die diastratische Variation der Termini problematischer zu sein. Man kann behaupten, dass wenn die diastratische Variation eines Terminus in einer Fachsprache berücksichtigt werden soll, so geht die Analyse über die Fachsprache (im Sinne von reiner Wissenschaftssprache) hinaus, zumal die Fachsprache die Kommunikation zwischen Experten voraussetzt. Bei der MÜ – und insbesondere bei einer SMÜ wie Google Translate – sollte aber die Tatsache berücksichtigt werden, dass die vertikale Schichtung eines Terminus den Output unvermeidlich beeinflusst. Ein Beispiel aus dem für die Studie ausgewählten Text ist „Rheumatische Erkrankungen", die auf Italienisch mit *reumatismi* von Google Translate übersetzt wurden. Eine Version, die nicht nur umgangssprachlich ist, sondern auch – und das ist oft der Fall – unklar.

Diese Ungenauigkeit ergibt sich aus der Tatsache, dass Google Translate nur statisch gesteuerte Entscheidungen trifft, und den Unterschied zwischen allgemeiner Sprache, Fachsprache und Subsprache nicht erkennen kann. Dieser letzte Begriff ist in der Sprachwissenschaft und in der NLP (Natural Language Processing) von besonderer Bedeutung.

> „We have found that the research papers in a given science subfield display such regularities of occurrence over and above those of the language as a whole that it is possible to write a grammar of the language used in the subfield, and that this specialized grammar closely reflects the informal structure of discourse in the subfield. We use the term *sublanguage* as for that part of the whole language which can be described by such a specialized grammar." (Sager 1982:9)

Ein Terminus kann innerhalb einer *sublanguage* eindeutig sein, und kann sich zugleich als polysemisch für die gesamte Fachsprache erweisen. Ein Beispiel dafür ist das Verb *entwickeln*, das sowohl in der allgemeinen Sprache als auch in der medizinischen Fachsprache *sich stufenweise herausbilden* bedeutet, aber in der *sublanguage* der Gynäkologie auch als der Austritt des Fötus aus dem Geburtskanal verstanden werden kann (vgl. Magris 1992).

In den im Rahmen dieser Studie verwendeten Text sind drei *sublanguages* erkennbar: Neben der Rheumatologie und der Molekularbiologie ist auch die „Patentsprache" als *sublanguge* zu sehen, da sie eine auf diese Domain beschränkte Verwendung von Grammatik und Lexik aufweist. Die Benutzung einer *sublanguage* ist für die maschinelle Übersetzung geeignet und das im Kapitel 5.1.1 erwähnte MÜ-System Météo ist ein Beispiel dafür. Google

Translate ist aber nicht in einer bestimmten Domain trainiert, und die jeweils die *sublanguages* auszeichnenden semantischen und syntaktischen Unterschiede stellen eher eine Problemquelle als eine erfolgreiche Lösung für die WSD dar.

Da Google Translate sich auf im Internet verfügbare Parallelkorpora und zielsprachige Korpora stützt, ist es besonders interessant zu testen, welche Qualität bei solchen Fachbereichen erzielt wird, deren Parallelkorpus nicht so breit gefasst ist – wie die Molekularbiologie – und in einem Sprachpaar (Deutsch – Italienisch), in dem die Übersetzung durch Englisch als Zwischensprache erfolgt.

Ein weiteres Merkmal des Textes ist, dass er durch das OCR gelesen wurde. Das führt natürlich zu Fehlern im Ausgangstext. Es wurde auf das Pre-Editing des Textes verzichtet, weil man einerseits testen wollte, wie solche Fehler wiedergegeben werden und andererseits, um eine reale Situation zu simulieren, in der Endbenutzer – die Ärzte – das Online-Tool Google Translate benutzten, um den Text verstehen zu können. Es war auch von besonderem Interesse zu testen, wie Google Translate auf die Besonderheiten dieses Textes reagiert und wie der sich daraus ergebende Output von den Annotatoren interpretiert wird.

5.2 Auswahl des MÜ-Systems: Überblick über Google Translate

Google Translate ist ein MÜ-Dienst von Google Inc. Die Übersetzung zwischen Ausgangs- und Zielsprache erfolgt direkt oder mittels Englisch als Zwischensprache (vgl. Boitet et al. 2009). Nach den letzten Abschätzungen, die von Barak Turovskz, Produkt-Manager von Google Inc., mitgeteilt wurden, wird Google Translate von mehr als 500 Millionen Anwendern benutzt. Es unterstützt 103 Sprachen und übersetzt pro Tag 100 Milliarden Wörter.[33] Verschiedene Evaluationsstudien haben gezeigt, dass Google Translate eine bessere Performance als andere öffentlich verfügbaren MÜ-Systeme erreicht, insbesondere mit der Sprachkombination Arabisch-Englisch und Chinesisch-Englisch auf der Ebene der Satzlänge, des Wortes, der Phrase und der syntaktischen Strukturen (Seljan, et al. 2011).

Die Idee, Google Translate als MÜ-System im Evaluationsprojekt zu benutzten, liegt daher in der Tatsache begründet, dass es eines der am meist verwendeten kostenlosen Übersetzungssysteme ist. Die Anwendung von Statistik in der Computerlinguistik ist aber über die letzten drei Jahrzehnte hinweg ein kontroverses Thema gewesen. Die Stellungnahmen Chomkys (1969) zum Thema fassen die Kernaussagen der Debatte zusammen: „It must be recognized that the notion of a 'probability of a sentence' is an entirely useless one, under any interpretation of this term“ (Chomsky 1969). Viele Experten aus den Bereichen der Künstlichen Intelligenz und der Computerlinguistik waren mit Chomskys Meinung einverstanden und die Statistik wurde jahrelang von der Computerlinguistik

[33] https://translate.googleblog.com/, Stand: 14.08.2016.

ausgeschlossen (vgl. Och 2001). Als Peter Brown et. al. im Jahr 1988 ihre Arbeit „A statistical approach to language translation" anlässlich der 12. Konferenz *International Conference on Computational Linguistics* (COLING) vorgestellt hatten, war das Publikum von diesen neuen Ansätzen überrumpelt. Harold Somers, der bei der Konferenz neben Peter Brown saß, erinnerte sich an diese Momente mit folgenden Worten: „The audience reaction was either incredulous, dismissive or hostile. Someone probably said 'Where's is the linguistic intuition?' to which to answer would have been 'Yes, that's the point, there isn't any'" (Way 2009:21).

Aber schon zum Zeitpunkt der Herausgabe ihrer Arbeit „The mathematics of statistical machine translation: parameter estimation" im Jahr 1993 (Brown et al.) hatten sich die SMÜ-Entwickler bereits durchgesetzt (vgl. Way 2009:18). Wie Way berichtet: „From this point on, SMT was mainstream, and no longer had to appeal to the reminder of the MT community to justify its acceptance; […]". (Way 2009:18)

Die SMÜ hat in den letzten Jahren an Dynamik gewonnen, sowohl in der Forschung als auch im Wirtschaftssektor. Allein im Zeitraum 2007-2010 wurden ca. fünftausend wissenschaftliche Arbeiten über SMÜ publiziert. Zugleich hat sich die SMÜ am Markplatz erfolgreich durchgesetzt: Von Language Weaver – dem ersten SMÜ-Unternehmen – bis zu Google Translate (vgl. Koehn 2010:xi), der im Oktober 2007 das Regelbasierte-System Systran für sein eigenes SMÜ-System aufgegeben hat (vgl. Garcia 2010).

Die statistische MÜ lernt aus zwei Typen von Daten: Zweisprachige Korpora und Korpora in der Zielsprache, das heißt, es setzt ein statistisches korpusbasiertes Lernverfahren ein. Die Parallelkorpora bestehen aus allieneirten Satzpaaren (in der Ausgangs-und Zielsprache). Aus diesen Übersetzungen der AS wird durch Zuweisung von Wahrscheinlichkeiten ein probabilistisches Übersetzungmodell trainiert, das die Wahrscheinlichkeiten berechnet, dass ein Wort *e* eine gute Übersetzung für das Wort *f* ist. Somit verstehen wir *e* als ein Wort in der Zielsprache und *e* als das entsprechende Wort in der Ausgangsprache (siehe Abbildung 12).

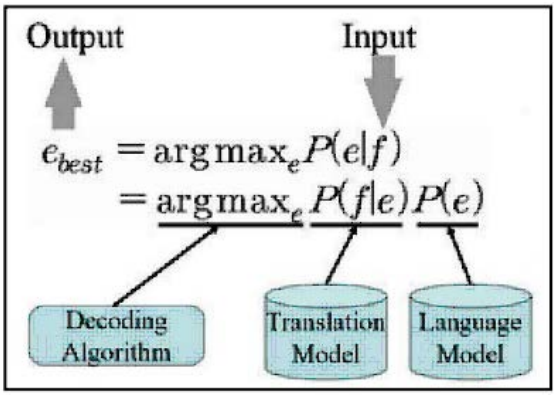

Abb. 12: Struktur einer SMÜ[34]

[34] Josef van Genabith, *Wie funktioniert maschinelle Übesertzung?* In: http://goo.gl/BWkkFt, Stand 14.08.2016.

Wenn man einen Satz übersetzt, dessen Wörter im Korpus nicht enthalten sind, geschieht die Übersetzung über die Suche nach der Wortfolge in der Zielsprache, welche nach dem erstellten Übersetzungsmodell die höchste Wahrscheinlichkeit hat. Das bedeutet, dass je mehr Daten es gibt, desto größer die Wahrscheinlichkeit ist, die richtige Übersetzung zu finden.

Beim Google Translate ist das Übersetzungsmodell nicht auf Wörter sondern auf Phrasen basiert. Phrasen-basierte Modelle sind Modelle, die kleine Wortgruppen übersetzten, diese Multi-Wort-Einheiten werden *phrases* genannt. Das Übersetzen von Wortgruppen statt einzelnen Wörtern ermöglicht eine bessere WSD (vgl. Koehn 2010:127f.).

Aus den zielsprachigen Korpora lernt das System die Sprachmodelle *P*(e). Sprachmodelle dienen dazu, einer bessere *fluency* zu erreichen und sind ein wesentlicher Bestandteil der statistischen Maschinellen Übersetzung. Sie berechnen die Wortfolgenwahrscheinlichkeit und beeinflussen u.a. die Wortauswahl, Wortumstellung (vgl. Koehn 2010:9f.).

5.3 Translation Edit Rate und Human-targeted Translation Edit Rate

Die automatischen Evaluationsmetriken, die für die vorliegende Studie ausgewählt wurden, sind die Translation Edit Rate (TER) und die Human-targeted Translation Edit Rate (HTER). Die TER wurde von Snover et. al. (2006) ausgearbeitet und im Jahr 2006 anlässlich der Konferenz der *Association for Machine Translation in the Americas* vorgestellt. Diese Metrik ist – zusammen mit METEOR und BLEU – eine der am meisten verwendeten Metriken für den Vergleich und das Ranking der Performance unterschiedlicher MÜ-Systeme im von *Association for Computational Linguistics* (ACL) veranstalteten *Workshops on Statical Machine Translation* (WMT).

Die TER stellt das Verhältnis von Fehlern in einem Segment A (Hypothese – HYP) zur Anzahl von Wörtern in einem Segment B (Referenzübersetzung – REF) dar. Es wird daher die Mindestzahl der Änderungen berechnet, die ein Post-Editor vornehmen sollte, um eine korrekte Humanübersetzung zu erzielen. Unter Referenzübersetzung versteht sich eine Humanübersetzung, während die Hypothese hingegen das Output des MÜ-Systems ist.

$$\text{TER} = \frac{\#edits}{avarange\ \#\ of\ refence\ words}$$

Die Fehler setzen sich aus *insertions*, *deletions*, *substitutions* und *shifts* von Wörtern bzw. Wortsequenzen zusammen. Ein *shift* stellt aufeinanderfolgende Wortsequenzen an eine andere Stelle der Hypothese. Alle Änderungen, inklusive *shifts* – unabhängig von der Wortanzahl der Sequenz oder des Abstands – werden mit 1 bemessen. Gerade aus diesem Grund unterscheidet sich die Metrik vom Word Error Rate (WER). Letztere erkennt umgestellte Wörter nicht und das *shift* verursacht eine Reihe von *insertions*, *deletions* und / oder *substitutions*. Interpunktionstokens werden als Wörter behandelt. Ein *shift* ist nur dann möglich, wenn die Wörter in der Hypothese und in der Referenzübersetzung gleich sind. Sollte das Wort sich auch nun wegen eines Buchstabens unterscheiden, dann kann das *shift* nicht erfolgen. Das gilt auch für Groß-und Kleinschreibung. Die *case sensitivity* kann aus dem Quellcode heraus aktiviert werden. In den Standardeinstellungen ist TER aber *case insensitive*, d. h. dass die Groß-und Kleinschreibung eventuell nur einen *shift* verursachen kann. Im Rahmen dieser Studie wurde die *case insensitive*-Einstellung beibehalten. Der Grund dafür ist, dass im Italienischen (die Zielsprache, die evaluiert wurde) die Groß-und Kleinschreibung die Bedeutung des Wortes bzw. des Segmentes nicht besonders beeinflusst – wie etwa im Deutschen – und die Änderung eines Buchstabes stellt daher keinen erheblichen (Post-Editing)Aufwand dar.
Im Folgenden wird einen Beispiel der Funktionsweise von TER angeführt.

> "REF: SAUDI ARABIA denied THIS WEEK information published in the AMERICAN new york times
> HYP: THIS WEEK THE SAUDIS denied information published in the new york times" (Snover et al. 2006:3)

HYP und REF haben dieselbe Bedeutung. Der einzige Fehler, der die Bedeutung des Satzes beeinflusst, ist, dass „AMERICAN" in der Hypothese fehlt. Der TER-Wert für dieses Segment entspricht 31%. Dieser Wert ergibt sich aus 4 Änderungen aufgeteilt auf die Wortanzahl der Referenzübersetzung. Als Änderungen gelten:

- 1 *Shift*: this week. Da die Wörter *this* und *week* aufeinanderfolgend sind, wurden sie als ein *shift* berechnet.
- 2 *Substitutions*: THE SAUDIS (HYP) statt SAUDI ARABIA (REF)
- 1 *Insertion*: AMERICAN

Die von TER berechneten Änderungen stellen aber nicht die optimale Lösung dar.
Die optimale Berechnung der Edit Distance durch *shifts* ist ein NP-vollständiges Problem (vgl. Lopresti/Tomkins 1997), das heißt, dass es sich nicht effizient lösen lässt. Aus diesem Grund werden die folgenden Approximationen bzw. Restriktionen benutzt, um den TER-Wert zu berechnen. Die Änderungszahl wird in zwei Phasen berechnet:

1) Ein Greedy-Algorithmus such nach den *shift*, welche die Anzahl der Änderungen (*insertions, deletions, substituion*) am besten reduzieren können.
2) *Insertion, Deletion,* und *Substitution* werden dann durch dynamische Programmierung berechnet.

Die Werte werden bei allen Referenzübersetzungen berechnet und nur die besten (bzw. niedrigsten) Werte, werden als definitive Werte benutzt. Für die vorliegende Studie wurden zwei Referenzübersetzungen herangezogen Für einen Text kann es verschiedene korrekte Übersetzungen geben. Sollte nur eine Referenzübersetzung verwendet werden, werden dadurch andere mögliche korrekte Übersetzungen nicht berücksichtigt und infolge können die TER-Werte somit gefälscht werden. Finch et al. (2004) behaupten, dass die Korrelation-Rate einer Metrik mit der Anzahl der Referenzübersetzungen steigt und erst mit vier Referenzübersetzungen stabil bleibt. Nach vier Referenzübersetzungen gibt es keine signifikante Steigerung. Coughlin (2003) zeigt, dass auch mit nur einer Referenzübersetzung zuverlässige Ergebnisse erzielt werden können, wenn die Testdaten aus mindestens 500 Sätzen bestehen oder die Domain sehr fachlich ist.

Da in der hier durchgeführten Studie die Domain sehr fachlich ist (Patent aus der Molekularbiologie und Rheumatologie), der Text aus 121 Segmenten besteht und diese Studie keine statistischen Ansprüche hat, wurden zwei Referenzübersetzungen als geeignet angesehen. Referenz 2 ist eine Paraphrase der Referenz 1.

Der TER-Algorithmus (*tercom*) ist quelloffen und ist auf der Website der University of Maryland verfügbar.[35] Das Java-Programm hat keine grafische Benutzeroberfläche, sondern nur die Kommandozeile. Eine Webschnittstelle wurde zur Segmentierung des Textes und zur graphischen Darstellung der von *tercom* eingeholten Informationen erstellt.[36]

Abb. 13: TER-Webschnittstelle

[35] http://www.cs.umd.edu/~snover/tercom/
[36] http://ter.panadigital.it/

In den drei schwarzen Fenstern (Abbildung 13) werden jeweils Hypothese, Referenzübersetzung 1 und Referenzübersetzung 2 eingefügt. Dann wird die Segmentierungsregel ausgewählt (Punkt, Ausrufzeichen, Fragezeichen). Aus der Standardeinstellungen der Anwendung wurden weitere Segmentierungsregeln eingeführt, die ad hoc für dieses Arbeit erstellt wurden (beispielweise „kein neues Segment nach *et. al.* oder *ca.*). Nur wenn HYP, REF 1 und REF 2 die gleiche Segmentanzahl haben, ist es möglich, die TER zu berechnen. Sollte das nicht der Fall sein, wird der Fehler „segment nummer mismatsch“ angezeigt. Nach der erfolgreichen Segmentierung kann man durch das Klicken auf den Button „TER it“ die Werte berechnen.

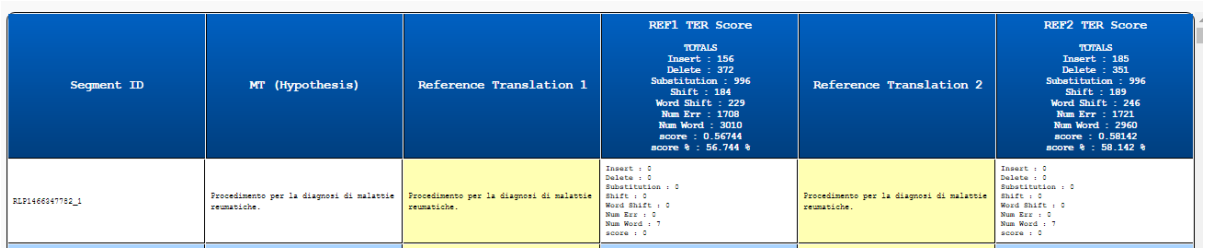

Abb. 14: TER-Webschnittstelle – Werte

Wie in der Abbildung 14 zu sehen ist, erstellt das Programm eine Tabelle mit den jeweiligen TER-Werten. Es werden die Gesamtwerte der ganzen Texte (REF1 und REF2) gezeigt, sowie auch die Werte jedes Segments und welche Änderungen vorgenommen wurden. Die Referenzübersetzung mit den besten (niedrigsten) Werten wird in Grün angezeigt. Die Ergebnisse der Evaluierung können dann in Excel heruntergeladen werden (Excel-Button).

Wie wir in dem letzten Beispiel gesehen haben, werden Synonyme nicht berücksichtig und als Fehler erkannt. Änderungen werden alle mit 1 bemessen, wobei einige Fehler die Bedeutung des Satzes mehr als anderen beeinflussen (beispielweise eine Negation). Diese Werte geben daher nicht wieder, ob die Hypothese annehmbar ist. Eine derartige Entscheidung kann nur durch das menschliche Wissen getroffen werden.

Aus diesen Gründen haben Snover et al. eine Metrik ausgearbeitet, welche berücksichtigt, dass unterschiedliche von TER signalisierte Fehler nicht von einem Annotator als solche erkannt werden. Diese Methode heißt HTER (Human-targeted Translation Edit Rate) oder *human-in the-loop-evalutation* und besteht aus einem Verfahren, um gezielte Referenzen (*targeted reference*s - TARG) zu erstellen.

Wie Snover et al. (2006) erklären:

> „In order to accurately measure the number of edits necessary to transform the hypothesis into a fluent English sentence with the same meaning as the references, one must do more than measure the distance between the hypothesis and the current references. Specifically, a more successful approach is one that finds the closest possible reference to the hypothesis from the space of all possible fluent references

> that have the same meaning as the original references. We then compute the minimum TER using this single targeted reference as a new human reference.“ (Snover 2006:5)

Wie im Kapitel 4 beschrieben, korreliert HTER gut mit der menschlichen Auswertung (basierend auf *fluency* und *adequacy*) und gilt als eine der beste Metriken, um die menschliche Bewertung zu simulieren. Daher wurde HTER (zusammen mit TER) als Evaluationsmethode für diese Studie ausgewählt.

Die Erstellung der *targeted reference* folgt einem bestimmten Verfahren, das eine inhaltliche Invarianz von HYP und TARG durch nur minimale erforderliche Änderungen gewährleistet. Bei diesem Verfahren werden den Annotatoren vier REFs und ein HYP vorgelegt. Eine der REF ist als *closest match* identifiziert. Wie schon oben erklärt, wurden im Rahmen dieser Studie nur zwei REFs verwendet, weil der Text sehr fachlich war und wenig Spielraum für syntaktische und semantische Änderung zuließ.
Um ein TARG zu erstellen, gilt es, folgende Schritte zu folgen (vgl. Snover 2006:13):

1. TER berechnen
2. *Closest match* lesen
3. HYP lesen
4. Sollte die HYP der Bedeutung der *closest REF* nicht entsprechen, ist es möglich, die anderen drei REF zu lesen.
5. Die *closest REF* so ändern, dass sie der HYP so nahe wie möglich kommt, ohne dabei die ursprüngliche Bedeutung zu verändern.
6. TER berechnen
7. Schritte 5-6 wiederholen, falls man glaubt, dass noch einige Änderungen vornehmen zu müssen, um bessere Werte zu erreichen.

Im Folgenden sind einige Beispiele des Verfahrens zur Erstellung von TARG angeführt:

Beispiel 1:

HYP: **RECLAMI** (testo OCR potrebbe contenere errori).

Best REF: **RIVENDICAZIONI** (**il** testo OCR potrebbe contenere errori).

TARG: **RIVENDICAZIONI** (testo OCR potrebbe contenere errori).

Das Wort “RECLAMI” ist in diesem Kontext falsch. Da es sich um einen Fehler handelt, der die Bedeutung des Satzes beeinflusst, wurde die (richtige) Übersetzung der REF – „RIVENDICAZIONI“ – beibehalten. Der Artikel „il“ (dt. „der“) wurde aber (wie in der

HYP) ausgelassen, weil man nach den Richtlinien Snovers einen gewissen Grad an Flexibilität bei Bestimmungswörter haben sollte (vgl. Snover 2006:16).

Beispiel 2:

HYP: 3) Polipeptide secondo la rivendicazione 2, caratterizzato dal fatto che presenta un residuo di arginina supplementare **ad** almeno due delle posizioni.

Best REF: 3) Polipeptide secondo la rivendicazione 2, caratterizzato dal fatto che presenta un residuo di arginina in più **in** almeno due delle posizioni.

TARG: 3) Polipeptide secondo la rivendicazione 2, caratterizzato dal fatto che presenta un residuo di arginina supplementare **ad** almeno due delle posizioni

Wie Snover et al. (2006) klarstellen, sollen keine strengen grammatikalischen Regeln beachtet werden: „So although the string, 'He left today from Baghdad' may seem less natural than, 'He left Baghdad today' both strings will be considered acceptable." (Snover et al. 2006:15)
Diesem Beispiel folgend, wurde die Präposition *ad* behalten, obwohl *in* die richtige Präposition gewesen wäre.

Beispiel 3:

HYP: Altri marcatori sierologici, come gli anticorpi anti-citrullina (PCC) o il punteggio HAQ iniziale, utilizzato per valutare le abilità nella vita quotidiana, o la radiografia o la **tomografia computerizzata (CT) -BiId** danno in forma in anticipo solo **piccoli affioramenti** e sono non solo abbastanza significativo al fine di valutare in che modo la prognosi del paziente sarà.

Best REF: Altri marker sierologici, come l'anticorpo anti-citrullina (CCP) o il punteggio iniziale del questionario HAQ, che serve a valutare la capacità di svolgere le attività quotidiane, oppure le immagini della Tomografia Computerizzata (TC) nello stadio iniziale della malattia forniscono **informazioni limitate** che da sole non sono abbastanza esaustive da poter formulare una prognosi del paziente.

