Dissertationen aus der Philosophischen Fakultät II
der Universität des Saarlandes

Optimierung kommerzieller Translation-Memory-Systeme durch Integration morphosyntaktischer Analyseverfahren

Melanie Weitz

universaar
Universitätsverlag des Saarlandes
Saarland University Press
Presses Universitaires de la Sarre

Melanie Weitz

Optimierung kommerzieller Translation-Memory-Systeme durch Integration morphosyntaktischer Analyseverfahren

universaar
Universitätsverlag des Saarlandes
Saarland University Press
Presses Universitaires de la Sarre

D 291
Dissertation zur Erlangung des akademischen Grades eines Doktors der Philosophie der Philosophischen Fakultäten I und II der Universität des Saarlandes
Dekan: Prof. Dr. Roland Marti

Berichterstatter: Prof. Dr. Johann Haller
Prof. Dr. Josef van Genabith
Prof. Dr. Uwe Reinke

Tag der letzten Prüfungsleistung: 09.01.2017

Universitätsverlag des Saarlandes
Saarland University Press
Presses Universitaires de la Sarre

Postfach 151150, 66041 Saarbrücken

ISBN gedruckte Ausgabe: 978-3-86223-245-1
ISBN Online-Ausgabe: 978-3-86223-246-8
URN: urn:nbn:de:bsz:291-universaar-1675

Projektbetreuung *universaar*: Matthias Müller, Verena Wohlleben

Satz: Melanie Weitz
Umschlaggestaltung: Julian Wichert

Herstellung über: readbox unipress in der readbox publishing GmbH
http://unipress.readbox.net

Gedruckt auf säurefreiem Papier

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Für Jan

Zusammenfassung

Translation Memorys zählen seit ihrer Einführung Anfang der 90er Jahre zu den bedeutsamsten Werkzeugen im Bereich der computergestützten Übersetzung. Umso verwunderlicher ist die Tatsache, dass in den meisten auf dem Markt verfügbaren Translation-Memory-Systemen lediglich ein Vergleich der Zeichenketten zwischen einem neu zu übersetzenden Segment mit einem im Translation Memory gespeicherten ausgangssprachlichen Segment durchgeführt wird, um ähnliche Übersetzungseinheiten aufzufinden. Linguistische Übereinstimmungen bleiben dabei häufig unberücksichtigt, sodass bedeutungsgleiche bzw. -ähnliche Segmente mit unterschiedlicher (morpho-)syntaktischer Struktur nicht oder nur mit einem geringeren Ähnlichkeitswert ausgegeben werden.

Diese Arbeit leistet einen Beitrag zu bereits bestehenden Forschungsbemühungen, bei denen linguistisches Wissen anstelle des konventionellen Zeichenkettenvergleiches zur Ermittlung ähnlicher im Translation Memory gespeicherter ausgangssprachlicher Segmente zum Einsatz kommt. Dazu wurde ein Prototyp in Form eines Plug-ins entwickelt, das in ein kommerzielles Translation-Memory-System eingebunden wird und Ergebnisse des morphosyntaktischen Analyseprogramms MPRO liefert.

Für die Erstellung des Plug-ins wurde u. a. ein Algorithmus zur Ermittlung der längsten gemeinsamen zusammenhängenden Teilzeichenketten unter Anwendung von generalisierten Suffix-Arrays und Arrays, die die Länge der längsten gemeinsamen Präfixe enthalten, ausgearbeitet. Die Datenstruktur der generalisierten Suffix-Arrays wurde durch die Verfasserin dieser Arbeit insofern erweitert, dass zunächst alle identischen sich wiederholenden längsten gemeinsamen zusammenhängenden Teilzeichenketten aus dem generalisierten Suffix-Array als mögliche Match-Partner identifiziert werden, um anschließend – unter Berücksichtigung der geringsten Positionsdifferenz und der Segmentzugehörigkeit – die bestmögliche Übereinstimmung zwischen zwei Segmenten aufzufinden. Ebenso wurden Kriterien zur Filterung der Einträge im selbst konzipierten Translation Memory definiert sowie ein Proximitätsmaß zur Ermittlung von Ähnlichkeitswerten, die dem menschlichen Ähnlichkeitsempfinden nahekommen sollen, erstellt.

Die umfangreiche Evaluierung des entwickelten Systems, die sich in Effektivitäts- und Effizienzmessungen gliedert, demonstriert das Ausmaß der in dieser Arbeit erreichten linguistischen Optimierung kommerzieller zeichenkettenbasierter Translation Memorys.

Abstract

Since their launch at the beginning of the 1990s, translation memories have been one of the most important tools in the field of computer-aided translation. It is all the more surprising that most of the commercially available translation memory systems still compare strings of characters between a source segment stored in the translation memory and the segment to be newly translated in order to retrieve translation units which are considered similar. In most of these systems, linguistic similarities are not taken into account so that semantically identical or similar segments whose (morpho)syntactic structure differs from that of the segment to be translated are not retrieved or only with a lower similarity value as expected.

This dissertation contributes to existing research efforts which pursue the use of linguistic knowledge instead of the conventional string comparison in order to retrieve similar source segments stored in the translation memory. For that purpose, a prototype was developed in the form of a plug-in that is implemented into a commercially available translation memory system and that provides results of the morphosyntactic analysis tool MPRO.

In order to develop the plug-in, an algorithm for finding the longest common substring was designed by using generalized suffix and longest common prefix arrays. The data structure of generalized suffix arrays was enhanced by the author of this thesis by firstly identifying the whole set of identical repeating longest common substrings within the generalized suffix array as potential match partners of each other. By then considering the lowest difference in word position and the segment ID of the identified longest common substrings, the best matching segment is retrieved from the database. Furthermore, filter criteria as well as a proximity measure were defined to search the developed translation memory and to provide similarity values that come close to the human perception of similarity.

Extensive evaluation was done by measuring both the effectiveness and the efficiency of the designed system, which demonstrate the extent of the linguistic optimization of commercially available translation memories reached in this study.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AS	ausgangssprachlich
AS_{iMem}	im iMem-TM gespeichertes ausgangssprachliches Segment
AS_{neu}	neu zu übersetzendes ausgangssprachliches Segment
AS_{ref}	als Referenzmaterial verwendetes ausgangssprachliches Segment
AS_{TM}	im Translation Memory gespeichertes ausgangssprachliches Segment
AT_{neu}	neu zu übersetzender ausgangssprachlicher Text
CAT	computer-aided translation, computer-assisted translation, computergestützte Übersetzung
CBMT	context-based machine translation, kontextbasierte maschinelle Übersetzung
DA	depth array, Tiefen-Array
EBMT	example-based machine translation, beispielbasierte maschinelle Übersetzung
FAHQT	fully automatic high quality translation, vollautomatische qualitativ hochwertige Übersetzung
GLCPA	generalisiertes LCP-Array
GSA	generalized suffix array, generalisiertes Suffix-Array
HAMT	human-aided machine translation, human-assisted machine translation, durch den Menschen unterstützte maschinelle Übersetzung
iMem	Intelligente Translation Memorys
KBMT	knowledge-based machine translation, wissensbasierte maschinelle Übersetzung
LCA	lowest common ancestor, niedrigster gemeinsamer Vorfahr
LCP	longest common prefix, längstes gemeinsames Präfix
LCS	longest common substring, längste gemeinsame Teilzeichenkette

MAHT	machine-aided human translation, machine-assisted human translation, maschinengestützte Humanübersetzung
MÜ	maschinelle Übersetzung
POS	part of speech, Wortart
RBMT	rule-based machine translation, regelbasierte maschinelle Übersetzung
SMT	statistical machine translation, statistische maschinelle Übersetzung
TM	Translation Memory
TMS	TM-System, Translation-Memory-System
ÜE	Übersetzungseinheit
ZS	zielsprachlich
ZS_{iMem}	im iMem-TM gespeichertes zielsprachliches Segment
ZS_{neu}	neu erstelltes zielsprachliches Segment, neue Übersetzung eines ausgangssprachlichen Segmentes
ZS_{ref}	als Referenzmaterial verwendetes zielsprachliches Segment
ZS_{TM}	im Translation Memory gespeichertes zielsprachliches Segment

Vorwort

Die vorliegende Arbeit entstand im Rahmen meiner Tätigkeit als Projektmitarbeiterin im Forschungsprojekt *iMem – Intelligente Translation Memorys durch computerlinguistische Optimierung*, das an der Fachhochschule Köln in der Zeit von Juni 2009 bis Juli 2013 durchgeführt wurde. Das Masterstudium *Terminologie und Sprachtechnologie*, das ich zuvor an der Fachhochschule Köln absolviert hatte, weckte in mir die Neugier, herauszufinden, was eigentlich im Hintergrund der verschiedenen Werkzeuge im Bereich der computergestützten Übersetzung abläuft, um gespeicherte Humanübersetzungen weiterzuverarbeiten bzw. als Retrieval-Ergebnis zu erhalten. Mit diesem Forschungsprojekt ergriff ich also die Möglichkeit, in die Untiefen der Translation-Memory-Technologie einzutauchen.

Ich danke all denen, die mich auf diesem langen Weg unterstützt haben. Mein Dank gilt zunächst meinem Doktorvater Prof. Dr. Johann Haller, der mich trotz seiner zum damaligen Zeitpunkt bevorstehenden Emeritierung als Doktorandin angenommen und weiterhin in seinem Ruhestand betreut hat. Dabei habe ich insbesondere seine unkomplizierte Art, seine stets zügigen Beantwortungen meiner Fragen und seine Unterstützung bei bürokratischen Problemen sehr zu schätzen gewusst.

Des Weiteren bin ich Prof. Dr. Josef van Genabith für seine fachlichen Anmerkungen im Bereich der maschinellen Übersetzung und Computerlinguistik sowie für seine Bereitschaft, als weiterer Berichterstatter zu fungieren, zu großem Dank verpflichtet.

Ebenso danke ich Prof. Dr. Uwe Reinke, der mir überhaupt erst ermöglicht hat, das Forschungsprojekt anzutreten, das mir die Grundlagen für meine Promotion gelegt hat, und der mir in linguistischen und Translation Memorys betreffenden Fragestellungen stets beratend zur Seite stand.

Im Bereich der Projektumsetzung danke ich meinem Mann Jan Weitz von ganzem Herzen. Ohne seine Bereitwilligkeit, die Programmierung meines entwickelten Algorithmus, Proximitätsmaßes und Plug-ins zu übernehmen, wäre es mir nicht möglich gewesen, diese Arbeit abzuschließen.

Daneben danke ich meiner Kollegin Melanie Opfer für das umfassende Korrekturlesen meiner Arbeit. Insbesondere ihr und meinem Kollegen Peter Lammers gilt weiterhin mein Dank, da sie mir bei Fragen bezüglich der Evaluierung meines Systems beiseitestanden.

Zuletzt danke ich allen Juroren sowie allen Befragten, die den Online-Fragebogen ausgefüllt haben und auf diese Weise zu der Fertigstellung der vorliegenden Arbeit beigetragen haben.

Teilergebnisse der vorliegenden Arbeit wurden in folgendem Aufsatz vorab publiziert: Weitz, Melanie (2017): „Improving retrieval performance of translation memories using morphosyntactic analyses and generalized suffix arrays“. In: *Machine Translation*. Dordrecht: Springer Science+Business Media B.V., DOI: 10.1007/s10590-017-9193-3.

Köln, 08.06.2017 Melanie Weitz

1 Einleitung

1.1 Problemstellung

Bereits Anfang der 90er Jahre fanden integrierte Übersetzungssysteme, auch *Translation-Memory-Systeme* (*TM-Systeme, TMS*) genannt, auf dem Übersetzungsmarkt Verbreitung (vgl. Kuhns 2007: 4). Seitdem zählen sie zu den zentralen Werkzeugen im Bereich der *computergestützten Übersetzung* (*computer-aided translation, computer-assisted translation, CAT*) (vgl. Somers 2003a: 31, Lagoudaki 2006: 3, Bowker/Barlow 2008: 2).

Den Kern des TM-Systems bildet das *Translation Memory* (*TM*). In den meisten Fällen handelt es sich dabei um einen Übersetzungsspeicher, in dem ausgangssprachliche Segmente (AS_{TM}) mit ihren zielsprachlichen Entsprechungen (ZS_{TM}) archiviert werden. Dies kann während des Übersetzungsprozesses oder im Zuge der Übersetzungsvorbereitung erfolgen. Im Laufe des Übersetzungsprozesses können mithilfe spezieller Retrieval-Mechanismen identische oder ähnliche gespeicherte Segmentpaare aufgefunden und für die Übersetzung des neu zu übersetzenden Textes (AT_{neu}) wiederverwendet werden (vgl. Simard/Langlais 2001: 335). TMs können demnach zu Systemen aus dem Bereich des Information Retrievals gezählt werden. Gemäß Manning et al. (2008) bedeutet der Begriff *Information Retrieval*:

> „finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)." (Manning et al. 2008: 1)

Bezüglich des Funktionsumfangs und der Retrieval-Leistung kommerzieller TMs[1] konnten in den letzten 20 Jahren zahlreiche Weiterentwicklungen beobachtet werden (vgl. Lagoudaki 2006: 1). So zählt der Einsatz von Fuzzy-Matching, das nicht nur das Auffinden identischer, sondern auch im geringen Maße voneinander abweichender AS-Segmente auf Satzebene (z. B. im Falle von orthografischen Varianten) ermöglicht, zu den Fortschritten der 90er Jahre (vgl. Reinke 2004: 37f.). Da jedoch mit Fuzzy-Technologie auf Satzebene

[1] Der Begriff *kommerziell* beschreibt in dieser Arbeit TM-Systeme, die auf dem Markt erworben werden können und dem Zwecke des Gelderwerbs für den Übersetzer dienen (im Gegensatz zu TMs, die lediglich zu Forschungszwecken entwickelt wurden und nicht der breiten Masse zur Verfügung stehen). Auch Open-Source-TM-Systeme werden in dieser Arbeit dem Begriff *kommerziell* zugeschrieben, da sie auch dem Gelderwerb dienen können. Kommerzielle TM-Systeme werden in Kapitel 3.2 näher betrachtet.

nicht immer die gewünschten Recall-Resultate erzielt werden können, werden seit wenigen Jahren auch Algorithmen zur Identifikation identischer Satzfragmente innerhalb eines Satzes, sogenannter Subsegmente, eingesetzt (vgl. Gotti et al. 2005). Allerdings werden „zur Ermittlung von ähnlichen AS Segmenten […] in erster Linie […] einfache Mechanismen zum Vergleich von Zeichenketten [verwendet]“ (Reinke 2004: 115). Aufgrund dessen können AS-Segmente bzw. Subsegmente, die eine Bedeutungsgleichheit bzw. -ähnlichkeit mit dem neu zu übersetzenden ausgangssprachlichen Segment (AS_{neu}) aufweisen, (morpho-)syntaktisch jedoch unterschiedlich strukturiert sind, nicht oder nur mit einem geringeren Ähnlichkeitswert (auch *Match-Wert* genannt) im Vergleich zum menschlichen Ähnlichkeitsempfinden ausgegeben werden (vgl. Reinke 2013: 40, Macklovitch 2000: o.S., Mitkov/Corpas 2008: o.S., Gupta/Orăsan 2014: 4). Dieses Phänomen äußert sich insbesondere im Falle von Paraphrasierung, Verwendung anderer morphosyntaktischer Merkmale (z. B. Unterschiede in Numerus, Kasus, Wortart), Zerlegung von Derivationen und Komposita sowie Umstellung von Teilsätzen in einem Satz (vgl. Kuhns 2007: 5). So wird z. B. den nachfolgend verglichenen AS-Sätzen bei der Übersetzung mit dem kommerziellen zeichenkettenbasierten TM-System SDL Trados Studio 2009 nur ein Match-Wert von 32 % zugewiesen, obwohl trotz der Phrasenvertauschung die Bedeutung beider Sätze gleich bleibt:

AS_{TM}: Nach der Epilation empfehlen wir die Verwendung einer Feuchtigkeitscreme.
AS_{neu}: Wir empfehlen die Verwendung einer Feuchtigkeitscreme nach der Epilation.

Auf Forschungsseite wird daher seit einigen Jahren versucht, die Optimierung der Retrieval-Leistung von TM-Systemen voranzutreiben, indem der zeichenkettenbasierte Vergleich der Datensätze durch den Einsatz linguistischen Wissens abgelöst werden soll (siehe Kapitel 3.3.2). Zwar konnten bereits erste Erfolge erzielt werden, aber trotz der Forschungsbemühungen haben linguistische Verfahren bisher nur in geringem Maße Einzug in kommerzielle TM-Systeme gehalten (siehe Kapitel 3.2.2).

1.2 Das iMem-Forschungsprojekt

Im Zeitraum von Juni 2009 bis Juli 2013 wurde am Institut für Informationsmanagement (IIM) der Fachhochschule Köln[2] das vom Bundesministerium für Bildung und Forschung im Rahmen des Programms *Forschung an Fachhochschulen* geförderte Forschungsprojekt *iMem*[3] *– Intelligente Translation Memorys durch computerlinguistische Optimierung* durchgeführt. Als Projektpartner konnten das Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e. V. an der Universität des Saarlandes (IAI) sowie die Trados GmbH, Stuttgart gewonnen werden.

Das Ziel des iMem-Forschungsprojektes ist die Verbesserung der Retrieval-Leistung kommerzieller TM-Systeme durch den Einsatz morphosyntaktischer Analyseverfahren. Neben der Optimierung des Recalls steht die Ermittlung von Match-Werten im Vordergrund, die mit dem menschlichen Ähnlichkeitsempfinden korrespondieren. Zu diesem Zweck wurde ein Plug-in für die Sprachkombination Deutsch-Englisch (im weiteren Verlauf dieser Arbeit *iMem-TM* genannt) entwickelt, das in ein bestehendes kommerzielles TM-System eingebunden wird und Ergebnisse morphosyntaktischer Analysen der im iMem-TM gespeicherten ausgangssprachlichen Segmente (AS_{iMem}) sowie der AS_{neu} liefert. Da das Retrieval von 100 %-Matches[4] sowie die Identifikation von No-Matches verhältnismäßig simpel ist, liegt das Wiederauffinden unscharfer Übereinstimmungen, sogenannter Fuzzy-Matches, im Fokus des iMem-Forschungsprojektes.

Die Aufgabe der Verfasserin dieser Arbeit bestand dabei in der Identifikation der für den Vergleich eines AS_{neu} mit einem AS_{iMem} notwendigen morphosyntaktischen Merkmale, in der Erarbeitung eines Algorithmus zur Ermittlung der *längsten gemeinsamen Teilzeichenketten* (*longest common substrings, LCS*)[5] zwischen einem AS_{neu} und einem AS_{iMem} einschließlich der Bestimmung von Arbeitsschritten zur Vorfilterung, in der Erstellung und Auslotung eines Proximitätsmaßes zur Ausgabe von dem menschlichen Ähnlichkeitsempfinden entsprechenden Match-Werten, in der Umsetzung des entwickelten Modells in Form eines Plug-ins mit Konstruktion einer relationalen Datenbank als Speicher sowie in der Evaluierung des entwickelten Systems.

---

2 Am 1. September 2015 umbenannt in *Technische Hochschule Köln* (TH Köln).

3 *iMem* ist die Abkürzung für *Intelligente Translation Memorys*.

4 Der Begriff *Match* bezeichnet ein aufgefundenes AS_{TM} bzw. AS_{iMem}, das ähnlich oder identisch zum AS_{neu} ist. Siehe Kapitel 2.2.1.4 zur Erläuterung der verschiedenen Match-Arten.

5 Eine Definition des Begriffes *longest common substring* wird in Kapitel 2.2.1.2 gegeben.

1.3 Methode

Im Folgenden wird die Konzeption des eigens erstellten Systems grob beschrieben. Detaillierte Informationen zu den einzelnen Schritten des Übersetzungsprozesses mit dem System sowie Erläuterungen fachspezifischer Benennungen werden an dieser Stelle nicht gegeben, sondern es wird dafür auf die folgenden Kapitel verwiesen.

Bei dem durch die Verfasserin dieser Arbeit selbst erstellten System, d. h. dem iMem-TM, handelt es sich um eine relationale Datenbank, die als Speicher für die Ergebnisse der morphosyntaktischen Analyse der AS_{iMem} und der AS_{neu}[6] verwendet wird. Als morphosyntaktisches Analyseprogramm wird exemplarisch das am IAI entwickelte Programm *MPRO*[7] verwendet. Als Beispiel für ein kommerzielles TM-System dient SDL Trados Studio 2009[8] mit seiner Trefferanzeige und seinem integrierten, zweigeteilten Übersetzungseditor zur Darstellung des AS_{neu} und seiner zielsprachlichen Entsprechung (*ZS_{neu}, neu erstelltes zielsprachliches Segment*). Die Anbindung der relationalen Datenbank an das kommerzielle TM-System erfolgt über ein Plug-in.

Die Übersetzung eines AS_{neu} sowie das Anzeigen der aufgefundenen Übersetzungseinheiten finden in der Übersetzungsumgebung des kommerziellen TM-Systems statt. Bei der Aktivierung eines AS_{neu} im Übersetzungseditor werden das kommerzielle TM und das iMem-TM parallel nach Matches durchsucht. Der Suchalgorithmus des kommerziellen TMs bleibt dabei unverändert.

Für das Retrieval mithilfe des iMem-TMs wird hingegen die AS-Seite der im iMem-TM gespeicherten Übersetzungseinheiten sowie der AS_{neu} mittels MPRO morphosyntaktisch analysiert. Des Weiteren wurden Arbeitsschritte zur Vorfilterung des iMem-TMs erarbeitet sowie ein auf der Datenstruktur der *generalisierten Suffix-Arrays* (*generalized suffix array*, *GSA*) und *longest-common-prefix-Arrays* (*LCP-Arrays*)[9] basierender Algorithmus zur Ermittlung der LCS in Form der Basiswörter zwischen einem AS_{iMem} und einem AS_{neu} entwickelt. Für die Ermittlung der LCS zwischen einem AS_{iMem} und

---

[6] Die Ergebnisse der morphosyntaktischen Analyse des AS_{neu} werden zunächst in einem Cache vorgehalten. Erst wenn der Übersetzer die neu erstellte Übersetzungseinheit in das iMem-TM übernimmt, wird sie dort als neuer Datensatz gespeichert.

[7] Das Akronym *MPRO* steht für *Morphologisches Programm* (vgl. Maas 1998: o.S.).

[8] Wie SDL Trados Studio 2009 beinhalten auch die Nachfolgeversionen SDL Trados Studio 2011, SDL Trados Studio 2014 und SDL Trados Studio 2015 keine Funktionen zur linguistischen Optimierung der Retrieval-Leistung.

[9] Eine Definition der Begriffe *(generalisiertes) Suffix-Array* und *LCP(-Array)* findet sich in Kapitel 2.2.1.2.

AS_{neu} wurde der bestehende Algorithmus zur Erstellung eines GSAs erweitert, indem zunächst alle identischen sich wiederholenden LCS innerhalb des GSAs identifiziert und daraufhin, unter Berücksichtigung ihrer geringsten Positionsdifferenz und ihrer Segmentzugehörigkeit, die am besten matchenden LCS zwischen den zwei Segmenten herausgefiltert werden. Dabei wird der Fokus auf morphologische Unterschiede und syntaktische Paraphrasen gelegt. Lexikalische Paraphrasen bleiben hingegen unberücksichtigt.

In dem nachfolgenden Beispiel werden die am besten matchenden LCS zwischen zwei verglichenen Sätzen dargeboten. Gleich umrandete Phrasen symbolisieren dabei die miteinander gematchten Wörter. Bei Wiederholungen desselben Wortes (worunter auch Satzzeichen und einzelne Kompositumsbestandteile fallen) werden diejenigen Wörter gematcht, die die geringste Positionsdifferenz zueinander aufweisen (in Abbildung 1: ,).

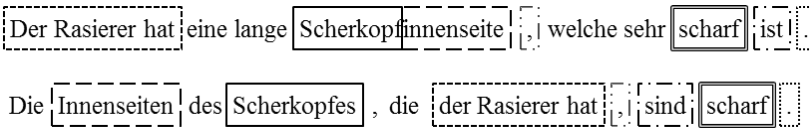

Abbildung 1: Beispiel für einen Vergleich zweier Sätze mit Berücksichtigung morphosyntaktischer Unterschiede sowie von Wortwiederholungen (Weitz 2017)

Mittels des durch die Verfasserin dieser Arbeit im Rahmen des iMem-Forschungsprojektes selbst entwickelten Proximitätsmaßes (nachfolgend auch *iMem-Proximitätsmaß* genannt) wird ein finaler Match-Wert ermittelt, der Aufschluss darüber gibt, wie ähnlich das AS_{neu} den aufgefundenen AS_{iMem} ist.[10]

In der Trefferanzeige erscheinen sowohl alle im kommerziellen TM-System als auch alle im iMem-TM aufgefundenen Übersetzungseinheiten. Welcher Treffer aus welchem System stammt, wird durch eine entsprechende Kennzeichnung in der Trefferanzeige verdeutlicht. Der Übersetzer kann den Match, der für die Übersetzung des AS_{neu} am geeignetsten erscheint, in die ZS-Seite des Übersetzungseditors des kommerziellen TM-Systems einfügen. Auf diese Weise ist eine konsistentere, schnellere und kostengünstigere Übersetzung im Vergleich zu der nicht computergestützten Übersetzung möglich.

Die Problematik beim parallelen Anzeigen der Treffer aus beiden Systemen besteht darin, dass die Match-Werte des iMem-TMs nicht direkt mit den

[10] Die detaillierte Beschreibung der einzelnen Schritte des Algorithmus findet sich in Kapitel 4.2, die Berechnung des Proximitätsmaßes in Kapitel 4.3 und die praktische Anwendung des iMem-TMs beim Übersetzungsprozess in Kapitel 5.2.2.

Match-Werten von SDL Trados Studio 2009 vergleichbar sind, da beide Systeme mit unterschiedlichen Algorithmen arbeiten: Einerseits stellt das iMem-TM nur linguistisch optimiertes Retrieval zur Verfügung, andererseits arbeitet SDL Trados Studio 2009 nur auf Basis des Zeichenkettenvergleiches, wobei sein Match-Algorithmus zudem nicht zugänglich ist. Dies hat auch zur Folge, dass gesetzte Schwellenwerte für die Trefferanzeige (siehe Kapitel 3.1.2.2) relativ sind und die Gegebenheiten eines jeden Algorithmus beim Setzen eines Schwellenwertes berücksichtigt werden müssten.

Um dennoch eine umfangreiche Evaluation des iMem-TMs im Vergleich zu SDL Trados Studio 2009 durchführen zu können, wird angenommen, dass die Match-Werte sowie die Schwellenwerte für die Trefferanzeige beider Systeme vergleichbar sind. Dazu wird der in der Übersetzungspraxis üblicherweise gewählte Schwellenwert von 70 % (vgl. Bruckner/Plitt 2001: 63, SDL plc. 2009–2011: 19, Bloodgood/Strauss 2014: 203) für beide Systeme gesetzt, da die Auffassung besteht, dass bis zu diesem Schwellenwert noch brauchbare Übersetzungseinheiten angezeigt werden (vgl. Vanallemeersch/ Vandeghinste 2014: 94). Ebenso werden für beide Systeme Untersuchungen für einen Schwellenwert von 30 % durchgeführt, um ein umfassendes Retrieval-Ergebnis – unterhalb der 70 %-Grenze – zu erhalten.

1.4 Aufbau der Arbeit

Die Arbeit gliedert sich in zwei Teile: Der erste Teil (Kapitel 2 und 3) befasst sich mit den theoretischen Hintergründen, die für das Verständnis der in dieser Arbeit beschriebenen Entwicklung eines linguistisch optimierten TMs notwendig erscheinen, während im zweiten Teil (Kapitel 4 bis 8) die praktische Phase des iMem-Forschungsprojektes, d. h. die Beschreibung des selbst entwickelten linguistisch optimierten TMs dargelegt wird.

Zunächst wird in Kapitel 2 der Ähnlichkeitsbegriff in der Linguistik definiert, wobei zwischen der Ähnlichkeit aus der Sicht des Menschen und der Ähnlichkeit aus informatischer Sicht unterschieden wird. Um die Ähnlichkeit zweier zu vergleichender AS-Segmente/AS-Sätze/AS-Phrasen etc. durch den Computer zu ermitteln, muss ein sogenanntes String-Matching erfolgen, auf das im Unterkapitel 2.2.1 genauer eingegangen wird. In Verbindung hiermit werden die Segmentierungsmöglichkeiten eines AT_{neu}, die für das iMem-TM relevanten Datenstrukturen und Proximitätsmaße sowie die unterschiedlichen Match-Arten in TMs vorgestellt. Da das Matching auf Basis semantischer Beziehungen zum derzeitigen Kenntnisstand noch nicht in TMs angewendet wird, erfolgt ebenso eine Beschreibung dieses Match-Verfahrens; demnach

soll es als Anregung zur weiteren Forschung im Bereich der Integration semantischer Analysen in TMs dienen.

Kapitel 3 beinhaltet einen Überblick über die Grundlagen der TM-Technologie. Dazu wird die Entstehungsgeschichte von TMs aufgegriffen, gefolgt von der Beschreibung der für diese Arbeit notwendigen Komponenten eines TM-Systems. Im Zuge dessen wird auf die unterschiedlichen TM-Ansätze sowie auf die Voraussetzungen für den Einsatz eines TMs eingegangen. Weitere Komponenten von TM-Systemen werden kurz aufgelistet.

Da die Entwicklung von TMs aus den Unzulänglichkeiten maschineller Übersetzungssysteme hervorgegangen ist, wird im Unterkapitel 3.1.3 das Gebiet der *maschinellen Übersetzung* (*MÜ*) und deren Berührungspunkte mit TMs angerissen, indem Begriffe sowie die Voraussetzungen und Anwendungsgebiete für den Einsatz eines MÜ-Systems erläutert werden. Daran knüpft die Beschreibung aktueller Trends in der Forschung zur MÜ an, bei der unterschiedliche Ansätze und Systemvarianten aufgeführt werden.

In den darauffolgenden Unterkapiteln werden Beispiele kommerzieller TM-Systeme und Forschungsprojekte zur Optimierung von TMs vorgestellt, indem jeweils zwischen nicht linguistisch optimierten und linguistisch optimierten TM-Systemen unterschieden wird.

In Kapitel 4 wird ein Modell eines linguistisch optimierten TMs vorgeschlagen. Dazu wird kurz auf die dafür notwendigen Komponenten eingegangen, gefolgt von der Beschreibung des durch die Verfasserin dieser Arbeit weiterentwickelten Algorithmus.

Der Algorithmus wird schrittweise erläutert: Zunächst werden sechs selbst definierte Arbeitsschritte zur Vorfilterung erläutert, um die Anzahl der AS_{iMem}, mit denen das AS_{neu} verglichen werden muss, so gering wie möglich zu halten. Auf diese Weise sollen die Rechenzeiten akzeptabel gestaltet und der Übersetzer nicht mit unbrauchbaren Übersetzungsergebnissen überflutet werden. Anschließend wird die Erstellung der GSAs beschrieben, mit deren Hilfe die LCS in Form der Basiswörter zwischen den beiden zu vergleichenden Segmenten ermittelt werden sollen.

Im Anschluss daran wird das durch die Autorin dieser Arbeit selbst entwickelte Proximitätsmaß zur Ermittlung des Match-Wertes zwischen einem AS_{neu} und einem AS_{iMem} vorgestellt, wobei die Berechnung einer jeden dafür notwendigen Variablen erläutert wird.

In Kapitel 5 wird sich mit der Realisierung des in Kapitel 4 vorgeschlagenen Modells befasst. Zunächst werden die dafür konkret verwendeten Komponenten aufgeführt, wobei insbesondere auf das morphosyntaktische Analyseprogramm MPRO sowie auf das selbst konzipierte iMem-TM eingegangen wird. Daraufhin wird die praktische Anwendung des iMem-TMs

dargelegt, indem die Benutzeroberfläche und der Übersetzungsprozess beschrieben werden. Dabei werden u. a. die MPRO-Merkmale aufgelistet, die für eine linguistische Optimierung eines kommerziellen TMs als notwendig erachtet werden.

Das Kapitel endet mit der Nennung von Unterschieden zwischen dem iMem-TM – inklusive des selbst entwickelten Algorithmus und Proximitätsmaßes – und anderen linguistisch optimierten TMs, die sowohl kommerzieller als auch forschungsbezogener Natur sind.

Kapitel 6 handelt von der Evaluierung des iMem-TMs. Das iMem-TM wird evaluiert, indem sowohl seine Effektivität als auch seine Effizienz gemessen wird.

Die Effektivitätsmessung gliedert sich in vier Teile: Zunächst werden Evaluierungsmaße aus dem Information Retrieval für sortierte Retrieval-Ergebnisse berechnet, was sowohl für das iMem-TM als auch für SDL Trados Studio 2009 auf Grundlage eines selbst erstellten Korpus geschieht. Die Maße werden für beide Systeme für einen vordefinierten Schwellenwert der Trefferanzeige von 30 % sowie von 70 % ermittelt und verglichen.

Daraufhin wird mittels einer vorher durchgeführten Relevanzbestimmung untersucht, wie viele relevante und nicht relevante Übersetzungseinheiten sowie welche linguistischen Unterschiede in spezifischen Match-Wert-Bereichen des jeweiligen Systems vorkommen.

Mithilfe eines selbst erstellten Online-Fragebogens werden danach die Match-Werte, die für ausgewählte AS_{neu}-AS_{TM}-Segmentpaare und AS_{neu}-AS_{iMem}-Segmentpaare ermittelt wurden, dem menschlichen Ähnlichkeitsempfinden gegenübergestellt: Die befragten Personen mussten beurteilen, ob die Match-Werte ihrem Ähnlichkeitsempfinden entsprechen. Die Ergebnisse der Befragung werden statistisch ausgewertet.

Neben den oben genannten Effektivitätsmessungen, die einen starken Bezug zum Vergleich der Match-Werte haben, wird der Nachbearbeitungsaufwand der besten Matches beider Systeme für ausgewählte im SDL- und iMem-TM gespeicherte zielsprachliche Entsprechungen eines AS_{TM} bzw. AS_{iMem} zur Erstellung der Übersetzung eines AS_{neu} durch unabhängige Juroren bewertet. Auf diese Weise werden nicht die Match-Werte, sondern die ausgegebenen Übersetzungseinheiten berücksichtigt.

Für die Effizienzmessung werden die Antwortzeit und der Speicherplatzbedarf verschieden großer iMem-TMs untersucht.

Kapitel 7 beinhaltet die Diskussion, in der eine kritische Auseinandersetzung mit dem iMem-TM erfolgt. Es werden sowohl positive Aspekte des Systems erläutert als auch Verbesserungsvorschläge aufgeführt.

Abschließend wird in Kapitel 8 ein Ausblick über Erweiterungsmöglichkeiten des derzeitigen Prototyps des iMem-TMs sowie über den zukünftigen Einsatz linguistisch optimierter TMs in der Übersetzungsbranche gegeben.

2 Der Ähnlichkeitsbegriff in der Linguistik

2.1 Das menschliche Ähnlichkeitsempfinden

Ähnlichkeit spielt eine zentrale Rolle in unserem alltäglichen Leben. Sie steuert unser Verhalten und Denken, indem wir Neues mit bereits gemachten Erfahrungen vergleichen. Durch das ständige Vergleichen gelingt es uns, Dinge bzw. Situationen zu klassifizieren, zu beurteilen oder aus ihnen zu lernen.

Der Begriff *Ähnlichkeit* unterliegt keiner eindeutigen Definition. Zwar wird unter *Ähnlichkeit* bzw. *ähnlich* im Allgemeinen „in bestimmten Merkmalen übereinstimmend" (Dudenredaktion 2006: 115) verstanden, doch um das menschliche Ähnlichkeitsempfinden zu beschreiben, ist diese allgemeine Definition zu weit gefasst bzw. unvollständig formuliert. Ob zwei oder mehr Objekte, Personen, Sachverhalte etc. als ähnlich empfunden werden, hängt vielmehr von der subjektiven Wahrnehmung des Einzelnen ab – basierend auf zuvor angeeignetem Weltwissen (vgl. Schmitt 2006: 119f.).

Die verschiedensten Fachgebiete beschäftigen sich mit dem Ähnlichkeitsbegriff und ziehen unterschiedliche Aspekte für die Ähnlichkeitsbestimmung heran. So gelten beispielsweise in der Psychologie Objekte etc. als ähnlich, wenn sie ähnliche Reize beim Menschen auslösen (vgl. Schmitt 2006: 215). Es besteht Konsens darüber, dass bei der Ähnlichkeitsempfindung nicht nur die Ermittlung gemeinsamer Merkmale, sondern auch der sich unterscheidenden Merkmale eine essenzielle Rolle spielt (vgl. Gentner/Markman 1994: 152). Der Grad der Ähnlichkeit hängt von der Gewichtung ab, die den gemeinsamen und sich unterscheidenden Merkmalen zugewiesen werden:

> „Naturally, an increase in the measure of the common features increases similarity and decreases difference, whereas an increase in the measure of the distinctive features decreases similarity and increases difference. However, the relative weight assigned to the common and the distinctive features may differ in the two tasks." (Tversky 1977: 339)

Bevor jedoch eine Ermittlung der Gemeinsamkeiten und Unterschiede durchgeführt werden kann, muss festgelegt werden, unter welchen Aspekten die zu vergleichenden Objekte etc. verglichen werden sollen. Demnach können z. B. zwei gleichförmige, verschiedenfarbige Objekte entweder unter dem Aspekt der Formgebung oder unter dem Aspekt der Farbgebung als ähnlich

empfunden werden (vgl. Gentner/Markman 1995: 114). Für tiefer gehende Erklärungen zur Ähnlichkeitsempfindung und -berechnung sowie zu verschiedenen Ähnlichkeitsmodellen in der Psychologie sei auf Sjöberg (1975), Tversky (1977), Tversky und Gati (1978) sowie auf Gentner und Markman (1994, 1995) verwiesen.

Auch in der Linguistik können unterschiedliche Betrachtungsweisen von zu vergleichenden Texten, Sätzen, Wörtern etc. in derselben Sprache zu einem unterschiedlichen Ähnlichkeitsempfinden führen. Vor allem die Einordnung desselben Wortes, Satzes etc. in unterschiedliche sprachliche Ebenen kann häufig Unterschiede im Ähnlichkeitsgrad hervorrufen. Nachfolgend sind Beispiele aufgeführt, bei denen aufgrund der Einordnung in unterschiedliche sprachliche Ebenen eine Ähnlichkeit (oder sogar Identität) und Unähnlichkeit besteht:

- *Synonyme Benennungen:* In lexikalischer Hinsicht unterscheiden sich Wörter voneinander, in semantischer Hinsicht gleichen sie sich jedoch.
  Beispiel: Lauch – Porree
- *Quasisynonyme Benennungen:* In lexikalischer Hinsicht unterscheiden sich Wörter voneinander, in semantischer Hinsicht ähneln sie sich jedoch. Quasisynonyme Benennungen sind in bestimmten Kontexten austauschbar, obwohl sie nicht genau denselben Begriff repräsentieren (vgl. E DIN 2342 2004–2009: 11).
  Beispiel: Tasche – Beutel
- *Homonyme Benennungen:* Entgegengesetzt zu synonymen Benennungen, d. h., in lexikalischer Hinsicht gleichen sich Wörter, in semantischer Hinsicht unterscheiden sie sich jedoch voneinander; auch Mehrdeutigkeit bzw. Ambiguität genannt.
  Beispiel: Boxer (Sportler) – Boxer (Hunderasse)
- *Vertauschung von Satzgliedern:* In syntaktischer Hinsicht unterscheiden sich die Sätze voneinander, in semantischer Hinsicht ähneln oder gleichen sie sich jedoch.
  Beispiel: Heute regnet es. – Es regnet heute.
- *Unterschiedlicher Kontext*: In syntaktischer Hinsicht gleichen sich die Sätze, in pragmatischer Hinsicht unterscheiden sie sich jedoch.
  Beispiel: Das Wetter ist heute ja super. (Fröhliche Stimme; das Wetter ist wirklich gut.) –
  Das Wetter ist heute ja super. (Ironische Stimme; das Wetter ist eigentlich schlecht.)

Die oben aufgeführten Beispiele gelten nur stellvertretend für eine Reihe solcher Paarungen. Auffällig ist jedoch, dass die Beurteilung der Ähnlichkeit auf

semantischer und pragmatischer Sprachebene wesentlich für das menschliche Ähnlichkeitsempfinden ist (vgl. Reinke 1999: 105).

Semantische Ähnlichkeit kann dabei in Bedeutungsähnlichkeit und konzeptuelle Ähnlichkeit aufgegliedert werden. Während sich die Bedeutungsähnlichkeit mit der Austauschbarkeit von Benennungen oder ganzer Phrasen, Teilsätze etc. beschäftigt, umfasst die konzeptuelle Ähnlichkeit die Kategorisierung der Inhalte (vgl. Reinke 1999: 105). Insbesondere die Bedeutungsähnlichkeit spielt für diese Arbeit eine große Rolle, da – wie bereits erwähnt – das Ziel darin besteht, bedeutungsgleiche bzw. -ähnliche AS_{TM} beim Vergleich mit einem AS_{neu} ausfindig zu machen und die Match-Werte dem menschlichen Ähnlichkeitsempfinden anzugleichen.

Ein linguistisches Phänomen, das der Bedeutungsähnlichkeit zugeordnet werden kann, ist das Paraphrasieren. Paraphrasen führen häufig zu niedrigen Match-Werten in TMs und erschweren auch die Beurteilung der Ähnlichkeit zweier Sätze durch den Menschen, weswegen an dieser Stelle dieses linguistische Phänomen kurz erwähnt werden soll. Reinke (2004: 252ff.) unterscheidet zwei Arten von Paraphrasen: Paraphrasen ohne Informationsverlagerung und Paraphrasen mit Informationsverlagerung. Im Falle ersterer Paraphrasenart bleibt trotz Austausch bestimmter Einheiten die Information der zu vergleichenden Sätze identisch; der Kontext bleibt erhalten. Bei Paraphrasen mit Informationsverlagerung erfolgt jedoch eine Verschiebung des Kontextes. Beide Arten von Paraphrasen können dabei auf unterschiedlichen sprachlichen Ebenen auftreten. Eine ausführliche Beschreibung dieses und anderer linguistischer Phänomene, die das menschliche Ähnlichkeitsempfinden beeinflussen (z. B. Mehrdeutigkeiten), findet sich in Reinke (2004: 235ff.).

2.2 Ähnlichkeit aus informatischer Sicht

Im Gegensatz zum menschlichen Ähnlichkeitsempfinden, bei dem vornehmlich die semantische und pragmatische Sprachebene zur Ähnlichkeitsbeurteilung herangezogen wird, steht bei den Retrieval-Mechanismen der meisten auf dem Markt verfügbaren kommerziellen TM-Systeme die formale Ähnlichkeit zwischen einem AS_{neu} und einem AS_{TM} im Vordergrund (vgl. Reinke 1999: 105). Formale Ähnlichkeit beschreibt – in Bezug auf TMs – die Ähnlichkeit der Zeichenabfolge zweier zu vergleichender AS-Segmente.

Die nachfolgenden Unterkapitel konzentrieren sich auf das Suchverfahren des String-Matchings, da dieses in kommerziellen TM-Systemen vorherrscht. Es sei jedoch darauf hingewiesen, dass durchaus auch andere Suchverfahren, wie beispielsweise die semantische Suche, existieren.

2.2.1 String-Matching

Hinter einem Zeichenkettenvergleich steckt stets ein herstellerspezifischer String-Matching-Algorithmus. Er ist wesentlicher Bestandteil für die Konkurrenzfähigkeit eines jeden TMS-Herstellers und ist demnach nicht öffentlich zugänglich (vgl. Sikes 2007: 41), weswegen an dieser Stelle keine genauen Angaben über verwendete String-Matching-Algorithmen kommerzieller TM-Systeme gemacht werden können. Dennoch wird angenommen, dass den meisten herstellerspezifischen String-Matching-Algorithmen die Edit Distance (siehe Kapitel 2.2.1.3) zugrunde liegt (vgl. Bloodgood/Strauss 2014: 202). Generell kann jedoch festgehalten werden, dass zwischen dem exakten String-Matching, bei dem eine spezifische Zeichenkette aufgefunden werden soll, und dem approximativen String-Matching (auch *Fuzzy-Matching* oder *Unscharfe Suche* genannt) unterschieden wird, bei dem auch ähnliche Zeichenabfolgen ermittelt werden können.

String-Matching kann in unterschiedlichen Kontexten angewendet werden, z. B. im medizinischen Bereich durch das Auffinden eines bestimmten Abschnitts auf einer DNA-Sequenz oder im Bereich der Linguistik durch das Auffinden eines bestimmten Wortes innerhalb eines Dokumentes oder eines identischen oder ähnlichen AS_{TM} zu einem AS_{neu}.

Gotti et al. (2005) ziehen die Art des String-Matchings als Merkmal heran, um zwischen drei verschiedenen Generationen von TMs zu unterscheiden: Während TM-Systeme der ersten Generation lediglich ganze Segmente als Übersetzungseinheiten verarbeiten und nur exakte Übereinstimmungen als Übersetzungsergebnisse liefern, verfügen TM-Systeme der zweiten Generation über Retrieval-Mechanismen, die es ermöglichen, auch unscharfe Übereinstimmungen zwischen einem AS_{neu} und einem AS_{TM} aufzufinden. Der Großteil der auf dem Markt verfügbaren TM-Systeme gehört der zweiten Generation an. Gotti et al. (2005) beschreiben weiterhin TM-Systeme der dritten Generation, die nicht nur ganze Segmente, sondern auch Subsegmente als Übersetzungseinheiten berücksichtigen.

Wie bei Gotti et al. (2005) ersichtlich, kann String-Matching – je nach Zweck der Suche – unterschiedlich granulär erfolgen. Suchmaschinen liefern beispielsweise Dokumente, die einen oder mehrere Suchbegriffe enthalten, Rechtschreibprüfprogramme überprüfen hingegen einzelne Wörter auf Richtigkeit und bei der beispielbasierten MÜ kommt String-Matching auf Phrasenebene zum Einsatz (vgl. Forster 2006: 3ff.). String-Matching sowohl auf Satz- als auch auf Phrasenebene findet hingegen u. a. in TMs statt – abhängig vom jeweiligen TM.

2.2.1.1 Segmentierung

Die Sätze, Phrasen etc., die zwischen einem TM und einem AT_{neu} gematcht werden, werden *Segmente* genannt. Esselink (2000) definiert den Begriff *Segment* wie folgt:

> „A segment is a text element, which is considered by the application as the smallest translatable unit, defined by periods, semi-colons, and hard returns. […] Translation memory tools usually allow the user to change and customize segmentation rules […]" (Esselink 2000: 362f.)

Bei der Übersetzung mit einem TM-System wird also der AT_{neu} durch herstellerspezifische Segmentierungsregeln, die vom Anwender verändert oder ergänzt werden können, in einzelne Segmente zerlegt. Im Zuge der Übersetzungsvorbereitung oder während des Übersetzungsprozesses werden die AS_{neu} mit ihren Übersetzungen im TM gespeichert. Dadurch entstehen Segmentpaare, die häufig auch als Übersetzungseinheiten[11] bezeichnet werden; sie setzen sich meistens aus AS-Sätzen mit ihren durch einen Humanübersetzer angefertigten ZS-Entsprechungen zusammen. In diesem Zusammenhang merkt Bowker (2002) Folgendes an:

> „However, not all text is written in sentence form. Headings, list items, and table cells are familiar elements of text, but they may not strictly qualify as sentences. Therefore, many TM systems allow the user to define other units of segmentation in addition to sentences. These units can include sentence fragments or even entire paragraphs." (Bowker 2002: 94)

Eine Übersetzungseinheit muss demnach nicht zwingend ein vollständiger Satz sein, sondern kann auch aus Segmenten unterhalb der Satzebene oder aus einem Zusammenschluss mehrerer Sätze bestehen. Jedoch erfordert das Auffinden von Segmenten unterhalb der Satzebene einen erheblichen Aufwand, wie Macken (2009) erläutert:

> „Translation memory systems working at the sub-sentential level face more challenges than sentence-based systems. In order to suggest matches at a sub-sentential level, the systems must be able to align source and target chunks (a non-trivial task); and must be able to identify (fuzzy) matches at sub-sentential level and have a mechanism to score multiple sub-sentential matches and select the best match." (Macken 2009: 201)

[11] Für eine ausführliche Diskussion über den Begriff *Übersetzungseinheit* siehe Reinke (2004: 177ff.).

Zur Segmentierung können spezielle Computerprogramme, sogenannte Tokenizer, Satzsegmentierer und Parser, eingesetzt werden. Tokenizer teilen Zeichenketten in kleinere Einheiten auf. Cordts (2012: 21) beschreibt sowohl Tokenizer, die Zeichenketten in linguistische Einheiten (meistens Wörter) aufteilen und Tokenizer, die keine linguistischen Einheiten erzeugen. Bei Ersteren kann durch Segmentierungsregeln eine Trennung der Zeichenkette z. B. beim Erscheinen von Leerzeichen oder bestimmten Interpunktionszeichen erfolgen. Keine linguistischen Einheiten entstehen hingegen, wenn die Zeichenkette gemäß einer definierten, gleich bleibenden Anzahl an Zeichen der Länge *N*, sogenannte N-Gramme, segmentiert wird. Diese Art von Tokenizer wird als N-Gram-Tokenizer bezeichnet.

Satzsegmentierer zerlegen Texte hingegen in ganze Sätze (vgl. Quasthoff 1998: 19). Dazu müssen diejenigen Zeichen definiert werden, die ein Satzende kennzeichnen. Ebenso muss eine Liste von Abkürzungen, die z. B. einen Punkt enthalten, hinterlegt werden, um falsche Satzsegmentierungen zu vermeiden.

Bei einem Parser handelt es sich dagegen um ein „Programm, das eine Zeichenreihe in ihre syntaktischen Bestandteile zerlegt und erkennt, ob die Zeichenreihe vorgegebenen Syntaxregeln entspricht oder nicht“ (Lang 2006: 95). Dem Parser liegt also eine Grammatik zugrunde, nach deren Regeln die Zeichenkette segmentiert wird. Neben den Syntaxparsern gibt es auch Phonologie-, Morphologie-, Semantik- und Textparser.

2.2.1.2 Datenstrukturen

Effizientes String-Matching ist abhängig von der Form, in der die zu durchsuchenden Daten strukturiert sind. Cormen et al. (2001) definieren den Begriff *Datenstruktur* wie folgt:

> „A ***data structure*** is a way to store and organize data in order to facilitate access and modifications. No single data structure works well for all purposes […]“ (Cormen et al. 2001: 8, Hervorhebung im Original)

Gängige Datenstrukturen sind Suffix-Bäume und Suffix-Arrays (vgl. Koehn/ Senellart 2010a). Sie ermöglichen das schnelle Auffinden von u. a. der längsten sich wiederholenden Abfolge von Zeichen innerhalb einer längeren Zeichenkette oder der längsten gemeinsamen Abfolge von Zeichen (d. h. des LCS)[12] zwischen mehreren Zeichenketten (vgl. Sung 2010: 59ff.).

[12] Es wird zwischen *longest common substring* und *longest common subsequence* unterschieden. Während der Begriff *longest common substring* die längste gemeinsame Kette aufeinander-

Im Gegensatz zur linguistischen Definition, bei der ein Präfix ein Wortbildungsmorphem vor einer Basis und ein Suffix ein Wortbildungsmorphem nach einer Basis bezeichnet (vgl. Lohde 2006: 14), sind die Begriffe *Präfix* und *Suffix* in der Informatik definiert als „ein beliebig langes Anfangs- [...] und Endstück eines Wortes“ (Erk/Priese 2008: 28), wobei ein Wort wiederum definiert ist als „eine endliche, eventuell leere Folge von Buchstaben“ (Erk/ Priese 2008: 27). Abhängig davon, ob der Suffix-Baum bzw. das Suffix-Array aus einem Wort oder einem Satz erstellt wird, bestehen ihre Suffixe aus einzelnen Zeichen oder ganzen Wörtern.

In dieser Arbeit wird mit der Bezeichnung *Suffix* ausschließlich auf die informatische Definition des Begriffes Bezug genommen, während der Begriff *Präfix* je nach Kontext entweder der informatischen oder linguistischen Definition unterliegt.

Suffix-Bäume

Ein Suffix-Baum T für eine aus n Zeichen bestehende Zeichenkette S ist ein sogenannter gerichteter Baum, der von einer Wurzel ausgeht und n Blätter besitzt (vgl. Gusfield 1997: 90). Die Merkmale eines Suffix-Baumes werden u. a. in Gusfield (1997: 90ff.) und Aluru (2004: 29-1) beschrieben.

Haben zwei Knoten eines von einer Wurzel ausgehenden Baumes einen gemeinsamen, vorangehenden Knoten und ist dieser gemeinsame Knoten zudem der tiefste Knoten des gemeinsamen Pfades, spricht man vom *niedrigsten gemeinsamen Vorfahren* (*lowest common ancestor, LCA*) (vgl. Cormen et al. 2001: 521). In einem Suffix-Baum entspricht das *längste gemeinsame Präfix* (*longest common prefix, LCP*) mehrerer Suffixe der Länge der Kantenbeschriftung, d. h. der Anzahl der Tokens, auf dem Pfad von der Wurzel bis zum LCA (vgl. Sung 2010: 62).

Nachfolgend ist exemplarisch ein Suffix-Baum für die Zeichenkette *titicacasee* grafisch dargestellt. Ein leeres Zeichen *$* wird an die Zeichenkette gehängt, um ihr Ende zu kennzeichnen und somit vorzubeugen, dass ein Suffix auch gleichzeitig ein Präfix eines anderen Suffixes sein kann. Das leere Zeichen kann durch jedes beliebige Zeichen symbolisiert werden, sofern es nicht in der vorangehenden Zeichenkette erscheint (vgl. Gusfield 1997: 91).

Die Ziffern am Ende eines Blattes repräsentieren die Positionen des Suffixes innerhalb der Zeichenkette. Während beispielsweise der LCA für die Suffixe 1 und 3 der Knoten *v* ist und das daraus resultierende LCP über die

folgender Zeichen beschreibt, ist der Begriff *longest common subsequence* als die längste gemeinsame Kette von nicht unbedingt aufeinanderfolgenden Zeichen definiert (vgl. Sung 2010: 61).

Länge 2 verfügt, repräsentiert der Knoten *w* den LCA für die Suffixe 2 und 4 mit einem dazugehörenden LCP der Länge 1.

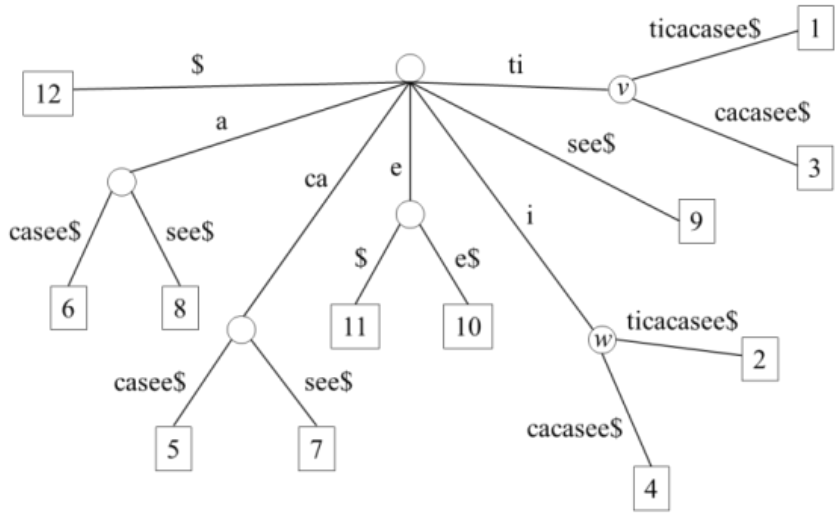

Abbildung 2: Suffix-Baum für die Zeichenkette *titicacasee*

Weiner präsentierte 1973 erstmals die Datenstruktur der Suffix-Bäume. Durch seinen Algorithmus kann die Erstellung von Suffix-Bäumen in linearer Zeit[13] erfolgen, jedoch erfordert sie einen großen Speicherplatzbedarf (siehe Weiner 1973). Im Laufe der Jahre wurden Weiterentwicklungen seines Algorithmus durch McCreight, Ukkonen und Farach hinsichtlich der Optimierung der Speicherkomplexität vorgenommen (siehe McCreight 1976, Ukkonen 1995, Farach 1997).

Suffix-Arrays

Um das Problem des großen Speicherplatzbedarfes eines Suffix-Baumes im Falle von großen Datenmengen zu lösen, wurde von Manber und Myers im Jahr 1990 eine neue Datenstruktur, das sogenannte Suffix-Array, vorgestellt (siehe Manber/Myers 1993).

Ein Suffix-Array ist eine alphabetisch aufsteigend sortierte Liste aller n Suffixe einer Zeichenkette S (vgl. Gusfield 1997: 149). Die geringe Speicherkomplexität kommt dabei durch die Speicherung der Suffixe in Form eines

[13] Dies entspricht der Zeitkomplexität $O(n)$.

Indexes zustande (vgl. Stehouwer/van Zaanen 2010: 505). Ein Suffix-Array kann auf zwei Arten konstruiert werden: entweder unmittelbar oder durch die vorherige Erstellung eines Suffix-Baumes, der anschließend in ein Suffix-Array konvertiert wird (vgl. Abouelhoda et al. 2002: 32, Stehouwer/van Zaanen 2010: 505).

In Tabelle 1 werden zum einen alle Suffixe mit ihren ursprünglichen Positionen innerhalb der Zeichenkette *titicacasee* präsentiert und zum anderen wird das entsprechende Suffix-Array zur selben Zeichenkette dargestellt – bestehend aus der alphabetischen Sortierung der Suffixe und ihren dazugehörigen Positionsnummern.

Auch im Falle von Suffix-Arrays wird das leere Zeichen *$* der Zeichenkette hinzugefügt. In der alphabetischen Sortierung entspricht es dem in der alphabetischen Reihenfolge kleinsten Zeichen und steht folglich an erster Stelle der sortierten Liste (vgl. Sung 2010: 72).

Position	Suffix
1	titicacasee$
2	iticacasee$
3	ticacasee$
4	icacasee$
5	cacasee$
6	acasee$
7	casee$
8	asee$
9	see$
10	ee$
11	e$
12	$

Suffix-Array

Position	Suffix
12	$
6	acasee$
8	asee$
5	cacasee$
7	casee$
11	e$
10	ee$
4	icacasee$
2	iticacasee$
9	see$
3	ticacasee$
1	titicacasee$

Tabelle 1: Suffixe mit ihren Positionen sowie dazugehöriges Suffix-Array für die Zeichenkette *titicacasee*

Um die Suche in einem Suffix-Array nach einer spezifischen Zeichenkette zu beschleunigen und den Aufbau eines Suffix-Baumes in einem Suffix-Array abbilden zu können, kann das Suffix-Array um zusätzliche Listen ergänzt werden. Die daraus resultierende neue Datenstruktur wird *enhanced suffix array* genannt und kann jeden auf Suffix-Bäumen basierenden Algorithmus in derselben Zeitkomplexität wie die eines Suffix-Baumes ersetzen (vgl. Abouelhoda et al. 2004: 53).

Einer der zusätzlichen Indizes[14] ist das sogenannte LCP-Array, das ebenfalls erstmals 1990 von Manber und Myers eingeführt wurde. Das LCP-Array gibt die Länge der LCPs zwischen zwei aufeinanderfolgenden Suffixen an, d. h. zwischen einem Suffix und seinem unmittelbar vorangehenden Suffix innerhalb des Suffix-Arrays (vgl. Manber/Myers 1993: 935, Aluru 2004: 29-2, Stehouwer/van Zaanen 2010: 506). Ein LCP-Array kann während der Erstellung des Suffix-Arrays errechnet werden (vgl. Kasai et al. 2001: 181).

In der nachfolgenden Tabelle wird das LCP-Array für die Zeichenkette *titicacasee* aufgeführt. Mithilfe der Berechnung des LCP-Arrays ist ein schnelles Auffinden der längsten sich wiederholenden Abfolge von Zeichen möglich. So besitzen beispielsweise die zwei im Suffix-Array aufeinanderfolgenden Suffixe *cacasee$* und *casee$* das gemeinsame Präfix *ca*, was einem LCP der Länge 2 entspricht.

Suffix-Array mit LCP-Array

Position	LCP	Suffix
12	-1	$
6	0	acasee$
8	1	asee$
5	0	cacasee$
7	2	casee$
11	0	e$
10	1	ee$
4	0	icacasee$
2	1	iticacasee$
9	0	see$
3	0	ticacasee$
1	2	titicacasee$

Tabelle 2: Suffix-Array und LCP-Array für die Zeichenkette *titicacasee*

Generalisierte Suffix-Arrays

Mit dem bisher beschriebenen Suffix-Array bzw. Suffix-Baum lassen sich die längsten sich wiederholenden Teilzeichenketten innerhalb einer Zeichenkette auffinden. Da beim Retrieval-Vorgang in einem TM jedoch ein AS_{neu} mit einem AS_{TM} und somit *zwei* Segmente miteinander verglichen werden und im iMem-Forschungsprojekt die Idee verfolgt wird, die LCS zwischen beiden

[14] Ein weiterer Index ist die Kind-Tabelle (child table), die 2002 von Abouelhoda et al. eingeführt wurde und ein schnelles Durchsuchen des Suffix-Arrays ermöglicht. Da dieser Index jedoch für die Forschungszwecke des iMem-Forschungsprojektes nicht benötigt wird, sei der interessierte Leser für mehr Informationen zum Thema Kind-Tabelle auf Abouelhoda et al. (2002) sowie Abouelhoda et al. (2004) verwiesen.