TARG:Altri marcatori sierologici, come gli anticorpi anti-citrullina (PCC) o il punteggio HAQ iniziale, utilizzato per valutare le abilità nella vita quotidiana, o la **tomografia computerizzata (CT)** nello stadio iniziale danno solo **piccoli indizi** e sono non da soli abbastanza significativi al fine di valutare in che modo la prognosi del paziente sarà.

In diesem Beispiel wurden zwei Fehler korrigiert. Ein Wort in der HYP wurde nicht übersetzt und auf Deutsch gelassen. In der TARG wurde es gelöscht (das Wort war in der HYP überflüssig). Solche Änderungen sind sehr intuitiv und die Annotatoren verlieren dadurch nicht viel Zeit. Bei anderen Änderungen sollen hingegen die Annotatoren nach der besten Lösung suchen, um an der HYP so wenig wie möglich zu ändern bzw. die REF bestmöglich

an die HYP anzupassen. Ein Beispiel dafür ist *piccoli affioramenti (*dt. „ kleine Aufschlüsse"), dessen Bedeutung in diesem Kontext im Italienischen kaum verständlich ist. Die korrekte Version wäre *informazioni limitate* (dt. „beschränkte Informationen"). Auch wenn man *piccoli* beibehält und nur *affioramenti* durch *informazioni* ersetzen möchte, würde die Genus-Kongruenz nicht übereinstimmen *(informazioni* ist feminin plural, *piccoli* maskulin plural). Aus diesem Grund wurde eine Variante ausgewählt, die zwar kein Synonym des Wortes *informazioni* darstellt, aber die Bedeutung des Satzes nicht beeinträchtig und dabei dasselbe Genus wie *piccoli* aufweist. Solchen Überlegungen führen zu einem größeren Zeitaufwand und entsprechen nicht unbedingt niedrigen TER-Werten. Wie in Kapitel 5.3 gezeigt wird, entsprechen die logischen Überlegungen des Annotatoren nicht dem Verfahren, das TER verwendet, um die bestmögliche Lösung zu finden.

5.4 Post-Editing

Eine der zwei menschlichen Evaluationsmethoden ist das Post-Editing. Die Entscheidung, diese Methode zu wählen, kam einerseits aus der Korrelation mit TER und HTER, die das Post-Editing-Effort berechnen, und andererseits gründete sie sich darauf, dass das Post-Editing ein wachsender Sektor der Übersetzungsindustrie ist.

Die zwei Teilnehmer der Studie sind zwei Übersetzerinnen mit einer Übersetzungsausbildung, die keine Erfahrungen in Post-Editing aufwiesen. Diese Prämisse ist besonders wichtig in Hinsicht auf den Ablauf und die Ergebnisse der Studie. Wie im Kapitel 5 erwähnt, hat der Post-Editor noch keine bestimmte Rolle im Übersetzungsbereich. In der Regel sind Post-Editoren Übersetzer. Das führt dazu, dass die Post-Editoren zu Beginn des Post-Editings zeigen, dass sie noch immer an den Qualitätskriterien der Humanübersetzung festhalten. Wie Allen berichtet spielen die psycho-soziologischen Aspekte auch eine wichtige Rolle:

> „We must also take into account the psycho-social issues of the translation process. Editors […] have often accumulated years of experience, something which is certainly not trivial by any means. Such experts are possibly plagued by the "red pen syndrome" which implies that any work-related document is subject to being edited with visible red ink, that the corrections should be made as quickly as possible, and higher levels of comment indicate higher productivity on the part of the editor/reviser who has reviewed the document." (Allen 2003:305)

Die Tatsache, dass die zwei Post-Editoren immer nur als Übersetzerinnen gearbeitet haben, wird in der Evaluierung der Ergebnisse berücksichtig werden (siehe Kapitel 6). Natürlich hängt die Anzahl der Änderungen an einem Text nicht nur von der Erfahrung der Übersetzer als Post-Editoren, sondern von vielen anderen Faktoren ab, die sich gegenseitig beeinflussen. Der wichtigste Faktor ist das Ziel des Post-Editings. Für diese Studie wurde der *outbund*

translation approach von Allen ausgewählt, in dem das MT-Output als Rohübersetzung verwendet wird, und so bearbeitet wird, dass die Endversion zur Veröffentlichung geeignet ist (vgl. Allen, 2003:303). Bei der Entscheidung der Post-Editing-Protokolle wurde auf die Begriffe *full-postediting* und *minimal post-editing* verzichtet. Der Grund dafür ist die fehlende deutliche Abgrenzung zwischen diesen zwei Post-Editing-Zielen / Tasks (vgl. Allen, 2003:304). Es wurden hingegen die von Wagner (1985) in Form von Do's and Dont's vorgeschlagenen Richtlinien herangezogen und an die Texttypologie angepasst. (Allen, 2003:311)

1) So viel wie möglich aus der Rohübersetzung übernehmen;
2) Stil ist nicht wichtig, also keine großen Änderungen vornehmen, wenn er prosaisch oder repetitiv ist.

Folgende Richtlinien wurden aber nicht übernommen:
3) Bei Problemen nicht zögern. Eventuell markieren und später darauf zurückkommen;
4) Keine zeitintensive Recherche durchführen;
5) Änderungen nur dann vornehmen, wenn diese unbedingt notwendig sind – z.B. nur Wörter oder Wortegruppen korrigieren, die (a) unsinnig (b) falsch oder (c) mehrdeutig sind.

Die Richtlinien 3-5 eignen sich nicht für das Ziel und die Textsorte der durchgeführten Post-Editings. Da der nachbearbeitete Text ein Patent ist und in einer strengen Fachsprache geschrieben wurde, würde sich jeder Terminologiefehler oder Mehrdeutigkeit auf den Inhalt auswirken und den Text unverständlich machen bzw. mit inhaltlichen Fehlern versehen. Die stilistischen Fehler sind aber nicht zu korrelieren.

Die Post-Editoren haben bereits medizinische Texte übersetzt, allerdings nicht aus dem Bereich der Molekularbiologie oder Rheumatologie. Dies bringt mit sich, dass sie mit dem Thema nicht hundertprozentig vertraut waren. Dafür wurde aber ein Glossar erstellt. Das Glossar enthielt die Termini in der Ausgangs-und Zielsprache, sowie auch Definition, Kontext und Quelle. Natürlich kann ein Glossar das fachliche Wissen des Übersetzers nicht ersetzen, es mindert aber die Effekte des ungenügenden Fachwissens und reduziert die Recherchezeit.

Während des Post-Editings wurde die Zeit aufgezeichnet, welche die Übersetzerinnen für das Nachbearbeiten des Segments aufgewendet haben – inklusive der Pausen für die Recherchearbeit. Pausen außerhalb des Post-Editing-Workflows wurden nicht aufgezeichnet. In einem weiteren Schritt haben die Post-Editoren eine Bewertung des MT-Outputs und Feedback im Form von Kommentaren gegeben. In den nächsten Absätzen werden der Post-Editing-Ablauf und dessen Aufzeichnung genauer beschrieben.

5.4.1 MateCat

Zur Aufzeichnung der Post-Editing-Zeit, der Änderungen und des Feedbacks bzw. Bewertung der Post-Editoren wurde das online CAT-Tool MateCat verwendet. MateCat wurde innerhalb eines von der Europäischen Union finanzierten Rechercheprojektes erstellt, an dem verschiedene Institutionen teilgenommen haben – u.a. das internationale Recherchezentrum Fondazione Bruno Kessler und die University of Edinburgh mit Phillip Koehn. MateCat ist ein innovatives CAT-Tool, das darauf abzielt, MÜ und CAT-Tools bestmöglich zu integrieren (Benutzung von MT, TM und Domain Adaptation), um den Post-Editing-Aufwand zu reduzieren. Hier werden nur die Features des Systems vorgestellt, die im Rahmen der Studie benutzt wurden.

Zunächst wurde das Projekt erstellt. MateCat benutzt unterschiedliche MT-Systeme, u.a. Google Translate, aber nur in Verbindung mit Microsoft Translator. Aus diesem Grund wurde das Google Translate-Output in Form von Translation Memory gespeichert und auf MateCat hochgeladen. Das zweisprachige Glossar wurde auch hochgeladen und war damit im Post-Editing-Verfahren direkt verwendbar. MateCat übernimmt automatisch die (öffentliche oder private) Translation Memories mit dem besten Match. In unserem Fall wird also die ad hoc eingerichtete TM mit dem MT-Output übernommen.

Abb. 15: MateCat – Post-Editing-Workbench

Die Übersetzung kann dann nacharbeitet werden. Bei der Nachbearbeitung wird die Zeit aufgezeichnet. Sollte das Segment nach dem Speichern nochmals bearbeitet werden, werden die einzelnen Editing-Zeitabschnitte zusammengerechnet.

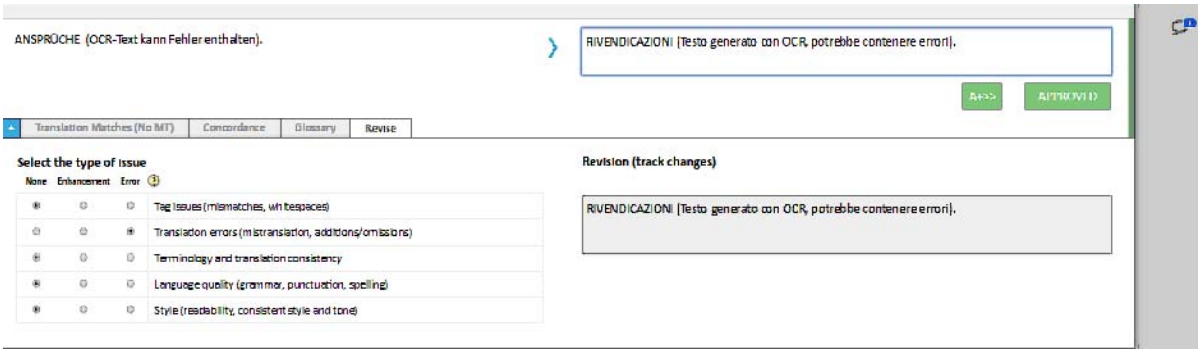

Abb. 16: Revision und Bewertung

Nach der Bearbeitung aller Segmente haben die Post-Editoren mit der Bewertung (Abbildung 16) begonnen. In dieser Phase wurden die Post-Editoren angehalten, die Segmente nicht weiter zu bearbeiten, da die Bearbeitungszeit im Revision-Modus nicht gespeichert wird. Im Kommentarfeld auf der rechten Seite konnten die Post-Editoren ein Feedback zur MÜ geben oder auf Problem hinweisen (z.B. “Nachbearbeitungszeit dieses Segment nicht berücksichtigen”).

Unten rechts konnten die Post-Editoren die MÜ-Output bewerten. Die Bewertungsrate besteht aus zwei Hauptkategorien: Verbesserung und Fehler. Die Unterkategorien sind: *Tag issues (mismatches, whitespaces), Translation errors (mistranslation, additions/omissions), Terminology and translation consistency, Language quality (grammar, punctuation, spelling), Style (readability, consistent style and tone).* Diese Klassifizierung grenzt sich von der klassischen Fehlertypologie ab und versucht eine tiefere Analyse aus der Perspektive des Post-Editors wiederzugeben. Hier konnten die Post-Editoren eine Bewertung des Outputs geben und indirekt die Störfaktoren signalisieren, die den Post-Editing-Aufwand beeinflusst hatten.

Die Ergebnisse der Bewertung wurden in einem Quality Report zusammengefasst. Die Zeitaufzeichnung und die vorgenommenen Änderungen wurden im Editing-Log gezeigt (Abbildung 17)

Words	Total Time-to-edit	Your avg secs/word	Your avg PEE
2538	02h:51m:50s	You're fast! 4.1s	30.68%

Editing Details

(Export detailed log as CSV)

Segment ID	Words	Time-to-edit (TTE)	Secs/Word	Post-editing effort (PEE)
307308962	25	02m:38s	6.4s	53%

Segment	In einer Ausführungsform des erfindungsgemäßen Verfahrens wird zu einem oben beschriebenen Polypeptid, das an einen Träger gebunden ist, als Probe die zu analysierende Körperflüssigkeit hinzugegeben.
Suggestion (TM - 100%)	In una forma di realizzazione del processo inventivo ad un polipeptide sopra descritto, che è legato ad un vettore, viene aggiunto come un campione da analizzare fluido corporeo.
Translation	In una forma di realizzazione del processo viene aggiunto come campione il fluido corporeo da analizzare al polipeptide sopra descritto, legato a un carrier.
Diff View	In una forma di realizzazione del processo ~~inventivo ad un~~ viene aggiunto come campione il fluido corporeo da analizzare al polipeptide sopra descritto, ~~che è~~ legato ~~ad~~ a un ~~vettore, viene aggiunto come un campione da analizzare fluido corporeo~~. carrier.
QA Issues	

Abb. 17: Editing-Log

Die erste Zeile enthält die Statistik des ganzen Post-Editing-Projekts: die gesamte Post-Editing-Zeit, Zeit pro Wort/Sekunde, und Post-Editing-Aufwand (PEE). Die unten stehende Tabelle zeigt die Details der einzelnen Segmente. Das PEE ist der Prozentsatz des veränderten Segments. Im Feld *Diff. View* werden durch die Track-Changes-Funktion die Änderungen gezeigt, welche die Post-Editoren am MÜ-Output vorgenommen haben. Das Editing-Log kann im CSV-Format für die weitere Verarbeitung der gesammelten Daten heruntergeladen werden.

5.5 *Gisting*, *fluency* und *adequacy*

Der vierte Teil der Evaluation ist die Bewertung der Endbenutzer von Google Translate. Die Endbenutzer, die an der Studie teilgenommen haben, sind zwei Ärzte. Die Domain des Textes wurde zusammen mit den Ärzten festgesetzt. Damit wurde vermieden, dass der Inhalt des Textes unverständlich war und es wurde die Variabel des mangelnden Vorwissens von fremden medizinischen Bereichen ausgeschlossen.

Es wurde eine reale Situation angenommen, in der die Ärzte den Text des Patents aus beruflichen Gründen verstehen möchten. Das Ziel der MÜ, das als Parameter für die Qualitätskriterien dient, ist das *gisting*. Das *gisting* ist die gängigste Verwendung eines ÜM-Systems – insbesondere von den online basierten Systemen. Unter *gisting* versteht man das Lesen eines Textes mit dem Ziel, die Kernaussagen zu verstehen (vgl. Koehn 2010:21).

Zur Evaluation des Outputs wurden zwei Qualitätskriterien ausgewählt: *fluency* und *adequacy*. Die zwei Kriterien beantworten folgende Fragen:

Fluency: Ist der Output flüssig? Dieser Aspekt bezieht sich auf die zielsprachliche Formulierungsgewandtheit, die grammatikalische Korrektheit und idiomatische Wörter und Ausdrücke.
***Adequacy*:** Wurde der Inhalt des Ausgangstextes korrekt wiedergegeben oder gibt es Auslassungen, falsche Informationen oder Informationen, die der Ausgangstext nicht enthielt? (vgl. Koehn 2010:218).

Für den Zweck dieser Evaluation ist *adequacy* relevanter als *fluency*. Ein Output, der grammatikalische Fehler enthält oder aus anderen Gründen nicht fließend lesbar ist, kann trotzdem alle im Ausganstext enthaltenen Informationen liefern.[37] Die Annotatoren haben den Output sowohl in Hinsicht auf das *adequacy* als auch auf die *fluency* bewertet, es wird aber nur der Parameter der inhaltlichen Korrektheit (*adequacy*) für den Vergleich mit den automatischen Evaluationsmethoden herangezogen, weil es das einzige Ziel der Endbenutzer ist, korrekte und vollständige Informationen aus dem Output zu gewinnen. Die *fluency*-Werte werden aufgezeichnet, um zu analysieren, wie sie mit der inhaltlichen Korrektheit korrelieren bzw. ob die *fluency* das Verstehen des Inhaltes beeinflusst hat.

Wie schon oben erwähnt, ist der Evaluationsplanung ein Vorgespräch mit den Ärzten vorangegangen, in welchem die Domain und der Text festgelegt wurden. Nach der Planung wurden den Ärzten genauere Auskünfte über die Ziele und die Methoden der Studie gegeben. Das zweiten Gespräch hat einen wichtigen Faktor hervorgehoben: Die nicht-sprachwissenschaftliche Denkweise, die zu weiteren Implikationen in der Evaluation führen

[37] Wie im Kapitel 3.1 beschrieben, sind *fluency* und *adequacy* nicht immer abgrenzbar. Eine Lösung dafür wird vom in Rahmen der Studie verwendeten Annotation-Tool Costa gegeben (vgl. Kapitel 5.5.1).

kann. Die Ärzte sind zwar “Sprachexperten” ihres Fachbereichs und daher diejenigen, die einen medizinischen Text am besten bewerten können, ihr kritisches Denken wurde allerdings nie im Bereich der Sprachwissenschaft angewandt und die Qualitätsparameter von *fluency* und *adequacy* eines MÜ-Outputs können von den Endbenutzern anders interpretiert werden, als von den Sprachwissenschaftlern, die diese Kriterien entwickelt haben. Die während des Gesprächs gelieferten Informationen wurden dann auch in Form eines Infoblattes ermittelt. Die Evaluationen wurden mittels eines Annotation-Tools aufgenommen: COSTA MT Evaluation Tool.

5.5.1 COSTA MT Evaluation Tool

Das COSTA MT Evaluation Tool (Chatzitheodorou/ Chatzistamatis 2013) ist ein Java-Programm für die menschlichen Evaluationen der Qualität eines MÜ-Outputs. Die Benutzung ist sehr einfach und intuitiv, und es eignet sich daher auch gut für Annotatoren mit nur geringen Computerkenntnissen. Die drei Hauptkriterien, die COSTA evaluiert, sind *fluency, adequacy* und die Fehlertypologie.

Die *fluency* und *adequacy* werden mittels einer 5-Punkte-Skala bewertet:

Fluency	***Adequacy***
5. Flawless language	5. All meaning
4. Good language	4. Most meaning
3. Non-native language	3. Much meaning
2. Disfluent language	2. Little meaning
1. Incomprehensible	1. None

Wie schon im Kapitel 3.1 erwähnt wurde, sind diese zwei Kriterien nicht immer voneinander abgrenzbar. Die Annotatoren können Schwierigkeit haben, welcher Fehler zu der jeweils einen oder anderen Kategorie gehört. COSTA hilft den Annotatoren bei der Unterscheidung, indem bei der Referenzübersetzung erst nur die Evaluation der *fluency* gezeigt wird.

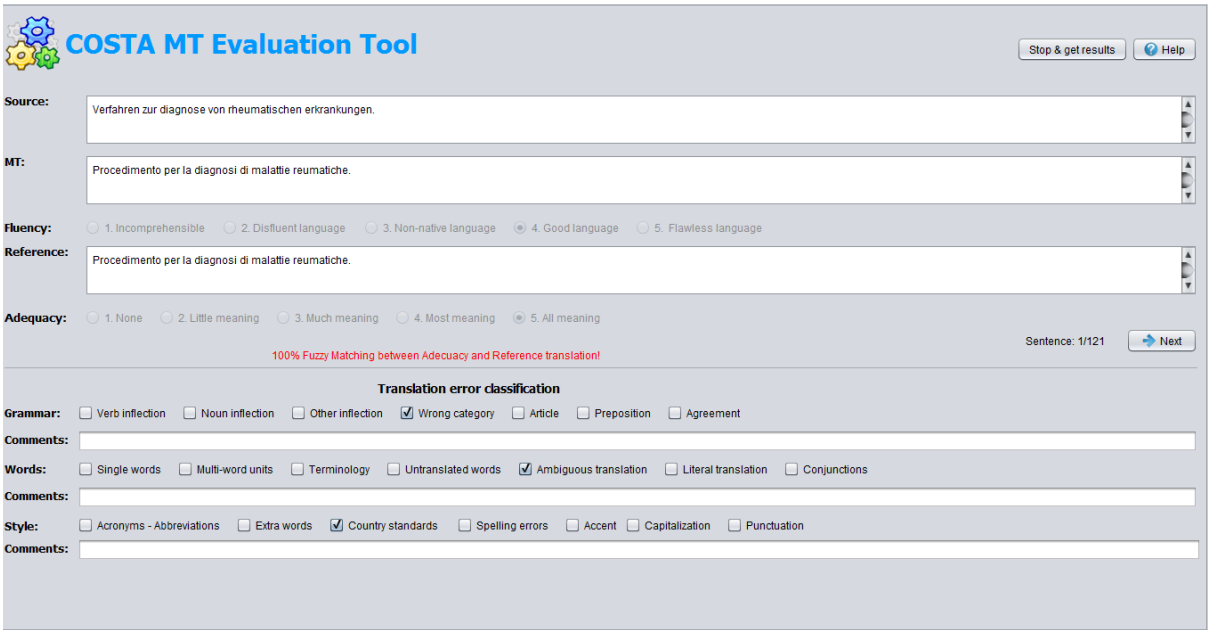

Abb. 18: COSTA MT Evaluation Tool

Die Evaluation der Übersetzungsfehler, die aus drei Hauptkategorien besteht (Grammatik, Wörter und Stil) ist fakultativ. Da die Fehlertypologie nicht als Evaluationsmethode der vorliegenden Arbeit gewählt wurde, bleibt den Annotatoren die Entscheidung über, die Fehler zu annotieren. Die Annotatoren wurden angewiesen, die Tabelle der Fehlertypologie nur dann auszufüllen, wenn sie einen Fehler signalisieren möchten, der das Verstehen des Satzes gestört hat. Für jede Fehlerkategorie gibt es ein Kommentarfeld, in dem die Annotatoren eventuell ein Feedback geben können. Es bleibt den Annotatoren überlassen, was sie mitteilen möchten. Es wurde ihnen aber erklärt, dass sie die Fehlertypologie nicht als strenge Kategorisierung gesehen wird (auch weil sie keine Sprachwissenschaftler sind), sondern dient ihnen als Input für eine kritische Analyse der Gründe, die zu einer besseren / schlechteren Evaluation geführt haben. Die Fehlertypologien wurden für die Ärzte ins Italienische übersetzt und anhand von Beispielen kurz erklärt.

6. Datenauswertung

Die im Rahmen dieser Studie gesammelten Daten wurden mit einem qualitativen Ansatz ausgewertet. Ziel dieser Arbeit ist nämlich, sich von den rein numerischen Daten zu distanzieren und so viele Informationen wie möglich über die Korrelation der in den letzten Kapiteln beschriebenen automatischen und menschlichen Evaluationsmethoden auszusagen. In einem ersten Schritt (Kapitel 6.1) werden TER und HTER-Werte verglichen und es wird versucht zu schätzen, ob die HTER-Werte und Edits-Klassifizierung den zur Erstellung der TARG-Referenz vorgenommenen Änderungen entsprechen und ob HTER bessere Vergleichswerte als TER erstellt.

Anschließend werden die (H)TER-Werte mit den Post-Editing-Daten vergleichen (Kapitel 6.2). Hier wird segmentweise analysiert, ob niedrige und höhere (H)TER-Werte durch einen niedrigen bzw. höheren Post-Editing-Aufwand gerechtfertigt werden können, und was genau den Aufwand verursacht hat. Dann werden die (H)TER Werte mit den Evaluationen der Ärzte verglichen (Kapitel 6.3). Diese letzte Methode weist einen niedrigeren Grad an Informativität auf, die aber teilweise durch die Kommentare der Studienteilnehmer ausgeglichen wird.

6.1 TER und HTER

Um eine initiale Analyse zu durchzuführen, wurden die gesamten TER und HTER-Werte berücksichtigt. Es wurden nämlich die beste Referenz und die TARG-Referenz verglichen. Die TER-Werte für die beste Referenz entsprechen 55,41%, während die Werte für die TARG 18,48% entsprechen. Das heißt, dass die Verbesserung der HYP zur TARG eine 67%ige Reduktion der Werte bewirkt hat. Das war natürlich ein vorhersehbares Ergebnis, da HTER gerade zum Zweck der Reduktion der TER-Werte (und eine vermutete konsequent höhere Korrelation mit dem Post-Editing-Aufwand) erstellt wurde. Nun werden wir anhand ausgewählter Segmente sehen, welche positiven und negativen Auswirkungen die Benutzung von HTER hat.

Im Kapitel 5.3 wurde erklärt wie eine TARG-Referenz zu erstellen ist. Das Problem der Reliabilität der TARG hängt mit ihrem Erstellungsverfahren zusammen. Die Annotatoren werden angewiesen, die HYP und die REF so ähnlich wie möglich zu halten, und schrittweise die TER zu berechnen, um die niedrigsten Werte zu erzielen. Die Änderungen, welche die Annotatoren vornehmen würden, würden aber nicht unbedingt niedrigeren TER-Werten entsprechen.