Segmenten zu identifizieren, wird eine Datenstruktur benötigt, die den Vergleich zweier Zeichenketten ermöglicht. Bei einer solchen Datenstruktur handelt es sich beispielsweise um ein *generalisiertes Suffix-Array* (*GSA*)[15]. Ein GSA verfügt über dieselben Eigenschaften wie ein einfaches Suffix-Array, jedoch mit dem Unterschied, dass die zu vergleichenden Zeichenketten gemeinsam in einer Tabelle abgebildet werden. Zum Zwecke des Auffindens sich wiederholender Zeichenketten kann das GSA um ein generalisiertes LCP-Array (hier auch *GLCPA* genannt) erweitert werden.

Generalisiertes Suffix-Array mit generalisiertem LCP-Array

ID	Position	LCP	Suffix
I	12	-1	$
II	8	0	#
I	6	0	acasee$
I	8	1	asee$
I	5	0	cacasee$
I	7	2	casee$
I	11	0	e$
II	7	1	e#
I	10	1	ee$
II	6	2	ee#
I	4	0	icacasee$
II	4	1	isee#
I	2	1	iticacasee$
II	2	3	itisee#
I	9	0	see$
II	5	3	see#
I	3	0	ticacasee$
II	3	2	tisee#
I	1	2	titicacasee$
II	1	4	titisee#

Tabelle 3: GSA und GLCPA für die Zeichenketten titicacasee und *titisee*

In Tabelle 3 wird ein GSA mit dem dazugehörigen GLCPA für die Zeichenketten *titicacasee* (ID I) und *titisee* (ID II) vorgestellt. Jede Zeichenkette verfügt über ein anderes leeres Zeichen (für ID I: $, für ID II: #), damit die Suffixe einer jeden Zeichenkette eindeutig bestimmt werden können (vgl. Sung 2010: 58). Verfügen zwei aufeinanderfolgende Suffixe über ein LCP mit

[15] Der Vergleich zweier Segmente kann auch über einen generalisierten Suffix-Baum erfolgen, dessen Konstruktion jedoch an dieser Stelle nicht weiter erklärt wird. Informationen dazu finden sich z. B. in Aluru (2004: 29-3f.).

mindestens der Länge 1 und stammen diese Suffixe aus unterschiedlichen Zeichenketten, so ist das LCP gleichzusetzen mit dem LCS zwischen diesen beiden Zeichenketten. Die längste gemeinsame zusammenhängende Teilzeichenkette zwischen den beiden Zeichenketten *titicacasee* und *titisee* ist demnach die Zeichenabfolge *titi* mit dem LCP der Länge 4.

GSAs und GLCPAs sind für das iMem-Forschungsprojekt von Interesse, da diese Datenstruktur eine schnelle Ermittlung des LCS zwischen zwei Segmenten ermöglicht und einen geringen Speicherplatz erfordert. Letzteres ist besonders von Vorteil, wenn mehrere sehr lange Segmente miteinander verglichen werden müssen.

Auch andere Forschungsgruppen haben sich bereits der Datenstruktur der Suffix-Arrays bedient, um Matches zu einem AS_{neu} aus den in einem TM gespeicherten AS-Segmenten zu ermitteln.

So konzipieren Callison-Burch et al. (2005) ein durchsuchbares TM, das – ähnlich einer bilingualen Konkordanzsuche – eine Liste aller Segmente ausgibt, in denen die gesuchte Benennung bzw. Phrase enthalten ist, jedoch mit dem Unterschied, dass die ZS-Entsprechungen markiert, gruppiert und nach Auftretenshäufigkeit sortiert werden. Wegen der Möglichkeit zur effizienten Speicherung des Indexes sowie zur schnellen Auffindung gesuchter Phrasen unterhalb der Satzebene wurde die Datenstruktur der Suffix-Arrays gewählt. Bei Eingabe einer gesuchten AS-Benennung/AS-Phrase können mithilfe von Suffix-Arrays die Anfangspositionen einer jeden im TM gespeicherten, korrespondierenden AS-Benennung/AS-Phrase innerhalb eines Segmentes bestimmt werden. Mittels Alignment-Methoden der statistischen MÜ kann anschließend die ZS-Entsprechung zur gesuchten AS-Benennung/AS-Phrase aufgefunden werden.

Koehn und Senellart (2010a) verwenden Suffix-Arrays aus Gründen der Zeit- und Speicherplatzkomplexität für das exakte Matching von N-Grammen zwischen einem AS_{neu} und einem AS_{TM}. Beim Match-Vorgang werden die Anfangs- und Endpositionen der N-Gramme im AS_{neu} und AS_{TM} sowie die Segment-ID des AS_{TM} erfasst, um die Edit Distance[16] im Falle eines Fuzzy-Matches zwischen dem AS_{neu} und einem AS_{TM} berechnen zu können. Filtermethoden, wie das Filtern nach Segmentlängenunterschieden, beschränken dabei die Anzahl der matchenden AS_{TM} auf ein Minimum und verbessern auf diese Weise die Leistung des Systems.

Vanallemeersch und Vandeghinste (2014) setzen Suffix-Arrays einerseits ein, um das Matching zwischen einem AS_{neu} und AS_{TM} zu beschleunigen, und

[16] Das Proximitätsmaß *Edit Distance* bzw. *Levenshtein-Distanz* wird in Kapitel 2.2.1.3 genauer beschrieben.

andererseits, um zu vermeiden, dass jedes AS_{neu} mit jedem im TM gespeicherten AS-Segment verglichen werden muss. Der Matching-Vorgang erfolgt auf die gleiche Weise wie in Koehn und Senellart (2010a) beschrieben.

Auch im Falle des iMem-Forschungsprojektes wird zunächst die Anzahl der potenziell matchenden AS_{iMem} mittels mehrerer Arbeitsschritte zur Vorfilterung begrenzt sowie die Positionen der einzelnen Wörter in den zu vergleichenden Segmenten erfasst. Unter Verwendung von GSAs und GLCPAs werden die LCS zwischen dem AS_{neu} und den herausgefilterten AS_{iMem} ermittelt, wobei jedoch nicht die Wortformen der einzelnen Wörter, sondern deren Basiswörter miteinander verglichen werden. Entgegen dem von Callison-Burch et al. (2005) konzipierten durchsuchbaren TM werden im Prototyp des iMem-TMs die ZS-Entsprechungen der ermittelten LCS nicht markiert.

2.2.1.3 Proximitätsmaße

Eine Methode, um die Ähnlichkeit bzw. Unähnlichkeit zwischen zwei Objekten durch ein Computerprogramm auszudrücken, ist die Berechnung von Proximitätsmaßen. Analog zum menschlichen Ähnlichkeitsempfinden werden auch im Falle der Ähnlichkeits-/Unähnlichkeitsberechnung durch den Computer die Merkmale, die die Objekte innehaben, auf ihre Gemeinsamkeiten und Unterschiede untersucht (vgl. Runte 1999: 1). Proximitätsmaße werden in Ähnlichkeitsmaße und Distanzmaße (auch Unähnlichkeitsmaße genannt) unterteilt. Während das Ähnlichkeitsmaß dadurch charakterisiert ist, dass sein Wert umso höher ist, je *ähnlicher* sich die Objekte sind, ist der Wert des Distanzmaßes umso größer, je *unähnlicher* sich die Objekte sind (vgl. Runte 1999: 1). Ähnlichkeits- und Distanzmaße können durch verschiedene mathematische Berechnungen ineinander umgewandelt werden, die in Brosius (1998: 690) genauer erläutert werden.

Ferner wird zwischen zeichenbasierten und tokenbasierten Ähnlichkeits- und Distanzmaßen unterschieden. Bei zeichenbasierten Proximitätsmaßen wird die Ähnlichkeit bzw. Unähnlichkeit ermittelt, indem die einzelnen Zeichen der Zeichenketten nacheinander verglichen werden. Tokenbasierte Proximitätsmaße hingegen geben die Ähnlichkeit bzw. Unähnlichkeit gemessen am Vergleich der einzelnen Tokens (z. B. Wörter) wieder (vgl. Somers 2003b: 19ff., Cordts 2012: 13ff.). Welches Proximitätsmaß eingesetzt wird, ist letztlich von den Datentypen abhängig, die miteinander verglichen werden sollen.

Es existieren zahlreiche Proximitätsmaße, von denen nachfolgend jedoch nur die für diese Arbeit interessanten genauer erörtert werden. Eine umfangreiche Auflistung mit detaillierter Erklärung verschiedener Proximitätsmaße

findet sich u. a. in Cordts (2012: 13ff.), Schmitt (2006: 222ff.), Reinke (2004: 193ff.) sowie Somers (2003b: 19ff.).

Levenshtein-Distanz

Das wohl bekannteste Distanzmaß ist die *Levenshtein-Distanz*, auch *Edit Distance* oder *Editierdistanz* genannt, die Mitte der 60er Jahre vom russischen Mathematiker Vladimir I. Levenshtein entwickelt wurde. Mit ihr wird die kleinste Anzahl an Editieroperationen – bestehend aus Hinzufügungen, Löschungen und Ersetzungen – ermittelt, die benötigt wird, um eine Zeichenkette in eine andere umzuwandeln (siehe Levenshtein 1966).

Den drei Editieroperationen wird eine Kostenfunktion zugeschrieben. Die Kosten betragen dabei für jede Editieroperation 1. Bei Varianten der Levenshtein-Distanz werden dagegen den einzelnen Editieroperationen unterschiedliche Kosten zugewiesen (vgl. Cordts 2012: 13f.). Die Levenshtein-Distanz ist nicht nur auf einzelne Zeichen einer Zeichenkette, sondern auch auf ganze Wörter anwendbar (vgl. Sikes 2007: 41). Häufig wird sie im Bereich der Rechtschreibprüfung oder bei der Plagiatserkennung eingesetzt. Aber auch auf TMs wurde bereits die Levenshtein-Distanz bzw. Varianten von ihr in verschiedenen Forschungsbemühungen angewendet.

So bedienen sich Planas und Furuse (1999) der Edit Distance für ein TM, das aus mehreren miteinander verknüpften Ebenen mit unterschiedlichen Informationsgehalten besteht[17]. Das Proximitätsmaß wird für jede der für den Vergleich eines AS_{neu} mit einem AS_{TM} herangezogenen Ebenen berechnet. Dabei betragen die Kosten für Hinzufügungen und Löschungen den Wert 1, während sich die Kosten für Gleichheit auf 0 belaufen. Als Resultat dieses mehrschichtigen Vergleichens auf Basis der Edit Distance können höhere Recall-Werte bei nur leicht verringerten Precision-Werten und somit bessere Ergebnisse als durch das konventionelle zeichenkettenbasierte TM-System Trados Translator's Workbench verzeichnet werden.

In dem im EURAMIS-Projekt entwickelten TM kommt eine Variante der Levenshtein-Distanz als Proximitätsmaß zum Einsatz. Die Kosten betragen für Ersetzungen der Groß- und Kleinschreibung am Wortanfang 1, für Varianten eines Buchstabens (z. B. für einen Umlaut) 2 und für alle anderen Editieroperationen 4 (vgl. Blatt 1998: 98). Blatt (1998: 98) führt ebenso den Gedanken an, dass Löschungen geringere Kosten zugewiesen werden könnten als Hinzufügungen, da der Bearbeitungsaufwand im Falle von Löschungen häufig geringer ist als im Falle von Hinzufügungen.

[17] Eine genauere Beschreibung der verschiedenen Schichten des TMs von Planas und Furuse findet sich in Kapitel 3.3.2.

Zur Berechnung der minimalen Anzahl an Editieroperationen wird sowohl im mehrschichtigen TM von Planas und Furuse als auch im TM des EURAMIS-Projektes die Methode der Dynamischen Programmierung angewendet (siehe Blatt 1998: 97f., Planas/Furuse 1999: 336, 2000: 623ff.). Dynamische Programmierung ist ein häufig verwendetes Verfahren, wenn es darum geht, die optimale Lösung eines Problems zu finden. Dazu wird das Problem in kleinere Teilprobleme gegliedert, zu diesen Teilproblemen jeweils eine Lösung gefunden, um letztlich die einzelnen Teillösungen wieder zur Lösung des ursprünglichen Problems zusammenzufügen (vgl. Cormen et al. 2001: 323). Die Darstellung des Algorithmus kann in tabellarischer Form erfolgen. Details zur Funktionsweise der Dynamischen Programmierung finden sich in Gusfield (1997: 215ff.) und Cormen et al. (2001: 323ff.).

Die Zeitkomplexität bei der Dynamischen Programmierung beträgt $O(nm)$. Da sich die Zeitkomplexität zur Berechnung der LCS mithilfe von Suffix-Arrays jedoch nur auf $O(n)$ beläuft, wurde sich im Falle des iMem-Forschungsprojektes für diese Datenstruktur entschieden. Der Unterschied in der Zeitkomplexität kommt vor allem beim Vergleich sehr langer Segmente, wie es z. B. beim DGT-TM (siehe Kapitel 6.1) der Fall ist, zu tragen.

Longest common substring als Ähnlichkeitsmaß

Ein Ähnlichkeitsmaß, das ebenso mittels Dynamischer Programmierung oder durch Suffix-Bäume bzw. Suffix-Arrays berechnet werden kann (siehe Kapitel 2.2.1.2), ist der longest common substring. Mit ihm wird die Anzahl an aufeinanderfolgenden Tokens gemessen, die zwei oder mehr Zeichenketten miteinander gemeinsam haben. Die Ähnlichkeit der Zeichenketten ist dabei umso größer, je länger die gemeinsame Teilzeichenkette ist.

Greedy String Tiling

Bei diesem von Wise im Jahr 1993 präsentierten Ähnlichkeitsmaß werden die längsten gemeinsamen Teilzeichenketten zweier zu vergleichender Zeichenketten auf Basis des Karp-Rabin-Algorithmus[18] ermittelt. Das Ergebnis liefert demnach nicht nur die Länge einer gemeinsamen Teilzeichenkette, sondern ggf. mehrerer gemeinsamer, sich nicht überschneidender Teilzeichenketten. Dabei müssen die LCS nicht in der gleichen Reihenfolge innerhalb beider Zeichenketten auftreten, sondern es können auch Positionswechsel der Teilzeichenketten berücksichtigt werden. Längere LCS werden gegenüber

[18] Beim Karp-Rabin-Algorithmus werden die Hash-Werte zweier Zeichenketten miteinander verglichen (siehe Karp/Rabin 1987).

kürzeren LCS bevorzugt, da auch beim Greedy String Tiling die Ähnlichkeit umso größer ist, je länger die gemeinsame Teilzeichenkette ist.

Dazu wird zunächst der längste LCS zwischen den beiden Zeichenketten ermittelt. Die Zeichen dieses LCS werden markiert und können dadurch nicht mehr für die Ermittlung weiterer LCS verwendet werden. Auf diese Weise wird sichergestellt, dass sich wiederholende Zeichenfolgen innerhalb einer Zeichenkette nur einmal gematcht werden. Anschließend wird ein weiterer Durchlauf des Algorithmus zur Berechnung des nächstkürzeren LCS gestartet, wobei jedoch nur die nicht markierten Zeichen berücksichtigt werden. Dieser Vorgang wird so oft wiederholt, bis der kürzeste LCS gefunden wurde. Die Mindestlänge, die ein LCS lang sein darf, wird zuvor definiert und kann entweder eine Länge von 1 oder größer betragen. Letztlich erhält man eine der Länge nach absteigend sortierte Auflistung aller LCS zwischen zwei Zeichenketten. Eine detaillierte Erläuterung zum Greedy String Tiling findet sich in Wise (1993).

Angle of Similarity

Das Konzept des Angle of Similarity wurde 1992 von Carroll veröffentlicht und ist ein Distanzmaß, das auf trigonometrischen Berechnungen basiert. Es kann z. B. beim Vergleich eines Segmentes mit einer Datenbank, in der viele Segmente gespeichert sind, angewendet werden.

Dieses Proximitätsmaß setzt sich aus zwei zu ermittelnden Werten zusammen: einem absoluten Wert, der stark von eventuellen Segmentlängenunterschieden abhängt, sowie einem relativen Wert, bei dem Segmentlängenunterschiede nicht berücksichtigt werden.

Zunächst wird die Distanz (absoluter Wert) zwischen zwei zu vergleichenden Segmenten ermittelt. Dazu wird den einzelnen Wörtern eine Gewichtung je nach Relevanz zugeteilt (z. B. werden Schlüsselwörter stärker gewichtet als Funktionswörter). Zudem werden die Wörter beider Segmente einer morphologischen Analyse unterzogen, wobei ebenso Hinzufügungen, Löschungen und Ersetzungen von Wörtern berücksichtigt werden. Bei den analysierten Merkmalen handelt es sich u. a. um den Wortstamm, die Wortart und den Numerus. Die Segmente werden anschließend anhand der Ergebnisse dieser Analyse miteinander verglichen. Treten Unterschiede in den Merkmalen auf, werden – je nach Relevanz des Merkmals – Kosten berechnet. Alle ermittelten Kosten werden letztlich miteinander addiert. Treffen mehrere Regeln auf ein Wort zu, so gilt stets diejenige, die die geringsten Kosten verursacht.

Beim zweiten Wert (relativer Wert) handelt es sich um den eigentlichen Winkel. Zur Berechnung dieses Winkels werden die Kosten der beiden

Segmente, die beim jeweiligen Vergleich mit dem Nullsatz[19] unter Anwendung der im ersten Schritt beschriebenen Kostenberechnung ermittelt wurden, auf jeweils einer Seite eines Dreiecks abgetragen. Die dritte Seite des Dreiecks ist die Distanz zwischen den beiden zu vergleichenden Segmenten. Unter Anwendung einer Sinusfunktion wird der Winkel zwischen den beiden Segmenten errechnet. Beträgt der Winkel weniger als 30°, werden die Segmente als ähnlich angesehen.

Um die Rechenzeit akzeptabel zu halten, werden zuvor potenziell ähnliche Segmente aus der Datenbank herausgefiltert. Als potenziell ähnlich werden diejenigen Datenbanksegmente erachtet, die mindestens zwei Schlüsselwörter mit dem neuen Segment gemein haben. Siehe Carroll (1992) für eine weiterführende Beschreibung dieses Proximitätsmaßes.

Das iMem-Proximitätsmaß bedient sich Ideen aus den verschiedenen Ansätzen: Das AS_{neu} und jedes herausgefilterte AS_{iMem} werden einer morphosyntaktischen Analyse unterzogen und mit den ermittelten Werten definierter Merkmale (Numerus, Genus etc.) annotiert. Die LCS zwischen den Basiswörtern des AS_{neu} und der AS_{iMem} werden ermittelt, wobei lange LCS gegenüber kurzen LCS bevorzugt werden; sich wiederholende Zeichenfolgen innerhalb einer Zeichenkette bleiben zudem unberücksichtigt. Die morphosyntaktischen Merkmale der LCS zwischen dem AS_{neu} und den AS_{iMem} werden verglichen und im Falle von Unterschieden werden Kosten zugeteilt. Letztlich wird berücksichtigt, ob und wie viele Löschungen bzw. Hinzufügungen stattfinden müssen, um aus einem AS_{iMem} das AS_{neu} zu erzeugen.

2.2.1.4 Match-Arten in Translation Memorys

Beim String-Matching in TMs wird ein AS_{neu} mit einem AS_{TM} verglichen. Kann ein identisches oder ähnliches AS_{TM} aufgefunden werden, handelt es sich um einen Match.

Die Ähnlichkeit zwischen einem AS_{neu} und einem AS_{TM} wird durch den Match-Wert ausgedrückt. Bei diesem Wert handelt es sich um eine Prozentzahl, die je nach TM-System einer spezifischen Berechnungsgrundlage unterliegt (vgl. Sikes 2007: 41 sowie Kapitel 2.2.1). Definierbare prozentuale Abzüge für beispielsweise eine unterschiedliche Formatierung oder automatische Anpassungen können den eigentlichen Match-Wert nachträglich verringern.

Im Folgenden werden die verschiedenen Arten von Matches erläutert, die abhängig vom jeweiligen TM-System existieren und jeweils einen

[19] Unter dem Nullsatz ist ein leeres Segment zu verstehen, d. h. ein Segment ohne jegliche Wörter (vgl. Carroll 1992: 20).

unterschiedlichen Grad der Ähnlichkeit zwischen einem AS_{neu} und einem AS_{TM} widerspiegeln.

100 %-Match

Ein *100 %-Match*, auch *Exact-Match* genannt, ist ein Match, bei dem das AS_{neu} sowohl hinsichtlich der Zeichenabfolge als auch der Formatierung identisch mit einem AS_{TM} ist (vgl. Bowker 2002: 96, Across Systems GmbH 2014: 946).

Moderne TM-Systeme bieten die Option, auch mehrere 100 %-Matches anzuzeigen, wenn z. B. dasselbe AS_{TM} auf verschiedene Weise übersetzt wurde. Dennoch kann ein 100 %-Match in manchen Situationen nicht die passende Übersetzung liefern, z. B. wenn der Kunde Anforderungen an die Übersetzung stellt, die den im TM gespeicherten Übersetzungseinheiten widersprechen, oder wenn der Kontext des 100 %-Matches nicht mit demjenigen des vorhergehenden oder nachfolgenden Segmentes übereinstimmt (vgl. Bowker 2002: 97).

Kontext-Match

Einige TMS-Hersteller haben in den letzten Jahren auf die Problematik des Kontextverlustes in datenbankbasierten TM-Systemen reagiert und eine neue Match-Art eingeführt: den *Kontext-Match* bzw. *101 %-Match*[20].

Ein Kontext-Match ist ein 100 %-Match, der ebenso hinsichtlich seines Kontextes übereinstimmt. Als Kontext ist dabei das vorangehende und nachfolgende Segment zu verstehen. Da die Übersetzungseinheiten im TM durch IDs miteinander verknüpft sind, können das vorangehende und das nachfolgende Segment eines AS_{TM} bestimmt werden (vgl. Across Systems GmbH 2014: 506). Beim Vergleich eines AS_{neu} mit einem AS_{TM} wird folglich geprüft, ob sowohl das vorangehende als auch das nachfolgende Segment eines AS_{neu} der Reihenfolge der im TM gespeicherten Übersetzungseinheiten entspricht (vgl. ATRIL Language Engineering 1993–2003: 573, Across Systems GmbH 2014: 506).

Je nachdem, welche Faktoren bei der Überprüfung des Kontextes beachtet werden, können verschiedene Varianten von Kontext-Matches auftreten.

[20] Je nach TM-System kann sich die Terminologie für die jeweiligen Match-Arten unterscheiden: In Déjà Vu X Professional wird z. B. *perfect match* synonym zu *exact match* verwendet und ein Kontext-Match wird *guaranteed match* genannt (vgl. ATRIL Language Engineering 1993–2003: 572f.). In Across v6 werden No-Matches als *Kein Match* bezeichnet (vgl. Across Systems GmbH 2014: 950).

So verwendet SDL Trados[21] u. a. zusätzlich den PerfectMatch, der dann auftreten kann, wenn ein AS_{neu} mit kompletten, zuvor übersetzten Dokumenten anstelle eines TMs verglichen wird (vgl. SDL plc. 2009–2011: 37). Ein weiteres Beispiel bietet Across mit seinem Kontext- und Struktur-Match, bei dem zwischen dem AS_{neu} und dem AS_{TM} neben dem Kontext auch das Strukturattribut, d. h. die Information zum Segmentursprung innerhalb des Dokumentes (z. B. Überschrift, Tabelle, Schaltfläche), übereinstimmen muss (vgl. Across Systems GmbH 2014: 950).

Full-Match

Bowker (2002) differenziert darüber hinaus zwischen einem *Full-Match* und einem 100 %-Match und definiert erstere Match-Art wie folgt:

> „A full match occurs when a new source segment differs from a stored TM unit only in terms of so-called variable elements, which are sometimes referred to as 'placeables' or 'named entities'. Variable elements include numbers, dates, times, currencies, measurements, and sometimes proper names." (Bowker 2002: 98)

Solche *Placeables*, d. h. „platzierbare und lokalisierbare Elemente" (Azzano et al. 2011: 124), haben die Eigenschaft, dass sie die Übersetzung des restlichen Segmentes nicht oder nur in geringem Maße beeinflussen und dadurch von modernen TM-Systemen beim Retrieval ignoriert werden (vgl. Bowker 2002: 98, Azzano et al. 2011: 125). Ferner ist oft eine Funktion zur automatischen Anpassung der Placeables in TM-Systemen integriert, wodurch zum einen schneller übersetzt werden kann und zum anderen höhere Match-Werte ausgegeben werden (vgl. Azzano et al. 2011: 125). Die Tatsache, dass TM-Systeme automatisch kleinere Angleichungen im Falle von Placeables vornehmen können, bedeutet jedoch nicht, dass TMs maschinelle Übersetzungssysteme sind (vgl. Azzano 2009: 21).

Sofern keine Abzüge innerhalb des TM-Systems für Unterschiede in den Placeables festgelegt wurden, kann das aufgefundene AS_{TM} sogar in einen 100 %-Match resultieren.

Fuzzy-Match

Stimmt das AS_{neu} teilweise mit einem AS_{TM} überein, handelt es sich um einen *Fuzzy-Match* (vgl. ATRIL Language Engineering 1993–2003: 573). Beim Fuzzy-Matching werden komplette Segmente miteinander verglichen, wobei

[21] Die verschiedenen auf dem Markt verfügbaren TM-Systeme werden in diesem Unterkapitel lediglich erwähnt. Genauere Beschreibungen erfolgen in den entsprechenden Unterkapiteln von Kapitel 3.

der Match-Wert die Ähnlichkeit der vollständigen Segmente repräsentiert (vgl. Bowker 2002: 103). Für das iMem-Forschungsprojekt sind vor allem Fuzzy-Matches von großem Interesse, da dem AS_{neu} ähnliche AS-Segmente im TM ausfindig gemacht werden sollen und untersucht werden soll, ob der durch das TM-System berechnete Match-Wert das menschliche Ähnlichkeitsempfinden widerspiegelt.

Abhängig vom gewählten Schwellenwert werden mehr oder weniger Treffer in der Trefferanzeige ausgegeben. Dabei werden die Ergebnisse absteigend nach ihrem Match-Wert sortiert, sodass der Übersetzer einen schnellen Überblick über die ähnlichsten Treffer erhält (vgl. Bowker 2002: 101). In jedem Fall ist die Übernahme der ZS-Entsprechung eines Fuzzy-Matches mit einem gewissen Maß an Überarbeitung verbunden, um eine inhaltliche Übereinstimmung mit dem AS_{neu} zu erzielen (vgl. Bowker 2002: 100).

Subsegment-Match

Im Gegensatz zum Fuzzy-Matching werden beim Subsegment-Matching Einheiten unterhalb der Satzebene für den Vergleich eines AS_{neu} mit den AS_{TM} herangezogen. Die Begründung für die Entwicklung dieser neuen Match-Art kann sowohl in Macklovitch und Russell (2000: 139) als auch in Schäler (2001: 51) gefunden werden: Zum einen können Segmente in ihren Subsegmenten wertvolle Informationen für die Übersetzung des AS_{neu} enthalten, die bei Anwendung des Fuzzy-Matchings auf Satzebene oft nicht aufgefunden werden können. Zum anderen ist die Wahrscheinlichkeit höher, einen 100 %-Match zwischen kurzen Phrasen zu finden als zwischen kompletten, langen Segmenten. Demnach wird für jedes Subsegment ein individueller Match-Wert ermittelt, der unabhängig von der Ähnlichkeit zwischen dem gesamten AS_{neu} und AS_{TM} ist. Die meisten kommerziellen TM-Systeme unterstützen das Subsegment-Matching, darunter Déjà Vu X, memoQ, Systeme der SDL-Trados-Studio-Reihe in Form der AutoSuggest-Funktion[22] sowie Similis.

No-Match

Der Begriff *No-Match* wird verwendet, falls kein identisches oder ähnliches AS_{TM} gefunden werden konnte bzw. falls der Match-Wert eines AS_{TM} unter dem voreingestellten Schwellenwert liegt (vgl. Across Systems GmbH 2014:

[22] Die AutoSuggest-Funktion ist eine zusätzliche Nachschlagefunktion innerhalb der Systeme der SDL-Trados-Studio-Reihe. Ist die AutoSuggest-Funktion aktiviert, werden während des Übersetzens über eine inkrementelle Suche Übersetzungsvorschläge angezeigt.

950). In der Trefferanzeige werden keine Matches dargestellt; der Übersetzer muss das AS_{neu} selbst übersetzen.

Term-Match
Obwohl diese Match-Art keine Treffer aus dem TM zum Gegenstand hat, wird sie der Vollständigkeit halber trotzdem kurz erläutert: Bei einem *Term-Match* handelt es sich um einen Match, bei dem ein Terminus aus dem AS_{neu} mit einem in der Terminologiedatenbank gespeicherten Terminus übereinstimmt (vgl. Bowker 2002: 154). Sofern die Terminologieerkennungskomponente im TM-System aktiviert ist, werden die AS-Termini mit ihren ZS-Entsprechungen in einer gesonderten Trefferanzeige dargestellt. Je nach System und Einstellung werden die gefundenen Termini ebenso im Übersetzungseditor markiert. Ein Match-Wert wird nicht zugewiesen.

In Tabelle 4 werden die unterschiedlichen Match-Arten nochmals zusammengefasst.

Match-Art	**Beschreibung**
100 %-Match	Vollständige Übereinstimmung eines AS_{neu} mit dem AS_{TM}
Kontext-Match	100 %-Match mit zusätzlicher Übereinstimmung des vorangehenden und nachfolgenden Segmentes
Full-Match	Unterschiede nur in Placeables zwischen AS_{neu} und AS_{TM}
Fuzzy-Match	Teilweise Übereinstimmung eines AS_{neu} mit dem AS_{TM}
Subsegment-Match	Übereinstimmung eines AS_{neu} mit dem AS_{TM} unterhalb der Satzebene
No-Match	Keine Übereinstimmung eines AS_{neu} mit dem AS_{TM}
Term-Match	Übereinstimmung eines Terminus aus dem AS_{neu} mit einem in der Terminologiedatenbank gespeicherten Terminus

Tabelle 4: Match-Arten in TM-Systemen

2.2.2 Matching auf Basis semantischer Beziehungen

Auch wenn in den derzeit verfügbaren kommerziellen TM-Systemen die semantische Sprachebene nicht für die Ähnlichkeitsberechnung herangezogen wird, existieren dennoch Anwendungen, die versuchen, semantische Beziehungen zwischen bedeutungstragenden Wörtern innerhalb einer Sprache abzubilden. Bei solchen Anwendungen handelt es sich beispielsweise um lexikalisch-semantische Wortnetze bzw. Thesauri.