Diese Argumentation kann anhand folgender Beispiele verdeutlicht werden:

HYP: He came and brought a **cookie at 5pm**. **(Beispiel 4)**
REF: He came at **5pm** and brought a **cake.**
TARG: He came and brought a **cake at 5pm**.

HYP vs. REF: Aus dieser Gegenüberstellung ergeben sie sich 1 *shift* (**at 5pm**) and 1 *substitution* (**cake-cookie**). Diese sind die Änderungen, die ein Annotator bei der Bearbeitung des Segments berücksichtig würde, um die HYP an die REF anzupassen.
Das sind dagegen die von TER berechneten Werte:

Insert 1; Delete1, Substitution 1, Shift 1 = 4 Edits / 8 Wörter
TER-Wert: 50%.

HYP vs. TARG: Die Änderung, die von TER erkannt wurde, entspricht der potentiellen Änderung eines Annatatoren.

Substitution 1 = 1 Edit / 8 Wörter
TER-Wert = 12,5%

Das zeigt, dass TER in der ersten Gegenüberstellung (HYP vs. TARG) andere Edits machen würde als jene, welche ein Humanannotator vornehmen würde. Im zweiten Fall stimmen aber menschliche und automatische Änderungen überein. Zumindest aber erkennt TER die TARG als *best match*, indem er der TARG niedrigere Werte als REF zuweist. Das ist aber nicht immer der Fall:

HYP: yesterday I asked her out. **(Beispiel 5)**
REF: yesterday I asked her out **on a date**.
TARG: I asked her out **yesterday**.

HYP vs. REF: Auf einen ersten Blick besteht der Unterschied nur in der Einfügung dreier Wörter (**on a date**). Aus den TER-Werten ergibt sich aber eine andere Berechnung:

Delete 3, Substitution 1 = 4/ 8 Wörter = 50%

HYP vs. TARG: Aus der Berechnung des TER-Wertes ergibt sich, dass der TARG sogar ein höherer Prozentsatz als der REF zugewiesen wurde. Das heißt, dass die TARG laut der Berechnung sogar schlechter ist als die REF. In Wirklichkeit unterscheidet sich die TARG von der HYP nur in einem Shift.

Insert 1, Delete 1, Substitution 1 = 3 / 5 Wörter = 60%

Das zeigt, dass TER und damit auch HTER ein Grundproblem aufweisen. Die Änderungen, die sievornehmen, können nicht den Äderungen eines Post-Editors entsprechen bzw. diese simulieren, und zwar weder die Typologie noch die Anzahl. Dieses Grundproblem des TER ergibt sich aus dem Greedy-Algorithmus, welcher versucht, die beste Lösung zu finden (vgl. Kapitel 5.3). Wie genau der Algorithmus funktioniert, ist nur durch eine genaue Analyse des Sourcecodes zu verstehen. Hier will man aber nicht auf eine solche Analyse eingehen, zumal das Ziel dieser Arbeit auf eine Black-Box Evaluation beschränkt ist (vgl. Kapitel 2.2.2) und sich nur mit den von TER gelieferten Ergebnisse auseinandersetzt.

Diese unterschiedlichen Berechnungen der Äderungen zwischen TER und Annotatoren führen aber zu einer paradoxe Situation, in der bei der Erstellung der TARG, die Annotatoren Änderungen durchführen, welche zwar die TER-Werte reduzieren, aber möglicherweise nicht den tatsächlichen Änderungen, die ein Post-Editor vornehmen könnte, entsprechen. Diese Annahme, die schon mit Hilfe der letzten beiden Beispiele veranschaulicht wurde, wird natürlich anhand des Vergleichs mit dem Post-Editor-Aufwand im Rahmen dieser Arbeit überprüft. Die Vorgehensweise nach den von Snover (2006:13) vorgeschlagenen Richtlinien zur Erstellung der TARG-Referenz, das heißt, die Segmente zu bearbeiten, um niedrigste Werte zu erreichen, stellt einen weiteren Fall der von Way bezeichneten „case of the tail wagging the dog“ (vgl. Way 2009:30) dar.

Die Situation, die im Beispiel 5 vorgestellt wurde – d.h. dass die beste Referenz von der TER als schlechte erfasst wurde – ist in der hier durchgeführten Studie nicht vorgekommen. Die TARG-Segmente haben durch den ganzen Text bessere (niedrigere) Werte als die REF erhalten (außer drei Segmente mit den gleichen Werten in der TARG und best-REF). Obwohl die Richtigkeit der Edits bzw. der TER-Werte diskutierbar ist, kommt es dank der Verwendung einer TARG-Referenz zu Segmenten, die der HYP besser ähneln und zumindest statistisch gesehen eine bessere Korrelation als jene der TER mit den menschlichen Werten aufweisen.

Ein großes Erfolgsbeispiel von HTER zeigt sich im Segment 35:

(Beispiel 6)

HYP: Inoltre, l'invenzione riguarda una composizione farmaceutica contenente uno dei polipeptidi, nonché l'uso dei polipeptidi per la preparazione di un medicamento per la profilassi e / o il trattamento di malattie reumatiche.
best-REF: Inoltre, l’invenzione riguarda un farmaco che contiene uno dei polipeptidi e l’utilizzo dei polipeptidi nella creazione di un farmaco per la profilassi e/o il trattamento di malattie reumatiche.
TARG: s. HYP

Dieser Hypothese wurde ein TER-Wert von 54% zugewiesen. Die HYP wurde aber eins-zu-eins für die REF übernommen und der Wert ist daher von 54% auf 0% gesunken.

Noch ein Beispiel zeigt einen ca. 67%igen Unterschied zwischen TARG und best-REF:

(Beispiel 7)

HYP: Nei primi mesi della malattia, questo sta procedendo molto rapidamente.
Best-REF: Durante i primi mesi, la malattia procede molto velocemente. (100%)
TARG: Nei primi mesi della malattia, questa procede molto rapidamente. (33%)

Trotz dass HYP und REF bereits auf einen ersten Blick sehr ähnlich sind, weißt die TER einen Wert von 100% auf, das heißt, der Post-Editor könnte sogar den ganzen Satz löschen und ihn erneut schreiben. Der von der TARG abgemilderte Wert ist hingegen realistischer. Der Satz ist nämlich gut verständlich und kann mit zwei Änderungen auch grammatikalisch korrekt sein.

Das zeigt, dass die beste Referenzübersetzung und die Hypothese beide inhaltlich (auf jedem Fall nicht stilistisch, lexikalisch oder syntaktisch) potentiell richtig sein können und trotzdem höhere TER-Werte bekommen. HTER stellt daher eine Lösung zu diesem Problem dar. Doch die Tatsache, dass genau das Gegenteil passieren kann (wie im Beispiel 5) lässt die Frage über die Reliabilität von TER offen.

6.2 HTER und Post-Editing-Aufwand

Eines der Vergleichskriterien für den Post-Editing-Aufwand ist die Zeit, die für die Bearbeitung des Segmentes verwendet wurde. Die Post-Editing-Zeit ist sehr subjektiv und kann natürlich von unterschiedlichen Faktoren abhängen, die über die Qualität des MÜ-Outputs hinausgehen.
Diese Faktoren können sein:

(a) Ziel des Post-Editings: Bei einem *minimal* Post-Editing-Auftrag sollen die Post-Editoren so wenig Zeit wie möglich für die Bearbeitung des Segments aufwenden. Bei *full*-Post-Editings wird dagegen von den Post-Editoren verlangt, dass sie einen Text erstellen (vgl. Allen, 2003:304). Wie auch Allen (2003) betont, kann bei einem *full*-Post-Editing die Nützlichkeit des Post-Editings im Vergleich zu einer Übersetzung hinsichtlich der Zeit und Kosten diskutierbar sein, denn ein Übersetzer könnte eventuell einen qualitativ hochwertigeren zielsprachigen Text liefern, und dabei schneller sein als beim Post-Editing.

Diese Problematik kommt gerade in der hier durchgeführten Studie zum Ausdruck. Die Ergebnisse der Post-Editoren haben gezeigt, dass bei Fachtexten die Richtigkeit des

Inhalts derartig ausschlaggebend ist, dass die Anzahl der Wörter oder Segmente, die passabel bzw. nicht zu bearbeiten sind, drastisch sinkt.

(b) Vertraulichkeit des Post-Editors mit dem Thema. Wie schon im Kapitel 5 erklärt wird, ist dieser Faktor für diese Studie relevant.

(c) Vertraulichkeit des Post-Editors mit dem Post-Editing-Ablauf und Qualitätskriterien.

(d) Persönlicher Rhythmus.

Abgesehen vom persönlichen Rhythmus, der schwierig zu bestimmen ist, haben die anderen Faktoren die Ergebnisse der durchgeführten Studie stark beeinflusst. Für beide Post-Editoren[38] handelt es sich um den ersten Post-Editing-Auftrag. Das hat sich in den Ergebnissen gezeigt, indem sie die im Kapitel 5.4 beschriebenen Richtlinien für das Post-Editing unterschiedlich interpretiert bzw. umgesetzt haben. Das hat sich darin gezeigt, dass zwei unterschiedliche Post-Editing-Ziele angestrebt wurden. Zumal es nicht immer möglich ist, ein *minimal* von einem *full*-Postediting zu trennen, sind die zwei Post-Editoren tendenziell in diese zwei Richtungen gegangen. Der Post-Editor 1 (PE1) hat ein geringes Post-Editing durchgeführt, weniger recherchiert, weniger überprüft. Gerade im Sinne eines Post-Editings hat er die Übersetzung von Google Translate, wo immer das möglich war, einer langen Recherche vorgezogen.

Der Post-Editor 2 (PE2) hat mehr Zeit für die Recherche und die Überprüfung der von Google Translate vorgeschlagenen Termini aufgewendet, und wahrscheinlich auch für ein besseres Verstehen des Ausgangstextes.
Diese Unterschiede zeigen sich in der Zeit, die sie für den ganzen Post-Editing-Auftrag verwendet haben:

PE1-Zeit→ 02h:51m:50s
PE2-Zeit→05h:44m:45s
Zeit-Unterschied → + 50%

PE2 hat genau die doppelte Zeit gebraucht, um den Auftrag durchzuführen. Ein so großer Unterschied könnte ein Minusfaktor für eine Metaevaluation sein, da die inter-Annotatoren-Korrelation nur schwer berechenbar ist. Im Rahmen dieser Studie, die sich von einer solchen statischen Auswertung distanziert hat, haben die zwei unterschiedlichen Ergebnisse den Vorteil, dass mehrere Aspekte desselben Phänomens zeigen können, was genau für eine qualitative Studie sehr wünschenswert ist.

[38] Es sei darauf hingewiesen, dass zum Zwecke der besseren Lesbarkeit ausschließlich die männliche Form verwendet wird. Die Teilnehmerinnen der Studie sind aber zwei Post-Editorinnen.

Aus den im Kapitel 6.1 erklärten Gründen werden als Hauptvergleichsparameter die HTER-Werte verwendet. Die TER-Werte werden zum reinen Vergleich angezeigt, aber nicht berücksichtigt. Die TER-Werte des gesamten Post-Editing-Outputs zeigen einen Unterschied von 8%.
PE1→39%
PE2→47%

Die ausgewählten Beispiele wurden wegen ihrer Informativität ausgewählt bzw. wegen den Schlussfolgerungen, die daraus abgeleitet werden können. Die Beispiele werden in abnehmender Reihenfolge der HTER-Werte angeführt.

Segmente (Beispiel 8)	HTER (Best REF)	PE-Zeit	PE-Zeit/Wort[39]
HYP: Gli anticorpi sono indicati per questa ragione oggi come gli anticorpi anti-cheratina (AKA)			
TARG: Gli anticorpi sono indicati per questa ragione oggi come gli anticorpi anti-cheratina (AKA)	0% (73%)		
PE1: ~~Gli~~ Per tale motivo, gli anticorpi sono ~~indicati per questa ragione~~ fino a oggi considerati come gli anticorpi anti-cheratina (AKA)	46%	1m:02s	4,8s
PE2: ~~Gli~~ Per questo motivo gli anticorpi sono indicati ~~per questa ragione~~ oggi come ~~gli~~ anticorpi anti-cheratina (AKA).	50%	1m:42	7,9s

Aus diesem ersten Beispiel (Beispiel 8) wird ersichtlich, dass sich die TER-Werte des TARG (d.h. HTER) sehr stark von den PE-TER-Werten unterscheidet. Die Zeit der zwei Post-Editoren weist leichte Unterschiede auf. Trotzt der Track-Change-Funktion, ist es schwierig zu sagen, wie viele Änderungen (im Sinne von Edits) tatsächlich vorgenommen wurden, denn zwei oder drei Wörter könnten beispielweise zusammen gelöscht worden sein. Wenn aber nur die eingefügten bzw. gelöschten Wörter berücksichtig werden, so sieht man, dass der PE1

[39] Es sei darauf hingewiesen, dass die PE-Zeit/Wort nicht vom MateCat stammt, weil das CAT-Tool in der Berechnung von Zeit/Wort, die Wörterzahl des Ausgangssegments berücksichtigt, und sich daher nicht die PE-Zeit, sondern die Übersetzungs-Zeit ergibt. Die PE-Zeit wurde hier extra berechnet.

mehr am Segment geändert hat. Trotzdem wurde dem Segment von PE2 ein höherer TER-Wert zugewiesen.

Segmente (Beispiel 9)	HTER (Best REF)	PE-Zeit	PE-Zeit/Wort
HYP: una ulteriore forma di realizzazione, il polipeptide comprende inoltre almeno una delle posizioni 3, 20, 33, 36, 37, 94, 165, 361, 399 o 426 rispetto alla sequenza nativa, un residuo di leucina supplementare, preferibilmente in posizioni 33, 36 e / o 37.			
TARG: In una ulteriore forma di realizzazione, il polipeptide comprende inoltre almeno una delle posizioni 3, 20, 33, 36, 37, 94, 165, 361, 399 o 426 rispetto alla sequenza nativa, un residuo di leucina supplementare, preferibilmente in posizioni 33, 36 e / o 37.	0% (35%)		
PE1: In una ulteriore forma di realizzazione, il polipeptide ~~comprende~~ presenta inoltre in almeno una delle posizioni 3, 20, 33, 36, 37, 94, 165, 361, 399 o 426 un residuo di leucina supplementare rispetto alla sequenza nativa, ~~un residuo di leucina supplementare,~~ preferibilmente ~~in~~ alle posizioni 33, 36 ~~e / o~~ e/o 37.	19%	1m:02s	1,5s
PE2: In una ulteriore ~~forma~~ modalità di realizzazione, il polipeptide ~~comprende inoltre~~ presenta, inoltre, in almeno una delle posizioni 3, 20, 33, 36, 37, 94, 165, 361, 399 o 426 ~~rispetto alla sequenza nativa,~~ un residuo di leucina ~~supplementare,~~ in più rispetto alla sequenza allo stato nativo, preferibilmente ~~in~~ nelle posizioni 33, 36 ~~e / o~~ e/o 37.	31%	4m:31s	5.7s

Dieses Beispiel (Beispiel 9) zeigt, dass der PE2 74% mehr gebraucht hat, um das Post-Editing durchzuführen. Die Edits weisen aber keine besonderen Unterschiede zwischen den PEs auf. An dieser Stelle soll darauf hingewiesen werden, dass der TER-Wert der *best*-REF besser mit dem Wert des PE2 korreliert. Allerdings korreliert er nicht gut mit den PE1-Werten. Diese Wert-Unterschiede können im Rahmen dieser (qualitativen) Studie nicht als repräsentativ angenommen werden. Es sei aber daran erinnert, dass es sich um einen Fachtext

handelt, und die sich nicht so sehr unterscheidenden Edits, vier unterschiedliche TER-Werte aufweisen.

Segmente (Beispiel 10)	HTER (Best REF)	PE-Zeit	PE-Zeit/Wort
HYP: o E 'estremamente sorprendente, per indicare che polipeptidi dell'invenzione dimostrano di essere antigeni altamente specifici e altamente sensibili per la diagnosi di anticorpi nei fluidi corporei di pazienti con malattie reumatiche, in particolare le malattie infiammatorie delle articolazioni e del sistema muscolo-scheletrico, in particolare 5 preferibilmente di artrite reumatoide.			
TARG: È estremamente sorprendente indicare che i polipeptidi dell'invenzione dimostrano di essere antigeni altamente specifici e altamente sensibili per la diagnosi di anticorpi nei fluidi corporei di pazienti con malattie reumatiche, in particolare le malattie infiammatorie delle articolazioni e del sistema muscolo-scheletrico, in particolare preferibilmente di artrite reumatoide.	15% (58%)		
PE1: o ~~E 'estremamente sorprendente, per indicare~~ è estremamente sorprendente che i polipeptidi ~~dell'invenzione dimostrano di essere~~ oggetto dello studio si presentino come antigeni altamente specifici e altamente sensibili per la diagnosi di anticorpi nei fluidi corporei di pazienti con malattie reumatiche, in particolare le malattie infiammatorie delle articolazioni e del sistema muscolo-scheletrico, in particolare 5 ~~preferibilmente~~ di artrite reumatoide.	27%	1m:02s	1,26s
PE2: ~~o E 'estremamente sorprendente, per indicare~~ È estremamente sorprendente che i polipeptidi ~~dell'invenzione dimostrano di~~ relativi all'invenzione si dimostrino essere antigeni altamente specifici e altamente sensibili per la ~~diagnosi~~ diagnostica di anticorpi ~~nei fluidi~~ in liquidi corporei di pazienti con malattie reumatiche, in particolare ~~le~~ affetti da malattie infiammatorie delle articolazioni e ~~del sistema muscolo-scheletrico, in particolare 5 preferibilmente di artrite~~ dell'apparato motorio, e soprattutto dell'artrite reumatoide.	56 %	05m:39s	6,91s

Dieses Beispiel (Beispiel 10) zeigt, wie die PEs mit dem von OCR verursachten Fehler im Ausgangstext umgehen. P1 korrigiert die Fehler nicht, er gibt aber im Kommentarfeld an, dass es im Ausgangstext einen möglichen Fehler gibt. Diese zwei Fehler, die nur von PE2 berücksichtigt wurden, rechtfertigten allerdings nicht den so großen zeitlichen Unterschied. Abgesehen von der Änderungszahl kann man behaupten, dass die Art der von PE2 vorgenommen Änderungen einen möglicherweise höheren kognitiven Aufwand verursacht hat, indem PE2 beispielweise über den leichten Unterschied zwischen *diagnosi* und *diagnostica* nachgedacht hat oder den Unterschied möglicherweise recherchiert hat. Dasselbe gilt für *fluidi* und *liquidi*. Solche Fehler können mehr Zeit beanspruchen als das Löschen von OCR-Fehlern. Aus diesem Grund sollten die Änderungen nicht nur in einem Zeit-Zahl-Verhältnis gesehen werden, sondern es soll auch der mögliche kognitive Aufwand berücksichtigt werden.

Segmente (Beispiel 11)	HTER (Best REF)	PE-Zeit	PE-Zeit/Wort
HYP: Ad esempio, il derivato peptidico di essere un polipeptide retro / inverso, vale a dire un polipeptide inversa dei polipeptidi sopra descritti, la ("speculare") dei polipeptidi di D-amminoacidi è prodotto secondo un'immagine speculare, un polipeptide retro avente una sequenza "reverse", così come un polipeptide retro-inverso riflette un dei polipeptidi sopra descritti e comprende anche una sequenza "inversa".			
TARG: Ad esempio, il derivato peptidico può essere un polipeptide retro / inverso, vale a dire un polipeptide inverso dei polipeptidi sopra descritti, che è prodotto secondo un'immagine speculare la ("mirror image") dei polipeptidi di D-amminoacidi, un retro polipeptide avente una sequenza "inversa", così come un polipeptide retro-inverso che riflette uno dei polipeptidi sopra descritti e comprende anche una sequenza "inversa".	22% (74%)		
PE1: ~~Ad~~ Per esempio, il derivato peptidico ~~di~~ può essere un polipeptide ~~retro / inverso,~~ retro/inverso, vale a dire un polipeptide ~~inversa~~ inverso dei polipeptidi sopra ~~descritti, la ("speculare")~~ descritti prodotto come immagine speculare("mirror image") dei polipeptidi di ~~D-amminoacidi è prodotto secondo un'immagine speculare, un polipeptide retro avente~~ amminoacidi della serie D, che presenta una sequenza ~~"reverse", così come~~ "inversa" e un polipeptide retro-	52%	04m:03s	7,36 s

inverso ~~riflette un~~ e che presenta un'immagine riflessa dei polipeptidi sopra descritti e ~~comprende~~ anche una sequenza "inversa".			
PE2: Ad esempio, il derivato peptidico ~~di~~ può essere un polipeptide ~~retro / inverso, vale a dire~~ retro/inverso, ovvero un polipeptide ~~inversa~~ inverso dei polipeptidi ~~sopra descritti,~~ ~~la~~ descritti sopra, il quale viene prodotto in maniera che riproduca l'immagine speculare (~~"speculare")~~ "mirror image") dei polipeptidi ~~di D-amminoacidi è prodotto secondo un'immagine speculare,~~ tramite amminoacidi della serie D, un retro-polipeptide, che presenta una sequenza "inversa", e un polipeptide ~~retro avente una sequenza "reverse", così come un polipeptide retro-inverso riflette un~~ retro-inverso, che rappresenta l'immagine speculare dei polipeptidi descrittisopra ~~descritti~~ e ~~comprende anche~~ che presenta pure una sequenza "inversa".	64%	10m:48s	19,64s

Auch in diesem Fall (Beispiel 11) zeigt die best-REF eine größere Korrelation mit den PE1- und PE2-Werten. Dieses Segment stellte für die PEs eine besondere Herausforderung dar, weil es inhaltlich und syntaktisch komplex war und ohne größere Kenntnisse dieses spezifischen Themas nicht schnell bearbeitet werden könnte.

PE1 gibt im Kommentarfeld an, dass er versucht hat, das komplizierte Segment bestmöglich umzuschreiben, bis es einen Sinn ergab, aber für einen genaueres Post-Editing sollte man eine längere Recherchearbeit durchführen. PE2 zeigt durch den größeren Zeitaufwand, dass das Segment besondere Schwierigkeiten verursacht hat. In der Fehler-Kategorisierung signalisiert er zwei Fehler und zwar jeweils in der Kategorie Translation errors (mistranslation, additions/omissions) und Language quality (grammar, punctuation, spelling).

Segmente (Beispiel 12)	HTER (Best REF)	PE-Zeit	PE-Zeit/Wort
HYP: Un altro scopo della presente invenzione è un frammento della menzionata sopra Polipeptidi, la vimentina nativo avendo No. SEQ ID 1 è derivato e l'almeno contiene una regione con almeno un residuo di arginina e che presenta reattività contro le malattie associate con autoanticorpi reumatoide.			

TARG: Un altro scopo della presente invenzione è un frammento del sopra menzionato polipeptide, derivaot da vimentina nativa avendo No. SEQ ID 1 e che almeno contiene una regione con almeno un residuo di arginina e che presenta reattività contro autoanticorpi associati con malattie reumatiche.	36% (66%)		
PE1: Un altro ~~scopo~~ oggetto della presente ~~invenzione~~ scoperta è un frammento ~~della menzionata~~ di uno dei polipeptidi sopra ~~Polipeptidi, la~~ menzionati, derivato dalla vimentina ~~nativo avendo No. SEQ ID 1 è derivato e l'almeno~~ nativa con n° 1, che contiene almeno una regione con ~~almeno~~ minimo un residuo di arginina e che presenta una reattività contro ~~le~~ gli autoanticorpi associati alle malattie ~~associate con autoanticorpi reumatoide.~~ reumatiche.	60%	3m:15s	4,4s
PE2: Un ~~altro scopo~~ ulteriore oggetto della presente invenzione è un frammento ~~della menzionata sopra Polipeptidi,~~ del suddetto polipeptide, derivato dalla vimentina allo stato nativo con la ~~vimentina nativo avendo No.~~sequenza SEQ ID ~~1 è derivato e l'almeno contiene~~ NO. No.1, che comprende almeno una regione con almeno un residuo di arginina e che presenta reattività ~~contro le~~ ad autoanticorpi associati a malattie ~~associate con autoanticorpi reumatoide.~~ reumatoidi.	52%	1m:54	2,5s

Aus dem Kommentar des Post-Editors1 geht Folgendes hervor: „In so langen Sätzen verliert sich Google Translate, in dem das MÜ-System die Wörter falsch umstellt. Google T. unterscheidet nicht zwischen Subjekt und Objekt. Das führt zu Schwierigkeiten beim Post-Editing."

In diesem Fall kommt der Aufwand des Post-Editors1 auch durch eine höhere Bearbeitungszeit zum Tragen. Post-Editing-Fehler Kategorie: Er hat in der Kategorie ‚ranslation errors (misttranslation, additions/omitions)' eine Verbesserungsänderung (*enhancement*) erkannt und in der Kategorie ‚style (readability, consistent style and tone)', ein Fehler erkannt.

Wie schon früher erwähnt wurde, war das Kommentarfeld auch dafür bestimmt, mögliche technische Probleme, die die Zeitaufzeichnung beeinflussen würden, zu signalisieren. Das ist genau der Fall bei PE2, welcher im Kommentarfeld angibt, dass aufgrund einer technischen Störung, die Zeitaufzeichnung länger sein könnte als sie es tatsächlich war.

Natürlich würde man in einer statistischen Studie, solche Segmente und auch die jeweiligen HYP, TARG, und PE1 vor der statischen Bearbeitung löschen, allerdings gilt das nicht für diese Studie. Darüber hinaus zeigt die Methode des Kommentarfelds, dass die Zeit, die für die Bearbeitung eines Segments verwendet wird, nicht unbedingt der Anzahl der Änderungen entspricht, sondern auch anderen Faktoren, die einen kognitiver Aufwand berücksichtigen, wie bei der PE1.

Segmente (Beispiel 13)	HTER (Best REF)	PE-Zeit	PE-Zeit/Wort
HYP: Questa ipotesi potrebbe essere confutata da noi da stati arricchiti da differenziale immunoaffinità cromatografia sequenze immunologicamente vimentin reattiva varianti aver mutato da monociti umani.			
TARG: Questa ipotesi è stata da noi confutata: tramite cromatografia differenziale per immunoaffinità sono state arricchite varianti di vimentina immunologicamente con sequenze mtate di monociti umani.	72% (90%)		
PE1: ~~Questa ipotesi potrebbe~~ Potremmo confutare tale ipotesi, in quanto tramite immunoaffinità-cromatografia differenziale, le varianti di vimentina immunologicamente reattive con sequenze mutate monociti umani possono essere~~confutata da noi da stati arricchiti da differenziale immunoaffinità cromatografia sequenze immunologicamente vimentin reattiva varianti aver mutato da monociti umani.~~ arricchite.	100%	8m:14s	20s
PE2: ~~Questa ipotesi potrebbe essere confutata da noi da stati arricchiti da~~ Abbiamo potuto confutare questa ipotesi, dato che è stato possibile arricchire le varianti di vimentina immunologicamente reattive con sequenze mutate di monociti umani, attraverso cromatografia differenziale per immunoaffinità ~~cromatografia sequenze immunologicamente vimentin reattiva varianti aver mutato da monociti umani.~~ .	93%	9m:59s	25s

PE2 gibt im Kommentarfeld an, dass der MÜ-Output kaum verständlich ist. Die Track-Changes-Funktion zeigt, dass beide Post-Editoren ganze Satzglieder gelöscht und neu geschrieben haben. Besonders soll darauf hingewiesen werden, dass sich beide PE dafür entschieden haben, dass es günstiger (in Hinsicht auf die Zeit und – unbewusst – auf den

kognitiven Aufwand) als die Bearbeitung einzelner Satzelemente gewesen wäre. Es würde sich daher lohnen, die kognitiven Prozesse (anhand einer höheren Datenzahl) zu analysieren und diese womöglich mit den Fehlern korrelieren.

Segmente (Beispiel 14)	HTER (Best REF)	PE-Zeit	PE-Zeit/Wort
HYP: metodi di rilevamento preferite sono nel processo dell'invenzione, un metodo radioimmunologico, un Chemolumineszenzimmunoassay, un saggio immuno-blot o un immunodosaggio enzimatico, ad esempio un test ELISA, messo in discussione.			
TARG: Metodi di rilevamento preferiti sono nel processo dell'invenzione, sono un metodo radioimmunologico, un un dosaggio immunologico chemiluminescente, un saggio immuno-blot o un immunodosaggio enzimatico, ad esempio un test ELISA.	34% (75%)		
PE1: I metodi di ~~rilevamento preferite sono~~ analisi principali nel processo ~~dell'invenzione, un metodo~~ di studio sono il dosaggio radioimmunologico, ~~un Chemolumineszenzimmunoassay, un saggio immuno-blot~~ il dosaggio immunologico chemiluminescente, il western blot o ~~un immunodosaggio enzimatico, ad~~ il dosaggio immunoenzimatico, per esempio ~~un test ELISA, messo in discussione.~~ l'ELISA.	93%	02m:09s	6,8s
PE2: Quali metodi di ~~rilevamento preferite sono nel processo dell'invenzione,~~ analisi preferiti nell'iter relativo all'invenzione figurano un ~~metodo~~ dosaggio radioimmunologico, un ~~Chemolumineszenzimmunoassay,~~ dosaggio immunologico chemiluminescente, un ~~saggio immuno-blot~~ western-blot o un ~~immunodosaggio enzimatico,~~ dosaggio immunoenzimatico, ad esempio un ~~test ELISA, messo in discussione.~~ ELISA.	64%	05m:44s	18,10s

Aus diesem Beispiel wird ersichtlich, dass sich nicht nur die TER-Werte der TARG sehr stark von den PE-TER-Werten unterscheiden, sondern auch, dass der PE1, der einen höheren TER-Wert erhalten hat, weniger Zeit (-62%) für die Nachbearbeitung aufgebracht hat.

Diese Beispiele zeigen, dass die TER viele Faktoren nicht berücksichtigt (bzw. berücksichtigen kann), welche die PE-Zeit beeinflussen. Ein weiterer Faktor ist die

Reduktion der PE-Zeit wegen des Nachbearbeitens ähnlicher Segmente oder des Korrigierens wiederholter Fehler. Ein Beispiel dafür ist in den Segmenten ersichtlich, welche im Text als Liste der ‚Ansprüche' abgebildet sind. Viele dieser Segmente zeigen eine gleiche oder absteigende PE-Zeit.

6.3 TER und *gisting*

Wie schon in den letzten Kapiteln erklärt wurde, berechnet die TER nur den Post-Editing-Aufwand und ist daher aus theoretischer Sicht kein Indikator eines guten oder mangelhaften Inhalts des Zieltexts (*adequacy*- AD) und sagt auch nicht aus, ob die maschinelle Übersetzung flüssig lesbar ist (*fluency* - FL). In dieser Meta-Evaluation wird die Korrelation zwischen TER-Werten und *fluency* und *adequacy* analysiert. Da im Kapitel 6.2 hauptsächlich die HTER Werte einbezogen wurden, werden nun die Bewertungen mit den TER-Werten vergleichen. Auf Segmentebene haben die HTER-Werte in der letzten Meta-Evaluation nämlich keine bessere Korrelation mit der Post-Editing-Zeit im Vergleich zu TER gezeigt. Deswegen wurden dieses Mal die TER-Werte als Hauptvergleichsmethode ausgewählt. Darüber hinaus stellt die HTER eine Optimierung der TER im Sinne einer vermutlich besseren Korrelation mit dem Post-Editing-Aufwand dar, weil sie die Edits eines Posteditors so realitätsnah wie möglich darstellen können sollte. Für diese zweite Meta-Evaluation eignet sich daher auch aus methodologischer Sicht die TER besser als die HTER.

In einem ersten Schritt wurden nur die Bewertungen der Annotatoren in Betracht gezogen und die Daten der zwei Ärzte wurden anschließend verglichen, um eine genauere Analyse des Bewertungsprozesses widergeben zu können. Der erste Unterschied zwischen den zwei Annotatoren liegt in der Zahl der aufgezeigten Fehler. Der Annotator A hat 302 Fehler und der Annotator B 220 Fehler gefunden, das heißt -82 Fehler. Aus einer rein statistischen Fehleranalyse könnte man herleiten, dass Annotator B die Übersetzung für besser hält als Annotator A. Die durchgeführte Studie behandelt aber nicht – wir schon oft erklärt wurde – die kategorisierten Fehler als Analyseparameter für die Evaluation des Outputs. Es wird vielmehr versucht zu verstehen, auf welche Fehler die Annotatoren hingewiesen haben und wieso. Es soll hier nochmals in Erinnerung gerufen werden, dass die Annotatoren gebeten wurden, nur jene Fehler bekannt zu geben, die das Verstehen des Segments oder seine inhaltliche Korrektheit beeinträchtigt haben. Diese Aufgabenstellung stützt sich auf die Tatsache, dass einerseits die Ärzte keine linguistische Analyse des Textes durchführen können, weil sie nur das Fachwissen im Bereich Medizin und nicht im Bereich der Sprachwissenschaft besitzen, und andererseits die prozedurale Ebene im Rahmen dieser qualitativen Studie wichtiger ist als die Fehlerkategorisierung selbst. Es wurde nämlich herausgefunden, dass es keine konstante Korrelation zwischen gefundenen Fehlern und

Bewertung des Segmentes gibt. Diese Erkenntnis wird in den nächsten Absätzen ausführlicher erklärt.

Bevor näher auf die Meta-Evaluation auf Segmentebene eingegangen wird, soll hier noch ein interessanter Anhaltspunkt auf Textebene angesprochen werden, der zeigt, dass sich die für die Evaluation ausgewählten Methoden als erfolgreich erwiesen haben. Wie im Kapitel 5.1 erklärt wurde, sind die zwei Kriterien der FL und AD nicht immer klar voneinander abgrenzbar. Die Annotatoren können Schwierigkeiten bei der Zuordnung haben, welcher Fehler zu der jeweils einen oder anderen Kategorie gehört. Aus diesem Grund wurde das Annotation-Tool COSTA ausgewählt, das den Annotatoren bei der Unterscheidung hilft, indem die Referenzübersetzung erst nach der Evaluation der FL gezeigt wird. Diese Methode hat sich als erfolgreich erwiesen, weil die Annotatoren – ohne jeglichen Einfluss aus der inhaltlichen Korrektheit – die FL im Durchschnitt anders die AD bewertet haben. Die *fluency* hat bei Annotator A 67 Mal (von 121) einen höheren Wert als die AD (54/121) und bei Annotator B 73/121 erhalten. Das zeigt nicht nur, dass die zwei Parameter unterschieden werden können, sondern dass, obwohl ein Text flüssig lesbar ist (weniger syntaktische und morphologische Fehler), der Inhalt trotzdem korrekt sein kann, und umgekehrt, der Inhalt eines Text korrekt und verständlich sein kann, obwohl einige Fehler die Flüssigkeit beeinträchtigen. Natürlich ist der Extremfall möglich, indem die Fehler, die zu der FL zuzuordnen sind, so zahlreich sind, dass auch ein korrekter Inhalt nicht verstanden werden kann. Diese Fälle sind aber durch eine Analyse auf Textebene, die die Standard-Evaluationsmethode darstellt, nicht erkennbar. Es sollte vielmehr auf Segmentebene beobachtet werden, wann die Fehler der FL die AD beeinflussen haben können. Diese Fälle, in der die Fehlerart (und nicht nur unbedingt die Anzahl der Fehler) die Verständlichkeit des Satzes beeinflusst, können mit dem Extremfall beim Post-Editing (vgl. Beispiel 13) verglichen werden, in dem die Post-Editoren den MÜ-Output gelöscht haben, um den Satz erneut zu schreiben. Das stellt einen kompletten Misserfolg des Outputs dar, denn es hat den Zweck des Post-Editings nicht erfüllt. In der gleichen Art und Weise kann von Misserfolg des Outputs gesprochen werden, wenn der Zweck des *gisting* nicht erfüllt werden kann. Einige Beispiele des Einflusses von der *fluency* auf die *adequacy* auf Segmentebene werden im weiteren Verlauf gezeigt.

6.3.1 Analyse des Bewertungsprozesses

Um diese Hypothese der fehlenden Korrelation zwischen Fehlerzahl und Bewertung zu überprüfen, wurden die Segmente folgendermaßen analysiert: Es wurden die Segmente herangezogen, die mehr als 2 Fehler erhalten (d.h.- zwischen 3-8[40] Fehler) und mit den Bewertungen vergleichen. Es wurde entschieden, 2 Fehler als maximale Fehleranzahl für eine gute Übersetzung (d.h. mit höheren *fluency* und *adequacy*-Werten) zu nehmen, weil festgestellt werden konnte, dass in besseren Übersetzungen (mit 3- 5 Werte für FL und AD[41]) im Durchschnitt nicht mehr als 2 Fehler vorkamen.

Bei der Bewertung des **Annotators A** wurde nämlich herausgefunden, dass die Segmente mit mehr als 2 Fehlern niedrige Bewertungen von *fluency* und *adequacy* erhalten haben (Werte zwischen 1 und 2). Aus diesen Daten könnte man herleiten, dass zwischen der Anzahl von Fehlern und der Bewertung eine Korrelation besteht. Nur 11/121 (S9, S14, S23, S24, S55, S56, S64, S74, S75, S103, S119[42]) Segmente haben eine Bewertung zwischen 3 und 4 erhalten, allerdings nur bei den *fluency*[43]. Das bedeutet, dass trotz den Fehlern, die Segmente gut lesbar sind.

Da soeben festgestellt wurde, dass – ausgenommen dieser 11/121 Ausnahmen – die Anzahl der Fehler mit der Bewertung korreliert, könnte gleichfalls angenommen werden, dass die Fehler, die vom Annotator A festgestellt wurden, hauptsächlich der Kategorie der AD zuzuordnen sind. Es wurde daher die genaue Fehlerart bei den 11 Segmenten angesehen. Um die Korrelation zwischen Fehler und FL/AD am besten bewerten zu können, wurden die Fehler, die im COSTA zu finden sind, den zwei Kategorien zugeordnet:

Linguistic

- Verb inflection: Incorrectly formed verb, or wrong tense
- Noun inflection: Incorrectly formed noun
- Other inflection: Incorrectly formed adjective or adverb
- Wrong category: Category error (e.g. noun vs. verb)
- Article: Absent or unneeded article.
- Preposition: Incorrect, absent or unneeded preposition
- Agreement: Incorrect agreement between subject-verb, noun-adjective
- Past participle: agreement with preceding direct object, etc.

} *Fluency*

[40] 8 ist die maximale Anzahl von Fehlern, die festgestellt wurde.
[41] Zur Erinnerung: 3. Non-native language. 4. Good language, 5. Flawless language
[42] Der Text und die vollständigen Bewertungen der Segmente werden aus Lesbarkeitsgründen nicht immer angezeigt. Sie können aber im Anhang angesehen und verglichen werden.
[43] Die *adequacy* wurde nur in einem Fall – Segment 64 – mit 3 bewertet. Da es sich nun um einen Einzelfall handelt, wurde dieser nicht in Betracht gezogen.

Words

- Single words: Sentence elements ordered incorrectly
- Multi-word units: Incorrect translation of multi-word expressions and idioms
- Terminology: Incorrect terminology
- Untranslated words: Word not in dictionary
- Ambiguous translation: Ambiguous target language
- Literal translation: Word-for-word translation
- Conjunctions: Failure to reconstruct parallel constituents after conjunction or failure to identify boundaries of conjoined units

Adequacy

Style

- Acronyms - Abbreviations: Incorrect abbreviations, acronyms and symbols
- Extra words: Extra words in target language
- Country standards: Incorrect format of dates, addresses, currency etc.
- Spelling errors: Misspelled words
- Accent: Incorrect accents
- Capitalization: Incorrect upper or lower case
- Punctuation: Punctuation is incorrect, absent or unneeded

Fluency

Natürlich sind die Fehler, die von COSTA vorgeschlagen werden, sehr spezifisch und würden sich gut für eine Evaluation durch Fehleranalyse eignen. In diesem Fall haben die Ärzte – wie voraussehbar – nur von einer niedrigeren Anzahl von Fehlertypen Gebraucht gemacht. Dies war aber genau das Ziel der Evaluation, die darin bestand, zu verstehen, welche Fehler der Durchschnittsnutzer (im Sinne eines Nutzers, der keinen sprachwissenschaftlichen Hintergrund hat) als solche identifiziert.

Überraschenderweise gehören die meisten Fehler innerhalb der o.g. Segmente, die eine gute FL aufweisen, genau zu der Kategorie der FL. Ausgehend davon könnte die Hypothese formuliert werden, dass die Fehler, die den Inhalt (bzw. die *adequacy*) beeinflussen, ein größeres Gewicht haben, als die stilistischen bzw. grammatikalischen oder morphologischen Fehler. Diese Gegensätzlichkeit zwischen Fehlerzahl und Fehlkategorie könnte aber auch davon abhängen, dass für die Bewertung der FL mehrere Fehleroptionen als für die *adequacy* von COSTA angezeigt werden. Das könnte das Verhalten der Annotatoren unbewusst beeinflussen, in dem sie nach Fehlern in den Segmenten suchen, die von dem Tool vorgeschlagen wurden. Um diese zwei Annahmen zu bestätigen oder zu wiederlegen, ist aber die Betrachtung eines jeden Segmenttextes nicht ausreichend. Es nicht nämlich nicht immer möglich, durch eine Ex-post-Analyse den Bewertungsprozess der Annotatoren und den kognitiven Vorgang, der vom Annotator vorgenommen wurde, zu erklären. Denn es ist nicht ersichtlich, welches Wort als welcher Fehler erkannt wird. Um zu dieser Erkenntnis zu gelangen und eine tiefgreifendere Analyse durchführen zu können, sollte man eventuell über

ein Annotations-Tool verfügen, das es ermöglicht, die Fehler direkt im Text zu markieren und die Fehlerkategorie jedes Mal bestimmen zu können.

Bei **Annotator B** wurde dieselbe Fehleranzahl-Analyse durchgeführt und herausgefunden, dass obwohl er weniger Fehler (- 82) im ganzen Text festgestellt hat, 15 Segmente mit mehr als 2 Fehlern und guten *fluency*-Werten (12 Segmente mit *fluency* 3 und 3 Segmente mit *fluency* 4) gefunden wurden. Die *adequacy* hat dagegen in diesen 15 Fällen nicht den Wert 2 überschritten. Auch bei der Evaluation des Annotators B gehören die meisten Fehler zur Kategorie der *fluency*.

Es ist außerdem festzustellen dass es einige Segmente gibt, bei denen beide Annotatoren keine Fehler aufgezeigt haben – nämlich 18 Segmente bei Annotator A und 37 bei Annotator B. Wenn keine Fehler im Segment festzustellen sind, sollten theoretisch die Bewertungen bei den beiden Kriterien Werte von 4 bis 5 aufweisen können. Anders als gedacht zeigen einige dieser Segmente schlechtere Bewertungen:

Annotator A – Segmente mit Null Fehler **(Beispiel 15)**				
Segment-Nummer	Hypothese	Referenz	FL	AD
20	0 12) polipeptide secondo la rivendicazione 1, caratterizzato dal fatto che almeno un residuo di arginina è presente come residuo citrullina.	12) Polipeptide secondo la rivendicazione 1, caratterizzato dal fatto che almeno un residuo di arginina è presente sotto forma di residuo di citrullina.	4	3
91	Esempi preferibili di muteine di vimentina umana hanno una sequenza con la SEQ ID 2, 3, 4, 5, 6, 7, 8 o 9	Esempi preferibili per muteine della viementina umana hanno una sequenza con SEQ ID No:2, 3, 4, 5, 6, 7, 8 o 9.	3	3
95	Un esempio preferito di un frammento è il frammento 51-65 (C2).	Un esempio preferibile di frammento è il 51-65 (C2).	2	2
105	L'agente di diagnostica, il polipeptide o un frammento incluso in forma libera o carrier-bound.	Tale agente può comprendere un polipeptide o frammento in forma libera oppure legata al carrier.	2	1
96	La lunghezza del frammento è preferibilmente almeno 6, più preferibilmente almeno 8 amminoacidi fino a 120, preferibilmente fino a 100 e più preferibilmente fino a 50 aminoacidi.	La lunghezza del frammento è preferibilmente di almeno 6, ancor meglio almeno 8 amminoacidi fino a 120, preferibilmente fino a 100 e meglio se fino a 50 amminoacidi.	2	1

Annotator B – Segmente ohne Fehler **(Beispiel 16)**

Segment-Nummer	Hypothese	Referenz	FL	AD
58	Gli anticorpi sono indicati per questa ragione oggi come gli anticorpi anti-cheratina (AKA).	Per questo motivo, questi anticorpi sono stati denominati anticorpi anti-cheratina (AKA).	3	2
35	Inoltre, l'invenzione riguarda una composizione farmaceutica contenente uno dei polipeptidi, nonché l'uso dei polipeptidi per la preparazione di un medicamento per la profilassi e / o il trattamento di malattie reumatiche.	Inoltre, l'invenzione riguarda un farmaco che contiene uno dei polipeptidi e l'utilizzo dei polipeptidi nella creazione di un farmaco per la profilassi e/o il trattamento di malattie reumatiche.	3	2
99	Altri esempi di derivati peptidici sono pagina gruppo amminico e / o carbossilici modificati polipeptidi di un gruppo amminico, per esempio, Polipeptidi, ad esempio sono modificato con un acido carbossilico o un radicale alchilico, modificato o un gruppo di acido carbossilico con un gruppo amminico oppure un gruppo estereo.	Altri esempi di derivati petidici sono i polipeptidi di un gruppo amminico modificati nei gruppi laterali, domini N- e/o C-terminali, come ad esempio i polipeptidi che, per esempio, sono modificati con un acido carbossilico o un residuo alchilico o i polipeptidi modificati nel gruppo di acidi carbossilici con un gruppo amminico o un gruppo di esteri.	1	1
63	Anticorpi anti-filaggrina dei filtri di tipo IgG con una specificità superiore al 99% un marcatore altamente specifico per l'artrite reumatoide.	Gli anticorpi anti-filaggrina di tipo IgG rappresentano, con una specificità di oltre 99%, un marker ad elevata specificità per l'artrite reumatoide.	1	1

Die Bewertung dieser Segmente sind Beispiele mangelhafter Intra-Annotator-Übereinstimmung im Sinne einer geringeren Korrelation zwischen Fehler und Bewertung. Diese Unterschiede sagen aber nichts über die Inter-Annotator-Übereinstimmung bei der Bewertungen aus. Diese mangelhafte Intra-Annotator-Übereinstimmung Bewertung/Fehler erschwert das Begründen bzw. das Verstehen der Faktoren, die zu einer schlechten Bewertung geführt haben.