In lexikalisch-semantischen Wortnetzen werden zum einen die hierarchischen oder auch nicht hierarchischen Beziehungen eines Begriffes zu anderen Begriffen dargestellt (vgl. Kunze/Lemnitzer 2007: 114f.). Hierarchische

Begriffsbeziehungen werden in Abstraktionsbeziehungen (logische bzw. generische Beziehungen) und Bestandsbeziehungen (partitive Beziehungen) unterteilt (vgl. DIN 2331 1980: 2ff.). Dabei werden Begriffe entweder in Unterbegriffe gegliedert bzw. zu einem Oberbegriff verdichtet oder anderen Begriffen nebengeordnet. Nicht hierarchische Beziehungen sind hingegen Beziehungen zwischen Begriffen, die z. B. einen kausalen oder chronologischen Bezug aufweisen. Zum anderen werden in lexikalisch-semantischen Wortnetzen lexikalische Relationen wie Synonymie oder Antonymie dargestellt. Synonyme Benennungen werden dabei aufgrund ihrer Bedeutungsgleichheit als eine Einheit repräsentiert, d. h. in einem sogenannten Synset[23] zusammengefasst (vgl. Kunze/Lemnitzer 2007: 135f.).

Lexikalisch-semantische Wortnetze können zum Zwecke der Lesartendisambiguierung beispielsweise im Bereich der MÜ oder des Information Retrievals eingesetzt werden. Das bekannteste monolinguale Wortnetz ist das von der Princeton University entwickelte System WordNet für die englische Sprache. Weitere monolinguale Wortnetze sind z. B. GermaNet für das Deutsche oder WoNeF für das Französische. Darüber hinaus existieren multilinguale Wortnetze wie EuroWordNet, bei dem die Synsets verschiedensprachiger monolingualer Wortnetze über einen Index miteinander verknüpft sind (siehe Vossen 1997). Auf diese Weise können fremdsprachige Entsprechungen eines Synsets aufgefunden werden.

Solche Systeme können als Hintergrundmaterial zur Durchführung einer semantischen Suche verwendet werden, bei der Texte mit den im Hintergrundmaterial gespeicherten Ontologien annotiert werden. Somit wird der Inhalt einer Suchanfrage berücksichtigt, wodurch auch synonyme Benennungen und Paraphrasen aufgefunden werden können.

[23] Unter Synset ist die Repräsentation eines Begriffes zu verstehen, die mindestens aus einer lexikalischen Einheit besteht. Die lexikalischen Einheiten eines Synsets müssen immer derselben Wortart angehören (vgl. Kunze/Lemnitzer 2007: 114f.).

3 Translation-Memory-Systeme

3.1 Grundlagen

3.1.1 Historischer Abriss

Als sich 1966 durch das Erscheinen des vom Automatic Language Processing Advisory Committee (ALPAC) erstellten Berichtes zur Zweckmäßigkeit von MÜ-Systemen herausstellte, dass die qualitativ hochwertige vollautomatische MÜ von Texten an ihre Grenzen gestoßen war, zeichnete sich ein vorläufiges Ende der bis dahin stark geförderten Entwicklung von MÜ-Systemen ab. Stattdessen erkannte man, dass die maschinengestützte Humanübersetzung einerseits eine schnelle und kostengünstige alternative Methode der Übersetzung darstellte und andererseits mit ihr ebenfalls die Forderung qualitativ hochwertiger Übersetzungen erfüllt werden konnte (vgl. ALPAC 1966: 32).

Zwei Institutionen stellten diese Entwicklung bereits vor Veröffentlichung des ALPAC-Berichtes fest: die Europäische Gemeinschaft für Kohle und Stahl (EGKS) sowie der Übersetzungsdienst der Bundeswehr (heute Bundessprachenamt genannt).

Die EGKS führte 1963 die terminologische Datenbank DICAUTOM ein, die AS-Termini mit ihren humanübersetzten ZS-Entsprechungen in Form von „phraseologischen Einheiten“ (Europäische Kommission 2009: 66) enthielt. Obwohl dieses System eher den Zweck einer Terminologiedatenbank erfüllte, kann es ansatzweise als Vorreiter heutiger TMs verstanden werden, da mit ihm identische oder ähnliche ausgangssprachliche phraseologische Einheiten mit ihren ZS-Entsprechungen ausgegeben wurden (vgl. ALPAC 1966: 27).

Von 1964 bis 1965 entwickelte der Übersetzungsdienst der Bundeswehr unter Friedrich Krollmann das System „textbezogene Fachwortliste“ (Hoffmann 1982: 186), das 1966 als Übersetzungshilfe eingesetzt und ab 1975 auch LEXIS genannt wurde. Das System war eine multilinguale, begriffsorientierte Terminologiedatenbank, in der Termini isoliert und nicht als phraseologische Einheiten betrachtet wurden (vgl. McNaught 1980: 297).

Als Erweiterung der textbezogenen Fachwortliste sah Krollmann im Jahr 1971 die Entwicklung einer linguistischen Datenbank vor. Sie sollte u. a. über gespeicherte Übersetzungsarchive verfügen (vgl. Krollmann 1971: 118f.), die Parallelen zu heutigen TMs aufwiesen:

> „[…] via descriptors or keywords, large batches of text could automatically be searched for particular passages and then be displayed on video screens as an aid to the translator; […] For revised new editions of translations only the changed passages

> would have to be retyped. Insertion of changes and corrections into the old text would automatically be done by computer […]" (Krollmann 1971: 123)

Daneben machte Krollmann auch auf den Gebrauch eines Mehrbenutzersystems aufmerksam, mit dem mehrere Benutzer über ihre lokalen Rechner gleichzeitig auf Daten zugreifen sollten (vgl. Hutchins 1998: 292f.). Die Umsetzung seiner Vorschläge konnte allerdings aufgrund der damaligen technischen Begebenheiten erst zu einem späteren Zeitpunkt erfolgen.

Anfang der 70er Jahre beabsichtigte Erhard O. Lippmann ebenfalls den Gebrauch eines Mehrbenutzersystems zur Unterstützung des Humanübersetzers durch den Computer, u. a. durch den Einsatz von Wörterbüchern während des Übersetzungsprozesses:

> „The system does not attempt to simulate the human translator by producing an autonomous translation via programmed algorithms; rather, it serves as an extension of the capabilities of the user, who is able to call on the resources of the computer as needed in the process of translation and get an immediate response. [...] In employing the system, the user can switch back and forth as many times as required among human translation, direct dictionary lookup, editing, printing, and system interrogation, and thereby achieve rapid iteration toward the desired goal, i.e., a finished translation." (Lippmann 1971: 10)

Die Unzulänglichkeiten damaliger Textverarbeitungsprogramme waren jedoch auch hier der Grund dafür, dass die Umsetzung von Lippmanns Ideen an ihre Grenzen stieß.

Eine Weiterentwicklung der von Krollmann vorgeschlagenen Übersetzungsarchive wurde 1979 von Peter Arthern vorgenommen. Dazu erstellte Arthern ein Konzept, das er „translation by text-retrieval" (Arthern 1979: 93) nannte. Er sah mit seinem System vor, alle AS-Texte zusammen mit ihren Übersetzungen in einer Datenbank zu speichern, wobei die Möglichkeit bestehen sollte, jede beliebige Textstelle in jeder Sprache in der Datenbank wiederfinden zu können (vgl. Arthern 1979: 94f.). Der Übersetzer sollte lediglich die Textpassagen übersetzen, die nicht im System enthalten waren. Die erstellte Übersetzung sollte schließlich zusammen mit dem Ausgangstext in der Datenbank gespeichert werden. Zudem schlug Arthern die Einbindung eines MÜ-Systems vor (vgl. Arthern 1979: 95). Auch dieses Modell konnte erst ein Jahrzehnt später in lauffähige Anwendungen umgesetzt werden.

Die Grundidee moderner TM-Systeme als integrierte Übersetzungssysteme kann auf Martin Kay durch das Erscheinen seines im Jahre 1980 verfassten Aufsatzes *The Proper Place of Men and Machines in Language Translation* zurückgeführt werden. In ihm beschreibt Kay die Einbindung

mehrerer Übersetzungshilfen in ein einziges System, die nach Bedarf vom Übersetzer aufgerufen werden können. Sein System, das er „The Translator's Amanuensis" (Kay 1980: 14) nannte, soll folgende Übersetzungswerkzeuge enthalten (vgl. Kay 1980: 14ff.): einen zweigeteilten Übersetzungseditor zur Darstellung des originalen und zu übersetzenden Textes, ein Wörterbuch zur Übersetzung unbekannter Wörter, eine Nachschlagefunktion zur Anzeige aller Vorkommnisse desselben Suchbegriffes im Text, eine Funktion zum Markieren unklarer Termini im zu übersetzenden Text sowie eine MÜ-Komponente. Zudem empfiehlt Kay zur Erkennung der Lemmata flektierter Wörter die Durchführung einer morphologischen Analyse durch den Computer.

Ein weiteres Übersetzungswerkzeug wurde 1981 von Alan Melby eingeführt. Dabei handelte es sich um ein bilinguales Konkordanzprogramm, mit dessen Hilfe einerseits Textfragmente (also AS-Einheiten mit ihren ZS-Entsprechungen unterhalb der Satzebene) zur Identifikation potenzieller Übersetzungen eines Suchbegriffes ausfindig gemacht werden konnten. Andererseits diente sein Programm der Überprüfung, ob ein bestimmter Terminus innerhalb eines Textes konsistent übersetzt wurde (vgl. Hutchins 1998: 9f.).

Zur gleichen Zeit arbeitete Melby, unabhängig von Kay, an der Erstellung einer „translator work station" (Melby 1982: 218). Melby beabsichtigte, wie Kay, dass der Übersetzer stets die Kontrolle über den Übersetzungsprozess behält; der Computer sollte lediglich als Unterstützung für den Übersetzer dienen. Dazu beschreibt Melby (vgl. 1982: 217ff., 1992: 147) einen dreistufigen Ansatz, bei dem auf unterschiedliche Übersetzungswerkzeuge (Terminologiedatenbanken, Übersetzungsarchiv, Konkordanzfunktion, MÜ-System etc.) zurückgegriffen werden soll.

Anfang der 80er Jahre setzte schließlich die Firma ALPS die in den letzten zwei Jahrzehnten identifizierten Anforderungen an ein integriertes Übersetzungssystem in ihrem kommerziellen „Translation Support System (TSS)" (Lonsdale 2007: 5) um, wobei es sich um ein multilinguales Textverarbeitungsprogramm mit Nachschlagefunktion handelte. Ferner ermöglichte es das System, Ausgangs- und Zieltext nebeneinander darzustellen, Formatierungen in den Zieltext automatisch zu übernehmen, Quelltextanalysen und Wortzählungen durchzuführen sowie Objekte im Text vollständig zu kopieren (vgl. Hutchins 1988: 27, 1998: 12). Zudem umfasste das System neben der MÜ-Komponente TransActive die Funktion AutoTerm zur automatischen Erstellung eines Glossars von AS-Termini mit ihren Übersetzungen sowie zum Speichern von Übersetzungen, die beim Auftreten eines identischen Segmentes automatisch wiederaufgefunden wurden (vgl. Good 1988: 89f., Hutchins 1998: 12f., Lonsdale 2007: 3ff.). Das Wiederauffinden wurde *repetitions processing* genannt und ermöglichte in seiner weiteren Entwicklung auch das

Auffinden von Fuzzy-Matches mit nur wenigen Unterschieden. Es kann als unmittelbarer Vorreiter heutiger TMs angesehen werden.

Eine weitere Übersetzungsmethode legte 1988 Brian Harris mit dem Konzept der Paralleltexte vor. Harris' Idee, Quell- und Zieltexte miteinander zu verlinken, brachte den Vorteil, dass der Kontext des gesuchten Wortes oder einer größeren Übersetzungseinheit beim Wiederauffinden mitberücksichtigt wurde (vgl. Hutchins 1998: 13f.). Dabei wurden beim Suchen von Übersetzungseinheiten innerhalb der Paralleltexte nicht nur identische, sondern auch ähnliche Ergebnisse angezeigt (vgl. Somers/Fernandez Diaz 2004: o.S.).

Mit dem technischen Fortschritt, den die Computerbranche Anfang der 90er Jahre erlebte, kamen schließlich vier kommerzielle TM-Systeme auf den Markt (vgl. Hutchins 1998: 15):

- Eurolang Optimizer, ein System, das als Nebenprodukt des Eurolang-Forschungsprojektes entstand, in dem primär die Entwicklung eines MÜ-Systems beabsichtigt worden war.
- TranslationManager/2 der IBM Corporation, in dem Lippmanns Ideen umgesetzt wurden.
- Transit der STAR AG, ein referenztextbasiertes Übersetzungssystem.
- Translator's Workbench von Trados. Dieses datenbankbasierte System beinhaltete als Erstes sowohl ein TM als auch eine Alignment-Komponente, mit der der Benutzer seine eigenen Übersetzungsarchive erstellen konnte.

In Abbildung 3 sind die bedeutsamsten Stationen der Geschichte von TM-Systemen aufgeführt.

Im Laufe der letzten 20 Jahre wurden weitere Forschungsprojekte durchgeführt, aus denen TM-Systeme entstanden sind, z. B. das System CTM am Japanischen Institut für Wissenschaft und Technik in Hokuriku oder das System ETOC am IBM Tokyo Research Laboratory (vgl. Reinke 2004: 42f.). Weitere TM-Systeme, die in den letzten 20 Jahren auf den Markt gebracht wurden, sind u. a. Across der Across Systems GmbH, MultiTrans der Multi-Corpora R&D Inc., Déjà Vu von ATRIL Language Engineering, Wordfast von Yves Champollion und memoQ von Kilgray Translation Technologies. Neben letzteren Systemen, deren TMs nur Zeichenketten vergleichen, wurden in den vergangenen Jahren auch TM-Systeme kommerzialisiert, deren TMs mittels linguistischer Analysen angemessenere Matches liefern sollen. Zu diesen linguistisch optimierten TMs zählen ZeresTrans der Zeres GmbH, das jedoch nicht mehr vertrieben wird, sowie Similis der französischen Firma Lingua et Machina (siehe Kapitel 3.2.2). Im Übrigen können einige Systeme als Add-on in ein Textverarbeitungsprogramm geladen werden, während

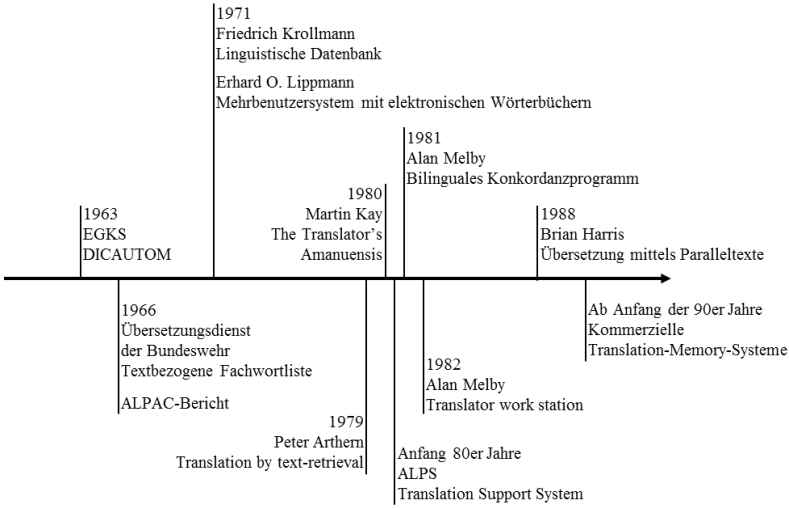

Abbildung 3: Bedeutsamste Stationen der Geschichte von TM-Systemen

integrierte Übersetzungssysteme ihren eigenen Übersetzungseditor zur Verfügung stellen. Des Weiteren existieren Open-Source-TM-Systeme, die ebeso über Funktionen zur linguistischen Optimierung verfügen können. Zu ihren bekanntesten Vertretern gehören OmegaT und OpenTMS bzw. OpenTM2 (siehe Kapitel 3.2.2).

3.1.2 Komponenten eines Translation-Memory-Systems

Waren zu Beginn der Entwicklung von TM-Systemen nur wenige übersetzungsorientierte Anwendungen realisierbar, so bestehen moderne kommerzielle integrierte Übersetzungssysteme aus zahlreichen Komponenten. Je nach Hersteller können die einzelnen Bestandteile mehr oder weniger umfangreich sein. Für welches TM-System sich ein Übersetzer entscheidet, hängt letztlich von seinem individuellen Bedarf ab. Im Folgenden werden die für diese Dissertation relevanten Komponenten von TM-Systemen genauer beschrieben sowie weitere Bestandteile der Vollständigkeit halber kurz umrissen.

3.1.2.1 Übersetzungseditor

Ein essenzieller Bestandteil eines TM-Systems ist der Übersetzungseditor, in dem die Übersetzung eines Dokumentes segmentweise erstellt werden kann. Mithilfe des Übersetzungseditors werden folglich die AS_{neu} mit ihren zugehörigen ZS_{neu} dargestellt.

Generell besteht im Übersetzungseditor die Möglichkeit, das AS_{neu} manuell zu übersetzen, durch Aktivierung des AS_{neu} das TM automatisch zu durchsuchen, eine Übersetzung aus dem TM in die ZS-Seite des Übersetzungseditors einzufügen, diese zu bearbeiten und die neu erstellte Übersetzungseinheit in das TM zu speichern. Abhängig vom gewählten TM-System können Segmente u. a. auch für die Übersetzung gesperrt, mit Kommentaren versehen, in kleinere Segmente geteilt oder miteinander verschmolzen werden. Je nach System gibt es Unterschiede in der Einbindung des Übersetzungseditors (vgl. Zerfaß 2002: 3):

Art der Einbindung	**Merkmale/Funktionsweise**	**Systeme (Beispiele)**
Eigener, integrierter Übersetzungseditor	Konvertierung in das systemeigene Format sowie Segmentierung und tabellarische Darstellung des AT_{neu}.	Systeme der SDL-Trados-Studio-Reihe, Across, memoQ, Déjà Vu, Transit
Kein eigener Übersetzungseditor	Ein Textverarbeitungsprogramm fungiert als Übersetzungseditor, in das das TM als Plug-in geladen wird. Der AT_{neu} muss nicht in das systemeigene Format des TM-Systems konvertiert werden.	Versionen der SDL Trados Translator's Workbench bis einschließlich Version 8, Similis
Eigener Übersetzungseditor für spezifische Dateiformate	Systemeigener Übersetzungseditor für getaggte Dateien, wodurch XML-Elemente geschützt sind. Systemfremdes Textverarbeitungsprogramm für andere Dateiformate.	SDL Trados 2007 mit seiner Komponente TagEditor

Tabelle 5: Unterschiede in der Einbindung des Übersetzungseditors

3.1.2.2 Trefferanzeige

Die Trefferanzeige ist ein Fenster, in dem ähnliche oder identische AS_{TM} absteigend nach ihrem Match-Wert sortiert dargestellt werden. Abhängig vom TM-System können mehr oder weniger Einstellungen für die Trefferanzeige vorgenommen werden. So können beispielsweise Unterschiede zwischen dem AS_{neu} und dem AS_{TM} markiert, der Text für eine bessere Lesbarkeit formatiert oder die Anzeige auf bestimmte Segmente beschränkt werden.

Durch einen vom Anwender definierbaren Schwellenwert wird die Trefferanzeige dahin gehend reduziert, dass nur die brauchbarsten Übersetzungseinheiten dargestellt werden und der Übersetzer nicht mit unbrauchbaren Treffern überflutet wird. Die Wahl des richtigen Schwellenwertes ist nicht immer einfach, wie Bowker (2002) aufzeigt:

> „If the sensitivity threshold is set too high (e.g., a minimum of 95 percent similarity), there is a risk that the TM system will produce 'silence': potentially useful partial

> matches will not be retrieved. However, if the sensitivity threshold is set too low (e.g., a minimum of 10 percent similarity), there is a risk that the TM system will produce 'noise': the suggested translations that are retrieved will be too different from the new source-text segment and therefore will not be helpful. When the threshold is very low, a match may be made on the basis of very general words (e.g., 'the', 'and', 'of", 'to') and the overall content of the retrieved segment may contain little of value for helping the translator to translate the new segment." (Bowker 2002: 99f.)

In der Übersetzungspraxis hat sich ein Schwellenwert für die Trefferanzeige von 70 % als sinnvoller Kompromiss zwischen Silence und Noise etabliert. Trotzdem können auch Matches mit einem geringeren Ähnlichkeitswert nützliche Informationen für die Übersetzung des AS_{neu} beinhalten (vgl. O'Brien 1998: 117).

3.1.2.3 Translation Memory

Das TM bildet den Hauptbestandteil eines TM-Systems. Es wird dazu verwendet, bereits humanübersetztes, identisches oder ähnliches Textmaterial bei der Übersetzung eines AT_{neu} wiederaufzufinden. Je nach System können die zu übersetzenden Segmente komplette Sätze, Subsegmente oder ganze Textabschnitte sein. Werden mit dem AS_{neu} identische oder ähnliche AS_{TM} gefunden, werden dem Übersetzer die gefundenen Übersetzungseinheiten mit ihren jeweiligen Match-Werten in der Trefferanzeige präsentiert – vorausgesetzt, der Match-Wert der Übersetzungseinheit ist gleich oder höher als der definierte Schwellenwert.

Durch die Wiederverwendung aufgefundener Übersetzungseinheiten wird die Arbeit des Übersetzers erleichtert sowie eine schnellere, konsistentere und kostengünstigere Übersetzung gewährleistet. Dennoch ist die Benutzung eines TMs auch mit Nachteilen verbunden, von denen einige in Esselink (2000: 367) aufgeführt sind.

Es wird zwischen zwei Ansätzen der Datenspeicherung und -wiederverwendung unterschieden: dem datenbankbasierten und referenztextbasierten Ansatz. Passend dazu formulieren Macklovitch und Russell (2000) zwei Definitionen (eine engere und eine weitere) von TMs:

> „[…] a translation memory (abbreviated henceforth as TM) is a particular type of translation support tool that maintains a database of source and target-language sentence pairs, and automatically retrieves the translation of those sentences in a new text which occur in the database.
>
> The broader definition regards TM simply as an archive of past translations, structured in such way as to promote translation reuse. This definition, notice, makes no assumptions about the manner in which the archive is queried, nor about the linguistic units that are to be searched for in the archive." (Macklovitch/Russell 2000: 137)

Während die erste Definition den datenbankbasierten Ansatz beschreibt, trifft die zweite, weiter gefasste Definition auch auf den referenztextbasierten Ansatz zu. Im Großteil der auf TMs bezogenen Fachliteratur findet sich oft jedoch nur die Beschreibung des datenbankbasierten Ansatzes. Aus diesem Grund werden im Folgenden beide Ansätze mit ihren generellen Unterschieden zueinander erläutert, auch wenn im weiteren Verlauf dieser Arbeit ausschließlich datenbankbasierte TM-Systeme behandelt werden.

3.1.2.3.1 Datenbankbasierter Ansatz

Beim datenbankbasierten Ansatz, den die meisten TMS-Hersteller verfolgen, ist der Übersetzungsspeicher eine Datenbank, mit der ein AS_{neu} verglichen wird. Vor dem Übersetzungsprozess sollte überlegt werden, wie viele Datenbanken erstellt werden sollen[24]. Einige TM-Systeme bieten darüber hinaus die Möglichkeit, mehrere Datenbanken als Referenz für die Übersetzung eines Dokumentes einzubinden; der Benutzer kann dabei wählen, welches TM er nur als Nachschlagewerk verwenden und in welchem TM er neue Übersetzungseinheiten abspeichern möchte. TM-Systeme, die den datenbankbasierten Ansatz verfolgen, sind z. B. SDL Trados, Across, memoQ und Déjà Vu.

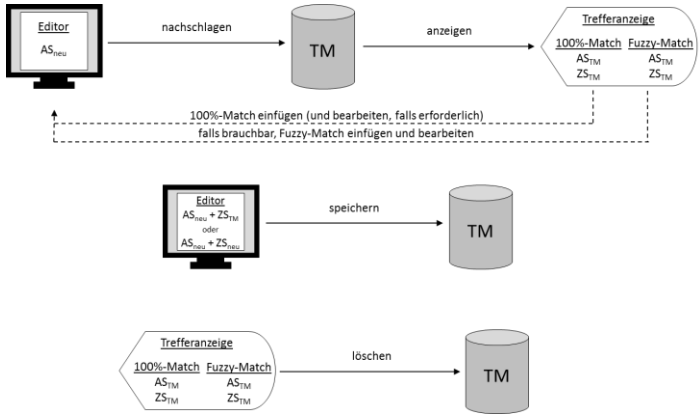

Abbildung 4: Datenbankbasiertes TM: Nachschlagen, Anzeigen, Einfügen, Bearbeiten, Speichern sowie Löschen einer Übersetzungseinheit

[24] Manche Übersetzer speichern alle Übersetzungseinheiten in ein einziges TM, ungeachtet verschiedener Auftraggeber, Projekte etc., während andere Übersetzer beispielsweise für jeden Kunden eine separate Datenbank anlegen. Der Übersetzer hat dabei die Möglichkeit, dem TM bzw. dem Übersetzungsprojekt bestimmte Metadaten (Datum, Kundenname, Projektname etc.) zuzuweisen, mithilfe derer spezifische Übersetzungseinheiten unter Anwendung von Filteroptionen extrahiert werden können.

In Abbildung 4 werden die generellen Funktionen des Nachschlagens, Anzeigens, Einfügens, Bearbeitens, Speicherns und Löschens einer Übersetzungseinheit im Falle eines datenbankbasierten TMs aufgezeigt: Werden durch Nachschlagen des AS_{neu} im TM ein oder mehrere identische oder ähnliche AS_{TM} aufgefunden, werden dem Übersetzer die gefundenen Übersetzungseinheiten abhängig vom definierten Schwellenwert in der Trefferanzeige – nach Match-Wert absteigend sortiert – angezeigt. Der Übersetzer kann daraufhin eine der vorgeschlagenen Übersetzungen in die ZS-Seite des Übersetzungseditors unverändert einfügen, zur Bearbeitung einfügen oder nicht einfügen. Werden keine Übersetzungseinheiten im TM gefunden, muss der Übersetzer das AS_{neu} manuell übersetzen. Die neu erstellten Übersetzungseinheiten können im TM abgespeichert werden. Im Gegenzug können bereits bestehende Übersetzungseinheiten z. B. über die Trefferanzeige wieder aus dem TM gelöscht werden.

Nachfolgend werden die Vor- und Nachteile des datenbankbasierten Ansatzes vergleichend gegenübergestellt:

Vorteile	Nachteile
• ÜEs[25], die während des Übersetzungsprozesses in das TM gespeichert werden, stehen unmittelbar für die weitere Übersetzung desselben Dokumentes zur Verfügung. • Wiederholt sich ein neu gespeichertes AS_{TM} im AT_{neu}, wird die Textstelle automatisch durch die Übersetzung ersetzt[26]. • Neue ÜEs werden sofort bei der Fuzzy-Suche mitberücksichtigt.	• Verlust des Kontextes, in dem sich die ÜE ursprünglich befand, aufgrund von Segmentierung des Textes in einzelne ÜEs und deren isolierter Speicherung in der Datenbank (vgl. Bowker 2002: 97). • Gegebenenfalls daraus resultierender enormer Zeitaufwand für die Korrektur fehlerhafter, bei der Vorübersetzung automatisch eingefügter ZS-Entsprechungen. • Tatsächlicher Nutzen des TMs hängt stark von der Qualität der gespeicherten ÜEs und vom Umfang der Datenbank ab: Je mehr ÜEs im TM vorhanden sind und je sauberer die AS-Segmente ihren ZS-Entsprechungen beim Alignment zugeordnet werden, desto höher ist die Wahrscheinlichkeit, einen brauchbaren Match zu erhalten. • Zusammentragen ausreichend vieler und fehlerloser ÜEs erfordert Zeit, die in der heutigen Arbeitswelt selten zur Verfügung steht (vgl. Gervais 2002: o.S.). • Zeitaufwendige und kostspielige TM-Pflege.

Tabelle 6: Vor- und Nachteile des datenbankbasierten Ansatzes

[25] ÜEs = Übersetzungseinheiten

[26] Diese Funktion wird in vielen TM-Systemen Auto-Propagate-Funktion genannt.

Der Kontextverlust stellt insbesondere bei referentiellen Mehrdeutigkeiten[27] ein Problem dar. So kann beispielsweise das englische Personalpronomen *it* im Deutschen je nach Kontext mit *er*, *sie* oder *es* übersetzt werden, wie das nachfolgende Beispiel demonstriert:

EN: *It* should be cleaned.
DE: *Sie* sollte gereinigt werden. (= die Bürste)
DE: *Er* sollte gereinigt werden. (= der Rasierer)

Die Schwierigkeit des Kontextverlustes und die damit einhergehende zeitaufwendige Korrektur fehlerhafter, bei der Vorübersetzung automatisch eingefügter Übersetzungen schildert Gervais (2002) detailliert:

> „Since TM systems maintain a database of isolated sentences, they lose the surrounding context within which the original sentence was used. The lack of style and usage context results in additional time-consuming translation review and editorial rework because translations built from isolated sentences are more likely to contain inconsistencies or errors. Further, by automatically 'pre-translating', TM systems blindly reuse sentences that might not fit the context of the new project, resulting in poor-quality translations. To improve quality, the translator and others must spend extra time and effort to review, edit and correct – resulting in lost productivity." (Gervais 2002: o.S.)

3.1.2.3.2 Referenztextbasierter Ansatz

Beim referenztextbasierten Ansatz wird ein komplettes AS-Dokument mit seiner Übersetzung im TM-System aligniert und das Ergebnis zum Zwecke der Referenz gespeichert. Das Übersetzungsarchiv besteht folglich nicht aus einer mit Segmentpaaren gefüllten Datenbank, sondern aus miteinander verknüpften bilingualen Texten, die vor dem Übersetzen vom Anwender ausgewählt werden müssen, damit diese als Nachschlagewerk fungieren können. Es können auch mehrere Dokumentpaare als Referenz herangezogen werden. Soll ein Text neu übersetzt werden, wird das definierte Referenzmaterial nach dem AS_{neu} ähnlichen oder identischen Übersetzungseinheiten durchsucht.

Die neueste Generation referenztextbasierter TM-Systeme erstellt zudem während des Übersetzungsprozesses ein temporäres TM, das dazu dient, die während des Übersetzungsprozesses neu erstellten Übersetzungseinheiten vorübergehend zu speichern sowie einmal abgefragte oder neu gespeicherte Übersetzungseinheiten schnell erneut abzurufen und anzuzeigen (vgl. STAR AG 2012: 228). Ebenso können mittels des temporären TMs sich wiederholende Textstellen unmittelbar durch ihre Übersetzungen ersetzt werden.

[27] Eine ausführliche Beschreibung des Problems referentieller Mehrdeutigkeiten und weitere Beispiele finden sich in Reinke (2004: 247ff.).

Nach Beendigung des Übersetzungsprozesses steht das temporäre TM allerdings nicht mehr zur Verfügung. Sollen die Übersetzungseinheiten aus dem aktuellen Übersetzungsauftrag für nachfolgende Aufträge genutzt werden, muss die soeben erstellte Übersetzung mit dem AS-Dokument aligniert und als Referenzmaterial in das Übersetzungsprojekt eingebunden werden. Zu den Vertretern des referenztextbasierten Ansatzes zählen vor allem die Systeme Transit und MultiTrans.

In den nachfolgenden Grafiken werden das Nachschlagen, Anzeigen, Einfügen, Bearbeiten und Speichern einer Übersetzungseinheit in einem referenztextbasierten TM dargestellt. Die Vorgehensweise gleicht derjenigen des datenbankbasierten Ansatzes, jedoch mit dem Unterschied, dass anstelle einer beständigen Datenbank bilinguale, alignierte Textpaare bzw. das temporär erzeugte TM als Nachschlagewerke herangezogen werden. Ein Löschen von Übersetzungseinheiten aus dem Referenzmaterial ist nicht möglich.

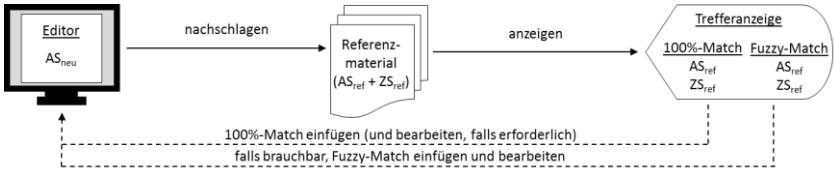

Abbildung 5: Referenztextbasiertes TM: Nachschlagen, Anzeigen, Einfügen und Bearbeiten einer Übersetzungseinheit

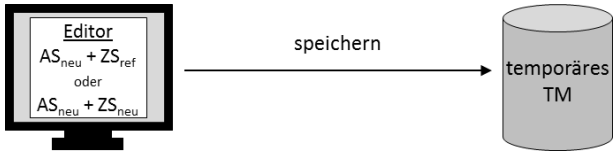

Abbildung 6: Referenztextbasiertes TM: Speichern einer Übersetzungseinheit in ein temporäres TM

Wie der datenbankbasierte Ansatz verfügt auch der referenztextbasierte Ansatz über Vor- und Nachteile, die in der nachfolgenden Tabelle aufgeführt werden:

Vorteile	Nachteile
• Kein Kontextverlust durch Heranziehen der vollständigen Dokumente als Referenzmaterial. Die Segmente bleiben somit in ihrer ursprünglichen Reihenfolge und ihrem Stil erhalten (vgl. MultiCorpora R&D Inc. 2011: o.S.). • Keine zeitaufwendige Erstellung einer Datenbank, die genügend ÜEs enthält, damit sie für die Übersetzung brauchbar ist (vgl. Gervais 2002: o.S.). • Abwesenheit einer Datenbank resultiert in – Wegfall zeit- und kostenintensiver TM-Pflege, – Wegfall rechtlicher Fragen, die z. B. das Urheberrecht der Datenbank (d. h., ob sie dem Kunden oder Übersetzer gehört) oder die Datenbanksicherheit betreffen würden (vgl. Lagoudaki 2006: 3, Smith 2009).	• Während des Übersetzungsprozesses neu erstellte ÜEs stehen nicht unmittelbar zur weiteren Übersetzung des AT_{neu} zur Verfügung (sofern die Erstellung eines temporären TMs nicht möglich ist). Erst nach Abschluss der Übersetzung können Ausgangs- und Zieltext als Referenzmaterial für den nächsten Übersetzungsauftrag eingesetzt werden. • Eventuell Entstehung von Inkonsistenzen, falls der AT_{neu} auf mehrere Übersetzer aufgeteilt wird. • Gute Dateistruktur notwendig, damit alle für die Übersetzung benötigten Dateien schnell aufgefunden und nicht versehentlich vergessen werden (vgl. Sack-Kastl 2007: 13).

Tabelle 7: Vor- und Nachteile des referenztextbasierten Ansatzes

3.1.2.3.3 Voraussetzungen für den Einsatz eines Translation Memorys

Um die Leistungsfähigkeit eines TMs voll ausnutzen zu können, sollte das ausgangssprachliche Textmaterial einen hohen Grad an Wiederholungen aufweisen. Lagoudaki (2006: 16) führt in ihrer Studie zu TM-Systemen an, dass technische Texte (Bedienungsanleitungen, Handbücher, Online-Hilfen etc.) für die Übersetzung mit einem TM prädestiniert sind. Grund dafür sind häufige Redundanzen, die simple Satzstruktur und der große Umfang an Terminologie, was für diese Textsorte charakteristisch ist. Texte aus dem Fachgebiet Recht sind hingegen wegen ihres geringeren Grades an Wiederholungen und ihrer komplexen Satzstrukturen weniger für die Bearbeitung mit einem TM geeignet.

Andererseits müssen die zu übersetzenden Texte elektronisch und in einem vom TM bearbeitbaren Dateiformat vorliegen, beispielsweise im Word-, Excel- oder PowerPoint-Format. Kundentexte werden häufig im PDF-Format bereitgestellt, jedoch eignet sich dieses Format selten für die Übersetzung mit einem TM. Obwohl einige TM-Systeme über ein PDF-Konvertierungsprogramm verfügen, kann nie ausgeschlossen werden, dass der Text fehlerhaft konvertiert wird. Dies kann dazu führen, dass die anschließende Segmentierung des Textes in kleinere Übersetzungseinheiten durch das TM ebenso fehlerhaft ist.

3.1.2.4 Weitere Komponenten eines Translation-Memory-Systems

TM-Systeme bieten neben dem Übersetzungsspeicher, dem Übersetzungseditor und der Trefferanzeige noch weitere Bestandteile, um effizientes Arbeiten zu gewährleisten. Da diese zusätzlichen Bestandteile jedoch nicht unmittelbar zum Forschungsgegenstand dieser Arbeit zählen, wird nur ein kurzer Überblick über die relevantesten weiteren Komponenten gegeben (Tabelle 8). Bowker (2002), Esselink (2000), Zerfaß (2002), Somers (2003c), Massion (2005), Seewald-Heeg (2005) sowie Azzano (2009) sind Beispiele für weiterführende Literatur, in der sich mit den Komponenten integrierter Übersetzungssysteme detaillierter auseinandergesetzt wird.

Komponente	**Funktion**
Terminologieerkennungskomponente	Zur automatischen Suche in einer mit dem TM-System verknüpften mono- oder multilingualen Terminologiedatenbank und Darstellung von Termini, die während des Übersetzungsprozesses im AS_{neu} aufgefunden werden.
Terminologieextraktionskomponente	Zur Filterung mono- oder bilingualer Termini aus einem Dokument mit Möglichkeit zur anschließenden Speicherung in einer Terminologiedatenbank.
Bilinguale Konkordanzsuche	Zur manuellen Suche eines im TM gespeicherten Segmentes oder Subsegmentes bis hin zu einzelnen Worten.
Alignment-Programm	Zur Zuordnung bereits übersetzter AS-Segmente zu ihren ZS-Entsprechungen. Die daraus resultierenden Übersetzungseinheiten können anschließend in ein TM übernommen werden.
Analysefunktion	Zur Ermittlung der Wortanzahl des AT_{neu} sowie der Anzahl an Matches zwischen dem AT_{neu} und dem TM.
Funktion zur automatischen Vorübersetzung	Zur automatischen Übersetzung der AS_{neu} mit ZS_{TM}, vorausgesetzt, die aufgefundenen Übersetzungseinheiten verfügen über einen entsprechend hohen Match-Wert.
Projektmanagementkomponente	Zur Preiskalkulation, Überwachung des Übersetzungsprozesses und Rechtevergabe.
Filteroptionen	Zum Exportieren von Übersetzungseinheiten oder terminologischen Einträgen aus dem TM-System sowie zum Importieren von Übersetzungseinheiten oder terminologischen Einträgen in dasselbe oder in ein anderes TM-System.
Qualitätssichernde Maßnahmen	Zur Überprüfung der ZS_{neu} auf sprachliche und stilistische Richtigkeit.
MÜ-Komponente	Moderne TM-Systeme bieten eine zusätzliche Komponente zur automatischen Übersetzung eines AS_{neu} an, für den Fall, dass kein Match im TM gefunden wird.

Tabelle 8: Weitere Komponenten eines TM-Systems mit jeweiliger Funktionsbeschreibung

3.1.3 Maschinelle Übersetzung und Berührungspunkte mit Translation Memorys

3.1.3.1 Begriffserklärungen

Die Tatsache, dass TM-Systeme eine MÜ-Komponente integriert haben können, deutet darauf hin, dass ein essenzieller Unterschied zwischen einem TM und einem MÜ-System besteht. Esselink (2000) verdeutlicht diesen Unterschied wie folgt:

> „Translation Memory (TM) should not be confused with machine translation (MT). The major difference is that in machine translation a computer translates the text, whereas in translation memory systems, a computer only stores translated sentences. Where an MT system tries to *replace* a translator, a TM system *supports* and assists the translator with the translation tasks." (Esselink 2000: 394, Hervorhebung im Original)

Während mit einem TM Humanübersetzungen verarbeitet werden, werden bei der MÜ neue Übersetzungen durch das System generiert[28].

Übersetzungsform	**Merkmale/Grad des menschlichen Eingreifens**
Traditionelle Humanübersetzung	Klassische Form des Übersetzungsberufes ohne Zuhilfenahme maschinengestützer Übersetzungswerkzeuge.
Machine-aided human translation (MAHT)	Unterstützung des Übersetzers durch eine Übersetzungssoftware, in der Humanübersetzungen weiterverarbeitet bzw. verwaltet werden (z. B. Rechtschreib- und Grammatikprüfprogramme, bilinguale Online-Wörterbücher, Terminologiedatenbanken). Es werden keine eigenen Übersetzungen durch das System erzeugt, weshalb auch TMs in dieser Sparte angesiedelt werden können.
Human-aided machine translation (HAMT)	Generierung von Übersetzungen durch eine Übersetzungssoftware. Die Aufgabe des Übersetzers besteht im interaktiven Übersetzen, Pre-Editing oder Post-Editing[29]. Der Post-Editing-Vorgang in der MÜ ähnelt der Nachbearbeitung eines ZS_{TM} zur Erstellung des ZS_{neu} und bildet somit einen Berührungspunkt zwischen MÜ und TM.
Fully automatic high quality translation (FAHQT)	Vollautomatische, qualitativ hochwertige Übersetzung ohne jegliches Eingreifen des Menschen, bei der das Wissen über sprachliche Besonderheiten zur Disambiguierung homonymer bzw. homografer Benennungen berücksichtigt werden soll. Trotz Forschungsprojekten zur wissensbasierten MÜ bleibt die FAHQT ein unerreichtes Ziel.

Tabelle 9: Übersetzungsformen mit ihren Merkmalen

[28] Dennoch können humanübersetzte Texte als Grundlage für die MÜ dienen, wie es beispielsweise bei der statistischen MÜ der Fall ist.

[29] Beim interaktiven Übersetzen hat der Anwender die Möglichkeit, in den laufenden Übersetzungsprozess einzugreifen. Beim Pre-Editing wird der AT_{neu} vor der MÜ nach bestimmten sprachlichen Kriterien durch den Übersetzer aufbereitet. Beim Post-Editing korrigiert der Übersetzer lediglich das Ergebnis der MÜ.

Hutchins und Somers (1992: 147ff.) teilen die Humanübersetzung und die MÜ in vier Bereiche ein (siehe Tabelle 9), wobei der Grad des menschlichen Eingreifens in den Übersetzungsprozess als Unterscheidungsmerkmal herangezogen wird. Diese vier Bereiche in Verbindung mit dem Grad der menschlichen und maschinellen Beteiligung am Übersetzungsprozess werden in Hutchins und Somers (1992: 148) auch grafisch dargestellt. Dabei sollte erwähnt werden, dass Hutchins und Somers (1992: 148) sowohl den Bereich der MAHT als auch der HAMT zu der computergestützten Übersetzung zählen, während andere Linguisten – wie Bowker (2002: 4) oder Jekat und Volk (2010: 654ff.) – nur die MAHT der CAT zuordnen.

3.1.3.2 Voraussetzungen für den Einsatz eines maschinellen Übersetzungssystems

Ähnlich wie die Arbeit mit einem TM erfordert auch der Einsatz eines MÜ-Systems bestimmte Voraussetzungen. Freigang (2000: 171ff.) nennt gleich mehrere Bedingungen für eine funktionierende MÜ. Dabei können folgende Parallelen zur Arbeit mit einem TM beobachtet werden:

- Der AT_{neu} muss in elektronischer Form vorliegen.
- Das Dateiformat muss vom System verarbeitbar sein.
- Beim AT_{neu} sollte es sich vorzugsweise um homogene Fachtexte handeln, die einen hohen Grad an Wiederholungen und Terminologie aufweisen.

Literarische Texte eignen sich weder für die Übersetzung mit einem TM noch für die MÜ, da diese Textsorte ein sehr gutes Ausdrucksvermögen erfordert, was MÜ-Systeme nicht bzw. nur in begrenztem Maße leisten können (vgl. Jekat/Volk 2010: 642, Toral/Way 2014: 174ff.).

Bei der MÜ kann stattdessen das Prinzip der kontrollierten Sprache eingesetzt werden, um brauchbare Übersetzungen zu produzieren – es sei denn, das Fachgebiet verfügt ohnehin über eine eindeutige Terminologie und definierte Syntax, wie es beispielsweise bei Wetterberichten der Fall ist[30]. Melby und Warner (1995) definieren den Begriff *kontrollierte Sprache* wie folgt:

> „A controlled language is essentially an artificially defined sublanguage in which authors learn to constrain their writing to conform to a set of rules about syntax and semantics when writing texts within a certain pragmatics (that is, for a given purpose, to a particular audience, and within a particular domain).“ (Melby/Warner 1995: 40)

[30] Ein bekanntes MÜ-System zur Übersetzung von Wetterberichten ist Météo, das 1976 von der Forschungsgruppe TAUM in Montréal entwickelt wurde. Eine Beschreibung Météos geben Hutchins und Somers (1992: 207ff.).

Das Ziel der kontrollierten Sprache besteht demnach darin, Texte zu disambiguieren, zu vereinfachen und auch für den Menschen verständlicher zu machen. Da bei diesem Prozess auf die Verwendung einer festgelegten Terminologie und definierter Stilrichtlinien geachtet wird, ist der AT_{neu} konsistenter und leichter wiederverwendbar, auch wenn dies bedeutet, die Arbeit des Autors bzw. des Übersetzers in kreativer Hinsicht einzuschränken. Dennoch werden durch den Einsatz kontrollierter Sprache nicht nur bei der MÜ bessere Ergebnisse erzielt, sondern ebenso bei der Übersetzung mit einem TM (vgl. Nyberg et al. 2003: 246ff.).

Unabhängig davon, ob die mit einem MÜ-System zu übersetzenden Dokumente zuvor durch die Anwendung kontrollierter Sprache optimiert wurden oder nicht, sollten in jedem Fall die AS-Texte orthografisch korrekt sein. Da MÜ-Systeme, im Gegensatz zu TMs, kein Fuzzy-Matching anwenden, sind Erstere auf die orthografische Richtigkeit eines Textes angewiesen, um Wörterbuchabfragen durchführen zu können (vgl. Freigang 2000: 172).

Letztlich muss jedoch abgeschätzt werden, ob sich die MÜ von Dokumenten lohnt. Zwar können mithilfe der MÜ große Textmengen in geringer Zeit übersetzt werden, allerdings ist eine brauchbare maschinell erzeugte Übersetzung immer mit einem gewissen zeitlichen und finanziellen Aufwand hinsichtlich des Pre- oder Post-Editings verbunden. Einen Text maschinell zu übersetzen, lohnt sich daher nur, solange der Aufwand für das Pre- oder Post-Editing geringer ist als derjenige, der für eine sofortige Humanübersetzung bzw. die Übersetzung mit einem TM erforderlich wäre (vgl. Freigang 2000: 173).

3.1.3.3 Einsatzgebiete für maschinelle Übersetzungssysteme

Obwohl sich die MÜ in ihrem derzeitigen Entwicklungsstand weder für die vollautomatische, qualitativ hochwertige Übersetzung noch für die Übersetzung von literarischen Texten eignet, finden sich doch sinnvolle Einsatzgebiete. So können Roh- bzw. Informativübersetzungen[31] dazu dienen, einem der Ausgangssprache nicht mächtigen Leser Anhaltspunkte über den Inhalt eines Textes zu geben. Ist der Inhalt der Rohübersetzung nützlich, kann der Leser entscheiden, ob er die maschinell erstellte Übersetzung nachbearbeiten bzw. eine qualitativ hochwertige Humanübersetzung anfertigen lässt.

Zwei weitere Einsatzgebiete führen Jekat und Volk (2010: 643) an: Zum einen kann die MÜ dabei behilflich sein, große Mengen an fremdsprachigen

[31] Der Begriff *Rohübersetzung* bzw. *Informativübersetzung* bezeichnet die reine MÜ ohne Pre-Editing, Post-Editing oder Verwendung kontrollierter Sprache (vgl. Hutchins/Somers 1992: 157).

Texten nach bestimmten Kriterien zu klassifizieren. Zum anderen kann mittels MÜ-Systeme eine Kommunikation zwischen Personen unterschiedlicher Sprache zustande kommen.

Des Weiteren finden MÜ-Komponenten heutzutage u. a. immer öfter Einzug in kommerzielle TM-Systeme (z. B. Systeme der SDL-Trados-Studio-Reihe mit ihrer statistischen MÜ-Komponente). Mithilfe der MÜ-Komponente kann eine Übersetzung eines AS_{neu} generiert werden, falls kein Match im TM gefunden wird. Auf diese Weise soll die Arbeit des Übersetzers beschleunigt und vereinfacht werden.

Umgekehrt können die in TMs gespeicherten Datensätze als Trainingsmaterial für statistische MÜ-Systeme herangezogen werden. Dieses Verfahren wird häufig in der MÜ-Praxis angewendet, wie es z. B. auch bei der CASMACAT Home Edition der Fall ist, einem kürzlich im Forschungsprojekt CASMACAT entwickelten Übersetzungswerkzeug (vgl. CASMACAT 2014: o.S.).

3.1.3.4 Aktuelle Trends in der MÜ-Forschung

In den letzten 60 Jahren wurden unterschiedliche Ansätze der MÜ entwickelt, die von regelbasierten Systemen bis hin zu statistischen Methoden oder gar hybriden Systemen reichen.

Die *regelbasierte maschinelle Übersetzung* (*rule-based machine translation, RBMT)* bildet dabei die traditionelle Form der MÜ. Sie beruht auf Regelwerken, die linguistisches Wissen verschiedener Ebenen erfordern: Morphologie, Syntax und Semantik (vgl. Hutchins/Somers 1992: 4f.). Es wird zwischen zwei Ansätzen unterschieden: dem direkten und dem indirekten Ansatz, wobei der indirekte Ansatz wiederum in den Transfer-Ansatz und den Interlingua-Ansatz unterteilt werden kann (vgl. Hutchins 1995: 218).

Eine Variante des Interlingua-Ansatzes ist die *wissensbasierte maschinelle Übersetzung* (*knowledge-based machine translation, KBMT*). Bei der KBMT werden Begriffe in einer Wissensdatenbank in Form einer Metasprache verwaltet, um die Bedeutung eines Satzes richtig zu übersetzen (vgl. Hutchins/Somers 1992: 313f., Stein 2009: 14). Die verschiedenen Methoden, Beispielübersetzungen sowie eine Darlegung der Vor- und Nachteile und Besonderheiten des jeweiligen regelbasierten Ansatzes sind in Hutchins und Somers (1992: 72ff.), Schmidt (1998: 133ff.), Stein (2009: 7ff.), Jekat und Volk (2010: 644ff.) sowie Carstensen (2012: 189ff.) beschrieben.

Aufgrund der Unzulänglichkeiten[32], die die RBMT mit sich bringt, wurden in den vergangenen Jahrzehnten neue Verfahren entwickelt, die die Vorteile von Korpora ausnutzen. Diese aktuellen Trends werden nachfolgend erläutert.

3.1.3.4.1 Korpusbasierte maschinelle Übersetzung

Da die RBMT auf den Einsatz linguistischen Wissens, zusammengefasst in aufwendigen Regeln, angewiesen ist, wurde die RBMT ab den 80er Jahren durch korpusbasierte Methoden verdrängt: die *beispielbasierte maschinelle Übersetzung* (*example-based machine translation*, *EBMT*), die *statistische maschinelle Übersetzung* (*statistical machine translation*, *SMT*) sowie die *kontextbasierte maschinelle Übersetzung* (*context-based machine translation*, *CBMT*).

Stein (2009: 12f.) führt an, dass korpusbasierte Methoden in ihrer reinen Form den Vorteil haben, nicht auf linguistisches Wissen zurückgreifen zu müssen, um geeignete Übersetzungen zu liefern. Durch die Verwendung von realen, humanübersetzten, miteinander alignierten Paralleltexten[33] entfällt die zeitaufwendige und kostspielige Erfassung linguistischer Regeln; brauchbare Ergebnisse stehen schon nach kurzer Zeit zur Verfügung.

Während sich also bei der RBMT auf die Produktion von Übersetzungen durch im Vorhinein erfasste Regeln einer Sprache konzentriert wird, bildet bei der korpusbasierten MÜ die Analyse von Humanübersetzungen den Ausgangspunkt:

> „Rule-based MT tends to focus exclusively on the *translation production* problem. […] Corpus-based methods, on the other hand, start from translations that have already been produced by humans and seek to discover their structure, completely or partially." (Isabelle 1993: 177, Hervorhebung im Original)

Neben dem genannten Vorteil, der für alle korpusbasierten Verfahren gilt, existieren weitere Vor- und Nachteile für jeden einzelnen Ansatz, die in Tabelle 10 gegenübergestellt werden:

---

[32] Zu den Unzulänglichkeiten zählen u. a. die schlechte Übersetzungsqualität aufgrund von Verwendung falscher Wörter und syntaktischer Strukturen der Zielsprache im Falle des direkten Ansatzes, die steigende Anzahl benötigter Transfer-Module bei jeder zusätzlich zu übersetzenden Sprache im Falle des Transfer-Ansatzes sowie die Schwierigkeit der Definition einer sprachunabhängigen Repräsentation der zu übersetzenden Sprachen im Falle des Interlingua-Ansatzes (vgl. Hutchins/Somers 1992: 72ff.). Eine weitere Schwierigkeit besteht im Umgang mit Mehrdeutigkeiten, für deren Disambiguierung oft ein komplexes linguistisches und Weltwissen erforderlich ist (vgl. Arnold 2003: 118ff.).

[33] Das Alignment paralleler Korpora kann auf verschiedenen Ebenen erfolgen, beispielsweise auf Basis einzelner Wörter, Sätze oder sogar ganzer Paragrafen (vgl. Lemnitzer/Zinsmeister 2006: 196).

SMT	EBMT	CBMT
• Funktioniert gut für Fachgebiete, für die ausreichend Textmengen zur Verfügung stehen. • Wenig geeignet für die unmittelbare Übersetzung aus einer wenig verbreiteten Sprache in eine andere wenig verbreitete Sprache, da große Parallelkorpora[34] für akzeptable Übersetzungen erforderlich sind, die für wenig verbreitete Sprachen oft nicht verfügbar sind[35]. • Gegebenenfalls Produktion fehlerhafter Übersetzungen bei strukturell stark voneinander abweichenden Sprachen. Durch die große Datenmenge können Fehlerquellen in den Parallelkorpora schwer lokalisiert und korrigiert werden (vgl. Stein 2009: 13).	• Besser geeignet für die unmittelbare Übersetzung aus einer wenig verbreiteten Sprache in eine andere wenig verbreitete Sprache, da die EBMT mit kleineren Korpora als die SMT auskommt. • In der Rekombinationsphase kann das sogenannte „boundary friction"-Problem[36] (Nirenburg et al. 1993: 48) auftreten.	• Auch anwendbar, falls keine Parallelkorpora für eine Sprachkombination gefunden werden können, da die CBMT mit einem monolingualen ZS-Korpus auskommt. • Monolinguales ZS-Korpus muss eine Größe zwischen 50 GB und 1 TB aufweisen (vgl. Carbonell et al. 2006: 19). • Erheblicher Zeitaufwand für das Anlegen eines bilingualen Vollformenlexikons für jede Sprachkombination (vgl. Carbonell et al. 2006: 19).

Tabelle 10: Vor- und Nachteile korpusbasierter Verfahren der MÜ

[34] Zu solchen Korpora gehört beispielsweise das Hansard- oder Europarl-Korpus. Beide Korpora werden häufig aufgrund ihres Volumens für Forschungszwecke im Bereich Computerlinguistik und MÜ herangezogen. Während das Hansard-Korpus Protokolle des Kanadischen Parlamentes in Englisch und Französisch umfasst, beinhaltet das Europarl-Korpus die Aufzeichnung gesprochener parlamentarischer Debatten in elf europäischen Sprachen (vgl. Linguistic Data Consortium 1992–2010, Lemnitzer/Zinsmeister 2006: 116f.).

[35] Soll aus einer wenig verbreiteten Sprache in eine andere wenig verbreitete Sprache übersetzt werden (z. B. Litauisch nach Baskisch), kann die Übersetzung über eine weitverbreitete Pivot-Sprache erfolgen (z. B. Englisch), da eine größere Menge an Paralleltexten zwischen weitverbreiteten und wenig verbreiteten Sprachen existiert als zwischen zwei wenig verbreiteten Sprachen.

[36] Im Falle von stark flektierenden Sprachen muss das System je nach Kasus die betroffenen Wortarten richtig flektieren. Um bei diesem Prozess grammatisch korrekte Ergebnisse erzielen zu können, sind entweder eine hinterlegte Grammatik oder statistische Verfahren notwendig.

3.1.3.4.1.1 Beispielbasierte maschinelle Übersetzung

Im Jahr 1984 veröffentlichte Nagao erstmals Ergebnisse seiner Forschung zur EBMT[37], die er selbst als „machine translation by example-guided inference" oder „machine translation by the analogy principle" (Nagao 1984: 179) bezeichnet. In seinem Aufsatz führt er die Funktionsweise der EBMT auf:

> „(1) Man does not translate a simple sentence by doing deep linguistic analysis, rather, (2) Man does the translation, first, by properly decomposing an input sentence into certain fragmental phrases (very often, into case frame units), then, by translating these fragmental phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference […]." (Nagao 1984: 178f.)

Die EBMT durchläuft demnach mehrere Phasen, die in der Literatur als Matching, Alignment und Rekombination bezeichnet werden. Die Funktionsweise dieser Phasen wird z. B. in Somers (2003d: 514ff.) und Forster (2006: 24ff.) detailliert beschrieben. Zudem findet sich eine den Input, Output, Verarbeitungsprozess und die Evaluierung betreffende tabellarisch vergleichende Übersicht zwischen der EBMT und TMs in Reinke (2004: 149).

Insbesondere in der Matching-Phase besteht eine große Gemeinsamkeit zwischen der EBMT und TMs: Bei beiden Werkzeugen wird ein AS_{neu}[38] mit einer Datenbank, bestehend aus alignierten AS- und ZS-Einheiten, verglichen, mit dem Ziel, ein oder mehrere AS-Segmente zu finden, die dem AS_{neu} möglichst ähnlich sind (vgl. Somers/Fernandez Diaz 2004: o.S.). Wird ein identisches AS-Segment in der Datenbank gefunden, wird sowohl beim EBMT- als auch beim TM-System die gefundene Übersetzungseinheit ausgegeben. Können nur ähnliche Segmente ermittelt werden, liegt der wesentliche Unterschied beider Systeme darin, welche Instanz die weitere Verarbeitung der gefundenen Übersetzungseinheiten vornimmt:

> „in a TMS it is up to the user to decide what to do with the matches, whereas in EBMT the system must operate automatically." (Somers/Fernandez Diaz 2004: o.S.)

Eine weitere zentrale Frage bei beiden Systemen ist die der Granularität der Segmente. Die Ähnlichkeit zwischen einem gespeicherten AS-Segment und

---

[37] Der Grundgedanke, MÜ auf Basis von realen Übersetzungsbeispielen durchzuführen, wurde erstmals 1981 durch Nagao vorgetragen; die gedruckte Variante wurde jedoch erst drei Jahre später veröffentlicht (vgl. Somers 2003b: 6).

[38] Auch bei der EBMT handelt es sich dabei meist um Sätze, es sind jedoch auch andere Einheiten, z. B. Fragmente unterhalb der Satzebene, denkbar.

einem AS_{neu} sowie die Qualität der Übersetzung hängen laut Nirenburg et al. (1993) von der Länge der Segmente ab:

> „The longer the matched passages, the lower the probability of a complete match […]. The shorter the passages, the greater the probability of ambiguity (one and the same *S'* can correspond to more than one passage *T'*) and the greater the danger that the resulting translation will be of low quality […]" (Nirenburg et al. 1993: 48)

In diesem Zusammenhang stellt sich auch die Frage, wie groß eine Datenbank sein soll, um gute Ergebnisse erzielen zu können. Generell gilt, je größer die Datenbank ist, desto größer ist die Wahrscheinlichkeit, einen Match zu erhalten bzw. eine Übersetzung guter Qualität zu generieren (vgl. Bowker 2002: 108)[39]. Allerdings besteht die Annahme, dass ab einer bestimmten Anzahl an Textbeispielen die Übersetzungsqualität in der EBMT nicht mehr steigt (vgl. Somers 2003b: 11). Gleichermaßen sollten bei TMs nicht jegliche Übersetzungseinheiten in einer großen Datenbank gesammelt, sondern vielmehr für unterschiedliche Fachgebiete eine separate Datenbank angelegt werden, um Homonyme und daraus resultierende falsche Matches zu vermeiden (vgl. Bowker 2002: 108).

EBMT-Systeme sind u. a. Gaijin, EDGAR, ReVerb, Metis-II und MTB2. Für eine genauere Darstellung dieser Systeme siehe Reinke (2004: 271ff.) und Forster (2006: 27).