Die interessanteste Erkenntnis kommt aber aus den Kommentaren, die die Annotatoren geschrieben haben. Im Folgenden werden einige ausgewählte Beispiele aufgezeigt:

Annotator A (Beispiel 17)					
Segment-Nr.	Hypothese	Kommentar	FL	AD	Fehler
38	L'artrite reumatoide è una Autoimmunkrankeit in cui mantenere i	Assenza del verbo che	1	1	5

	meccanismi di difesa del corpo umano erroneamente endogena cartilagine articolare per estranea e ostile e li attaccano.	spiega il senso della frase			
44	Tuttavia, ad oggi dalla tecnica anteriore, senza rilevamento affidabile e sensibile di artrite reumatoide dal periodo di tempo noto.	Assenza del verbo	1	1	1
69	Scopo della presente invenzione fornire nuovi polipeptidi per la rilevazione di associati con malattie reumatiche, artrite reumatoide particolare associato per fornire anticorpi che un sensibile e specifica diagnosi, la classificazione e la prognosi di malattie reumatiche, in particolare di dolori articolari permettere e il sistema muscolo-scheletrico.	Errata costruzione della frase	2	1	6

Annotator B (Beispiel 18)					
Segment-Nummer	Hypothese	Kommentar	FL	AD	Fehler
37	Un test di laboratorio che permette questo dolore una tensione muscolare innocuo, l'artrosi o il più comune e più grave delle malattie, l'artrite reumatoide (RA) a te, non è ancora noto.	Manca il verbo. Elenco di parole senza senso.	1	1	2
38	L'artrite reumatoide è una Autoimmunkrankeit in cui mantenere i meccanismi di difesa del corpo umano erroneamente endogena cartilagine articolare per estranea e ostile e li attaccano.	Assoluta mancanza di nesso logico	1	1	2
44	Tuttavia, ad oggi dalla tecnica anteriore, senza rilevamento affidabile e sensibile di artrite reumatoide dal periodo di tempo noto.	Manca il verbo. Parole inserite a caso	1	1	0
46	Secondo i criteri della ACR del fattore reumatoide dell'indicatore sierologico precedenza base per la diagnosi di artrite reumatoide (RA).	Manca il verbo.	1	1	0
105	L'agente di diagnostica, il polipeptide o un	Manca il verbo, la frase	1	1	0

	frammento incluso in forma libera o carrier-bound.	non ha senso			

Aus den Beispielen 17 und 18 wird ersichtlich, dass einige Segmente niedrige Bewertungen bekommen haben und teilweise auch weniger Fehler aufweisen, schwerwiegende Mangel hatten, die die Bedeutung bzw. das Verstehen des Satzes so gestört haben, dass die Annotatoren diese hervorheben wollten. Es soll hier in Erinnerung gerufen werden, dass den Annotatoren die Entscheidung überlassen wurde, was sie in den Kommentaren schreiben wollen bzw. es wurden dazu keine Richtlinien vorgegeben. Überraschenderweise haben die Kommentare dieselben Fehlerarten bzw. Störfaktoren gezeigt. Wenn die Kommentare der Annotatoren in eine „sprachwissenschaftliche Sprache“ übersetzt werden, kann daraus entnommen werden, dass die meisten Kommentare der beiden Annotatoren zu zwei Kategorien gehören: fehlendes Verb und falsche Satzstellung. Beide Fehler beeinflussen stärker als andere die Bewertung des Segments, indem sie die logische Verbindung zwischen den Wörtern beeinträchtigen. Das bestätigt nochmal die Annahme, gemäß welcher unterschiedliche Fehler unterschiedliche Einflüsse auf das Verstehen des Satzes haben, und das spiegelt sich in der Bewertung wider. Einige dieser Fehler, wie beispielweise ein fehlendes Verb, könnten die einzigen Fehler in einem Satz sein und trotzdem das Verstehen des Satzes so stark beeinflussen, dass das Segment mit 1 bewertet wird.
Aus diesem ersten Teil der Analyse können folgende Schlussfolgerungen abgeleitet werden:

1. Zu viele Fehlerkategorien können die Annotatoren überfordern und den Bewertungsprozess der Annotatoren beeinflussen.

2. Die Anzahl der Fehler und die Bewertungen korrelieren nicht immer, weil einerseits oft mehrere Fehler auf ein einziges Wort zurückführen sind (zum Beispiel ein Wort könnte einen terminologischen und einen oder mehrere grammatikalische Fehler enthalten) und andererseits einige Fehler – wie beispielweise ein fehlendes Verbs und Satzstellung – stärkere Störfaktoren als andere darstellen.

3. Die Fehler, die bei jedem Segment im Raster signalisiert werden, können nicht immer eindeutig auf die einzelnen Wörter zurückgeführt werden. Die Fehler sind daher kein zuverlässiges Mittel zur Erforschung des kognitiven Evaluationsprozesses der Ärzte gewesen. Um zu dieser Erkenntnis zu gelangen und eine tiefgreifendere Analyse durchführen zu können, sollte man eventuell über ein Annotations-Tool verfügen, das es ermöglicht, die Fehler direkt im Text zu markieren und die Fehlerkategorie im Einzelfall bestimmen zu können.

6.3.2 Korrelation TER – *fluency*-und *adequacy*-Werte

Die Korrelation zwischen TER-Werten und der Bewertung von *adequacy* und *fluency* wird auch qualitativ erforscht. Da es sich aber nur um numerische Werte handelte, wurde es erforderlich, ein numerisches System zu finden, um den Vergleich und die Analyse der zwei Skalen zu ermöglichen. Da die Skala der FL und AD 5 Werte enthält, wurden die TER-Werte (0-110%) in 5 Intervalle eingeteilt:

TER-Intervalle:

1. 0% – 20%
2. 21% – 40%
3. 41% – 60%
4. 61% – 80%
5. 81%– 100%

Natürlich wird nicht erwartet, dass es eine strenge und direkte Korrelation zwischen beispielweise dem TER-Intervall 1 (beste TER-Werte) und einem AD-und-FL-Wert von 5 (beste Bewertung) gibt. Die (Toleranz)grenzen zwischen den Intervallen sind fließend. Diese Einteilung wird ausschließlich dafür verwendet, um die Analyse übersichtlich und nachvollziehbar zu gestalten. Pro Intervall wurden die Bewertungen beider Annotatoren miteinbezogen.

1. TER-Werte von 0% bis 20%

Zum ersten Intervall gehören nur 5/121 Segmente. Diese niedrige Zahl ist nicht überraschend, weil im Kapitel 5.3 erklärt wurde, dass selbst die kleinsten Unterschiede zwischen Hypothese und Referenz die Berechnung des Algorithmus stark beeinflussen können. Die TER setzt daher in diesem Intervall einen Output voraus, der im Sinne des Ähnlichkeitsgrades zur Referenzübersetzung beinahe perfekt ist. Es wird daher in diesem Intervall erwartet, dass die Segmente mit guten bis sehr guten Werten (von 4 bis 5) von den Annotatoren bewertet werden.
Dieses Intervall enthält 5 Segmente und 4 davon haben Werte von 4 bis 5. Nur das Segment Nr. 11 hat einen Wert von 4 erhalten. Dass das Segment-Nr. 11 nicht mit den TER-Werten korreliert, wird von beiden Annotatoren bestätigt. Annotator A weist für dieses Segment FL 2 und AD 2 zu, erkennt aber nur ein Fehler (Präposition). Annotator B weist 4 Punkte der FL und 3 der AD zu und findet zwei Fehler (Präposition und Terminologie).

Hypothese:

3) Polipeptide secondo la rivendicazione 2, caratterizzato dal fatto che presenta un residuo di arginina supplementare ad almeno due delle posizioni.

Referenz:
3) Polipeptide secondo la rivendicazione 2, caratterizzato dal fatto che presenta un residuo di arginina in più in almeno due delle posizioni.

(Beispiel 19: Segment 11**)**

Die zwei Segmente zeigen aber keine besonderen Unterschiede außer der falschen Präposition (*ad* statt *in*), insbesondere keine Fehler, die die Bewertung des Annotators A rechtfertigen könnten. Da die Gründe für solch eine strenge Bewertung im Übersetzungstext nicht zu finden sind, könnte man diese als Begründung ableiten, dass das Segment am Anfang des Textes bzw. der Evaluation stand und die Intra-Übereinstimmung beider Annotatoren noch keine Konstanz hatte. Das heißt, sie hatten am Anfang des Textes noch zu wenige Sätze evaluiert, um konsistent einschätzen zu können, was sie für eine gute oder schlechte Übersetzung hielten. Aus diesem Beispiel kann man erschließen, dass die Subjektivität der Evaluationen, die sich in diesem Beispiel als Extremfall zeigt, es oft auf Segmentebene unmöglich macht, eine Korrelation zu finden.

2. TER-Werte von 21% bis 40%

Ab dem zweiten Intervall variieren die Bewertungen von FL und AD stärker als in dem ersten Intervall. Dieses Ergebnis war aber voraussehbar, denn in dieser mittleren Schicht könnten die Fehler von jeder Art sein und die Bewertungen unterschiedlich beeinflussen. Obwohl die TER voraussetzt, dass eine Korrelation zwischen besseren TER-Werten und besserer Bewertung besteht, sind die Daten in diesem Intervall so unterschiedlich, dass das empirische Erkennen eines Pfades (der dann auch statistisch erforscht bzw. überprüft werden kann) nicht möglich ist. Es wurden daher Segmente mit ähnlichen Werten verglichen und die dazugehörenden Ausnahmenfälle festgestellt.

(Beispiel 20)			**Annotator A**			**Annotator B**		
Seg.-Nr.	Hypothese	TER	FL	AD	Fehler	FL	AD	Fehler
88	In un'altra forma di realizzazione, il polipeptide si trova almeno un residuo di arginina come residuo citrullina prima, ad esempio in almeno una delle posizioni 4, 12, 23, 28, 36, 45, 50, 64, 71, 100, 320, 364 o 378.	28%	**1**	**1**	5	**1**	**1**	4
30	DESCRIZIONE (testo OCR potrebbe contenere errori).	29%	4	**5**	0	5	**5**	0
31	Un metodo per la diagnosi di malattie reumatiche.	29%	4	**4**	1	5	**4**	0

66	Gli anticorpi anti-filaggrina non sono correlati con l'età, il sesso o la durata della malattia.	29%	**4**	**5**	0	**4**	**5**	0
87	Per esempio, il polipeptide ha almeno uno, due, tre o quattro posizioni su un residuo di tirosina.	29%	**2**	1	4	**2**	2	2
39	Circa 1 persona su 100 soffre nei paesi dell'Europa occidentale per l'artrite reumatoide.	31%	**4**	**3**	1	**4**	**3**	0

Aus diesem Ausschnitt der Evaluation kann festgestellt werden, dass obwohl die Segmente kleine bis keine Unterschiede in den TER-Werten aufzeigen, hingegen die Bewertungen der FL und AD stark schwanken können. Der Übergang von Segment 88 zu 30 und von 67 zum 87 verdeutlicht ganz klar die fehlende Konsequenz. Aus diesen Segmentreihen wird darüber hinaus besonders ersichtlich, dass die Evaluationen der zwei Annotatoren sehr gut miteinander korrelieren[44].

Ab einem TER-Wert von ca. 35% fangen die Bewertungen an, stabiler zu werden, indem sie nur selten 3 Punkte überschreiten und außerdem nur in der FL. Zwischen 39% und 40% sind die Bewertungen wieder sehr unterschiedlich. Wie schon festgestellt wurde, ist somit kein konstantes indirekt proportionales Verhältnis zwischen den TER-Werten und Bewertungen in diesem Intervall festzustellen.

3. TER-Werte von 41% bis 60%

In diesem Intervall beginnen die Werte leicht zu sinken. Bei Annotator A sind die Bewertungen der *fluency* im Durchschnitt leicht höher als jene der AD (in 16 von 31 Segmenten ist die FL ≥3). Die *adequacy* bleibt im Schnitt zwischen 1 und 2, mit einigen Ausnahmen: In insgesamt 31 Segmenten in diesem Intervall ist die AD 3 und in 2 Segmenten 4. Obwohl die Werte insgesamt beginnen zu sinken, gibt es weiterhin zahlreiche Ausnahmen auf der Segmentebene.

(Beispiel 21)			Annotator A		Annotator B	
Seg.-Nr.	Hypothese	TER	FL	AD	FL	AD
29	5 21) Un kit diagnostico secondo la rivendicazione 19 o 20, caratterizzato dal fatto che il supporto è DNA,	46%	**1**	**1**	**1**	**1**

[44] Die Werte, die bei Annotator A und B gleich sind, sind fett formatiert.

	RNA, medicamente polimeri accettabili, biopolimeri o snythetische proteine.					
45	La diagnosi di artrite reumatoide si basa sui criteri di classificazione del ACR (American College of Rheumatology).	47%	**4**	**4**	**4**	**4**
73	Queste varianti mutanti di vimentina nativo differiscono da vimentina nativo per la presenza di residui di arginina aggiuntivi ed eventualmente altri differenze di sequenza.	47%	2	**2**	3	**2**
85	Per esempio, il polipeptide ha almeno uno, due o tre posizioni su un residuo di treonina.	47%	**2**	**2**	**2**	**2**
25	17) peptide Derivato secondo la rivendicazione 16, caratterizzato dal fatto di essere scelto fra Retro / o inverso polipeptidi e peptidi ciclici.	48%	2	1	3	2
36	Le malattie reumatiche, in particolare dolori alle articolazioni e il sistema muscolo-scheletrico sono tra le più comuni malattie in Germania.	48%	**4**	**3**	**4**	**3**

Auch dieses Beispiel zeigt eine starke Inter-Annotator-Übereinstimmung. Und genau wie im zweiten Intervall können die Bewertungen steigen und dann plötzlich sinken. Aus diesen und auch aus anderen Segmente kann noch etwas festgestellt werden: Die Segmente, deren Inhalt allgemeiner ist als die anderen, haben bessere Bewertungen bekommen. Das könnte von einer Kombination aus zwei Faktoren abhängen: Die Performance von Google Translate bei allgemeinen Texten besser als bei sehr fachlichen Texte und die allgemeinen Segmente, deren Inhalt einfacher ist und auch explizit formuliert ist, sind einfacher zu verstehen, auch in dem Fall, in welchem einige Fehler die Bedeutung leicht beeinflussen.

4. TER-Werte von 61% bis 80%

In diesem Intervall sind auch die Bewertungen sehr unterschiedlich. Es wird zum Beispiel deutlich, dass viele Segmente B (Segment-Nr. 56, 67, 140, 43, 101), die einen 64%er Wert haben, vom Annotator Bewertungen von 4 (FL) und 3 (AD) erhalten haben. Danach ist ein Abfall bis auf 2-1 am Segment-Nr. 69 zu beobachten, dem ein Wert von 65% zugewiesen wurde. Die Bewertungen des Annotators A entsprechen (aber nicht völlig) jenen des Annotators B. Die Bewertungen, in dem keine Inter-Annotator-Überseinstimmung erkennbar ist, sind die Bewertungen des Segments 43.

(Beispiel 22)	Annotator A	Annotator B

Seg.-Nr.	Hypothese	Referenz	TER	FL	AD	FL	AD
43	Reumatologi tenta di utilizzare la stretta finestra di tempo tra l'insorgenza della malattia e la comparsa di danno strutturale.	I reumatologi cercano di utilizzare al meglio il breve lasso di tempo che intercorre tra l'inizio della malattia e l'insorgere di danni strutturali alle articolazioni.	64%	**2**	**1**	**4**	**3**

Es ist nicht immer einfach, die Bewertungen bzw. die Bewertungsunterschiede zu begründen – wie bereits auf den vorangehenden Seiten festgestellt werden konnte. Das Segment-Nr. 43 ist nämlich nicht besonders gut lesbar, aber fast alle Informationen sind nach Annotator B gegeben. Wenn der fiktive Zweck dieses MÜ-Outputs berücksichtigt wird – d.h. das *gisting* –, kann gesagt werden, dass für den Annotator B das Ziel erfüllt wurde (mit einem Wert von 3 – das Segment liefert also genügende Informationen), während Annotator A das Segment als unverständlich und mit mangelnden Informationen beschreibt.

(Beispiel 23)				Annotator A		Annotator B	
Seg.-Nr.	Hypothese	Referenz	TER	FL	AD	FL	AD
60	Così, la filaggrina proteina basica è stata identificata come l'antigene bersaglio.	Ciò ha portato ad identificare la proteina basica fillagrina come antigene bersaglio.	75%	**4**	**4**	**4**	**4**

In dem Beispiel 23 wird dagegen ersichtlich, dass beide Annotatoren genau dieselben Bewertungen gegeben haben. Die TER-Werte sind aber hoch (75%) und sollten daher auf einen schlechten Output hinweisen.

5. TER-Werte von 81% bis 100%

Dieses Intervall enthält 13 Segmente. In diesem Intervall wird ein Pfad besonders ersichtlich, der sich auch in anderen Segmenten durch den ganzen Text gezogen hat. Die längeren Segmente, denen einen höherer TER-Wert zugewiesen wurde, sind mit entsprechenden niedrigen Bewertungen bewertet worden. Aber die kürzeren Segmente haben dagegen höhere Bewertungen bekommen.

(Beispiel 24)	Annotator A	Annotator B

Seg.-Nr.	Hypothese	Referenz	TER	FL	AD	FL	AD
53	Altri marcatori sierologici, come gli anticorpi anti-citrullina (PCC) o il punteggio HAQ iniziale, utilizzato per valutare le abilità nella vita quotidiana, o la radiografia o la tomografia computerizzata (CT) -BiId danno in forma in anticipo solo piccoli affioramenti e sono non solo abbastanza significativo al fine di valutare in che modo la prognosi del paziente sarà.	Altri marker sierologici, come l'anticorpo anti-citrullina (CCP) o il punteggio iniziale del questionario HAQ, che serve a valutare la capacità di svolgere le attività quotidiane, oppure le immagini della Tomografia Computerizzata (TC) nello stadio iniziale della malattia forniscono informazioni limitate che da sole non sono abbastanza esaustive da poter formulare una prognosi del paziente.	81%	3	1	1	1
109	Quindi o, dal momento che i polipeptidi contenenti più siti di legame dell'anticorpo, entrambi gli anticorpi monomerici e multimeriche vengono efficacemente legati.	Dal momento che i polipeptidi hanno diversi punti di interazione agli anticorpi, possono essere legati in modo efficiente sia anticorpi in forma di monomeri che di multimeri.	81%	1	1	2	1
64	Gli anticorpi sono in linea di principio rilevabili precoce e precedono la malattia clinica.	Gli anticorpi, in linea di principio, possono essere identificati in fase iniziale e riescono a suggerire prima di altri un quadro clinico.	82%	4	3	4	2
49	Generalmente, livelli elevati di fattori reumatoidi sono associati con una malattia più grave.	In generale, un innalzamento della concentrazione del fattore reumatoide viene associato con un grave decorso della malattia.	94%	4	3	4	3
3	**SINTESI.**	**ESTRATTO.**	**100%**	**5**	**4**	**4**	**5**
8	**I reclami.**	**Rivendicazioni.**	**200%**	**4**	**4**	**5**	**3**

Der Grund für diesen extremen Unterschied wurde schon im Kapitel 5.3 erklärt. Dabei wurde die Situation erläutert, in welcher kürzere Hypothese-Segmente bessere TER-Werte als die von den Annotatoren geänderten Referenzen (TARG) erhalten haben. Es konnte daher eine paradoxe Situation bei der Erstellung der TARG beobachtet werden, die Annotatoren

Änderungen durchführen, welche zwar die TER-Werte reduzieren, aber möglicherweise nicht den tatsächlichen Änderungen, die ein Post-Editor vornehmen könnte, entsprechen.

Dieses Verhalten von TER mit kürzen Segmente kann auch hier beobachten werden. Eine ähnliche paradoxe Situation stellen nämlich die Segmente 3 und 8 dar. Diese sind eigentlich sehr gut bewertet worden (4 bis 5), aber die TER-Werte sind jeweils 100% und 200%. Dieser letzte Wert sollte eigentlich gar nicht vorkommen, zumal die TER-Werte von 0 bis 100% gehen. Die Berechnung des Segments 8 lässt sich aber folgendermaßen erklären: In der Hypothese gibt es zwei Edits, die Referenz enthält aber nur ein Wort (2/1=2). Wenn viele solcher Fälle in einem Text vorkommen, könnte die gesamte TER-Evaluation gefälscht werden.

6.4 Feedback der Annotatoren

Am Ende der Evaluation wurde ein kurzes Gespräch mit den Teilnehmern durchgeführt, in dem sie gefragt wurden, ob der Output für die jeweiligen Zwecke geeignet war bzw. ob er die Qualitätskriterien erfüllt hat. In der Studie wurden die Vorgehensweise und die Bewertungen der Annotatoren unter die Lupe genommen und unterschiedliche Hypothesen über ihren kognitiven Aufwand und die Gründe ihrer Bewertungen formuliert. Dieser letzte Schritt der Meta-Evaluation dient dazu, den Annotatoren eine Stimme zu geben.

Post-Editor A ist der Ansicht, dass die MÜ teilweise gute Entwürfe erstellt hat, aber die Segmente oft so viele Fehler enthielten, dass es aufwändiger war, den Satz zu korrigieren, als diesen erneut zu schreiben bzw. zu übersetzen. Post-Editor A hebt auch ein wichtiges Thema hervor: das Vertrauen. Der Text, der übersetzt wurde, war sehr fachlich und, obwohl ein Glossar vorhanden war, war eine lange Recherchearbeit notwendig. Aber auch im Fall, in dem einige Termini potentiell korrekt maschinell übersetzt wurden, konnte man eine Recherche nicht ausschließen, um die Termini zu überprüfen. Man kann sich auf die Übersetzung von Google Translate nicht einfach verlassen. Da die Recherche und Überprüfungsarbeit so zeitaufwendig ist, würde eine menschliche Übersetzung vermutlich dieselbe Zeit beanspruchen.

Post-Editor B gibt zu, dass er nicht so viel Recherchearbeit betrieben hat. Bei einigen Segmenten wäre sie wirklich notwendig gewesen, weil man nie weiß, ob die von Google vorgeschlagenen Termini korrekt sind und ohne Kontrolle übernommen werden können. Er hat dem MÜ-Output einfach vertraut, weil es sonst zeitaufwendig gewesen wäre. Er arbeitet als In-House Übersetzer in einem Übersetzungsbüro und, obwohl Google Translate einige Segmente wirklich gut übersetzt hat, wäre bei dem Stand der Dinge ein Post-Editing-System

in einem Übersetzungsbüro nicht umsetzbar. Die Qualitätsansprüche der Übersetzungen in einem Übersetzungsbüro sind so hoch, dass die nachbearbeitete Übersetzung nie für einen realen Übersetzungsauftrag verwendet werden könnte.

Beide **Annotatoren** sind der Meinung, dass das Output so viele unverständliche Übersetzungen beinhaltete, dass es ohne Referenzübersetzung kaum möglich gewesen wäre, den Text zu verstehen und die richtigen Informationen zu finden.

7. Fazit

Die in der vorliegenden Masterarbeit durchgeführte Studie, welche die Korrelation auf Segmentebene zwischen automatischen und menschlichen Evaluationsmethoden untersuchte, kommt zu folgenden Ergebnissen:

- Die von TER identifizierten Fehler, die die Basis für die Berechnung bilden, entsprechen auch bei einfachen Sätzen, in denen nur 1-2 Fehler zu finden sind, nicht unbedingt der Anzahl und der Art der von einem Post-Editor identifizierten Fehler.
- Die TER-Werte der MÜ-Segmente, die HTER- Werte und die TER-Werte der von den Post-Editoren nachbearbeiteten Segmente weisen große Unterschiede auf. Die HTER-Werte korrelieren nicht mit den TER-Werten der nachbearbeiteten Segmente. Es gibt auch Extremfälle, in denen dem Segment ein HTER von 0% zugewiesen wurde und von Post-Editoren hingegen stark verändert wurde (ab TER von 40/50%).
- Oft korrelieren TER-Werte der MÜ-Segmente mit den TER-Werten der nachbearbeiteten Segmente besser als HTER.
- Obwohl höhere TER-Werte niedrigerer Post-Editing-Zeit und einem geringeren Prozentsatz des korrigierten Segments entsprechen sollten, haben Post-Editoren oft weniger Zeit für die Nachbearbeitung des Satzes mit höheren TER-Werten aufgewandt.
- TER kann in der Berechnung nicht folgende Faktoren berücksichtigen, welche die Post-Editing-Zeit beeinflussen: (a) Reduktion der PE-Zeit wegen des Nachbearbeitens ähnlicher Segmente oder des Korrigierens wiederholter Fehler; (b) einen höheren Zeitaufwand, der nicht von der Anzahl der Änderungen abhängig ist, sondern vom kognitiven Aufwand, den einige Fehler verursacht haben; (c) einen höheren Zeitaufwand, der nicht von der Anzahl der Änderungen abhängig ist, sondern von der Recherche einzelner Wörter.
- Auch die Endbenutzer bestätigen, dass nur wenige Fehler den Satz unverständlich machen können. Einigen dieser Sätze wurden aber höhere TER-Werte zugewiesen.
- In den meisten TER-Wert-Intervallen, die der Meta-Evaluation der Korrelation zwischen TER-Werten und Endbenutzer-Bewertungen gedient haben, konnte kein konstantes indirekt proportionales Verhältnis zwischen den TER-Werten und Bewertungen festgestellt werden.

  Aus den Ergebnissen dieser detaillierten Analyse der Korrelation zwischen TER-Werten und der menschlichen Evaluationsmethode auf Segmentebene konnte

festgestellt werden, dass die Korrelation nicht vorhanden ist und aus den oben ausgeführten Gründen nicht vorhanden sein kann. Es kann durchaus möglich sein, dass die Werte auf Textebene (die dann normalisiert werden) eine Korrelation zeigen können. Die Ergebnisse dieser Arbeit bestärken die zu Beginn aufgestellte Hypothese bestärkt, dass die Korrelation der Werte nur teilweise bzw. zufallsmäßig erfolgen kann, zumal die Variablen, die TER nicht berücksichtigt bzw. berücksichtigen kann, zu viele sind, um eine treffende Aussage über eine mutmaßliche Korrelation machen zu können.