3.1.3.4.1.2 Statistische maschinelle Übersetzung

Obwohl bereits 1949 durch Warren Weaver erste Überlegungen zur SMT angestellt wurden, fand ihr Durchbruch erst Ende der 80er Jahre statt, als eine IBM-Forschungsgruppe reine statistikbasierte Versuche unternahm. Unter SMT wird das „Prinzip, dass sich die Wahrscheinlichkeit, mit der eine sprachliche Äußerung die Übersetzung einer anderen ist, aufgrund möglichst vieler ‚Belege' genau berechnen lässt" (Lenders 2012: 466) verstanden.

Für diese Berechnung werden zwei Modelle benötigt, die 1988 bzw. 1990 erstmals von Brown et al. vorgestellt wurden: das Übersetzungsmodell und das Sprachmodell.

Mithilfe des Übersetzungsmodells wird die Wahrscheinlichkeit ermittelt, dass ein Wort bzw. eine Wortsequenz in der Zielsprache die Übersetzung eines Wortes bzw. einer Wortsequenz der Ausgangssprache ist (vgl. Somers

[39] Heyn (1998: 128) nennt ein TM groß, wenn es zwischen 100.000 und 1.000.000 Übersetzungseinheiten enthält. Gemäß Somers und Fernandez Diaz (2004: o.S.) umfasst eine große EBMT-Datenbank über 700.000 Beispiele, während die kleinste lediglich 7 beinhaltet.

2003d: 516, Hearne/Way 2011: 211). Die Basis des Übersetzungsmodells bildet dabei ein auf Satzebene aligniertes Parallelkorpus.

Zur Bildung des Sprachmodells wird hingegen ein monolinguales ZS-Korpus benutzt. Das Sprachmodell gibt die Wahrscheinlichkeit an, dass eine Abfolge von Wörtern von einem Muttersprachler der Zielsprache geäußert wird (vgl. Koehn 2010: 181). Eine ausführliche Beschreibung des statistischen MÜ-Prozesses findet sich u. a. in Brown et al. (1988, 1990), Stein (2009: 9ff.), Koehn (2010) und Hearne und Way (2011).

Eines der bekanntesten SMT-Systeme ist Google Translate. Weitere Systeme sind das Open-Source-System Moses sowie das SMT-System der Firma Language Weaver[40]. Siehe Benjamin et al. (2003), Koehn et al. (2007) und Lenders (2012: 467) für mehr Informationen zu den oben genannten Systemen.

Statistische Verfahren finden sich auch in TM-Systemen wieder, z. B. beim Alignment, bei dem „auf der Basis von Absatz- und Satzlängenvergleichen – gemessen wird entweder die Zahl der Wörter oder die Zahl der Zeichen pro Absatz und Satz – die wahrscheinlichste Zuordnung von AS und ZS Segmenten ermittelt wird“ (Reinke 2004: 72). Unter Berücksichtigung spezifischer Indikatoren (z. B. Zahlen, Datumsangaben, Formatierungen) können die AS- und ZS-Segmente zudem besser einander zugeordnet werden (vgl. Reinke 2004: 72).

3.1.3.4.1.3 Kontextbasierte maschinelle Übersetzung

Ein weiterer bedeutsamer korpusbasierter Ansatz der MÜ ist die CBMT. Sie wurde 2002 von Eli Abir (siehe Abir et al. 2002) als Antwort auf die Nachteile, die die RBMT, EBMT und SMT mit sich bringen, konzipiert:

> „Traditional MT paradigms require either extensive transfer-rule writing by linguists and computer scientists or very large parallel (pre-translated) training corpora. The former can take person-decades to write, debug and perfect the set of rules for reasonable quality translation (e.g., at the SYSTRAN level), and acquiring parallel text for Statistical MT and Example-Based MT in sufficient quantity for comparable or better quality MT proves to be a daunting task even for major language pairs. For less- translated language pairs, accessible parallel text is simply non-existent in sufficient quantities. To address these serious challenges, CBMT is being developed as a corpus-based method that requires neither rules nor parallel corpora.“ (Carbonell et al. 2006: 19)

Stattdessen wird sich bei der CBMT eines großen monolingualen Korpus in der Zielsprache, eines bilingualen Vollformenlexikons sowie optional eines

[40] Language Weaver wurde 2010 von SDL aufgekauft.

kleineren monolingualen Korpus in der Ausgangssprache bedient. Das ZS-Korpus liegt in Form von N-Grammen vor. Das AS-Korpus kann zur Verbesserung der Übersetzungsqualität eingesetzt werden. Das ZS- und AS-Korpus müssen keine Paralleltexte sein, sodass auch maschinelle Übersetzungen von seltenen Sprachkombinationen angefertigt werden können.

Der AT_{neu} wird in sich überlappende N-Gramme (von zwischen vier und acht Wörtern) zerteilt. Jedes Wort eines ausgangssprachlichen N-Gramms wird daraufhin im bilingualen Vollformenlexikon zur Ermittlung der Übersetzung nachgeschlagen. Schließlich wird das ZS-Korpus nach denjenigen N-Grammen durchsucht, die die Übersetzungen der AS-Wörter/AS-Phrasen enthalten. Die zielsprachlichen N-Gramme, die die höchsten Überschneidungen aufweisen, bilden die wahrscheinlichste Übersetzung des AS-Segmentes.

Können nur wenige oder keine sich überschneidenden N-Gramme gefunden werden, können beide monolingualen Korpora nach Synonymen durchsucht werden. In Carbonell et al. (2006) wird diese Form der MÜ eingehend beschrieben.

3.1.3.4.2 Multi-Engine-Systeme

Da jeder MÜ-Ansatz seine Vor- und Nachteile hat, wird seit einigen Jahren untersucht, ob die Kombination verschiedener Verfahren bessere Ergebnisse erzielt als die Verwendung eines einzelnen Ansatzes. Eine Möglichkeit der Kombination besteht in der Konzeption eines Multi-Engine-Systems. Bei einem solchen System werden dieselben Arbeitsschritte in mehreren gleichen oder unterschiedlichen Übersetzungssystemen parallel durchlaufen. Aus allen generierten Übersetzungen wird letztlich die zutreffendste Übersetzung herausgefiltert (vgl. Carl/Way 2003: xxii). Ein Multi-Engine-System muss nicht nur aus MÜ-Systemen bestehen; es können auch andere Übersetzungstools, wie TMs, zum Einsatz kommen.

Der Pionier der Multi-Engine-Systeme ist das System Pangloss Mark III für die Sprachkombination Spanisch-Englisch (siehe Frederking et al. 1994, Frederking/Nirenburg 1994). In diesem System werden Übersetzungen durch das KBMT-System PANGLOSS, durch ein EBMT-System sowie durch ein Transfersystem, das auf verschiedenen bilingualen Wörterbüchern und Glossaren sowie auf morphologischen Analyse- und Synthesemodulen beruht, erzeugt.

Ein neueres Multi-Engine-System stellen Alegria et al. (2008) für die Sprachkombination Spanisch-Baskisch vor, bei dem ein EBMT-, SMT- und RBMT-System eingesetzt werden. Die beste Übersetzung wird gemäß einer

Hierarchie ausgewählt: Die oberste Priorität besitzt das EBMT-System. Falls der Satz nicht durch das EBMT-System abgedeckt wird, wird das Übersetzungsergebnis des SMT-Systems gewählt. Wird kein Ergebnis durch das SMT-System geliefert, wird der Output des RBMT-Systems als Übersetzung vorgeschlagen.

Auch im Bereich der Open-Source-Anwendungen finden sich Multi-Engine-Systeme wieder: Das webbasierte CAT-Tool Matecat, das aus dem gleichnamigen kürzlich abgeschlossenen Projekt EU FP7 hervorgegangen ist, gibt sowohl gefundene Übersetzungseinheiten aus TMs als auch durch das SMT-System Moses generierte Übersetzungen aus. Die Matches werden mit Informationen zu ihrem Ursprung sowie mit ihrem Match-Wert bzw. der Angabe zur Übersetzungsqualität markiert (vgl. Federico et al. 2014: 129ff.).

Weitere Multi-Engine-Systeme werden in van Zaanen und Somers (2005), Matusov et al. (2006) sowie in Macherey und Och (2007) beschrieben.

3.1.3.4.3 Hybride Systeme

Eine weitere Möglichkeit der Verknüpfungen mehrerer Übersetzungstools ist die Erstellung hybrider Systeme. Im Gegensatz zu Multi-Engine-Systemen wird bei hybriden Systemen jedoch jedem Subsystem ein spezifischer Arbeitsschritt im Übersetzungsprozess zugewiesen (vgl. Carl/Way 2003: xxif.). Die Idee besteht darin, die Nachteile des einen Subsystems durch die Vorteile des anderen Subsystems aufzuwiegen. Bereits 1993 merken Nirenburg et al. exemplarisch Folgendes an:

> „The hybridization route is chosen in the hope that the resulting systems will have fewer practical shortcomings than the pure rule-based systems (a high complexity of processing plus a high price of knowledge acquisition) or the pure EBMT systems (a very ungraceful degradation curve when matches are bad).“ (Nirenburg et al. 1993: 48)

Auch bei hybriden Systemen können verschiedene MÜ-Systeme auf unterschiedliche Weise miteinander kombiniert werden. Sánchez-Martinez et al. (2009) integrieren z. B. das EBMT-System MaTrEx in das transferbasierte Open-Source-RBMT-System Apertium. Mittels MaTrEx werden bilinguale, syntaktisch motivierte Übersetzungsbeispiele unterhalb der Satzebene aus parallelen Korpora ermittelt. Basierend darauf wird mittels Dynamischer Programmierung errechnet, welche Segmentierung der AS-Segmente unterhalb der Satzebene ein AS_{neu} am besten abdeckt. Die weitere Übersetzung wird durch Apertium durchgeführt.

Eine andere Idee verfolgen Chatterji et al. (2009), indem sie SMT mit RBMT für das Sprachenpaar Bengali-Hindi kombinieren. Die Ergebnisse des

statistischen Phrasenalignments werden durch Hinzuschaltung eines bilingualen Wörterbuches linguistisch angereichert. Die anschließend durch das SMT-System entstandenen Übersetzungen werden mithilfe weiterer linguistischer Regeln optimiert.

Smith und Clark (2009) verknüpfen hingegen EBMT mit SMT in zwei unterschiedlichen Ansätzen: Zunächst wird das AS_{neu} durch das EBMT-System gematcht. Im ersten Ansatz werden lediglich Zeichenketten verglichen, während im zweiten Ansatz Syntax-Bäume als Matching-Methode herangezogen werden. Mithilfe des SMT-Systems Moses werden noch unübersetzte Teile des Satzes übersetzt.

Wie bei Multi-Engine-Systemen können auch bei hybriden Systemen CAT- und MÜ-Systeme miteinander kombiniert werden. Dandapat et al. (2010) integrieren beispielsweise zwei TMs sowie das SMT-System Moses in ein EBMT-System. Das TM wird mit durch Moses alignierten Phrasen und Wörtern befüllt. In der Rekombinationsphase des beispielbasierten MÜ-Prozesses werden noch unübersetzte Satzfragmente zuerst mithilfe von Moses an die richtige Position im Satz platziert und daraufhin unter Verwendung der TMs übersetzt.

Im webbasierten Open-Source-System CASMACAT wird SMT in ein CAT-System integriert (siehe Koehn et al. 2013), wodurch u. a. nicht gematchte Wörter eines aus dem TM herrührenden Fuzzy-Matches durch das SMT-System übersetzt werden können.

He et al. (2011) sowie Ma et al. (2011) kombinieren ebenso ein SMT-System mit einem TM. Das TM wird für das Auffinden der ähnlichsten Subsegmente zwischen einem AS_{neu} und den AS_{TM} herangezogen. Die korrespondierenden ZSTM-Subsegmente werden für die Übersetzung des AS_{neu} verwendet, sofern der Klassifikator eine entsprechend hohe Übersetzungsqualität prognostiziert. Für die Klassifikation kommen u. a. linguistische Informationen zum Einsatz. Während diese bei Ma et al. (2011) auf syntaktische Abhängigkeitsbeziehungen beschränkt sind, bedienen sich He et al. (2011) zusätzlich lexikalischer und semantischer Merkmale sowie POS-Tags. Nicht aufgefundene Subsegmente werden durch das SMT-System übersetzt.

Die Forschungsbemühungen von Ma et al. (2011) und He et al. (2011) sind Beispiele für Systeme, bei denen versucht wird, die Übersetzungsergebnisse korpusbasierter MÜ durch Integration linguistischer Informationen zu verbessern. Neben der Kombination aus korpusbasierten mit regelbasierten MÜ-Systemen kann dies durch Integration morphosyntaktischer oder semantischer Analyseprogramme in korpusbasierte MÜ-Systeme erfolgen.

Singh und Bandyopadhyay (2010) versuchen ein EBMT-System für die Sprachrichtung Manipuri-Englisch durch Hinzuschaltung eines POS-Taggers, Stemmers, Named-Entity-Recognition-Moduls und Chunkers zu optimieren.

Aziz et al. (2011) annotieren die AS-Seite eines englisch-spanischen und englisch-deutschen Parallelkorpus mit semantischen Rollen, POS-Tags sowie Informationen zur Satzsegmentierung in linguistisch motivierte Phrasen. Im deutschen Korpus werden zudem Komposita in ihre Bestandteile zerlegt. Das daraus entstandene Übersetzungsmodell wird für die weitere Übersetzung durch ein SMT-System verwendet.

Shilon et al. (2012) befassen sich mit der Einbindung linguistischer Informationen in ein transferbasiertes SMT-System, um eine bessere Übersetzung von Präpositionen für das Sprachenpaar Arabisch-Hebräisch zu erzielen.

3.2 Kommerzielle Translation-Memory-Systeme

3.2.1 Nicht linguistisch optimierte Translation-Memory-Systeme

Obwohl die Retrieval-Leistung kommerzieller TM-Systeme seit längerer Zeit bemängelt wird (siehe z. B. Macklovitch/Russell 2000: 138ff.), setzt der Großteil der auf dem Markt verfügbaren TM-Systeme nach wie vor simple Verfahren zum Zeichenkettenvergleich zwischen dem AS_{neu} und den AS_{TM} ein. Zu ihnen zählen u. a. die TM-Systeme SDL Trados, Across, Déjà Vu, memoQ, Wordfast, Transit und MultiTrans.

Zur Illustration der Unzulänglichkeiten des zeichenkettenbasierten Vergleiches wird Tabelle 11 angeführt, in der jeweils ein AS_{neu} einem ähnlichen AS_{TM} gegenübergestellt wird. Exemplarisch werden die Ergebnisse des Matchings durch SDL Trados Studio 2009 SP3 – 9.1.2307.8 dargestellt. Die AS_{neu} und AS_{TM} stammen teilweise aus Bedienungsanleitungen der Firma Braun (siehe auch Kapitel 6.1) und teilweise wurden sie von der Verfasserin dieser Arbeit eigenständig hinzugefügt. Von den 20 eindeutigen AS_{neu} wurden sieben modifiziert und sechs hinzugefügt, während von den 19 eindeutigen AS_{TM} zehn modifiziert und ebenfalls sechs hinzugefügt wurden, um das Verhalten des TMs bei spezifischen, isolierten linguistischen Phänomenen beobachten zu können. Die restlichen sieben AS_{neu} und drei AS_{TM} wurden im Original beibehalten, um Szenarien aus der realen Übersetzungsarbeit wiederzugeben. Eine Auflistung der modifizierten, original beibehaltenen und eigenständig hinzugefügten AS_{neu} und AS_{TM} findet sich in Anhang A und B. Zur besseren Lesbarkeit wird lediglich der erste Fuzzy-Match aus der Trefferliste dargelegt – ungeachtet der 100 %-Matches.

	AS_{neu}	Match-Wert	AS_{TM}
1	Alternativ können Sie den Rasierer mit der mitgelieferten Reinigungsbürste reinigen:	99 %	Alternativ können Sie den Rasierer mit der mitgelieferten Reinigungsbürste reinigen.
2	Günstigste Umgebungstemperatur beim Laden:	93 %	Günstige Umgebungstemperatur beim Laden:
3	Dieses Gerät darf nicht in der Nähe von mit Wasser gefüllten Behältern, z.B. Badewanne, Dusche, Waschbecken, verwendet werden.	93 %	Dieses Gerät darf nicht in der Nähe von mit Wasser gefüllten Behältern, z.B. Dusche, Badewanne, Waschbecken, verwendet werden.
4	Einige praktische Tips	90 %	Einige praktische Tipps
5	Reinigung mit Wasser	90 %	Reinigen mit Wasser
6	Vergewissern Sie sich, dass das Batteriefach trocken und sauber ist, bevor Sie die Batteriefach-Abdeckung wieder schließen.	88 %	Vergewissern Sie sich, dass die Batteriefach-Abdeckung trocken und sauber ist, bevor Sie das Batteriefach wieder schließen.
7	Scherkopf und Klingenblock separat unter fließendem Wasser reinigen.	88 %	Scherkopf und Klingenblock separat unter fließendes Wasser halten.
8	Die Scherfolie darf nicht mit der Bürste gereinigt werden, da dieses zu Beschädigungen führen könnte.	82 %	Die Scherfolie darf nicht mit der Bürste gereinigt werden, da es die Scherfolie beschädigen könnte.
9	Reinigen unter Wasser	77 %	Reinigen mit Wasser
10	Die Bürsten werden gereinigt.	75 %	Die Bürste wird gereinigt.
11	Das Reinigen der Bürste.	75 %	Die Reinigung der Bürste.
12	Ein Reinigungsbürstchen wird mitgeliefert.	75 %	Eine Reinigungsbürste wird mitgeliefert.
13	Nach der Epilation empfehlen wir die Verwendung einer Feuchtigkeitscreme.	70 %	Nach der Epilation empfehlen wir, eine Feuchtigkeitscreme zu verwenden.
14	Halten Sie die Haut gestrafft.	67 %	Straffen Sie die Haut.
15	Die Reinigung der Bürste.	49 %	Die Bürste wird gereinigt.
16	Die Scherkopf-Innenseite reinigen.	47 %	Die Innenseite des Scherkopfes reinigen.
17	Wir empfehlen die Verwendung einer Feuchtigkeitscreme nach der Epilation.	32 %	Nach der Epilation empfehlen wir die Verwendung einer Feuchtigkeitscreme.
18	Wenn sich der Distanzkamm mit Haaren zusetzt, sollten Sie ihn abnehmen und säubern.	30 %	Sie sollten ihn abnehmen und säubern, wenn sich der Distanzkamm mit Haaren zusetzt.

Tabelle 11: Gegenüberstellung bedeutungsgleicher Segmente mit Match-Werten aus dem nicht linguistisch optimierten kommerziellen TM-System SDL Trados Studio 2009

In der Tabelle ist erkennbar, dass morphologische und lexikalische Unterschiede vorwiegend in einem höheren Match-Wert resultieren als syntaktische Unterschiede. Lediglich die Zerlegung von Komposita (Beispiel 16) liefert einen Match-Wert von unter 70 %. Dies ist insofern problematisch, als dass dieser Match bei einem voreingestellten Schwellenwert für die Trefferanzeige von 70 %, der üblicherweise von Übersetzern in der Übersetzungspraxis gewählt wird, nicht angezeigt werden würde, obwohl er eine Bedeutungsgleichheit mit dem AS_{neu} aufweist und somit für die Übersetzung brauchbar ist. Syntaktische Unterschiede resultieren hingegen beinahe ausschließlich in Match-Werten von unter 70 %. Nur im Falle einer Umstellung kurzer Phrasen in langen Segmenten (Beispiele 3 und 6) werden Matches über dem üblichen Schwellenwert ausgegeben. In den meisten Fällen müsste der Match-Wert höher angesiedelt sein, so auch bei Segmentpaar 18, in dem Haupt- und Nebensatz vertauscht sind, die Bedeutung jedoch identisch ist. Ebenso ist anzumerken, dass sich geringe Unterschiede umso stärker auf den Match-Wert auswirken, je kürzer die zu vergleichenden Segmente sind (Beispiele 4, 5, 9, 10 und 11). Trotz der Mängel muss zur Kenntnis genommen werden, dass in SDL Trados Studio 2009 zu jedem AS_{neu} ein Match ausgegeben wird.

Allerdings kann beim zeichenkettenbasierten Vergleich auch der gegenteilige Effekt eintreten: Segmentpaare, die hinsichtlich ihrer Zeichenabfolge ähnlich sind, jedoch unterschiedliche Bedeutungen innehaben, können mit einem zu hohen Match-Wert versehen sein. In der nachfolgenden Tabelle werden zwei Beispiele zu dieser Problematik aufgeführt. In der Spalte AS_{TM} *(aufgefunden)* werden die aufgefundenen Matches dokumentiert, während in der Spalte AS_{TM} *(erwartet)* die Matches (inklusive Match-Wert) angegeben werden, die gemäß dem menschlichen Ähnlichkeitsempfinden an erster Stelle in der Trefferanzeige aufgelistet werden sollten.

	AS_{neu}	**Match-Wert**	**AS_{TM} (aufgefunden)**	**AS_{TM} (erwartet)**
19	Die Maus ist klein.	93 %	Die Laus ist klein.	Die Mäuse sind klein. (75 %)
20	Die Frau ist nett.	66 %	Die Laus ist klein.	Die Frau, die an mir vorbei gegangen ist, ist nett. (38 %)

Tabelle 12: Gegenüberstellung nicht bedeutungsgleicher Segmente mit Match-Werten aus dem nicht linguistisch optimierten kommerziellen TM-System SDL Trados Studio 2009

In Tabelle 12 wird deutlich, dass linguistische Analysen nicht Bestandteil des Match-Algorithmus des getesteten TM-Systems sind und der zeichenkettenbasierte Vergleich durch geringe Unterschiede in der Zeichenabfolge auch zu unerwarteten Treffern führen kann: Es werden AS_{TM} an erster Stelle in der Trefferanzeige aufgeführt, die Bedeutungsunterschiede gegenüber dem AS_{neu} aufweisen. Die erwarteten AS_{TM}, die eine größere Bedeutungsähnlichkeit mit dem AS_{neu} besitzen, werden zwar auch in der Trefferanzeige angegeben, jedoch nur mit deutlich geringeren Match-Werten von 75 % (Beispiel 19) bzw. 38 % (Beispiel 20).

Dies bedeutet jedoch nicht, dass nicht linguistisch optimierte TM-Systeme nicht auch ihre Vorteile haben. So sind sie sprachunabhängig, d. h., ihr Angebot an bearbeitbaren Sprachen ist nicht auf diejenigen Sprachen reduziert, die beispielsweise durch ein morphosyntaktisches Analyseprogramm verarbeitet werden können. Ferner sind auch bei großen Datenmengen eine stabile Arbeitsumgebung und schnelle Reaktionszeiten gewährleistet, da keine aufwendigen computerlinguistischen Analysen durchgeführt werden müssen. McTait (2001) merkt dazu an:

> „their addition [i.e. of linguistic resources; M. W.] has consequences for portability and computational complexity. The more resources required, the less portable and the more complex the system." (McTait 2001: 23f.)

Um auf dem Übersetzungsmarkt konkurrieren zu können und den Bedarf an Werkzeugen für schnelle und kostengünstige Übersetzungen zu decken, wurden in den letzten 20 Jahren weitere Verbesserungen des Retrievals durch nicht linguistische Methoden erzielt. Zu diesen Verbesserungen zählen der bereits erwähnte Einsatz des Fuzzy-Matchings sowie die Einführung von Kontext-Matches und Placeables.

Ebenso finden sich Methoden der MÜ in TM-Systemen wieder. Je nach TM-System werden andere Verfahren verwendet. So bedient sich Déjà Vu X der EBMT, indem Fuzzy-Matches ohne Einwirkung des Menschen zu 100 %-Matches modifiziert werden können, vorausgesetzt, alle Subsegmente der zu generierenden Übersetzung sind im TM oder der Terminologiedatenbank genau enthalten (vgl. ATRIL Language Engineering 1993–2003: 147ff., Somers/Fernandez Diaz 2004: o.S., Azzano 2009: 26). Andere TM-Systeme, beispielsweise SDL Trados, verwenden hingegen eine MÜ-Komponente, die auf statistischen Analysen beruht (vgl. Azzano 2009: 27ff.).

3.2.2 Linguistisch optimierte Translation-Memory-Systeme

Neben den konventionellen, nicht linguistisch optimierten TM-Systemen kamen in den letzten 15 Jahren nur wenige linguistisch optimierte TM-Systeme auf den Markt. Diese Systeme setzen bzw. setzten linguistische Analysen zur Verbesserung der Retrieval-Leistung ein. Obwohl die Integration linguistischer Analysen in Übersetzungswerkzeuge mit Nachteilen verbunden ist (Sprachenabhängigkeit vom Analysewerkzeug, hohe Komplexität des Systems einhergehend mit einer unstabileren Arbeitsumgebung, eingeschränkte Portabilität), haben linguistisch optimierte TM-Systeme auch einen klaren Vorteil gegenüber nicht linguistisch optimierten TM-Systemen: die gleichzeitige Verbesserung von Precision und Recall (vgl. McTait 2001: 23). Neben Open-Source-Systemen, die die Möglichkeit zur linguistischen Weiterentwicklung bieten, konnte sich bis heute nur ein proprietäres linguistisch optimiertes TM-System nachweislich dauerhaft auf dem Markt behaupten.[41]

So gehört beispielsweise das erste TM-System dieser Art, ZeresTrans, das 1996 von der Zeres GmbH entwickelt wurde, zu denjenigen linguistisch optimierten kommerziellen TM-Systemen, die nicht mehr vermarktet werden. Demnach existieren auch nur wenige Informationen über die Funktionsweise dieses TM-Systems. Laut Reinke (2004: 57) werden die Übersetzungseinheiten in ZeresTrans morphosyntaktischen Analysen unterzogen und in drei verschiedenen Repräsentationen gespeichert:

- Als Folge von Wörtern in ihrer Originalform
- Als Folge von Wörtern in ihrer Grundform
- Als Folge von morphosyntaktischen Merkmalen

Jeder dieser Repräsentationen wird beim Retrieval eine bestimmte Gewichtung zugeordnet. Zur Auffindung von Subsegmenten wird auf den zusätzlich bestehenden Phrasenspeicher zurückgegriffen (vgl. Reinke 2004: 57).

Das von der finnischen Firma Master's Innovations Ltd. für die Sprachen Englisch, Finnisch und Schwedisch entwickelte TM-System Masterin zerlegt das AS_{neu} sowie die im TM gespeicherten Übersetzungseinheiten in Subsegmente durch Vergleich mit einer vorinstallierten Wissensdatenbank, die Informationen zu häufig auftretenden syntaktischen Strukturen enthält und während des Übersetzungsprozesses erweitert werden kann (vgl. Grönroos/Becks 2005: o.S., Lagoudaki 2008: 264). Im anschließenden Übersetzungsprozess

[41] Laut Lagoudaki (2008: 264) existieren zwei proprietäre linguistisch optimierte TM-Systeme auf dem Übersetzungsmarkt: Masterin und Similis. Allerdings konnte bei der Recherche zu Masterin nicht eindeutig bestätigt werden, dass dieses TM-System weiterhin vertrieben wird.

greift je nach Match-Art eine andere Retrieval-Methode (vgl. Grönroos/Becks 2005: o.S.):

- 100 %-Match: Masterin verhält sich wie ein konventionelles kommerzielles TM-System, indem es den 100 %-Match in den Übersetzungseditor einfügt.
- Fuzzy-Match: Das AS_{neu} und die Subsegmente der im TM gespeicherten Übersetzungseinheiten werden auf Basis ihrer syntaktischen Struktur miteinander verglichen. Existieren mehrere Subsegmente im TM mit derselben syntaktischen Struktur wie diejenige des AS_{neu}, wird eine Disambiguierung mittels semantischer Informationen sowie Informationen zur Verwendungshäufigkeit und zum Fachgebiet des Subsegmentes vorgenommen. Mithilfe eines integrierten Lexikons sowie eines Wortformengenerators wird anschließend dem Übersetzer ein Übersetzungsvorschlag angezeigt.
- No-Match: Eine MÜ-Komponente liefert eine Rohübersetzung, für den Fall, dass kein Match im TM gefunden wird.

Bei dem einzigen noch auf dem Markt nachweislich verfügbaren proprietären linguistisch optimierten TM-System handelt es sich um das Programm Similis der französischen Firma Lingua et Machina. Similis unterstützt die Sprachen Englisch, Deutsch, Französisch, Italienisch, Spanisch, Portugiesisch und Niederländisch und arbeitet wie Masterin auf Subsegentebene. Dazu werden die Übersetzungseinheiten einer linguistischen Analyse unterzogen und in syntaktische Einheiten (z. B. Nominal- oder Verbalphrasen) segmentiert (vgl. Planas 2005: o.S., Lingua et Machina 2006: 4, Macken 2009: 203). Die einzelnen Wörter der Subsegmente werden in einer weiteren Analyse mit ihrer Grundform und Wortart annotiert (vgl. Planas 2005: o.S.). Sind die Ausgangs- und Zielsprache ähnlich genug, kann auch einem AS-Subsegment seine ZS-Entsprechung mittels integrierter bilingualer Wörterbücher und dem Vergleich grammatischer Satzstrukturen zugeordnet werden (vgl. Planas 2005: o.S., Lingua et Machina 2006: 4).

Neben der Analyse und Segmentierung der Übersetzungseinheiten kann Terminologie aus den bilingualen Subsegmenten automatisch extrahiert, einander zugeordnet und während der Übersetzung angezeigt werden (vgl. Planas 2005: o.S., Lingua et Machina 2006: 43ff.).

	AS_{neu}	Match-Wert	AS_{TM}
1	Alternativ können Sie den Rasierer mit der mitgelieferten Reinigungsbürste reinigen:	95 %	Alternativ können Sie den Rasierer mit der mitgelieferten Reinigungsbürste reinigen.
2	Günstigste Umgebungstemperatur beim Laden:	85 %	Günstige Umgebungstemperatur beim Laden:
3	Dieses Gerät darf nicht in der Nähe von mit Wasser gefüllten Behältern, z.B. Badewanne, Dusche, Waschbecken, verwendet werden.	93 %	Dieses Gerät darf nicht in der Nähe von mit Wasser gefüllten Behältern, z.B. Dusche, Badewanne, Waschbecken, verwendet werden.
4	Einige praktische Tips	77 %	Einige praktische Tipps
5	Reinigung mit Wasser	75 %	Reinigen mit Wasser
6	Vergewissern Sie sich, dass das Batteriefach trocken und sauber ist, bevor Sie die Batteriefach-Abdeckung wieder schließen.	84 %	Vergewissern Sie sich, dass die Batteriefach-Abdeckung trocken und sauber ist, bevor Sie das Batteriefach wieder schließen.
7	Scherkopf und Klingenblock separat unter fließendem Wasser reinigen.	83 %	Scherkopf und Klingenblock separat unter fließendes Wasser halten.
8	Die Scherfolie darf nicht mit der Bürste gereinigt werden, da dieses zu Beschädigungen führen könnte.	80 %	Die Scherfolie darf nicht mit der Bürste gereinigt werden, da es die Scherfolie beschädigen könnte.
9	Reinigen unter Wasser	75 %	Reinigen mit Wasser
10	Die Bürsten werden gereinigt.	–	–
11	Das Reinigen der Bürste.	–	–
12	Ein Reinigungsbürstchen wird mitgeliefert.	–	–
13	Nach der Epilation empfehlen wir die Verwendung einer Feuchtigkeitscreme.	–	–
14	Halten Sie die Haut gestrafft.	71 %	Straffen Sie die Haut.
15	Die Reinigung der Bürste.	–	–
16	Die Scherkopf-Innenseite reinigen.	–	–
17	Wir empfehlen die Verwendung einer Feuchtigkeitscreme nach der Epilation.	93 %	die Verwendung einer Feuchtigkeitscreme
18	Wenn sich der Distanzkamm mit Haaren zusetzt, sollten Sie ihn abnehmen und säubern.	–	–

Tabelle 13: Gegenüberstellung bedeutungsgleicher Segmente mit Match-Werten aus dem linguistisch optimierten kommerziellen TM-System Similis

In Tabelle 13 sind Retrieval-Ergbnisse aufgeführt, die mit Similis V2.16.04 erzeugt wurden. Für diesen Test wurden dieselben Segmentpaare herangezogen wie für den Test mit dem nicht linguistisch optimierten kommerziellen TM-System[42]. Die Ergebnistabelle ist ebenso identisch aufgebaut, d. h., dass auch in dieser Tabelle immer nur der erste Fuzzy-Match aus der Trefferliste – ungeachtet der 100 %-Matches – dargestellt wird.

Fast alle Match-Werte aus dem Test liegen wider Erwarten unter den Match-Werten, die durch das nicht linguistisch optimierte kommerzielle TM-System angezeigt wurden[43]. Insbesondere der Unterschied im Interpunktionszeichen (Beispiel 1) fällt mit 5 % Abzug stark ins Gewicht. Lediglich Beispiel 14 weist einen etwas höheren Match-Wert auf als derjenige, der durch SDL Trados Studio 2009 ermittelt wurde. Zudem sind nur die Match-Werte aus Beispiel 3 bei beiden Systemen identisch.

Im Gegensatz zu SDL Trados Studio 2009 kann bei Similis auch nicht die Feststellung gemacht werden, dass die Match-Werte im Falle von morphologischen und lexikalischen Unterschieden häufiger im oberen Match-Wert-Bereich liegen als im Falle von syntaktischen Unterschieden.

Des Weiteren ist auffällig, dass bei mehr als einem Drittel der Beispiele kein Match (ausgenommen 100 %-Matches) ausgegeben wird, wobei es sich bei einem Großteil dieser Beispiele um kurze Segmente handelt.

Da Similis mit dem Retrieval von Subsegmenten wirbt, sollte angenommen werden, dass mehr Subsegment-Matches in Tabelle 13 angeboten werden; zumindest für den Fall, dass kein vollständiges Segment als Fuzzy-Match gefunden werden kann. Nur zu einem der 18 Segmente konnte ein Subsegment-Match aufgefunden werden (Beispiel 17), wobei nicht ersichtlich ist, warum dem Subsegment gemäß Planas (2005) nicht ein Match-Wert von 100 % zugeordnet wurde.

Bei den Treffern, die von der Terminologieerkennungskomponente angeboten werden, ist deutlich zu erkennen, dass die Übersetzungseinheiten linguistischen Analysen unterzogen werden (siehe Tabelle 14). Des Weiteren wird

---

[42] Da das TM und der AT_{neu} nur aus sehr wenigen Segmenten bestanden (siehe Kapitel 3.2.1), wurde auf ein Ranking verzichtet. In der Trefferliste von Similis wurde entweder kein Treffer ausgegeben oder derselbe erstplatzierte wie bei SDL Trados Studio 2009. Das einzige aufgefundene Subsegment in Similis rührt zudem aus demselben Treffer her, zu dem in SDL Trados Studio 2009 das vollständige Segment aufgefunden wurde.

[43] Auch im Falle von Similis sind die Match-Werte nicht direkt mit den Match-Werten von SDL Trados Studio 2009 vergleichbar, und zwar aufgrund der unterschiedlichen Algorithmen (Zeichenkettenvergleich vs. linguistische Optimierung). Um die Systeme dennoch vergleichen zu können, wird analog zum methodischen Vorgehen bei der Evaluierung des iMem-TMs angenommen, dass die Match-Werte vergleichbar sind.

ZS-Terminologie angezeigt, die nicht nur aus den alignierten Subsegmenten herrührt; vielmehr scheint es, dass zudem Wörterbücher für die verschiedenen Sprachkombinationen integriert sind. Trotz der vermeintlichen Übersetzungshilfe können Auffälligkeiten in der Analyse und Präsentation der AS- und ZS-Terminologie festgestellt werden. Einige dieser Auffälligkeiten werden in Tabelle 14 aufgeführt.

	AS-Terminus (original)	**AS-Terminus (von Similis angezeigt)**	**ZS-Terminus (von Similis angezeigt)**
A	Bürste	Bürste	brush, brush, crew cut
B	Bürsten	bürsten	brush, brush, crew cut
C	Scherkopf	Scher#Kopf	brain, brains, head, heads, mastermind
D	Umgebungstemperatur	Umgebungs#Temperatur	circle of acquaintances, environment, neighborhood, neighbourhood, parts, surroundings, temperature
E	Scherkopf-Innenseite	Scher#kopf^-Innen#Seite	edge, lateral, laterally, page, side
F	Badewanne	Bade#Wanne	basin, bath, bath, bathe, go for a swim, tub
G	Waschbecken	Wasch#Becken	basin, cymbals, do the laundry, pelvis, wash, washbasin, washbowl
H	Reinigungsbürstchen	Reinigungs#Bürste^}chen	dry cleaner's, purification
I	fing … an	fangen	capture, catch, tag

Tabelle 14: Auffälligkeiten der Terminologieerkennungskomponente in Similis

Einfache Substantive (Simplizia) werden als solche erkannt und in ihrer Grundform (Nominativ, Singular) angezeigt (Beispiel A). Gleicht jedoch die flektierte Form eines Simplex dem Infinitiv des zugehörigen Verbs (Beispiel B), wird das Wort nicht als Substantiv, sondern als Verb erkannt. Es scheint, dass sowohl die Übersetzung für das Substantiv als auch für das Verb ausgegeben wird, wobei nicht gekennzeichnet ist, welches der beiden Wörter *brush* die Übersetzung für die jeweilige Wortart ist.

Komposita hingegen werden in ihre Bestandteile zerlegt (Beispiele C bis H). Bei genauerer Betrachtung der Übersetzungen wird deutlich, dass nicht nur das Kompositum als Ganzes übersetzt wird (z. B. *washbasin*, *washbowl*), sondern zudem für jeden Kompositumsbestandteil isolierte Übersetzungsvorschläge ausgegeben werden (z. B. *edge*, *side*). In manchen Fällen werden

ebenfalls Derivationsbestandteile markiert (Beispiel H), die im Übersetzungsvorschlag jedoch keine Beachtung finden.

Bei Verben, die über ein abtrennbares Präfix verfügen (I) und mit diesem Präfix eine Verbklammer bilden, werden die beiden Bestandteile isoliert betrachtet und lediglich das erste Wort der Verbklammer im Wörterbuch nachgeschlagen. Die daraus resultierenden Übersetzungsvorschläge sind jedoch nicht verwertbar, da durch das Ignorieren des abgetrennten Präfixes ein Wort mit einer völlig unterschiedlichen Bedeutung nachgeschlagen wurde (*fangen* in der Bedeutung von *fassen* anstatt *anfangen* in der Bedeutung von *beginnen*; siehe Tabelle 14, I).

Generell lässt sich für alle Beispiele jedoch festhalten, dass die Nützlichkeit der Übersetzungsvorschläge fraglich ist (z. B. *brain*, *mastermind*, *page*, *pelvis*), was erkennen lässt, dass keine semantische Disambiguierung der Übersetzungsvorschläge erfolgt.

Eine linguistische Optimierung ist allerdings nicht nur bei proprietären TM-Systemen, sondern auch im Bereich der Open-Source-Anwendungen möglich. Durch ihren frei zugänglichen Quellcode besteht bei Open-Source-TM-Systemen – wie OmegaT oder OpenTM2 – generell die Möglichkeit, das System linguistisch anzureichern.

Im Falle von OmegaT wurde bereits ein Plug-in, der sogenannte OmegaT Tokenizer, entwickelt, mit dessen Hilfe flektierte Formen der Wörter in den AS- und ZS-Segmenten erkannt und die Wortformen somit auf ihren Wortstamm zurückgeführt werden können (vgl. Smolej o.J.: 107). Diese Vorgehensweise soll in besseren Fuzzy-Matches sowie in einer verbesserten Terminologieerkennung resultieren.

Eine ähnliche Funktion bietet OpenTM2: Mithilfe sprachunterstützender Dateien, die monolinguale morphologische Daten beinhalten, werden die AS-Wörter auf ihre Stammformen reduziert. Soll ein AS-Wort im Wörterbuch nachgeschlagen werden, wird die Stammform anstelle des flektierten Wortes gesucht. Des Weiteren werden bei germanischen Sprachen Komposita in ihre Bestandteile zerlegt und diese einzelnen Wörter separiert im Wörterbuch nachgeschlagen, sofern für das Kompositum kein Wörterbucheintrag existiert (vgl. OpenTM2 2010: o.S.).

In der nachfolgenden Tabelle werden die Methoden zur linguistischen Optimierung der oben beschriebenen Systeme zusammengefasst.

System	Methoden zur linguistischen Optimierung
ZeresTrans	• Speicherung der Wörter in ihrer Originalform, in ihrer Grundform und als Folge von morphologischen Merkmalen • Unterschiedliche Gewichtung der Repräsentationen • Vergleich mit Phrasenspeicher zum Auffinden von Subsegmenten
Masterin	• Vergleich mit vorinstallierter Wissensdatenbank zur Zerlegung des AS_{neu} und der Übersetzungseinheiten in Subsegmente • Vergleich syntaktischer Strukturen • Disambiguierung mittels semantischer Informationen sowie Informationen zur Verwendungshäufigkeit und zum Fachgebiet des Subsegmentes • Verwendung eines integrierten Lexikons sowie eines Wortformengenerators
Similis	• Segmentierung der Übersetzungseinheiten in syntaktische Einheiten • Annotation der Wörter der Subsegmente mit Grundform und Wortart • Mittels integrierter bilingualer Wörterbücher und Vergleich grammatischer Satzstrukturen ggf. Zuordnung von AS-Subsegmenten zu ihren ZS-Entsprechungen • Kompositazerlegung und Markierung von Derivationsbestandteilen bei der Terminologieerkennung
OmegaT	• Stammformreduktion flektierter Wörter in den AS- und ZS-Segmenten
OpenTM2	• Stammformreduktion flektierter Wörter in den AS-Segmenten • Kompositazerlegung bei germanischen Sprachen

Tabelle 15: Übersicht über die Methoden zur linguistischen Optimierung in ZeresTrans, Masterin, Similis, OmegaT und OpenTM2

3.3 Forschungsprojekte zur Optimierung von Translation Memorys

Seit Ende der 90er Jahre werden immer wieder Versuche unternommen, das Retrieval von TMs durch Einbindung zusätzlicher Werkzeuge zu verbessern. Die Forschungsunternehmungen laufen dabei in verschiedene Richtungen: Einerseits besteht die Möglichkeit, TMs durch Integration nicht linguistischer Ressourcen, z. B. durch MÜ-Komponenten, zu optimieren. Andererseits können aber auch linguistische Informationen in TMs eingebunden werden. Im Folgenden werden Experimente beider Forschungsrichtungen aufgezeigt.

3.3.1 Nicht linguistisch optimierte Translation Memorys

In Kapitel 3.1.3.4.3 wurden bereits hybride Systeme beschrieben, in denen MÜ-Systeme durch Hinzuschaltung von TMs optimiert wurden. Jedoch ist auch die andere Richtung der Hybridität möglich. Dabei ist anzumerken, dass insbesondere im Bereich der Integration von SMT-Systemen in TMs geforscht wird, da TMs als Trainingskorpora für SMT-Systeme dienen können und sich die beiden Systeme dadurch gut ergänzen. Der Einbindung beispielbasierter maschineller Verfahren in TMs wurde hingegen seltener nachgegangen, da die EBMT im Laufe der Jahre durch die SMT abgelöst wurde.

So verfolgen Koehn und Senellart (2010b) das Ziel, den Retrieval-Mechanismus eines TMs mit einem SMT-System zu ergänzen, sodass anstelle von Fuzzy-Matches vollständige Übersetzungen ausgegeben werden. In der besten im TM gefundenen Übersetzungseinheit werden zunächst die gematchten und nicht gematchten Wörter zwischen AS_{neu} und ZS_{TM} identifiziert. Dazu wird einerseits die Edit Distance zwischen AS_{neu} und AS_{TM} errechnet und andererseits die AS- und ZS-Wörter der gefundenen Übersetzungseinheit aligniert. Die nicht gematchten AS-Wörter werden durch das SMT-System Moses übersetzt, während die gematchten ZS-Satzteile der gefundenen Übersetzungseinheit beibehalten werden.

Zhechev und van Genabith (2010a, 2010b) verwenden hingegen Methoden zur automatischen Alignierung von Teilbäumen sowie ein SMT-System, um vollständige Übersetzungen anstatt Fuzzy-Matches zu erzeugen. Wird ein Fuzzy-Match mit einem ausreichend hohen Match-Wert aufgefunden, können mithilfe der alignierten Teilbäume gematchte und nicht gematchte Satzfragmente zwischen dem AS_{neu} und ZS_{TM} identifiziert werden. Mit dem SMT-System wird dann die finale Übersetzung produziert. Dabei werden zwei Ansätze verfolgt: Einerseits werden die nicht gematchten Satzfragmente ohne jegliche Kontextinformation herausgefiltert. Die durch das SMT-System übersetzten Satzfragmente werden danach mit den durch das TM gematchten Satzfragmenten in der Satzstellung der Zielsprache aneinandergereiht. Andererseits werden die mit dem TM gematchten Satzfragmente mit ihren ZS-Entsprechungen unter Berücksichtigung des Kontextes annotiert. Der sich in der Satzstellung der Zielsprache befindliche, annotierte, komplette AS_{neu} wird durch Moses übersetzt.

Auch Biçici und Dymetman (2008) setzen SMT zur Optimierung von Fuzzy-Matches eines TMs ein. Das phrasenbasierte SMT-System MATRAX wird mithilfe der TM-Datensätze trainiert. Dabei wird ein Wortalignment aller AS_{TM} und ZS_{TM} durchgeführt sowie eine Sammlung von alignierten Phrasen erstellt, die Lücken enthalten können. Wird ein Fuzzy-Match durch

Identifikation der longest common subsequences zwischen AS_{neu} und AS_{TM} gefunden, werden die gematchten Sequenzen des AS_{TM} mittels der Wortalignment-Informationen übersetzt. Die entstandene Übersetzung wird mit einer Gewichtung versehen und der Phrasensammlung hinzugefügt. Die abschließende Übersetzung wird durch MATRAX erstellt.

Obwohl He et al. (2010a, 2010b) ebenfalls ein SMT-System in ein TM einbinden, liegt ihr Bestreben in der Erstellung eines Multi-Engine-Systems anstelle eines hybriden Systems mit der Absicht, den Post-Editing-Aufwand für den Übersetzer zu minimieren.

Dara et al. (2013) greifen diese Idee auf, indem sie ein System entwickeln, das die passendste Übersetzung – entweder aus einem TM oder aus einem MÜ-System – vorschlägt.

In dem von Esplà-Gomis et al. (2011) entwickelten System können hingegen mehrere MÜ-Systeme integriert werden. Das Vorhaben besteht nicht darin, durch die MÜ-Komponenten automatische Übersetzungen zu generieren, sondern die ZS-Wörter der gefundenen Übersetzungseinheit, die eines Post-Editings bedürfen, hervorzuheben. Dazu wird jedes ZS-Wort der gefundenen Übersetzungseinheit mit Merkmalen versehen, die durch die MÜ-Systeme und ggf. weitere bilinguale Ressourcen erzeugt werden. Mithilfe der Merkmale kann klassifiziert werden, welche Wörter durch den Übersetzer bearbeitet werden müssen.

3.3.2 Linguistisch optimierte Translation Memorys

Hinsichtlich der linguistischen Optimierung von TMs finden sich in der Literatur sowohl theoretische Anregungen als auch praktische Untersuchungen. Häufig wird dabei eine Optimierung mittels Segmentierung von Segmenten in linguistisch motivierte Subsegmente in Betracht gezogen, da die Wahrscheinlichkeit höher ist, einen 100 %-Match zwischen kurzen Phrasen zu finden als zwischen kompletten, langen Segmenten (vgl. Schäler 2001: 51).

Das von Schäler (2001) erstellte hybride System besteht aus einem TM zur Identifikation von 100 %-Matches, einer MÜ-Komponente zur Übersetzung von No-Matches und einem Phrasenlexikon zur Verarbeitung von Fuzzy-Matches. Der Fokus seiner Forschung liegt in der Erstellung des Phrasenlexikons, das linguistisch motivierte Phrasen beinhaltet, die aus Übersetzungseinheiten des TMs herrühren. Dazu werden die AS- und ZS-Seite der Übersetzungseinheiten mithilfe zweier sprachenabhängiger morphologischer Wörterbücher, eines Parsers und einer Grammatik in Segmente unterhalb der Satzebene zerlegt, die mit linguistischen Merkmalen versehen und im Phrasenlexikon gesammelt werden. Ein AS_{neu}, zu dem nur ein Fuzzy-Match im

TM gefunden wurde, wird auf die gleiche Weise analysiert und annotiert. Die so erstellten Phrasen des AS_{neu} werden mit den im Phrasenlexikon gespeicherten Phrasen und deren linguistischen Merkmalen verglichen. Die Übersetzung des AS_{neu} wird schließlich erstellt, indem die gematchten ZS-Phrasen aneinandergefügt werden.

In dem von Simard und Langlais (2001) entwickelten TM werden die Übersetzungseinheiten sowie das AS_{neu} in syntaktische Subsegmente zerlegt. Dem Übersetzer werden letztlich die durch Methoden des Wortalignments gematchten ZS-Subsegmente angezeigt. Nicht gematchte Satzteile müssen manuell übersetzt werden.

Gotti et al. (2005) führen ein Experiment für die Sprachkombination Französisch-Englisch durch, bei dem die Übersetzungseinheiten eines TMs und ein AS_{neu} auf vier verschiedene Arten segmentiert und anschließend gematcht werden. Bei den Segmentierungs- bzw. Matching-Ebenen handelt es sich um komplette Sätze, beliebige Wortketten, linguistisch motivierte Subsegmente und Syntaxbäume. Letztere ermöglichen auch den Vergleich nicht aufeinanderfolgender Wörter. Die Testergebnisse zeigen, dass eine Segmentierung bzw. ein Matching auf Ebene der linguistisch motivierten Subsegmente sowie der Syntaxbäume das beste Verhältnis zwischen Recall und Precision liefert.

Vanallemeersch und Vandeghinste (2014) versuchen für das Sprachenpaar Englisch-Niederländisch, den Recall eines TMs durch Vergleich syntaktischer und lexikalischer Informationen zu verbessern. Der Einsatz von Suffix-Arrays beschleunigt dabei das Matching.

Neben der Möglichkeit, TMs durch die Segmentierung der Übersetzungseinheiten in linguistisch motivierte Subsegmente zu optimieren, weisen Macklovitch und Russell (2000: 141ff.) darauf hin, dass auch durch eine Generalisierung von Wortformen infolge von Stemming oder Lemmatisierung eine höhere Match-Rate erzielt werden kann. POS-Tagging hilft zudem bei der Disambiguierung von Homografen. Die verschiedenen Repräsentationen eines Satzes (Originalform der Wörter, Grundformen, POS-Tags etc.) können dann im TM gespeichert werden.

Eine solche mehrschichtige Struktur, die sogenannte TELA[44]-Struktur, beschreiben Planas und Furuse (1999, siehe auch Planas 1998). Sie schlagen acht verschiedene, miteinander verknüpfte Ebenen vor. Die Ebenen beinhalten zum einen Ergebnisse linguistischer Analysen (POS-Tags, Lemmata und Informationen zur Satzstruktur), zum anderen aber auch nicht linguistische

[44] Französisches Akronym für *Treillis Etagés et Liés pour le traitement Automatique* (vgl. Planas/Furuse 1999: 332).

Informationen (einzelne Zeichen des Segmentes, Wörter in ihrer Originalform, XML-Tags zur Identifikation von Layout-Attributen oder grafischer Elemente sowie Glossareinträge). Die Konzeption weiterer Ebenen, z. B. einer Ebene mit semantischen Informationen, ist ebenso denkbar. Je nach Anforderung können alle oder nur ausgewählte Ebenen für die Ähnlichkeitsberechnung zwischen einem AS_{TM} und dem AS_{neu} zum Einsatz kommen.

Rapp (2002) schlägt ebenfalls ein TM bzw. EBMT-System vor, in dem die Übersetzungseinheiten in Form von POS-Tags hinterlegt sind, die für das Matching mit einem AS_{neu} herangezogen werden. Über ein bilinguales Wortformenwörterbuch, das ebenso Informationen zu den Wortarten der einzelnen Wörter beinhaltet, kann eine syntaktische Disambiguierung der potenziellen Matches erfolgen. Durch die statistische Ermittlung der Häufigkeiten von gemeinsam auftretenden Wörtern soll eine semantische Disambiguierung stattfinden.

Flanagan (2014) setzt in seinem TM namens Lift, das exemplarisch in SDL Trados Studio 2014 integriert wurde, neben einem Tokenizer, Stoppwortlisten und bilingualen Wörterbüchern einen Stemmer und Lemmatisierer ein, um die Wörter zwischen AS- und ZS-Segmenten zu alignieren und somit den Recall von Subsegmenten zu verbessern.

Einen Ansatz zur Optimierung von TMs unter Hinzuziehung semantischer Ressourcen stellen hingegen Elita und Gavrila (2006) vor. Ihre Idee umfasst die Erstellung generalisierter Phrasen, die mit semantischen Informationen annotiert werden, um Homografen semantisch zu disambiguieren. Mit einer fachspezifischen Ontologie werden die aus einem bilingualen Korpus extrahierten generalisierten Phrasen mit fachspezifischen Benennungen annotiert.

Gupta und Orăsan (2014) integrieren semantisches Wissen in ein TM, indem sie Paraphrasen – bestehend aus einem oder mehreren Wörtern – der im TM gespeicherten Segmente mittels einer Paraphrasendatenbank identifizieren und die Ergebnisse im TM speichern.

Chatzitheodorou (2015) liefert einen weiteren Ansatz, bei dem sich Paraphrasen bedient wird. Die in den AS-Segmenten eines TMs enthaltenen Funktionsverbgefüge werden mithilfe von Wörterbüchern und Grammatiken ermittelt und daraufhin durch passende Paraphrasen ersetzt.

Auch im Forschungsvorhaben von Mitkov und Corpas (2008) werden die Segmente eines TMs hinsichtlich ihrer semantischen Struktur analysiert. Zudem kommen syntaktische Analysen zum Einsatz. Ihr Ziel ist dem des iMem-Forschungsprojektes ähnlich: eine Optimierung des Retrievals bedeutungsgleicher Sätze, die unterschiedliche syntaktische Strukturen aufweisen. Der Fokus ihrer Forschung liegt jedoch in der Identifikation textsortenspezifischer

rhetorischer Prädikate[45], mit denen die Segmente eines TMs annotiert werden. Segmente werden demnach als semantisch ähnlich erachtet, wenn die rhetorischen Prädikate übereinstimmen.

Das Augenmerk des iMem-Forschungsprojektes liegt dagegen auf der Identifikation der LCS zwischen einem AS_{neu} und einem AS_{iMem} in Form ihrer Basiswörter sowie der morphosyntaktischen Unterschiede zwischen den ermittelten LCS unter Zuhilfenahme morphosyntaktischer Analysen. Ebenso wird untersucht, ob aus dem AS_{iMem} Basiswörter gelöscht oder dem AS_{iMem} Basiswörter hinzugefügt werden müssen, um das AS_{neu} zu erhalten. Mithilfe eines selbst entwickelten Proximitätsmaßes wird die Ähnlichkeit zwischen einem AS_{neu} und einem AS_{iMem} berechnet und letztlich in einem Ähnlichkeitswert angegeben. In die Berechnung werden die Anzahl und Länge der LCS, die morphosyntaktischen Unterschiede sowie Hinzufügungen oder Löschungen einbezogen.

Auf den Vergleich semantischer Informationen wird im Prototyp des iMem-TMs verzichtet, da MPRO als hinzugeschaltetes morphosyntaktisches Analyseprogramm semantische Analyseergebnisse liefert, die für die Zwecke des iMem-Forschungsprojektes nicht ausreichend sind. Aus diesem Grund wird davon ausgegangen, dass zwei AS-Segmente bedeutungsgleich sind, wenn ihre Basiswörter mit ihren morphosyntaktischen Merkmalen identisch sind. Dennoch wäre die Ausweitung auf einen Vergleich semantischer Informationen denkbar, sofern MPRO diesbezüglich optimiert würde.

Im Gegensatz zu einigen der erwähnten Forschungsansätzen zur linguistischen Optimierung von TMs wird beim iMem-Forschungsprojekt versucht, die Komplexität des iMem-TMs so gering wie möglich zu halten. Dies geschieht durch die Verwendung von GSAs, mittels derer die LCS in nur einer Repräsentation, nämlich in Form ihrer Basiswörter, und zudem mit geringem Speicherplatzbedarf und Zeitaufwand ermittelt werden. Des Weiteren wird mit MPRO nur eine linguistische Ressource in den Analyseprozess eingebunden, was gleichzeitig zu einer stabileren Arbeitsumgebung führt.

[45] Ein rhetorisches Prädikat gibt die Funktion eines Satzteiles an, z. B. „*topic, background, methodology, solution* and *conclusion*“ (Mitkov/Corpas 2008, Hervorhebung im Original).

4 Modell eines linguistisch optimierten Translation Memorys

Die in dieser Arbeit vorgestellte Idee zur Modellierung eines linguistisch optimierten TMs gliedert sich in zwei Aspekte: Zum einen wird ein Algorithmus beschrieben, mit dem die LCS zwischen dem AS_{neu} und einem im linguistisch optimierten TM gespeicherten AS-Segment, basierend auf den Basiswörtern der einzelnen Wortformen, ermittelt werden. Zum anderen wird ein selbst definiertes Proximitätsmaß vorgestellt, bei dem die Anzahl und Länge der LCS, die morphosyntaktischen Unterschiede zwischen den gematchten Wörtern der LCS sowie Hinzufügungen und Löschungen zwischen dem AS_{neu} und dem im linguistisch optimierten TM gespeicherten AS-Segment berücksichtigt werden.

Bevor diese beiden Kernstücke detailliert erläutert werden, wird zunächst ein kurzer Überblick darüber gegeben, welche Komponenten für eine linguistische Optimierung von TMs überhaupt erforderlich sind.

4.1 Notwendige Komponenten zur linguistischen Optimierung von Translation Memorys

Um ein TM linguistisch anzureichern, sind zwei wesentliche Komponenten erforderlich: zum einen ein TM und zum anderen ein Programm zur computerlinguistischen Analyse. Je nachdem, in welchem Umfang die linguistische Optimierung erwünscht ist, muss das Analyseprogramm morphologische, syntaktische, semantische etc. Analysen durchführen können.

4.1.1 Programm zur computerlinguistischen Analyse

Ein Programm zur computerlinguistischen Analyse erfüllt „spezielle Verarbeitungsaufgaben als Komponente […] umfangreicher Sprachverarbeitungssysteme […]. Dieses kann die Erkennung von Wortarten oder Grundformen sein, von Namen und Daten der verschiedensten Art, die Erkennung der Sprache in einem Text oder die Herstellung einer normalisierten Textversion als Vorbereitung für weitere Verarbeitungsschritte“ (Lobin 2009: 99).

Für die linguistische Optimierung eines TMs ist ein Analyseprogramm erforderlich, das sowohl die Originalform als auch die Basis eines Wortes sowie grammatische Kategorien wie Genus, Numerus, Kasus, Tempus, Wortart etc. codiert. Ebenso sollte die Erkennung einzelner Kompositumsbestandteile möglich sein.

Ein solches System kann auf zwei Arten konzipiert sein: entweder als Aufführung aller Wortformen zu einem Wort inklusive der Ergebnisse der morphosyntaktischen Analyse in einem sogenannten Vollformenlexikon oder als regelgesteuertes System (vgl. Fitschen 2004: 21). Erstere Variante birgt den Nachteil, dass aufgrund des stetig wachsenden Wortschatzes eine vollständige Liste nicht realisierbar ist; Wortformen, die nicht erfasst wurden, werden nicht erkannt (vgl. Fitschen 2004: 22). Ein Beispiel für ein System, das auf dieser Grundlage arbeitet, ist CELEX Lexical Database[46].

Bei einem regelgesteuerten System hingegen sind bestimmte Regeln zur Flexion und Wortbildung hinterlegt. Unter Verwendung verschiedener Lexika (z. B. Stammlexikon, Lexemlexikon, Morphemlexikon) können alle Wortformen aus den gespeicherten Einträgen zusammengesetzt werden (vgl. Fitschen 2004: 22). Dies kann jedoch zu einer sogenannten Übergenerierung führen, bei der auch ungültige Wortformen durch die Aneinanderreihung beliebiger Morpheme entstehen können. Die Systeme GERTWOL, Morphy und MPRO[47] sind Beispiele regelgesteuerter Analyseprogramme.

4.1.2 Translation Memory

Das von der Verfasserin dieser Arbeit entwickelte linguistisch optimierte TM verfolgt den datenbankbasierten Ansatz, wie die meisten auf dem Markt verfügbaren TMs. Dennoch wäre es auch möglich, ein referenztextbasiertes TM linguistisch anzureichern. Entscheidend ist, dass ein Fuzzy-Match-Algorithmus existiert, um nicht nur identische, sondern auch ähnliche gespeicherte AS-Segmente aufzufinden. Ebenso muss die Möglichkeit bestehen, über eine Trefferanzeige die gefundenen Übersetzungseinheiten anzuzeigen, Übersetzungseinheiten im TM abzuspeichern und zu löschen sowie die gespeicherten ZS-Entsprechungen für die Übersetzung des AS_{neu} verwenden und bearbeiten zu können.