Selbst eine objektive Bemessung des Post-Editing-Aufwands lässt sich schwer durchführen. Eine Metrik, die diesen Aufwand berechnen kann, ist daher bis heute nicht denkbar. Mit der Benutzung der neusten Technologien, welche die Gründe des kognitiven Aufwandes (wie beispielweise das Eye-Tracking) berechnen, könnte man vielleicht zu objektiveren Ergebnissen kommen, die dann auch als Parameter für die analysierten Metriken dienen bzw. in die Metriken integriert werden können (beispielweise in Form von unterschiedlichen Gewichtungsfaktoren der Fehler).

Bibliographie

ACL, Association for Computational Linguistics. In: http://aclweb.org/, Stand 14.08.2016.

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

ALPAC (Automatic Language Processing Advisory Committee). 1966. Report of the ALPAC; Language and Machines: Computers in Translation and Linguistics. Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, D.C.

Arnold, Doug. 2003. Why *Translation is difficoult for computers*. In: Somers, H. L. (Hg.), *Computers and Translation: A Translator's Guide*. Amsterdam/Philadelphia: John Bejamins.

Arnold, Doug/ Balkan, Lorn/Humphreys, R.Lee/Meijer, Siety/Sadler, Louisa. 1994. *Machine Translation: An Introductory Guide*. Oxford: NCC Blackwell.

Banerjee, Satanjeev/Lavie, Alon. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, 65-72.

Bar-Hilel, Yehôšuaʿ. 1959. *Report on the state of machine translation in the United States and Great Britain*. Technical Report No.1. In: http://www.mt-archive.info/Bar-Hillel-1959.pdf, Stand: 14.08.2016.

Bar-Hilel, Yehôšuaʿ. 1960. The Present Status of Automatic Translation of Languages. In: *Advances in Computers*, 1960:1, 158-163.

Bennet, Winfield. S./Slocum, Jonathan. 1988. The LRC machine translation system. In: Slocum (Hg.), *Machine Translation Systems. Studies in Natural Language Processing*. Cambridge: Cambridge University Press.

Boitet, Christian/ Blanchon, Hervé/Seligman, Mark/Bellynck, Valérie. 2009. Evolution of MT with the Web. In: *Proceedings of the Conference Machine Translation 25 Years On, 1-13*. Cranfield: Bedfordshire.

Brown, Peter F./Della Pietra Stephen A./Della Pietra Vincent J./Mercer Robert L. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation. Association for Computational Linguistics*. In: http://www.aclweb.org/anthology/J93-2003, Stand 14.08.2016.

Callison-Burch, Chris/ Fordyce, Cameron/ Koehn, Philipp/Monz, Christof /Schroeder, Josh. 2007. (Meta-)Evaluation of Machine Translation. In: *Proceeding of the Second Workshop on Statistical Machine Translation*, 136-158.

Canelli, Maria/Grasso, Daniele/King, Maghi. 2000. Methods and Metrics for the Evaluation of Dictation Systems: A Case Study. In: *Proceedings of the 2nd LREC*, Athens, Greece.

Canepari, Michela. 2013. *Viaggio intersemiotico nel linguaggio della scienza. Band 1: Prospettive e teorie*. Roma: Nuova Cultura, Roma.

Carroll, John B. 1966. *An experiment in evaluating the quality of translations. Washington DC, National Academy of Science, ALPAC Report*, Anhänge 10 und 11, 67-78.

Cartoni, Bruno / Gesmundo, Andrea / Henderson, James/ Grisot, Cristina/ Merlo, Paola / Meyer, Thomas/Moeschler, Jacques/ Zufferey, Sandrine/ Popescu Belis, Andrei. 2011. *Improving MT coherence through textlevel processing of input texts*: *The COMTIS project.* In Tralogy I, Session 6, http://lodel.irevues.inist.fr/tralogy/index.php?id=78, Stand: 14.08.2016.

Chan, Yee Seng/Tou Ng, Hwee/Chiang, David. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, June 2007, 33-40.

Chatzitheodorou, Konstantinos/Chatzistamatis, Stamatis. 2013. COSTA MT Evaluation Tool: An Open Toolkit for Human Machine Translation Evaluation. In: *The Prague Bulletin of Mathematical Linguistics*, 2013:100, 83-89.

Chomsky, Noam. 1969. Quine's empirical assumptions. In: Davidson, Donald/Hintikka, Jaakko (Hg.), *Words and objections. Essays on the work of W. V. Quine.* Dordrecht: Reidel.

COSTA MT Evaluation Tool. In: https://code.google.com/archive/p/costa-mt-evaluation-tool/, Stand 14.08.2016.

Coughlin, Deborah. 2003. Correlating Automated and Human Assessments of Machine Translation Quality. In: *Proceedings of Machine Translation Summit IX: Machine Translation for Semitic Languages: Issues and Approaches, 23-27 September 2003, New Orleans, LA*, 63-70.

Culy, Christopher/Riehemann, Susanne Z. 2003. The Limits of N-Gram Translation Evaluation Metrics. In: *In Proceedings of the Ninth Machine Translation Summit*. New Orleans, Louisiana, USA.

Denkowski, Michael/Lavie, Alon. 2010. Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks. In: *Proceedings of the Ninth Biennial Conference of the Association for Machine Translation in the Americas*, Denver, Colorado.

Doherty, Stephen/O'Brien, Sharon/Carl, Michael. 2010. Eye tracking as an MT evaluation technique. In*: Machine Translation*, 2010:24(1), 1-13.

Dorr, Bonnie/Olive, Joseph/McCary, John/Christianson, Caitlin. 2011. Machine Translation Evaluation and Optimization. In: Olive, Joseph/McCary, John/Christianson, Caitlin. (Hg.), *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. London: Springer-Verlag.

EAGLES Evaluation Working Group. 1996. *EAGLES Evaluation of Natural Language Processing Systems. Final Report*, Center for Sprogteknologi, Copenhagen, Danemark.

Enkvist, Nils Erik. 1987. More about Text Strategies. In: Lörscher, W./Schulze R. (Hg.), *Perspectives on Language in Performance. To Honour Werner Hüllen on the Occasion of his Sixthieth Birthday*. Tübingen: Gunter Narr Verlag, 337-350.

Estelle, Josh/Khare, Rohit. Found in Translation. Going Global with the Translate API. In: http://goo.gl/63LviU, Stand: 14.08.2016.

Estrella, Paula/Popescu-Belis, Andrei/ Underwood, Nancy. 2005. Finding the System that Suits you Best: Towards the Normalization of MT Evaluation. In: *27th ASLIB International Conference on Translating and the Computer, 24-25 November 2005, London.*

Estrella, Paula/Popescu-Belis, Andrei/King, Maghi. 2008. *Improving quality models for MT evaluation based on evaluators' feedback*. In: http://www.lrec-conf.org/proceedings/lrec2008/pdf/236_paper.pdf, Stand: 14.08.2016, Stand 14.08.2016.

Estrella, Paula/Popescu-Belis, Andrei/King, Maghi. 2009. *The FEMTI guidelines for contextual MT evaluation: principles and resources*. In: http://andreipb.free.fr/textes/-LANS_8_2009_Estrella_FINAL.pdf, Stand: 14.08.2016.

FEMTI. In: http://www.issco.unige.ch:8080/cocoon/femti/st-home.html, Stand: 14.08.2016.

Figge, L. Udo. 1989. Fachsprache und maschinelle Übersetzung. In: Dahmen, Wolfgang u.a. (Hg.) *Technische Sprache und Technolekte in der Romania: Romanistisches Kolloquium II*, Tübingen: Gunter Narr Verlag.

Finch, Andrew/Yasuhiro Akiba/ Eiichiro Sumita. 2004. How Does Automatic Machine Translation Evaluation Correlate with Human Scoring as the Number of Reference Translations Increases? In*: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), 26−28 May 2004, Lisbon, Portugal,* 2019-2022.

Frederking, Robert/Nirenburg, Sergei. 1994. Three Heads are Better than One. In: *Proceedings of the Fourth Conference on Applied Natural Language Processing* (ANLP-94), Stuttgart, Germany.

Gambäck, Björn/Alshawi, Hiyan/Carter, David/Rayner, Manny. 1991. Measuring compositionality in trasfer-based machine translation systems. In: *Natural Language Processing Systems Evaluation Workshop*, Griffiss Air Force Base, NY. Rome Laboratory, Air Force System Command.

Garcia Ignacio. 2010. The proper place of professionals (and non-professionals and machines) in web translation. In: *Revista tradumatica*, 2010:8, https://ddd.uab.cat/pub/tradumatica/-15787559n8/15787559n8a2.pdf, Stand 14.08.2014.

Gotti, Maurizio. 1991. *I linguaggi specialistici. Caratteristiche linguistiche e criteri pragmatici*. Firenze: La Nuova Italia.

Guzmán, Francisco/Abdelali, Ahmed/Temnikova, Irina/Sajjad, Hassan/ Vogel, Stephan. 2015. How do Humans Evaluate Machine Translation. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, 17-18 September 2015, 457-466.

Halliday, M. A. K./ McIntosh, A./ Strevens, P. 1965. *The linguistic Sciences and Language Teaching*. London: Longmans.

Halliday, T. C./ Briss E. A. (Hg.). 1977. *The Evaluation and Systems Analysis of the Systran Machine Translation System*. Report RADC-TR-76-399, January 1977, Rome Air Development Center, Griffiss Air Force Base, New York.

Hirschman, Lynette/Mani, Inderjeet. 2003. Evaluation. In: Mitkov, Ruslan (Hg.), *Oxford Handbook of Computational Linguisti*c. Oxford: Oxford University Press, 414-447.

Hovy, Eduard H./ King, Margaret/Popescu-Belis, Andrei. 2002. Principles of Context-Based Machine Translation Evaluation. In: *Machine Translation*, 2002:17(1), 1-33.

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

Hutchins, John. 1996. ALPAC: the (in)famous report. In: *MT News International*, 1996:14, 9-12.

Hutchins, John. 1999. *Milestones no.6: Bar-Hillel and the nonfeasibility of FAHQT*. In: *International Journal of Language and Documentation no.1 (September 1999), 20-21*.

Hutchinson, Ben. 2004. Acquiring the meaning of discourse markers. In: *Proceedings of ACL 2004 (42nd Annual Meeting of the Association for Computational Linguistics)*. Barcelona, Spain.

ICD-10-GM, Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme . In: https://www.dimdi.de/static/de/klassi/icd-10-gm/index.htm, Stand: 14.08.2016.

ICF, Internationale Klassifikation der Funktionsfähigkeit, Behinderung und Gesundheit. In: http://www.dimdi.de/static/de/klassi/icf/, Stand: 14.08.2016.

ISO/IEC. (1999). ISO/IEC 14598-1:1999 (E) -- Information Technology -- Software Product Evaluation -- Part 1: General Overview. Geneva: International Organization for Standardization / International Electrotechnical Commission.

ISO/IEC. (2001). ISO/IEC 9126-1:2001 (E) -- Software Engineering -- Product Quality -- Part 1:Quality Model. Geneva: International Organization for Standardization / International Electrotechnical Commission.

Jelinek, Fred. 2004. Some of my Best Friends are Linguists, LREC 2004. In: http://www.lrec-conf.org/lrec2004/doc/jelinek.pdf, Stand: 14.08.2016.

King, Margaret. 1996. Evaluating Natural Language Processing Systems. In: *Communication of the ACM*, 29(1):73–79, January.

King, Margaret/Falkedal, Kirsten. 1990. Using test suites in evaluation of machine translation systems. In: *Proceedings of the 13th COLING*, Helsinki, Finland, 211-16.

Kit, Chunyu/Wong, Billy Tak-ming. 2015. Evaluation in Machine Translation and Computer-Aided translation. In: Chan, Sin-Wai (Hg.), *Routledge Encyclopedia of Translation Technology*. Oxon/New York: Routledge. 213-236.

Kittredge, Richard und Lehrberger, John (Hg.).1982. *Sublanguage: studies of language in restricted semantic domains.* Berlin/New York: Walter de Gruyter.

Knight, Kevin/Chander, Ishwar. 1994. Automated Postediting of Documents. In: *Proceedings of National Conference on Artificial Intelligence (AAAI)*, Seattle, Washington, 779-784.

Kocourek, Rostislav.1972. A Semantic Study of Terminology and its Application in Teaching the Technical Language. In: Fried Vilém (Hg.), *The Prague School of Linguistics and Language Teaching*. London: Oxford University Press.

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

LDOCE, Longman English Dictionary Online. In: http://www.ldoceonline.com/, Stand 14.08.2016.

Lehrberger, John/ Bourbeau, Laurent. 1988. *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation.* Amsterdam/Philadelphia: John Benjamins.

Llitjós, Ariadna Font/ Carbonell, Jaime G./Lavie, Alon. 2005. A framework for interactive and automatic refinement of transfer-based machine translation. In: *Proceedings of the 10th Annual Conf. of the European Association for Machine Translation (EAMT)*, Budapest, Hungary.

Lo, Chi-kiu/Wu, Dekai. 2011. MEANT: An Inexpensive, High-accuracy, Semi-automatic Metric for Evaluating Translation Utility via Semantic Frames. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-11)*, Portland, OR, 220-229.

Lopresti, Daniel/Tomkins, Andrew. 1997. Block edit models for approximate string matching. In: *Theoretical Computer Science*, 1997:181(1), 159-179.

Magris, Marella. 1992. *La traduzione del linguaggio medico: analisi contrastiva di testi in lingua italiana, inglese e tedesca*. In:*Traduzione, società e cultura, 2* (1992), pp. 1-82. Trieste: EUT Edizioni Università di Trieste.

MateCat. In: https://www.matecat.com/, Stand 14.08.2016.

Murata, Masaki/Uchimoto, Kiyotaka/Ma, Qing/Isahara, Hitoshi. 2001. *A machine-learning approach to estimating the referential properties of Japanese noun phrases*. In CICLING 2001, Mexico City, 142-153.

Naidenova, Xenia. 2010. *Machine Learning Methods for Commonsense Reasoning Processes: Interactive Models*. Hershey New York: Information Science Reference.

National Institute of Standards and Technology (NIST). 2009. *The 2009 NIST Open Machine Translation Evaluation Plan (MT09)*. In: http://www.itl.nist.gov/-iad/mig/tests/mt/2009/MT09_EvalPlan.pdf (Stand: 14.08.2016)

National Institute of Standards and Technology (NIST). 2010. The NIST Metrics for MAchine TRanslation 2010 Challenge (MetricsMATR10): Evaluation Plan.

Nomura, Hirosato. 1992. *JEIDA Methodology and Criteria on Machine Translation Evaluation*: Japan Electronic Industry Development Association (JEIDA).

O'Brien, Sharon/ Winther Balling, Laura/Carl, Michael/Michel,Simard/Specia, Lucia (Hg.). 2014. *Post-editing of Machine Translation: Processes and Applications*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Och, Franz Josef/ Ney, Hermann. 2001. What Can Machine Translation Learn from Speech Recognition? In*: Workshop MT 2010 -- Towards a Road Map for MT, Santiago de Compostela, Spain, September 2001*, 23-61.

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

Prunç, Erich. *Einführung in die Translationswissenschaft*. 2001. Bd 1: Orientierungsrahmen. Graz: Institut für Translationswissenschaft.

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

Reiß, Katharina und Vermeer, Hans J. 1991. Grundlegung einer allgemeinenTranslationstheorie. 2. Auflage. Tübingen: Niemeyer.

Rocca, G, Spampinato, L, Zarri, Gian Piero, & Black, William. (1994). COBALT: Construction, Augmentation and Use of Knowledge bases from Natural Language Documents. In: http://cordis.europa.eu/project/rcn/17214_it.html, Stand: 14.08.2016.

Sabatini, Francesco. 1999. “Rigidità-esplicitezza” vs “elasticità-implicitezza”: possibili parametri massimi per una tipologia dei testi. In: Skytte, Gunver/Sabatini, Francesco (Hg.), *Linguistica testuale comparativa. In memoriam Maria Elisabeth Conte.* Kopenhagen: Museum Tusculanum Press (=Etudes Romanes).

Sager, Naomi. 1982. Syntactic Formatting of Science Information. In: Kittredge Richard/ Lehrberger John (Hg.), *Sublanguage: Studies of Language in Restricted Semantic Domains,* Berlin: Walter de Gruyter, 9-26.

Scarton, Carolina/Specia, Lucia. 2016. A Reading Comprehension Corpus for Machine Translation Evaluation. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, May, 2016. Paris: ELRA, 3652- 3658.

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

Seljan, Sanja/Brkić, Marija/ Kučiš, Vlasta. 2011. Evaluation of Free Online Machine Translations for Croatian-English and English-Croatian Language Pairs. In: *Proceedings of the 3rd International Conference on the Future of Information* Sciences: INFuture2011-Information Sciences and e-Society, Zagreb, Croatia, 331-345.

Senez, Dorothy. 1998. Post-Editing service for machine translation users at the European Commission. In: *Translating and the Computer 20. Proceedings from Aslib conference, 12 & 13 November 1998.*

Snover, Matthew G./Madnani, Nitin/Dorr, Bonnie/Schwartz Richard. 2008. TERp System Description. In: *MetricsMATR workshop at AMTA.*

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

Snover, Matthew/ Dorr, Bonnie/Schwartz, Richard/Micciulla, Linnea/Makhoul, John. 2006. A study of translation edit rate with targeted human annotation. In: *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2006)*, Cambridge, Massachusetts.

Somers, Harold L./Wild, Elizabeth. 2000. Evaluating Machine Translation: The Cloze Procedure Revisited. In: *Proceedings of the 22nd International Conference on Translating and the Computer*, London, UK.

Stymne, Sara/ Danielsson, Henrik/ Bremin, Sofia/ Hu, Hongzhan/ Karlsson, Johanna/ Prytz Lillkull, Anna/ Wester, Martin. 2012. Eye Tracking as a Tool for Machine Translation Error Analysis. In: *Proceedings of the International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Taylor, Christopher. 1998. *Language to language, A Practical and Theoretical Guide for Italian/English Translators.* Cambridge: Cambridge University Press.

Taylor, Kathryn/ White, John. 1998. Predicting What MT Is Good for: User Judgements and Task Performance. In: *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas: Machine Translation and the Information Soup (AMTA-98)*, 28–31 October 1998, Langhorne, PA, 364–373.

Taylor, Wilson L. 1953. Cloze Procedure: A New Tool for Measuring Readability. In: *Journalism Quarterly* 30: 415–433.

TEMAA. 1996. TEMAA Final Report (No. LRE-62-070 (March 1996): Center fo Sprogteknologi, Copenhagen, Danemark.

Tercom. In: http://www.cs.umd.edu/~snover/tercom/, Stand 14.08.2016.

TER-Webschnittstelle. In: http://ter.panadigital.it/, Stand 14.08.2016.

Thurmair, Gregor. 2005. Automatic Means of MT Evaluation. In: *Proceedings of the ELRA-HLT Evaluation Workshop*, Malta, 1-2 December 2005.

Tomita, Masaru. 1992. Application of the TOEFL Test to the Evaluation of Japanese-English MT. In: *Proceedings of the AMTA Workshop on MT Evaluation*, San Diego, CA.

Tomita, Masaru/ Shirai, Masako/Tsutsumi, Junya/Matsumura, Miki/Yoshikawa, Yuki. 1993. Evaluation of MT Systems by TOEFL. In: *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation: MT in the Next Generation (TMI-93)*, 14–16 July 1993, Kyoto, Japan, 252–265.

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

Turian, Joseph P./Shen, Luke/Melamed, Dan I. 2003. Evaluation of machine translation and its evaluation. In: *Proceedings of the MT Summit IX*, New Orleans, LA.

UMDNS, Amtliche Nomenklatur zur Verschlüsselung von Medizinprodukten. In: http://www.dimdi.de/static/de/klassi/umdns/index.htm, Stand: 14.08.2016.

Van Slype, Georges. 1979. *Critical study of methods for evaluating the quality of machine translation. Final Report*. Bruxelles: Bureau Marcel Van Dijk.

Vieira Nunes, Lucas. 2014. Indices of cognitive effort in machine translation post-editing. In: Machine Translation, December 2014, Volume 28, Issue 3, 187-216.

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.

Wagner, Emma. 1985. Post-editing Systran — A Challenge for Commission Translators. In: *Terminologie & Traduction*, 1985:3.

Way, Andy. 2009. *A critique of Statistical Machine Translation, in:* Daelemans, Walter/ Hoste, Véronique (Hg.), *Evaluation of Translation Technology. Special issue of Linguistica Antverpiensia New Series – Themes in Translation Studies 8. 2009. 2.*

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

White, John S./Doyon, Jennifer B./ Talbott, Susan W. 2000. Task Tolerance of MT Output in Integrated Text Processes. *ANLP/NAACL 2000: Embedded Machine Translation Systems*, 9-16.

White, John S./O'Connell, Theresa A. 1994. Evaluation in the ARPA Machine Translation Program: 1993 Methodology. In: *Proceedings of the Workshop on Human Language Technology (HLT-94)*, Plainsboro, NJ, 134–140.

Wilss, Wolfram. 1977. *Übersetzungswissenschaft. Probleme und Methoden*. Stuttgart: Klett-Cotta.

Wilss, Wolfram. 1994. Grundkonzepte der Maschinellen Übersetzung. In: Fischer, Ingeborg/Freigang, Karl-Heinz/Mayer, Felix/Reinke, Uwe (Hg.), *Sprachdatenverarbeitung*

für Übersetzer und Dolmetscher : Akten des Symposiums zum Abschluß des Saarbrücker Modellversuchs, 28. - 29. September 1992, Hildesheim [u.a.]:Olms.

Wong, Billy Tak-ming/Kit, Chunyu. 2012. Extending Machine Translation Evaluation Metrics with Lexical Cohesion To Document Level. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea, 12–14 July 2012, 1060–1068.

WordNet®. In: http://wordnet.princeton.edu/, Stand 14.08.2016.

Anhang

Ausgangstext

Verfahren zur diagnose von rheumatischen erkrankungen.[45]
EP 1896497 A2 (Text aus WO2007000320A2)**.**
ZUSAMMENFASSUNG.

Die Erfindung betrifft Polypeptide, die mit Rheuma-assoziierten Autoantikörpern reagieren. Die Erfindung betrifft außerdem ein Diagnostikum, das eines dieser Polypeptide enthält, einen diagnostischen Kit, der dieses Diagnostikum enthält, sowie ein Verfahren zum in vitro-Nachweis von rheumatischen Erkrankungen. Des Weiteren betrifft die Erfindung ein Arzneimittel, das eines der Polypeptide enthält, sowie die Verwendung der Polypeptide zur Herstellung eines Arzneimittels zur Prophylaxe und/oder zur Behandlung von rheumatischen Erkrankungen.

ANSPRÜCHE (OCR-Text kann Fehler enthalten).

Ansprüche.

1) Polypeptid abgeleitet von nativem Vimentin mit der SEQ ID No. 1 , dadurch gekennzeichnet, dass es gegenüber der nativen Sequenz mindestens einen zusätzlichen Arginin-Rest enthält.

2) Polypeptid nach Anspruch 1 , dadurch gekennzeichnet, dass es in mindestens einer der Positionen 16, 17, 19, 41 , 58, 59, 60, 68, 76, 140, 142, 147, 363, 406 oder 452 einen zusätzlichen Arginin- Rest aufweist.

3) Polypeptid nach Anspruch 2, dadurch gekennzeichnet, dass es in mindestens zwei der Positionen einen zusätzlichen Arginin- Rest aufweist.

4) Polypeptid nach Anspruch 1 , dadurch gekennzeichnet, dass es gegenüber der nativen Sequenz in mindestens einer der Positionen 3, 20, 33, 36, 37, 94, 165, 361, 399 oder 426 einen zusätzlichen Leucin-Rest aufweist.

5) Polypeptid nach Anspruch 4, dadurch gekennzeichnet, dass es in mindestens zwei der Positionen einen zusätzlichen Leucin- Rest aufweist.

[45] Der vorgliegende Text ist ein Auszug aus dem Originaltext: https://www.google.com.ar/patents/-EP1896497A2?cl=de

6) Polypeptid nach Anspruch 1 , dadurch gekennzeichnet, dass es gegenüber der nativen Sequenz in mindestens einer der Positionen 21 , 41 , 43, 50, 54, 62, 64 oder 89 einen zusätzlichen Prolin- Rest aufweist.

7) Polypeptid nach Anspruch 6, 5 dadurch gekennzeichnet, dass es in mindestens zwei der Positionen einen zusätzlichen Prolin- Rest aufweist.

8) Polypeptid nach Anspruch 1 , o dadurch gekennzeichnet, dass es gegenüber der nativen Sequenz in mindestens einer der Positionen 24, 35 oder 99 einen zusätzlichen Threonin-Rest aufweist.

9) Polypeptid nach Anspruch 8, 5 dadurch gekennzeichnet, dass es in mindestens zwei der Positionen einen zusätzlichen Threonin- Rest aufweist.

10) Polypeptid nach Anspruch 1 , o dadurch gekennzeichnet, dass es gegenüber der nativen Sequenz in mindestens einer der Positionen 25, 39, 42, 49, 55 oder 67 einen zusätzlichen Tyrosin-Rest aufweist.

5 11) Polypeptid nach Anspruch 10, dadurch gekennzeichnet, dass es in mindestens zwei der Positionen einen zusätzlichen Tyrosin- Rest aufweist.

0 12) Polypeptid nach Anspruch 1 , dadurch gekennzeichnet, dass mindestens ein Arginin-Rest als Citrullin-Rest vorliegt.

13) Polypeptid nach Anspruch 12, dadurch gekennzeichnet, dass es in mindestens einer der Positionen 4, 12, 23, 28, 36, 45, 50, 64, 71 , 100, 320, 364 oder 378 einen Citrullin-Rest aufweist.

14) Polypeptid nach Anspruch 13, dadurch gekennzeichnet, dass es in mindestens zwei der Positionen einen Citrullin-Rest aufweist.

15) Fragment bestehend aus mindestens sechs Aminosäuren, abgeleitet von nativem Vimentin mit der SEQ ID No. 1 , dadurch gekennzeichnet, dass es mindestens einen Bereich mit mindestens einem zusätzlichen Arginin-Rest enthält und dass es eine Reaktivität gegen Rheuma- assoziierte Autoantikörper zeigt.

16) Peptid-Derivat eines Polypeptids oder Fragments nach einem der Ansprüche 1 bis 15.

17) Peptid-Derivat nach Anspruch 16, dadurch gekennzeichnet, dass es aus Retro- oder/und Inverso-Polypeptiden und cyclischen Peptiden ausgewählt ist.

18) Diagnostikum, dadurch gekennzeichnet, dass es ein Polypeptid oder Peptid-Derivat, wie in einem der Ansprüche 1 bis 17 definiert, enthält.

19) Diagnostischer Kit zur Verwendung zum Nachweis rheumatischer Erkrankungen, dadurch gekennzeichnet, dass er ein Diagnostikum wie in Anspruch 18 definiert enthält.

20) Diagnostischer Kit nach Anspruch 19, dadurch gekennzeichnet, dass er zum Nachweis von Rheumatoider Arthritis verwendet wird.

5 21) Diagnostischer Kit nach Anspruch 19 oder 20, dadurch gekennzeichnet, dass der Träger DNA, RNA, medizinisch verträgliche Polymere, snythetische Biopolymere oder Proteine sind.

BESCHREIBUNG (OCR-Text kann Fehler enthalten).

Verfahren zur Diagnose von rheumatischen Erkrankungen.

Beschreibung.

Die Erfindung betrifft Polypeptide, die mit Rheuma-assoziierten Autoantikörpern reagieren. Die Erfindung betrifft außerdem ein Diagnostikum, das eines dieser Polypeptide enthält, einen diagnostischen Kit, der dieses Diagnostikum enthält, sowie ein Verfahren zum in vitro-Nachweis von rheumatischen Erkrankungen. Des Weiteren betrifft die Erfindung ein Arzneimittel, das eines der Polypeptide enthält, sowie die Verwendung der Polypeptide zur Herstellung eines Arzneimittels zur Prophylaxe und/oder zur Behandlung von rheumatischen Erkrankungen.

Rheumatische Erkrankungen, insbesondere Schmerzen im Bereich der Gelenke und des Bewegungsapparates gehören zu den häufigsten Krankheiten in Deutschland. Ein Labortest, der es ermöglicht, diese Schmerzen einer harmlosen Muskelverspannung, einer Arthrose oder der häufigsten und schwersten der Erkrankungen, der Rheumatoiden Arthritis (RA), zuzuordnen, ist bisher nicht bekannt.

Die Rheumatoide Arthritis ist eine Autoimmunkrankeit, bei der die Abwehrmechanismen des menschlichen Körpers irrtümlich körpereigenen Gelenkknorpel für fremd und feindlich halten und diesen angreifen. Ungefähr 1 von 100 Menschen leidet in westeuropäischen Ländern an Rheumatoider Arthritis. In den ersten Monaten der Erkrankung schreitet diese sehr rasch voran.

Eine wesentliche Schlüsselstrategie in der modernen Rheumatologie ist daher der frühzeitige Einsatz von biologischen Arzneistoffen, die den Krankheitsverlauf modifizieren. Zahlreiche klinische Studien haben gezeigt, dass mit geeigneten Wirkstoffen, z.B. mit TNF-Antagonisten, sehr gute Therapieerfolge und Ansprechraten erzielt werden können, wenn diese bei Patienten bereits im Frühstadium eingesetzt werden. Rheumatologen versuchen, das schmale Zeitfenster zwischen dem Beginn der Krankheit und dem Auftreten von strukturellen Gelenkschäden zu nutzen. Bisher ist jedoch aus dem Stand der Technik kein zuverlässiger und sensitiver Nachweis der Rheumatoiden Arthritis in diesem Zeitfenster bekannt.

Die Diagnose der Rheumatoiden Arthritis erfolgt nach den Klassifikationskriterien des ACR (American College of Rheumatology). Gemäß den Kriterien des ACR ist der Rheumafaktor der bisher grundlegende serologische Indikator zur Diagnose der Rheumatoiden Arthritis (RA). Rheumafaktoren sind eine Teilgruppe von Immunglobulinen, die sich durch die

immunologische Kreuzreaktion gegen die Fc-Region von Immunglobulin G (IgG) auszeichnen.

Das Vorhandensein eines Rheumafaktors ist jedoch nicht auf Erkrankungen des rheumatischen Formenkreises (differenzialdiagnostische Anhaltspunkte) beschränkt, man findet Rheumafaktoren auch im Serum von Patienten mit Infektionserkrankungen, Hyperglobulinanämien, lymphoproliferativen B-ZeII- Erkrankungen und allgemein bei älteren Bevölkerungsschichten.

Im Allgemeinen werden erhöhte Konzentrationen von Rheumafaktoren mit einem schwereren Krankheitsverlauf assoziiert. Dabei korrelieren die Konzentrationen nicht mit dem Aktivitätsgrad und dem Therapieerfolg. Auf Basis der Konzentration von Rheumafaktoren kann keine sensitive und spezifische Prognose für den Beginn einer Rheumatoiden Arthritis getroffen werden. Gesunde Menschen haben eine erhöhte Rheumafaktorkonzentration ohne zu erkranken, Patienten ohne Rheumafaktoren haben dagegen eine sehr aggressive Form der Rheumatoiden Arthritis.

Andere serologische Marker, wie der anti-citrulline Antikörper (CCP) oder der initiale HAQ-Score, mit dem Fähigkeiten im Alltag beurteilt werden, oder das Röntgen- oder Computer-Tomografie (CT)-BiId geben bei der Frühform nur geringe Aufschlüsse und sind alleine nicht aussagekräftig genug, um beurteilen zu können, wie die Prognose des Patienten sein wird.

Zur Optimierung der bestehenden Klassifikationskriterien des ACR werden von der US-amerikanischen Rheumatologischen Fachgesellschaft sieben Klassifikationskriterien vorgeschlagen, die auf eine schlechte Prognose hindeuten:

(1) Morgensteifigkeit der Gelenke von mehr als einer Stunde, (2) Arthritis an drei oder mehr Gelenken, (3) Gelenkentzündung von mindestens drei Gelenkregionen zur gleichen Zeit, (4) Handgelenke oder Fingergelenke sind ebenfalls betroffen, (5) bilaterale Druckschmerzhaftigkeit von Metacarpophalangeal-Gelenken, (6) Erosionen im Röntgenbild, (7) Nachweis spezieller Rheumafaktoren und Anti-perinukleäre Faktor- Positivität (APF).

Autoantikörper gegen den so genannten Anti-perinukleären Faktor wurden erstmals von Young et al. bei Patienten mit Rheumatoider Arthritis beschrieben. Aufgrund ihrer spezifischen Reaktion gegen das verhornte Epithel des Stratum corneum auf Rattenoesophagusschnitten wurde lange Zeit Keratin als das entsprechende Antigen angesehen. Die Antikörper werden aus diesem Grund bis heute als Anti- Keratin-Antikörper (AKA) bezeichnet.

Spätere Untersuchungen haben darüber hinaus gezeigt, dass AKA oder APF auch durch Anti-Filaggrin-Antikörper erkannt werden. Somit wurde das basische Protein Filaggrin als Zielantigen identifiziert. Das 40 kDa-Protein aggregiert Zytokeratinfilamente und hilft mit, die intrazelluläre Fasermatrix der verhornten Zellen zu bilden.

Da APF, AKA und Anti-Filaggrin-Antikörper enthaltende Seren in gleicher Weise reagieren, sind diese Antikörpersysteme scheinbar identisch. Anti- Filaggrin-Antikörper vom Typ IgG stellen mit einer Spezifität von über 99 % einen hochspezifischen Marker für die Rheumatoide Arthritis dar. Die Antikörper sind prinzipiell früh nachweisbar und gehen dem klinischen Krankheitsbild voraus. In mehreren Studien konnten positive Korrelationen zu Schwere und Aktivität der Krankheit gefunden werden. Anti-Filaggrin- Antikörper korrelieren nicht mit Alter, Geschlecht oder Krankheitsdauer. Sie können in ca. 34 % der Rheumafaktor-negativen Patienten nachgewiesen werden und stellen hier eine wertvolle diagnostische Hilfe dar.

Mit heute gebräuchlichen Methoden sind die Antikörper jedoch nur in ca. 40 % der Fälle im Serum zu finden.

Aufgabe der vorliegenden Erfindung war es daher, neue Polypeptide zum Nachweis von mit rheumatischen Erkrankungen assozierten, insbesondere mit Rheumatoider Arthritis assoziierten, Antikörpern bereitzustellen, die eine sensitive und spezifische Diagnose, eine Klassifizierung und eine Prognose von rheumatischen Erkrankungen, insbesondere von Schmerzen im Bereich der Gelenke und des Bewegungsapparates ermöglichen.

Bei der Analyse der Antikörper-Bindung an natives Vimentin, d.h. der APF- Positivität oder anti-Sa-Reaktivität in Form von mutierten immunologisch reaktiven Varianten vorliegen. Diese Erkenntnis ist unerwartet, da nach dem bisherigen Stand der Technik davon ausgegangen wurde, dass Vimentin citrulliert sein muss, um immunologisch reaktiv zu sein. Diese Annahme konnte von uns widerlegt werden, indem durch differenzielle Immunaffinitäts- chromatographie immunologisch reaktive Vimentin-Varianten mit mutierten Sequenzen aus humanen Monocyten angereichert werden konnten. Diese mutierten Varianten von nativem Vimentin unterscheiden sich von nativem Vimentin durch das Vorhandensein zusätzlicher Arginin-Reste und gegebenenfalls weiterer Sequenzunterschiede. Sie reagieren mit humanen RA-assoziierten Antikörpern und weisen überraschenderweise eine höhere Spezifität und Sensitivität als die aus dem Stand der Technik bekannten citrullinierten Peptide auf.

Ein Gegenstand der Erfindung ist daher ein Polypeptid, das von nativem Vimentin mit der SEQ ID No. 1 abgeleitet ist und das sich gegenüber der nativen Sequenz durch das Vorhandensein von mindestens einem zusätzlichen Arginin-Rest unterscheidet.

Die zusätzlichen Arginin-Reste sind vorzugsweise durch Substitution anderer Aminosäurereste des nativen humanen Vimentins in die Sequenz eingefügt. Vorzugsweise weist das Polypeptid in mindestens einer der Positionen 16, 17, 19, 41 , 58, 59, 60, 68, 76, 140, 142, 147, 363, 406 oder 452 einen Arginin-Rest auf. Besonders bevorzugte Positionen sind 41 , 58, 59, 60 und/oder 68. Beispielsweise weist das Polypeptid in mindestens einer, zwei, drei oder vier Positionen einen zusätzlichen Arginin-Rest auf. In einer weiteren Ausführungsform weist das Polypeptid außerdem in mindestens einer der Positionen 3, 20, 33, 36, 37, 94, 165, 361 , 399 oder 426 gegenüber der nativen Sequenz einen zusätzlichen

Leucin-Rest auf, vorzugsweise an den Positionen 33, 36 und/oder 37. Beispielsweise weist das Polypeptid in mindestens einer, zwei, drei oder vier Positionen einen zusätzlichen Leucin-Rest auf.

In einer weiteren Ausführungsform weist das Polypeptid in mindestens einer Position, z.B. in einer der Positionen 21 , 41 , 43, 50, 54, 62, 64 oder 89, gegenüber der nativen Sequenz einen zusätzlichen Prolin-Rest auf, vorzugsweise an den Positionen 41 , 43, 50, 54 62, und/oder 64. Beispielsweise weist das Polypeptid in mindestens einer, zwei, drei oder vier Positionen einen Prolin-Rest auf.

In einer weiteren Ausführungsform weist das Polypeptid in mindestens einer Position, z.B. in einer der Positionen 24, 35 oder 99, gegenüber der nativen Sequenz einen zusätzlichen Threonin-Rest auf. Beispielsweise weist das Polypeptid in mindestens einer, zwei oder drei Positionen einen Threonin- Rest auf.

In einer weiteren Ausführungsform weist das Polypeptid in mindestens einer Position, z.B. in einer der Positionen 25, 39, 42, 49, 55 oder 67, gegenüber der nativen Sequenz einen zusätzlichen Tyrosin-Rest auf. Beispielsweise weist das Polypeptid in mindestens einer, zwei, drei oder vier Positionen einen Tyrosin-Rest auf.

In einer weiteren Ausführungsform liegt in dem Polypeptid mindestens ein Arginin-Rest als Citrullin-Rest vor, z.B. in mindestens einer der Positionen 4, 12, 23, 28, 36, 45, 50, 64, 71, 100, 320, 364 oder 378. Beispielsweise weist das Polypeptid in mindestens einer, zwei, drei oder vier der Positionen einen Citrullin-Rest auf. Andererseits kann das Polypeptid jedoch auch ein citrullinfreies Polypeptid sein. Bevorzugte Beispiele für Muteine des humanen Vimentin haben eine Sequenz mit der SEQ ID No. 2, 3, 4, 5, 6, 7, 8 oder 9.

Ein weiterer Gegenstand der Erfindung ist ein Fragment eines der o.g. Polypeptide, das von nativem Vimentin mit der SEQ ID No. 1 abgeleitet ist und das mindestens einen Bereich mit mindestens einem Arginin-Rest enthält und das eine Reaktivität gegen mit rheumatoiden Erkrankungen assoziierten Autoantikörpern zeigt. Vorzugsweise liegt das Fragment im Bereich der Positionen 10-145. Besonders bevorzugt liegt das Fragment im Bereich der Positionen 30-70. Ein bevorzugtes Beispiel eines Fragments ist das Fragment 51-65 (C2). Die Länge des Fragments beträgt vorzugsweise mindestens 6, besonders bevorzugt mindestens 8 Aminosäuren bis zu 120, vorzugsweise bis zu 100 und besonders bevorzugt bis zu 50 Aminosäuren.

Ein weiterer Gegenstand der Erfindung sind Peptid-Derivate der o.g. Polypeptide oder Fragmente. Beispielsweise kann das Peptid-Derivat ein Retro-/Inverso-Polypeptid sein, d.h. ein inverses Polypeptid der oben beschriebenen Polypeptide, das entsprechend einem Spiegelbild („mirror image") der Polypeptide aus D-Aminosäuren hergestellt wird, ein Retro-Polypeptid, das eine „umgekehrte" Sequenz aufweist sowie ein Retro- Inverso-Polypeptid, das ein Spiegelbild der oben beschriebenen Polypeptide ist und zudem eine „umgekehrte" Sequenz aufweist.

Weitere Beispiele von Peptid-Derivaten sind Seitengruppen-, amino- oder/und carboxyterminal modifizierte Polypeptide einer Aminogruppe, z.B. Polypeptide, die z.B. mit einer Carbonsäure oder einem Alkylrest modifiziert sind, oder die an einer Carbonsäuregruppe mit einer Aminogruppe oder einer Estergruppe modifiziert sind. Die Polypeptide und/oder Peptid- Derviate können auch als cyclische Peptide vorliegen.

Ein weiterer Gegenstand der Erfindung ist eine Nukleinsäure, die für ein oben beschriebenes Polypeptid kodiert. Als Nukleinsäuren kommen z.B. DNA und RNA, insbesondere cDNA infrage. Die Nukleinsäuren können zur rekombinanten Herstellung der Polypeptide in übliche eukaryontische oder prokaryontische Vektoren einkloniert und in geeigneten Wirtszellen exprimiert werden.

5 Ein weiterer Gegenstand der Erfindung ist ein Diagnostikum, das ein oder mehrere der oben beschriebenen Polypeptide oder deren Fragmente enthält. Das Diagnostikum kann das Polypeptid oder das Fragment in freier oder trägergebundener Form enthalten.

o Es ist als ausgesprochen überraschend zu bezeichnen, dass sich die erfindungsgemäßen Polypeptide als hochspezifische und hochsensitive Antigene für die Diagnostik von Antikörpern in Körperflüssigkeiten von Patienten mit rheumatischen Erkrankungen, insbesondere mit entzündlichen Erkrankungen der Gelenke und des Bewegungsapparates, besonders 5 bevorzugt von Rheumatoider Arthritis, erweisen. Bevorzugte Körperflüssigkeiten im Sinne der Erfindung sind Blut, Serum oder Plasma, besonders bevorzugt ist Serum.

Das erfindungsgemäße Diagnostikum weist eine Reihe von Vorteilen auf. So o können, da die Polypeptide mehrere Antikörperbindungsstellen enthalten, sowohl monomere als auch multimere Antikörper effizient gebunden werden. Ein weiterer Vorteil des mutierten Polypeptids ist, dass dieses die Bereitstellung eines Diagnostikums ermöglicht, das mit 99 % Spezifität und 85 % Sensitivität Patienten mit entzündlichen und chronischen 5 Erkrankungen der Gelenke und des Bewegungsapparates, insbesondere mit Rheumatoider Arthritis, identifizieren kann.

Aus dem Stand der Technik ist bisher kein vergleichbar spezifisches oder sensitives Diagnostikum bekannt, das den Nachweis von rheumatischen o Erkrankungen, insbesondere von Rheumatoider Arthritis unter Verwendung eines citrullinfreien Proteins oder Peptids ermöglicht. Ein weiterer Gegenstand der Erfindung ist ein diagnostischer Kit zur Verwendung zum Nachweis rheumatischer Erkrankungen, insbesondere von Rheumatoider Arthritis, der ein oben beschriebenes Diagnostikum enthält. Daneben kann der diagnostische Kit übliche Bestandteile, wie Puffer, Lösungsmittel und/oder Markierungsgruppen, enthalten.

Als Träger kommen Makromoleküle, wie DNA, RNA, medizinisch verträgliche Polymere, wie beispielsweise Polyethylen, PoIy D.L-Laktide, PoIy D,L-Laktid-co-glykolide, synthetische Biopolymere, wie beispielsweise Polylysine und Dextrane, und Proteine, wie beispielsweise Serumalbumin und Hämocyanin, infrage.

Ein weiterer Gegenstand der Erfindung ist ein Verfahren zum in vitro- Nachweis von rheumatischen Erkrankungen, insbesondere von Rheumatoider Arthritis, bei dem die Konzentration von Autoantikörpern in einer Körperflüssigkeit bestimmt wird. Das Verfahren erlaubt die Stellung einer Diagnose, die Klassifizierung und/oder die Bewertung des Schweregrades der Erkrankung. Als Nachweisreagenz dient das oben beschriebene Diagnostikum oder der oben beschriebene diagnostische Kit.

In dem erfindungsgemäßen Verfahren können als Nachweismethoden alle auf dem Gebiet der Diagnostik üblichen Methoden, wie (a) enzymologische Methoden, (b) Lumineszenz-basierende Methoden oder (c) radiochemische Methoden verwendet werden.

Als bevorzugte Nachweismethoden kommen in dem erfindungsgemäßen Verfahren ein Radioimmunoassay, ein Chemolumineszenzimmunoassay, ein Immuno-Blot-Assay oder ein Enzymimmunoassay, z.B. ein ELISA, infrage.

In einer Ausführungsform des erfindungsgemäßen Verfahrens wird zu einem oben beschriebenen Polypeptid, das an einen Träger gebunden ist, als Probe die zu analysierende Körperflüssigkeit hinzugegeben. Nach Inkubation der Probe werden die ungebundenen Bestandteile weggewaschen.

Bewertungen von *fluency* und *adequacy*

			PE1		PE2	
MÜ-Segment	**TER**[46]	**Seg.**	**FL**	**AD**	**FL**	**AD**
Procedimento per la diagnosi di malattie reumatiche.	0,00	1	5	5	5	5
Descrizione.	0,00	32	5	5	5	5
Dopo l'incubazione del campione, i componenti non legati vengono lavati via.	0,09	121	4	5	5	5
3) Polipeptide secondo la rivendicazione 2, caratterizzato dal fatto che presenta un residuo di arginina supplementare ad almeno due delle posizioni.	0,14	11	2	2	4	3
EP 1896497 A2 (testo dal WO2007000320A2).	0,17	2	4	5	5	5
14) Il polipeptide secondo la rivendicazione 13, caratterizzato dal fatto di comprendere un residuo citrullina in almeno due posizioni.	0,21	22	4	3	4	3
La lunghezza del frammento è preferibilmente almeno 6, più preferibilmente almeno 8 amminoacidi fino a 120, preferibilmente fino a 100 e più preferibilmente fino a 50 aminoacidi.	0,21	96	2	1	3	2
0 12) polipeptide secondo la rivendicazione 1, caratterizzato dal fatto che almeno un residuo di arginina è presente come residuo citrullina.	0,22	20	4	3	3	3
5 11) Polipeptide secondo la rivendicazione 10, caratterizzato dal fatto di comprendere un residuo di tirosina supplementare in almeno due delle posizioni.	0,23	19	4	3	4	4
Preferibilmente, il polipeptide ha almeno una delle posizioni 16, 17, 19, 41, 58, 59, 60, 68, 76, 140, 142, 147, 363, 406 o 452 a un residuo di arginina.	0,23	77	2	2	2	3

[46] Aufsteigend.

13) Il polipeptide secondo la rivendicazione 12, caratterizzato dal fatto di comprendere un residuo citrullina in almeno una delle posizioni 4, 12, 23, 28, 36, 45, 50, 64, 71, 100, 320, 364 o 378 a	0,24	21	4	4	4	3
2) Un polipeptide secondo la rivendicazione 1, caratterizzato dal fatto che un ulteriore arginina almeno una delle posizioni 16, 17, 19, 41, 58, 59, 60, 68, 76, 140, 142, 147, 363, 406 o 452 che ha resto.	0,24	10	1	1	2	1
5) Un polipeptide secondo la rivendicazione 4, caratterizzato dal fatto di comprendere un residuo di leucina supplementare ad almeno due delle posizioni.	0,27	13	2	2	4	2
Esempi preferibili di muteine di vimentina umana hanno una sequenza con la SEQ ID 2, 3, 4, 5, 6, 7, 8 o 9	0,27	91	3	3	4	3
In un'altra forma di realizzazione, il polipeptide si trova almeno un residuo di arginina come residuo citrullina prima, ad esempio in almeno una delle posizioni 4, 12, 23, 28, 36, 45, 50, 64, 71, 100, 320, 364 o 378.	0,28	88	1	1	1	1
DESCRIZIONE (testo OCR potrebbe contenere errori).	0,29	30	4	5	5	5
Un metodo per la diagnosi di malattie reumatiche.	0,29	31	4	4	5	4
Gli anticorpi anti-filaggrina non sono correlati con l'età, il sesso o la durata della malattia.	0,29	66	4	5	4	5
Per esempio, il polipeptide ha almeno uno, due, tre o quattro posizioni su un residuo di tirosina.	0,29	87	2	1	2	2
Circa 1 persona su 100 soffre nei paesi dell'Europa occidentale per l'artrite reumatoide.	0,31	39	4	3	4	3
Studi successivi hanno anche dimostrato che AKA o APF essere rilevato anche da anticorpi anti-filaggrina.	0,31	59	3	2	3	2
9) un polipeptide secondo la rivendicazione 8, 5 caratterizzato dal fatto di comprendere un residuo treonina aggiuntivo in almeno due delle posizioni.	0,32	17	4	3	4	3
Un esempio preferito di un frammento è il frammento 51-65 (C2).	0,33	95	2	2	4	3

In una ulteriore forma di realizzazione, il polipeptide comprende inoltre almeno una delle posizioni 3, 20, 33, 36, 37, 94, 165, 361, 399 o 426 rispetto alla sequenza nativa, un residuo di leucina supplementare, preferibilmente in posizioni 33, 36 e / o 37.	0,35	80	2	1	3	3
Per esempio il polipeptide ha almeno uno, due, tre o quattro posizioni ad un residuo di prolina.	0,35	83	2	2	3	2
L'invenzione riguarda anche un agente diagnostico che contiene uno di questi polipeptidi, un kit diagnostico che contiene questo uso diagnostico, come pure un metodo per il rilevamento in vitro di malattie reumatiche.	0,37	5	2	1	4	3
L'invenzione riguarda anche un agente diagnostico che contiene uno di questi polipeptidi, un kit diagnostico che contiene questo uso diagnostico, come pure un metodo per il rilevamento in vitro di malattie reumatiche.	0,37	34	2	2	2	2
Ad esempio, il polipeptide ha almeno uno, due, tre o quattro posizioni di un ulteriore residuo di arginina on.	0,37	79	3	2	2	2
Per esempio, il polipeptide ha almeno uno, due, tre o quattro posizioni su un residuo di leucina aggiuntivo.	0,37	81	2	1	3	2
In una ulteriore forma di realizzazione, il polipeptide ha almeno una posizione, ad esempio in una delle posizioni 25, 39, 42, 49, 55 o 67, rispetto alla sequenza nativa, un residuo tirosina supplementare.	0,39	86	1	1	2	2
15) frammento costituito da almeno sei aminoacidi, derivato da vimentina nativa con la SEQ ID 1, caratterizzato dal fatto di contenere almeno una regione con almeno un residuo di arginina addizionale, e che mostra una reattività contro reumatismi associati autoanticorpi.	0,39	23	4	2	4	3
I residui di arginina aggiuntivi sono preferibilmente inseriti per sostituzione di altri residui aminoacidici di vimentina umana nativa nella sequenza.	0,39	76	3	2	4	3
10) Un polipeptide secondo la rivendicazione 1, caratterizzato dal fatto che o viene confrontato con la sequenza nativa in almeno una delle posizioni 25, 39, 42, 49, 55 o 67 aventi un residuo di tirosina supplementare.	0,40	18	2	1	3	2

16) derivato peptidico di un polipeptide o un frammento di una qualsiasi delle rivendicazioni da 1 a 15	0,40	24	4	2	4	4
20) Un kit secondo la rivendicazione 19 diagnostica, caratterizzato dal fatto che è utilizzato per il rilevamento di artrite reumatoide.	0,40	28	2	1	2	2
In una ulteriore forma di realizzazione, il polipeptide ha almeno una posizione, ad esempio in una delle posizioni 21, 41, 43, 50, 54, 62, 64 o 89, in confronto con la sequenza nativa, un residuo di prolina supplementare, preferibilmente 64 nelle posizioni 41, 43, 50, 54 62, e / o.	0,40	82	2	1	1	1
Preferibilmente, il frammento è nella gamma di posizioni 10-145.	0,40	93	4	3	4	4
7) un polipeptide secondo la rivendicazione 6, 5 caratterizzato dal fatto di comprendere un ulteriore prolina residuo almeno due delle posizioni.	0,41	15	2	2	3	2
In una ulteriore forma di realizzazione, il polipeptide ha almeno una posizione, ad esempio in una delle posizioni 24, 35 o 99, in confronto con la sequenza nativa, un residuo di treonina supplementare.	0,41	84	1	1	3	2
Per esempio, il polipeptide ha almeno uno, due, tre o quattro posizioni di un residuo citrullina on.	0,41	89	2	1	3	2
1) polipeptide derivato da vimentina nativa con il n ° SEQ ID 1, caratterizzato dal fatto che comprende un residuo di arginina aggiuntivo rispetto alla sequenza nativa almeno.	0,41	9	4	2	3	2
4) Il polipeptide secondo la rivendicazione 1, caratterizzato dal fatto che viene confrontato con la sequenza nativa in almeno una delle posizioni 3, 20, 33, 36, 37, 94, 165, 361, 399 o 426 ha un residuo di leucina aggiuntivo.	0,43	12	4	3	3	2
RECLAMI (testo OCR potrebbe contenere errori).	0,43	7	4	3	5	3
5 21) Un kit diagnostico secondo la rivendicazione 19 o 20, caratterizzato dal fatto che il supporto è DNA, RNA, medicamente polimeri accettabili, biopolimeri o snythetische proteine.	0,46	29	2	1	1	1
La diagnosi di artrite reumatoide si basa sui criteri di classificazione del ACR (American College of Rheumatology).	0,47	45	4	4	4	4

Queste varianti mutanti di vimentina nativo differiscono da vimentina nativo per la presenza di residui di arginina aggiuntivi ed eventualmente altri differenze di sequenza.	0,47	73	2	2	3	2
Per esempio, il polipeptide ha almeno uno, due o tre posizioni su un residuo di treonina.	0,47	85	2	2	2	2
17) peptide Derivato secondo la rivendicazione 16, caratterizzato dal fatto di essere scelto fra Retro / o inverso polipeptidi e peptidi ciclici.	0,48	25	2	1	3	2
Le malattie reumatiche, in particolare dolori alle articolazioni e il sistema muscolo-scheletrico sono tra le più comuni malattie in Germania.	0,48	36	4	3	4	3
6) Un polipeptide secondo la rivendicazione 1, caratterizzato dal fatto che viene confrontato con la sequenza nativa in almeno una delle posizioni 21, 41, 43, 50, 54, 62, 64 o 89 aventi un residuo di prolina aggiuntivo.	0,49	14	4	2	4	2
18) un agente diagnostico, caratterizzato dal fatto di contenere un polipeptide o peptide derivato come definito in una qualsiasi delle rivendicazioni da 1 a 17.	0,50	26	3	1	4	3
L'agente diagnostico dell'invenzione ha una serie di vantaggi.	0,50	108	4	3	4	4
19) Un kit diagnostico per uso nella rilevazione di malattie reumatiche caratterizzato dal fatto che contiene come definito nella rivendicazione 18 per uso diagnostico.	0,52	27	2	1	2	1
8) Un polipeptide secondo la rivendicazione 1, caratterizzato dal fatto che o viene confrontato con la sequenza nativa in almeno una delle posizioni 24, 35 o 99 comprendente un residuo treonina aggiuntivo.	0,53	16	2	1	2	2
Inoltre, l'invenzione riguarda una composizione farmaceutica contenente uno dei polipeptidi, nonché l'uso dei polipeptidi per la preparazione di un medicamento per la profilassi e / o il trattamento di malattie reumatiche.	0,54	6	2	2	3	2
Inoltre, l'invenzione riguarda una composizione farmaceutica contenente uno dei polipeptidi, nonché l'uso dei polipeptidi per la preparazione di un medicamento per la profilassi e / o il trattamento di malattie reumatiche.	0,54	35	2	3	3	2

Altri esempi di derivati peptidici sono pagina gruppo amminico e / o carbossilici modificati polipeptidi di un gruppo amminico, per esempio, Polipeptidi, ad esempio sono modificato con un acido carbossilico o un radicale alchilico, modificato o un gruppo di acido carbossilico con un gruppo amminico oppure un gruppo estereo.	0,54	99	1	1	1	1
Gli acidi nucleici vengono come DNA e RNA, soprattutto cDNA in questione.	0,54	102	2	2	3	2
Gli acidi nucleici possono essere clonate per la produzione ricombinante di polipeptidi in normali vettori eucariotiche o procariotiche ed espressi in cellule ospiti adatti.	0,54	103	3	2	3	1
In una forma di realizzazione del processo inventivo ad un polipeptide sopra descritto, che è legato ad un vettore, viene aggiunto come un campione da analizzare fluido corporeo.	0,54	120	2	1	4	2
Un altro scopo dell'invenzione è un metodo per il rilevamento in vitro di malattie reumatiche, in particolare artrite reumatoide, in cui la concentrazione di autoanticorpi è determinato in un fluido corporeo.	0,55	115	4	2	4	4
Il metodo permette una diagnosi, classificazione e / o valutare la gravità della malattia.	0,56	116	3	2	4	3
Un altro vantaggio del polipeptide mutato è che questa consente l'erogazione di un agente diagnostico in grado di specificità del 99% e 85% dei pazienti sensibilità con malattie infiammatorie croniche e delle articolazioni 5 e il sistema muscolo-scheletrico, in particolare con artrite reumatoide, identificare.	0,57	110	1	1	2	1
Anticorpi anti-filaggrina dei filtri di tipo IgG con una specificità superiore al 99% un marcatore altamente specifico per l'artrite reumatoide.	0,57	63	3	1	1	1
Un altro scopo della presente invenzione è un kit diagnostico per l'utilizzo nella rilevazione di malattie reumatiche, in particolare artrite reumatoide, che contiene un agente diagnostico sopra descritto.	0,57	112	4	3	4	3

(1) rigidità mattutina delle articolazioni di più di un'ora, (2) artrite di tre o più articolazioni, (3) artrite di almeno tre regioni articolati Allo stesso tempo, (4) sono colpiti da polso o delle articolazioni delle dita, (5) la tenerezza bilaterale articolazioni metacarpo (6) erosioni sulle radiografie, (7) di rilevazione di fattori reumatoidi speciali e anti-perinucleare fattore di positività (APF).	0,58	55	3	1	2	1
Dal APF, alias e anti-filaggrina anticorpi contenenti siero reagiscono allo stesso modo, questi sistemi anticorpi sono apparentemente identiche.	0,58	62	2	1	1	1
o E 'estremamente sorprendente, per indicare che polipeptidi dell'invenzione dimostrano di essere antigeni altamente specifici e altamente sensibili per la diagnosi di anticorpi nei fluidi corporei di pazienti con malattie reumatiche, in particolare le malattie infiammatorie delle articolazioni e del sistema muscolo-scheletrico, in particolare 5 preferibilmente di artrite reumatoide.	0,58	106	3	1	1	1
fattori reumatoidi sono un sottoinsieme di immunoglobuline, che sono caratterizzati da reazione crociata immunologica contro la regione Fc di immunoglobulina G (IgG).	0,59	47	3	4	4	3
5 Un altro scopo della presente invenzione è un agente diagnostico, contenenti uno o più dei polipeptidi descritti sopra o loro frammenti.	0,62	104	2	2	4	3
L'artrite reumatoide è una Autoimmunkrankeit in cui mantenere i meccanismi di difesa del corpo umano erroneamente endogena cartilagine articolare per estranea e ostile e li attaccano.	0,62	38	1	1	1	1
Per ottimizzare i criteri di classificazione esistenti della ACR sono dalle società specializzata sette criteri di classificazione degli Stati Uniti Rheumatology proposte che indicano una prognosi infausta:	0,63	54	2	1	3	2
Gli autoanticorpi contro il cosiddetto fattore anti-perinucleare sono stati descritti da Young et al. descritto in pazienti con artrite reumatoide.	0,64	56	4	2	4	3
Essi possono essere rilevati in circa il 34% dei pazienti fattore negativo reumatoide e di fornire qui un ausilio diagnostico prezioso.	0,64	67	3	2	4	3

I polipeptidi e / o derivati peptidici della stessa, possono anche essere presenti come peptidi ciclici.	0,64	100	4	3	4	3
Reumatologi tenta di utilizzare la stretta finestra di tempo tra l'insorgenza della malattia e la comparsa di danno strutturale.	0,64	43	2	1	4	3
Un altro scopo della presente invenzione è un acido nucleico che codifica un polipeptide sopra descritto.	0,64	101	4	3	4	3
Scopo della presente invenzione fornire nuovi polipeptidi per la rilevazione di associati con malattie reumatiche, artrite reumatoide particolare associato per fornire anticorpi che un sensibile e specifica diagnosi, la classificazione e la prognosi di malattie reumatiche, in particolare di dolori articolari permettere e il sistema muscolo-scheletrico.	0,65	69	2	1	2	1
Uno scopo dell'invenzione è un polipeptide codificato da vimentina nativa con la SEQ ID No. 1 è derivato e confrontata con la sequenza nativa differisce dalla presenza di almeno un residuo di arginina supplementare.	0,65	75	3	1	3	2
Fluidi corporei preferiti secondo l'invenzione sono sangue, siero o plasma, più preferibilmente è siero.	0,65	107	2	2	4	3
Un altro scopo della presente invenzione è un frammento della menzionata sopra Polipeptidi, la vimentina nativo avendo No. SEQ ID 1 è derivato e l'almeno contiene una regione con almeno un residuo di arginina e che presenta reattività contro le malattie associate con autoanticorpi reumatoide.	0,66	92	1	1	1	1
La presenza di fattore reumatoide, tuttavia, non si limita a malattie di tipo reumatico (indizi diagnostici differenziali), a trovare fattori reumatoidi nel siero dei pazienti con malattie infettive, malattie linfoproliferative, Hyperglobulinanämien B-ZeII- e in generale nella popolazione anziana.A103	0,66	48	2	2	3	2
Più preferibilmente, il frammento è nella gamma di posizioni 30-70.	0,67	94	3	3	4	3
Un altro scopo della presente invenzione sono derivati peptidici della suddetta Polipeptidi o frammenti.	0,67	97	2	1	3	2

L'agente di diagnostica, il polipeptide o un frammento incluso in forma libera o carrier-bound.	0,67	105	2	1	1	1
Questo risultato è inaspettato, poiché si è ipotizzato nella tecnica presuppone che vimentina deve essere citrulliert essere immunologicamente reattive.	0,68	71	1	1	1	1
Le persone sane hanno a soffrire una elevata concentrazione di fattore reumatoide, senza, pazienti senza fattori reumatoidi altra parte hanno una forma molto aggressiva di artrite reumatoide.	0,71	52	1	1	3	1
Le concentrazioni non correlavano con il grado di attività e il successo terapeutico.	0,71	50	3	2	4	4
Numerosi studi clinici hanno mostrato che con composti attivi adatti, per esempio, antagonista del TNF, ottimi risultati terapeutici e tassi di risposta può essere raggiunto quando viene utilizzato in pazienti in una fase precoce.	0,72	42	2	1	2	1
L'invenzione si riferisce a polipeptidi che reagiscono con autoanticorpi reumatismi-associata.	0,73	4	2	3	3	3
L'invenzione si riferisce a polipeptidi che reagiscono con autoanticorpi reumatismi-associata.	0,73	33	4	3	4	3
Gli anticorpi sono indicati per questa ragione oggi come gli anticorpi anti-cheratina (AKA).	0,73	58	4	4	3	2
La proteina 40 kDa aggregati Zytokeratinfilamente e aiuta a formare la matrice di fibre delle cellule cornified intracellulari.	0,73	61	1	1	2	1
Con i metodi convenzionali oggi, gli anticorpi, tuttavia, si trovano solo nel 40% dei casi nel siero.	0,73	68	4	3	4	3
Particolarmente preferite sono le posizioni 41, 58, 59, 60 e / o 68.	0,73	78	4	2	4	4
Nel processo dell'invenzione possono (a) metodi enzymological, (b) metodi basati luminescenza o (c) metodi radiochimici sono utilizzati come metodi di rilevazione sono tutti di uso comune nel settore dei metodi diagnostici, come.	0,74	118	1	1	1	1
Secondo i criteri della ACR del fattore reumatoide dell'indicatore sierologico precedenza base per la diagnosi di artrite reumatoide (RA).	0,74	46	1	1	1	1

nessun agente diagnostico specifico o sensibili comparable è noto dalla tecnica precedente nota ad oggi, che permette l'individuazione di malattie reumatiche o, in particolare, di artrite reumatoide utilizzando una proteina citrullinfreien o peptide.	0,74	111	1	1	1	1
Inoltre, il kit diagnostico può contenere ingredienti convenzionali come tamponi, solventi e / o gruppi di marcatura contenenti.	0,74	113	2	2	3	2
Ad esempio, il derivato peptidico di essere un polipeptide retro / inverso, vale a dire un polipeptide inversa dei polipeptidi sopra descritti, la ("speculare") dei polipeptidi di D-amminoacidi è prodotto secondo un'immagine speculare, un polipeptide retro avente una sequenza "reverse", così come un polipeptide retro-inverso riflette un dei polipeptidi sopra descritti e comprende anche una sequenza "inversa".	0,74	98	1	1	1	1
Reagiscono con anticorpi RA-associati umani e sorprendentemente hanno una maggiore sensibilità e specificità rispetto conosciuta dai peptidi citrullinati dell'arte nota.	0,74	74	3	1	2	1
Come vettori sono macromolecole quali DNA, RNA, medicamente polimeri compatibili, quali polietilene, poli DL-lattide, poli D, L-lattide-co-glycolides, biopolimeri sintetici, come polilisina e destrani e proteine, come l'albumina sierica e emocianina, messo in discussione.	0,74	114	1	1	2	1
Così, la filaggrina proteina basica è stata identificata come l'antigene bersaglio.	0,75	60	4	4	4	4
metodi di rilevamento preferite sono nel processo dell'invenzione, un metodo radioimmunologico, un Chemolumineszenzimmunoassay, un saggio immuno-blot o un immunodosaggio enzimatico, ad esempio un test ELISA, messo in discussione.	0,75	119	3	1	2	1
Sulla base della concentrazione di fattori reumatoidi può nessuna previsione sensibile e specifico per l'inizio di una artrite reumatoide sono soddisfatte.	0,76	51	2	1	2	1
In diversi studi, una correlazione positiva con la severità e il tipo di malattia potrebbe essere trovato.	0,78	65	2	1	3	2
Tuttavia, ad oggi dalla tecnica anteriore, senza rilevamento affidabile e sensibile di artrite reumatoide dal periodo di tempo noto.	0,79	44	1	1	1	1

Altri marcatori sierologici, come gli anticorpi anti-citrullina (PCC) o il punteggio HAQ iniziale, utilizzato per valutare le abilità nella vita quotidiana, o la radiografia o la tomografia computerizzata (CT) -BiId danno in forma in anticipo solo piccoli affioramenti e sono non solo abbastanza significativo al fine di valutare in che modo la prognosi del paziente sarà.	0,81	53	3	1	1	1
Quindi o, dal momento che i polipeptidi contenenti più siti di legame dell'anticorpo, entrambi gli anticorpi monomerici e multimeriche vengono efficacemente legati.	0,81	109	1	1	2	1
Gli anticorpi sono in linea di principio rilevabili precoce e precedono la malattia clinica.	0,82	64	4	3	4	2
Una strategia chiave essenziale in reumatologia moderna applicazione pertanto precoce di farmaci biologici che modificano la progressione della malattia.	0,83	41	2	1	3	1
Grazie alla loro specifica reazione all'epitelio cheratinizzato dello strato corneo per Rattenoesophagusschnitten lungo cheratina è stato considerato corrispondente antigene.	0,83	57	1	1	1	1
Come reagente di rilevamento agente diagnostico descritto sopra o kit diagnostico sopra descritto viene utilizzato.	0,85	117	1	1	1	1
Nell'analisi di legame dell'anticorpo alla vimentina nativo, ad esempio la positività APF o anti-Sa-reattività in forma di mutante varianti immunologicamente reattive sono presenti.	0,87	70	1	1	1	1
Un test di laboratorio che permette questo dolore una tensione muscolare innocuo, l'artrosi o il più comune e più grave delle malattie, l'artrite reumatoide (RA) a te, non è ancora noto.	0,88	37	2	1	1	1
Questa ipotesi potrebbe essere confutata da noi da stati arricchiti da differenziale immunoaffinità cromatografia sequenze immunologicamente vimentin reattiva varianti aver mutato da monociti umani.	0,90	72	1	1	2	1
Generalmente, livelli elevati di fattori reumatoidi sono associati con una malattia più grave.	0,94	49	4	3	4	3
SINTESI.	1,00	3	5	4	4	5
Nei primi mesi della malattia, questo sta procedendo molto rapidamente.	1,00	40	2	1	2	1
D'altra parte, però, il polipeptide può essere un polipeptide citrullinfreies.	1,00	90	3	2	3	2
I reclami.	2,00	8	4	4	5	3

Abstract

Die maschinelle Übersetzung (MÜ) prägt den Alltag von Millionen von Menschen. Das sind nicht nur die Folgen der Globalisierung, sondern ist auch ein Zeichen dafür, dass die online verfügbaren MÜ-Systeme einen starken Aufschwung erleben und eine bessere Qualität aufweisen.

Wie gut sind denn die aktuellen MÜ-Systeme? Diese Frage lässt sich nicht einfach beantworten und ist die zentrale Frage der Evaluation der maschinellen Übersetzung, ein reges Forschungsgebiet, das sich parallel zur MÜ entwickelt hat. Traditionell wird die Qualität der MÜ von (menschlichen) Evaluatoren bewertet. Da die menschliche Evaluation kostspielig, zeitaufwendig und wenig objektiv ist, wurden automatische Methoden entwickelt, welche die Qualität eines Systems berechnen sollten.

Die Benutzung solcher Metriken ist aber sehr umstritten. Sie werden kritisiert, weil sie unzuverlässig sind und - vor allem auf Satzebene - nur bedingt zwischen guten und schlechten Übersetzungen unterscheiden. Ziel dieser Arbeit ist es, anhand eines praktischen Vergleichs von menschlichen und automatischen Evaluationsmethoden versuchen zu verstehen, was hinter den rein numerischen Daten steckt und ob die Werte der Metriken - insbesondere auf Segmentebene - mit menschlichen Evaluationen korrelieren.

Im zweiten Kapitel werden die Grundlagen der Evaluation der maschinellen Übersetzung erarbeitet. Es wird ein umfassender Überblick über die Merkmale, welche die Evaluation ausmachen, gegeben. Ferner wird ausgeführt, welche Qualitätskriterien in den vergangen Jahren verwendet wurden und noch heute die theoretische Basis der Evaluation bilden. Am Ende des Kapitels werden die Standards, die für die Software gelten, sowie ein Framework – das FEMTI – dargelegt, die diese Parameter mit den für eine Übersetzung geltenden Qualitätskriterien verbinden.

Das dritte Kapitel bietet einen Überblick über die menschlichen Evaluationsmethoden. Es werden die Vor-und Nachteile jeder Methode insbesondere im Hinblick auf die Objektivität und Inter-/Intra-Annotator-Übereinstimmung und Informativität der Ergebnisse dargelegt. Im vierten Kapitel werden die gängigsten automatischen Methoden ausgeführt. Es wird der Stand der Technik im Bereich der automatischen Evaluation vorgestellt und die Art und Weise, wie die Metriken wiederum evaluiert werden. Außerdem werden die Meta-Evaluation und ihre Rolle in der Wissenschaftsgemeinde diskutiert.

Das fünfte Kapitel enthält die Beschreibung der im Rahmen dieser Masterarbeit durchgeführten Studie. Es werden die einzelnen Bestandteile des Evaluationsdesigns

detailliert vorgestellt. Der Text, der für die Evaluation ausgewählt wurde, ist ein Fachtext aus dem Bereich der auf der Rheumatologie angewandten Molekularbiologie. Zuerst wird daher die Tauglichkeit von Fachtexten für die maschinelle Übersetzung insbesondere durch die WSD (Word Sens Disambiguation) unterstützt und die Merkmale der medizinischen Sprache werden in Bezug auf die maschinelle Übersetzung erklärt. In einem weiteren Schritt wird die Funktionsweise des ausgewählten MÜ-Systems - Google Translate - erklärt. Nach diesem einführenden Teil werden die Gegenstände der Meta-Evaluation vorgestellt, nämlich die Metriken TER (Translation Edit Rate) und HTER (Translation Edit Rate with Human Targeted Reference) und die menschlichen Evaluationsmethoden, nämlich der Post-Editing-Aufwand und die Bewertung von *fleuncy* und *adequcy*. Zudem werden die Methoden der praktischen Ausführung der Studie dargelegt, d.h. die Aufzeichnung von Post-Editing-Zeit- und Aufwand durch das Tool MateCat und die Benutzung des Tools COSTA für die Bewertung der Endbenutzer.

Im sechsten Kapitel werden die Ergebnisse ausgewertet. Es wird aufgezeigt, wie TER, HTER, Post-Eding-Zeit- und Aufwand und Bewertungen (*fluency* und *adequacy*) in Zusammenhang stehen. Die Analyse erfolgt durch einen segmentweisen Vergleich und wird anhand eines qualitativen Ansatzes durchgeführt.