Das entwickelte linguistisch optimierte TM ist ein eigenständiges Programm. Es besteht zum einen aus einer relationalen Datenbank, die als Speicher für die Ergebnisse der morphosyntaktischen Analyse dient, und zum anderen aus einem Algorithmus, mit dessen Hilfe dem AS_{neu} potenziell ähnliche gespeicherte AS-Segmente unter Zuhilfenahme morphosyntaktischer Merkmale aufgefunden und mit dem AS_{neu} gematcht werden. Durch ein selbst

[46] Fitschen (2004: 66ff.) liefert eine Beschreibung der deutschsprachigen Komponente des Systems.

[47] Detaillierte Informationen zur Funktionsweise von GERTWOL finden sich in Haapalainen und Majorin (1994: o.S.), von Morphy in Fitschen (2004: 23ff.) und von MPRO in Kapitel 5.1.2 dieser Arbeit. Weitere Systeme werden in Roth (2014: 94ff.) vorgestellt.

entwickeltes Proximitätsmaß wird den gematchten Segmenten ein Ähnlichkeitswert zugeschrieben. Das Anzeigen, Einfügen, Speichern, Löschen und Bearbeiten von Übersetzungseinheiten erfolgt über den Übersetzungseditor und die Trefferanzeige des kommerziellen TM-Systems. Dazu wird das linguistisch optimierte TM mithilfe eines Plug-in-Moduls in das kommerzielle TM-System eingebunden.

Der Grund für diese Vorgehensweise ist, dass der Übersetzer die Möglichkeit haben soll, Treffer aus beiden Systemen bzw. nur die linguistisch optimierten Treffer oder nur diejenigen des kommerziellen TMs angezeigt zu bekommen, ohne seine gewohnte Übersetzungsumgebung verlassen zu müssen. Ebenso soll der Übersetzer durch die (De-)Aktivierbarkeit des Plug-ins die Freiheit haben, zu entscheiden, ob linguistisch optimierte Übersetzungseinheiten dargeboten werden sollen oder nicht.

4.2 Algorithmus

Der nachfolgend beschriebene und im iMem-Forschungsprojekt entwickelte Algorithmus erläutert die Vorfilterung des linguistisch optimierten TMs zur Auffindung der N potenziell bestmatchenden, im linguistisch optimierten TM gespeicherten AS-Segmente sowie die Erstellung von GSAs zur Ermittlung der LCS zwischen einem AS_{neu} und jedem der potenziell matchenden Segmente. Die Vorfilterung wird lediglich einmal durchlaufen, die Erstellung des GSAs zur Ermittlung der LCS kann jedoch öfter erfolgen, abhängig von der Länge der zu vergleichenden Segmente und der Anzahl an matchenden Basiswörtern.

Voraussetzung für den Algorithmus ist, dass sowohl das AS_{neu} als auch jedes potenziell matchende gespeicherte AS-Segment zuvor durch ein morphosyntaktisches Analyseprogramm analysiert wurden. Komposita werden dabei in ihre Bestandteile zerlegt. Die LCS werden auf Grundlage der aus der morphosyntaktischen Analyse gewonnenen Basiswörter der einzelnen Wortformen der beiden zu vergleichenden Segmente ermittelt. Falls das morphosyntaktische Analyseprogramm mehrere Lesarten einer Wortform codieren kann, wird lediglich das Basiswort der ersten Lesart aus der Liste aller Lesarten dieser Wortform für den Vergleich der beiden Segmente herangezogen. Leerzeichen zwischen den einzelnen Basiswörtern teilen die Zeichenkette in linguistische Einheiten auf, sodass im GSA die vollständigen Basiswörter und nicht einzelne Buchstaben als zu vergleichende Einheiten betrachtet werden.

Der Unterschied zu anderen auf der Datenstruktur der GSAs basierenden Algorithmen besteht darin, dass beim im iMem-Forschungsprojekt

entwickelten Algorithmus identische sich wiederholende LCS innerhalb desselben Segmentes nicht miteinander gematcht werden. Stattdessen besteht die Herausforderung darin, zu berechnen, welche der identischen sich wiederholenden LCS des einen Segmentes Match-Partner eines LCS des anderen Segmentes ist. Dazu werden die Positionen der Suffixe herangezogen: Es wird angenommen, dass zwei Strings umso ähnlicher sind, je geringer die Differenz ihrer Positionen zwischen den Segmenten ist.

4.2.1 Vorfilterung

Damit die Ermittlung der LCS und die Anwendung des Proximitätsmaßes nicht mit jedem im linguistisch optimierten TM gespeicherten AS-Segment durchlaufen werden muss, die Rechenzeiten also akzeptabel gestaltet werden können und der Übersetzer nicht mit unbrauchbaren Übersetzungsergebnissen überflutet wird, erfolgt eine Vorfilterung, um die maximal N besten potenziellen Match-Partner für das AS_{neu} zu bestimmen. Je mehr Basiswörter des AS_{neu} in einem gespeicherten AS-Segment enthalten sind, desto wahrscheinlicher ist das gespeicherte AS-Segment einer der besten potenziellen Match-Partner.

Die Vorfilterung besteht aus verschiedenen Arbeitsschritten, die teilweise auf die Basiswörter des AS_{neu} und teilweise auf die Basiswörter der im linguistisch optimierten TM gespeicherten AS-Segmente angewendet werden. Bei den Arbeitsschritten handelt es sich um Normalisierungen der Daten, Datenbankabfragen sowie um die Filterung der Ergebnisse der vorangegangenen Arbeitsschritte. Die Auswahl der Filterkriterien wurde durch Carroll (1992) sowie durch Koehn und Senellart (2010a) motiviert. Die einzelnen Arbeitsschritte werden nachfolgend chronologisch aufgeführt:

1. *Identifikation aller eindeutigen Inhaltswörter im AS_{neu}:* Bei der Vorfilterung wurde sich auf die Inhaltswörter, d. h. auf die bedeutungstragenden Wörter, beschränkt, da die Untersuchung der Bedeutungsähnlichkeit zweier zu vergleichender Segmente beim iMem-Forschungsprojekt im Fokus steht. Dieses Vorgehen kann bereits bei Carroll (1992) beobachtet werden:

   > „The first function *(U(x)* assigns a cost to each word in the sentence. This indicates the importance of that word. So keywords (i.e. technical terms) may weIl be weighted more heavily than other (non-technical) content words. These, in turn, may be weighted more heavily than syntactic words (e.g. determiners. prepositions. auxiliary verbs).“ (Carroll 1992: 3)

Durch die morphosyntaktische Analyse kann für jedes Wort im AS_{neu} die Wortart codiert werden. Handelt es sich dabei um ein Substantiv, Verb, Adjektiv oder Adverb, wird das Wort als Inhaltswort markiert. Tritt die gleiche Basis eines Inhaltswortes mehrfach innerhalb des AS_{neu} auf, wird dieses Inhaltswort so gezählt, als ob es nur einmal im AS_{neu} auftreten würde, da ein Heranziehen aller Inhaltswörter (nicht nur der eindeutigen) zur Folge haben könnte, dass ein Datenbanksegment irrtümlich als zu ähnlich erachtet würde.[48]

2. Anwendung einer *vordefinierten Stoppwortliste:* Die Stoppwortliste enthält derzeit zwölf Verben, bei denen es sich um die Modalverben *brauchen, dürfen, können, mögen, müssen, sollen* und *wollen*, um die Hilfsverben *haben*, *sein* und *werden* sowie um die Handlungsverben *tun* und *machen* handelt. Wurde eines dieser Wörter im ersten Arbeitsschritt der Vorfilterung ermittelt, wird es aus der Liste der Inhaltswörter des AS_{neu} gelöscht.
3. *Bestimmung der Anzahl* der *im AS-Segment des linguistisch optimierten TMs vorkommenden Inhaltswörter des AS_{neu}:* In diesem Schritt erfolgt die Datenbankabfrage bzw. die Filterung der im linguistisch optimierten TM gespeicherten AS-Segmente. Jedes im linguistisch optimierten TM gespeicherte AS-Segment wird dahin gehend überprüft, ob es eine Mindestanzahl der eindeutigen Inhaltswörter des AS_{neu} aufweist. Ist dies der Fall, wird das gespeicherte AS-Segment für den nachfolgenden Filterschritt berücksichtigt. Bereits von Carroll (1992) wird dieses Vorgehen als Filterkriterium vorgeschlagen:

> „It was decided to use a system of keywords to both weight the metric definitions and to facilitate initial access of likely candidates for repetitions. By maintaining an index of sentences for each keyword and then only looking at sentences from the database with a threshold number of keywords in common with the current sentence we can greatly reduce the number of possible repetition sentences we have to consider." (Carroll 1992: 2)

Enthält ein gespeichertes AS-Segment weniger als die Mindestanzahl[49] der eindeutigen Inhaltswörter des AS_{neu}, wird es als potenzieller Match-Partner ausgeschlossen.

---

[48] Wenn beispielsweise im AS_{neu} dreimal dasselbe Inhaltswort auftaucht, im Datenbanksegment jedoch nur einmal, würde dieses Inhaltswort beim Vergleich beider Segmente dreimal gezählt anstatt nur einmal.

[49] Die im iMem-Forschungsprojekt verwendeten Werte für die Mindestanzahl an Inhaltswörtern, den maximalen Segmentlängenunterschied und die Anzahl an Einträgen in der sortierten Liste finden sich in Kapitel 5.2.2.1.

Da diese Filterung auf einem Zeichenkettenvergleich basiert, ist es nicht möglich, synonyme Inhaltswörter zwischen dem AS_{neu} und dem gespeicherten AS-Segment aufzufinden. Da jedoch nur eine bestimmte Anzahl an Inhaltswörtern in einem gespeicherten AS-Segment vorhanden sein muss, um als Match-Partner weiterhin berücksichtigt zu werden, besteht dennoch die Möglichkeit, dass paraphrasierte Segmente als Match-Partner identifiziert werden.

4. *Segmentlängenvergleich:* Die in Schritt 3 ermittelten potenziellen Match-Partner werden einem Segmentlängenvergleich unterzogen. Die Segmentlänge des gespeicherten AS-Segmentes darf dabei einen bestimmten Segmentlängenunterschied nicht über- oder unterschreiten, um als möglicher Match-Partner des AS_{neu} weiterhin infrage zu kommen. Andernfalls wird das gespeicherte AS-Segment für das weitere Vorgehen nicht mehr berücksichtigt.
5. *Sortierung der potenziellen Match-Partner:* Die bis zu diesem Schritt herausgefilterten potenziellen Match-Partner werden in einer Liste sortiert. Die Sortierung erfolgt absteigend gemäß der Anzahl der mit dem AS_{neu} übereinstimmenden Inhaltswörter. Folglich steht der Match-Partner mit den meisten übereinstimmenden Inhaltswörtern ganz oben in der Liste. Es liegt keine spezifische Sortierung vor für den Fall, dass mehrere potenzielle Match-Partner über die gleiche Anzahl an mit dem AS_{neu} übereinstimmenden Inhaltswörtern verfügen.
6. *Limitierung der potenziellen Match-Partner:* Enthält die sortierte Liste mehr als N Einträge, werden die überschüssigen, in der sortierten Liste zuletzt aufgeführten Einträge verworfen. Die Ermittlung der LCS und die Anwendung des Proximitätsmaßes werden demnach mit maximal N, hinsichtlich der übereinstimmenden Inhaltswörter zutreffendsten potenziellen Match-Partner durchgeführt.

4.2.2 Erstellung des GSAs

Um den Speicherplatzbedarf – insbesondere im Falle von sehr großen TMs – so gering wie möglich zu halten und ein schnelles Auffinden der LCS zwischen einem (sehr langen) AS_{neu} und einem (sehr langen) im linguistisch optimierten TM gespeicherten AS-Segment zu ermöglichen, wurde die Datenstruktur der GSAs gewählt, erweitert um zwei zusätzliche Listen.

4.2.2.1 Begriffserklärung

Das GSA verfügt neben der Information zur Segment-ID und zur Suffixposition über zwei weitere Arrays: das GLCPA (siehe Kapitel 2.2.1.2) und ein

zusätzliches Array, das in dieser Arbeit unter der Benennung *depth array* (*Tiefen-Array*, *DA*) eingeführt wird. Das GLCPA gibt die Länge des längsten gemeinsamen Präfixes zwischen zwei aufeinanderfolgenden Suffixen innerhalb des GSAs an – gleichgültig, ob die Suffixe aus demselben Segment oder aus unterschiedlichen Segmenten herrühren. Das DA dient hingegen dazu, die Länge des längsten gemeinsamen Präfixes zwischen zwei aufeinanderfolgenden Suffixen innerhalb des GSAs anzugeben, mit der Einschränkung, dass die aufeinanderfolgenden Suffixe aus zwei *unterschiedlichen* Segmenten stammen. Der Begriff *depth* (bzw. *Tiefe*) wurde dabei gewählt, um die Länge des längsten gemeinsamen Präfixes in einem generalisierten Suffix-Baum von der Wurzel bis zum LCA zwischen zwei *unterschiedlichen* Segmenten zu benennen.

4.2.2.2 Eigenschaften der Suffixe

Die zu vergleichenden Segmente werden – zunächst unsortiert – in ihre Suffixe zerlegt (Abbildung 7). Die Suffixe erfüllen dabei folgende Eigenschaften:

- Jedes Suffix ist mit der ID des Segmentes, aus dem es herrührt, sowie mit seiner Position im Segment markiert.
- Ein Suffix besteht aus den durch Leerzeichen separierten Basiswörtern des Segmentes. Das kürzeste Suffix verfügt über genau ein Basiswort, während das längste Suffix die Basiswörter des gesamten Segmentes enthält.
- Das Suffix kann nachträglich markiert werden, um festzustellen, ob es mit einem Suffix des anderen Segmentes gematcht wurde oder nicht.
- Jedes Basiswort des Suffixes kann ebenso markiert werden.

Segment A (id = 0)	Segment B (id = 1)
d g d d g d f e b a	d g d g c e b a
a	a
b a	b a
e b a	e b a
f e b a	c e b a
d f e b a	g c e b a
g d f e b a	d g c e b a
d g d f e b a	g d g c e b a
d d g d f e b a	d g d g c e b a
g d d g d f e b a	
d g d d g d f e b a	

Abbildung 7: Beispiel einer Zerlegung zweier zu vergleichender Segmente in ihre Suffixe. Die Buchstaben stehen stellvertretend für die Basiswörter der Wortformen. Gleiche Buchstaben bedeuten dabei gleiche Basiswörter.

Die Suffixe von Segment A (id = 0) und Segment B (id = 1) werden anschließend in einer gemeinsamen Liste alphabetisch aufsteigend sortiert, wobei keine Unterscheidung zwischen Groß- und Kleinschreibung erfolgt. Besitzen Segment A und Segment B das gleiche Suffix, wird eine weitere aufsteigende Sortierung nach Segment-ID vorgenommen. Es wird demzufolge ein GSA erstellt, in dem die Suffixe mit ihrer Segment-ID und Suffixposition aufgelistet werden. Alle Werte des GLCPAs und des DAs sind zu diesem Zeitpunkt unbestimmt. Dies entspricht dem Wert 0 (Abbildung 8).

id = 0	pos = 9	lcp = 0	depth = 0	a
id = 1	pos = 7	lcp = 0	depth = 0	a
id = 0	pos = 8	lcp = 0	depth = 0	b a
id = 1	pos = 6	lcp = 0	depth = 0	b a
id = 1	pos = 4	lcp = 0	depth = 0	c e b a
id = 0	pos = 2	lcp = 0	depth = 0	d d g d f e b a
id = 0	pos = 5	lcp = 0	depth = 0	d f e b a
id = 1	pos = 2	lcp = 0	depth = 0	d g c e b a
id = 0	pos = 0	lcp = 0	depth = 0	d g d d g d f e b a
id = 0	pos = 3	lcp = 0	depth = 0	d g d f e b a
id = 1	pos = 0	lcp = 0	depth = 0	d g d g c e b a
id = 0	pos = 7	lcp = 0	depth = 0	e b a
id = 1	pos = 5	lcp = 0	depth = 0	e b a
id = 0	pos = 6	lcp = 0	depth = 0	f e b a
id = 1	pos = 3	lcp = 0	depth = 0	g c e b a
id = 0	pos = 1	lcp = 0	depth = 0	g d d g d f e b a
id = 0	pos = 4	lcp = 0	depth = 0	g d f e b a
id = 1	pos = 1	lcp = 0	depth = 0	g d g c e b a

Abbildung 8: GSA mit GLCPA und DA: Zu Beginn des Algorithmus sind alle Werte des GLCPAs und des DAs auf 0 gesetzt.

4.2.2.3 Erstellung des GLCPAs und des DAs

Zur Ermittlung der Werte des GLCPAs und des DAs wird das GSA paarweise von oben nach unten durchlaufen: Zunächst werden die Basiswörter der Suffixe des ersten und zweiten Eintrags, dann des zweiten und dritten Eintrags, dann des dritten und vierten Eintrags etc. von links nach rechts auf Gleichheit geprüft (Abbildung 9).

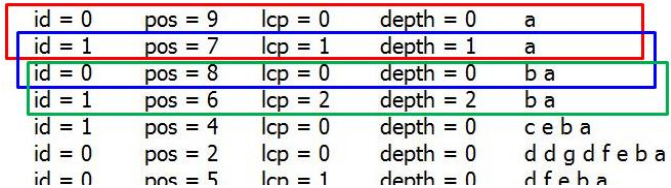

Abbildung 9: Paarweises Vergleichen der Suffixe im GSA von oben nach unten

Die Anzahl gleicher Basiswörter zwischen den beiden miteinander verglichenen Suffixen ist das LCP dieser Suffixe. Ist der LCP-Wert > 0, werden die Suffixe dahin gehend markiert, dass sie über x gleiche Basiswörter verfügen. Der LCP-Wert wird stets in die Spalte für das GLCPA (also $lcp = x$) in der Zeile des – in der paarweisen Betrachtung – zweiten Suffixes eingetragen. Bei diesem Vorgehen wird keine Unterscheidung getroffen, ob die miteinander verglichenen Suffixe aus demselben Segment oder aus unterschiedlichen Segmenten stammen. Dennoch spielen die LCP-Werte im weiteren Verlauf des Algorithmus eine bedeutsame Rolle, weswegen ihre Berechnung unerlässlich ist (siehe Kapitel 4.2.3.1). In Abbildung 10 ist das GLCPA vollständig erstellt. Die roten Markierungen um die LCP-Werte und zwischen den Buchstaben der Suffixe kennzeichnen die Anzahl gleicher Basiswörter zwischen zwei Suffixen.

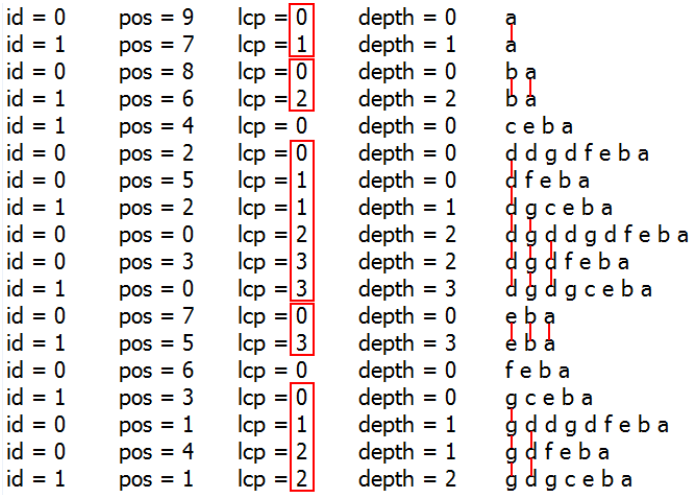

Abbildung 10: Vollständig erstelltes GLCPA

Für die Ermittlung der depth-Werte kommt die Betrachtung der Segment-IDs zum Einsatz: Stammen die miteinander verglichenen Suffixe aus unterschiedlichen Segmenten, entspricht der depth-Wert des – in der paarweisen Betrachtung – zweiten Suffixes dem LCP-Wert dieses zweiten Suffixes (Abbildung 11). Für den Fall, dass der LCP-Wert gleich 0 ist, beträgt der depth-Wert ebenfalls 0, da kein Suffix mit gleichem Basiswort vorausgeht.

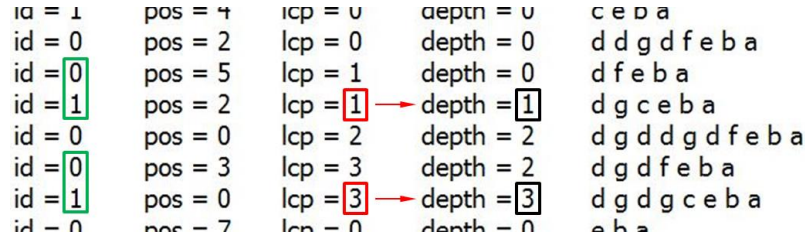

Abbildung 11: Erstellung des DAs: Stammen die Suffixe aus unterschiedlichen Segmenten, ist der depth-Wert gleich dem LCP-Wert.

Falls die miteinander verglichenen Suffixe aus demselben Segment herrühren, entspricht der depth-Wert des – in der paarweisen Betrachtung – zweiten Suffixes dem depth-Wert des ersten Suffixes (Abbildung 12). Diese Vorgehensweise liegt darin begründet, dass einerseits ein Segment nicht mit sich selbst gematcht werden kann; Wiederholungen innerhalb desselben Segmentes werden folglich nicht berücksichtigt. Andererseits kann sowohl das erste als auch das zweite Suffix des Segmentes A (id = 0) Match-Partner eines nachfolgenden Suffixes des Segmentes B (id = 1) sein.

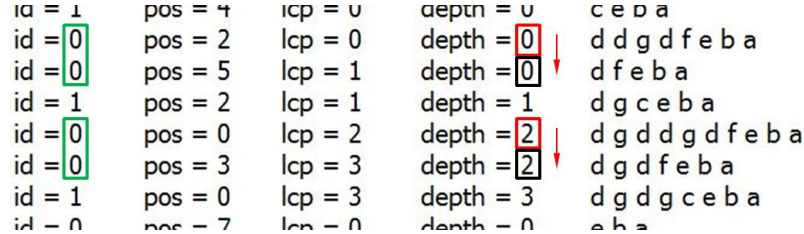

Abbildung 12: Erstellung des DA: Stammen die Suffixe aus demselben Segment, ist der depth-Wert gleich dem depth-Wert des vorangehenden Suffixes.

4.2.3 Ermittlung der LCS

Um herauszufinden, welche Suffixe des einen Segmentes (Segment A (id = 0)) Match-Partner der Suffixe des anderen Segmentes (Segment B (id = 1)) sind

und welche Basiswörter innerhalb der matchenden Suffixe die LCS zwischen den miteinander verglichenen Segmenten sind, müssen zunächst Blocks definiert werden. Aus diesen Blocks werden wiederum sich wiederholende Basiswörter entfernt.

4.2.3.1 Erstellung der Blocks

Der Sinn der Blockerstellung besteht darin, diejenige Menge aller Suffixe zusammenzufassen, die stets mit dem höchsten im GSA vorkommenden depth-Wert die längste Teilzeichenkette gemein hat. Zum Zeitpunkt der Blockerstellung wird noch nicht unterschieden, ob die im Block zusammengefassten Suffixe aus demselben Segment oder aus unterschiedlichen Segmenten stammen. Entscheidend ist zunächst nur, dass alle Suffixe, die über den spezifischen LCS verfügen, im weiteren Verlauf des Algorithmus zusammen weiter betrachtet werden können. Pro Durchlauf des Algorithmus kann immer nur ein Block erstellt und die LCS darin ermittelt werden.

Der Startpunkt zur Erstellung eines Blocks ist stets der Eintrag, dessen depth-Wert am höchsten ist, da dieser Eintrag den längsten LCS innehat. Existieren mehrere Einträge mit demselben depth-Wert (Abbildung 13), wird stets der letzte Eintrag im GSA mit diesem depth-Wert als Startpunkt ($depth_{Start}$) für die Blockerstellung gewählt.

id = 0	pos = 9	lcp = 0	depth = 0	a
id = 1	pos = 7	lcp = 1	depth = 1	a
id = 0	pos = 8	lcp = 0	depth = 0	b a
id = 1	pos = 6	lcp = 2	depth = 2	b a
id = 1	pos = 4	lcp = 0	depth = 0	c e b a
id = 0	pos = 2	lcp = 0	depth = 0	d d g d f e b a
id = 0	pos = 5	lcp = 1	depth = 0	d f e b a
id = 1	pos = 2	lcp = 1	depth = 1	d g c e b a
id = 0	pos = 0	lcp = 2	depth = 2	d g d d g d f e b a
id = 0	pos = 3	lcp = 3	depth = 2	d g d f e b a
id = 1	pos = 0	lcp = 3	depth = 3	d g d g c e b a
id = 0	pos = 7	lcp = 0	depth = 0	e b a
id = 1	pos = 5	lcp = 3	depth = 3	e b a
id = 0	pos = 6	lcp = 0	depth = 0	f e b a
id = 1	pos = 3	lcp = 0	depth = 0	g c e b a
id = 0	pos = 1	lcp = 1	depth = 1	g d d g d f e b a
id = 0	pos = 4	lcp = 2	depth = 1	g d f e b a
id = 1	pos = 1	lcp = 2	depth = 2	g d g c e b a

Abbildung 13: Blockerstellung: Einträge mit den höchsten depth-Werten. Der höchste depth-Wert wird als Startpunkt bevorzugt. Bei gleich großen depth-Werten wird stets der letzte Eintrag im GSA mit diesem depth-Wert als Startpunkt für die Blockerstellung gewählt.

Nach der Bestimmung des Startpunktes wird das GLCPA so lange nach oben durchlaufen, bis der LCP-Wert kleiner $depth_{Start}$ ist. Alle diese Einträge, d. h. diejenigen mit LCP ≥ $depth_{Start}$[50] sowie der erste Eintrag mit LCP < $depth_{Start}$, werden als ein Block markiert (Abbildung 14).

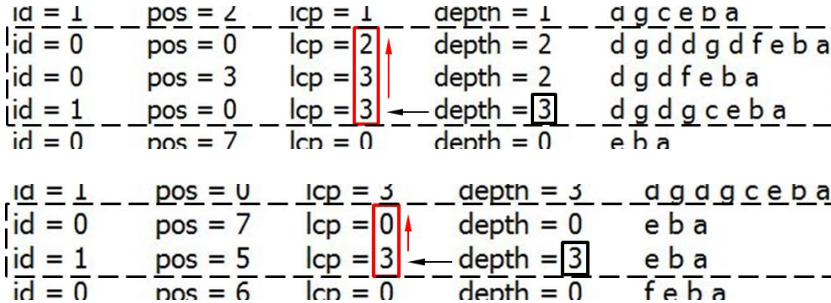

Abbildung 14: Blockerstellung: GLCPA so lange nach oben durchlaufen, bis LCP < $depth_{Start}$. In diesem Beispiel ist $depth_{Start}$ = 3.

4.2.3.2 Auffinden der LCS im Block

Falls der Block lediglich zwei Suffixe beinhaltet, sind diese beiden Suffixe gegenseitige Match-Partner. Der höhere der beiden depth-Werte gibt dabei die Länge des LCS an (Abbildung 15). Die Basiswörter in diesen beiden Suffixen werden bis zu dieser Länge als LCS markiert.

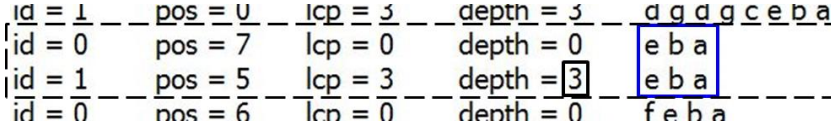

Abbildung 15: Auffinden der LCS im Block: Verfügt ein Block nur über zwei Suffixe, sind diese Suffixe Match-Partner voneinander. Die Länge des LCS entspricht dem höheren depth-Wert der beiden Suffixe.

Umfasst der Block jedoch mehr als zwei Suffixe, werden mittels Dynamischer Programmierung diejenigen Suffixe als Match-Partner bestimmt, deren Positionsdifferenz am geringsten ist. Dazu werden alle Positionen der Suffixe, die aus Segment A (id = 0) herrühren, mit den Positionen der Suffixe aus Segment B (id = 1) verglichen (Abbildung 16 und 17).

[50] Treten Wiederholungen in beiden Segmenten auf, kann es auch vorkommen, dass der LCP-Wert größer als $depth_{Start}$ ist.

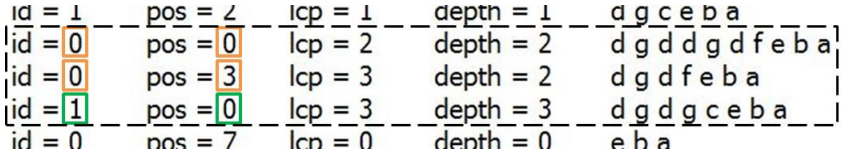

Abbildung 16: Auffinden der LCS im Block: Verfügt der Block über mehr als zwei Suffixe, müssen die LCS über die Positionsdifferenz ermittelt werden. Die geringste Positionsdifferenz wird bevorzugt.

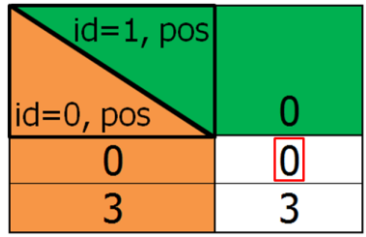

Abbildung 17: Auffinden der LCS im Block: Dynamische Programmierung zur Ermittlung der geringsten Positionsdifferenz. Die farbigen Felder beinhalten die Positionen im entsprechenden Segment. Die weißen Felder enthalten die absolute Positionsdifferenz. Die rot umrandete Positionsdifferenz markiert die besten Match-Partner.

Diejenigen Suffixe, die die geringste Positionsdifferenz aufweisen und zudem aus unterschiedlichen Segmenten stammen, sind gegenseitige Match-Partner. Der höhere der beiden depth-Werte steht dabei für die Länge des LCS (Abbildung 18). Die Basiswörter in den gematchten Suffixen werden bis zu dieser Länge als LCS markiert.

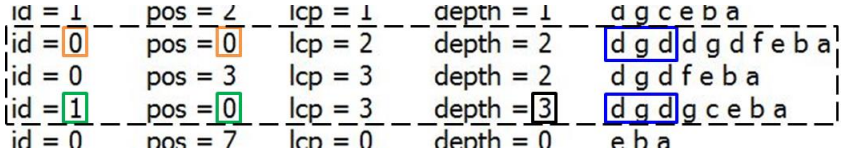

Abbildung 18: Auffinden der LCS im Block: Die Suffixe unterschiedlicher Segmente mit der geringsten Positionsdifferenz bilden den LCS dieses Blocks. Die Länge des LCS entspricht dem höheren depth-Wert der matchenden Suffixe.

Falls das Suffix länger als der depth-Wert ist, werden diejenigen Basiswörter, die den markierten Basiswörtern nachfolgen, entfernt. Übrig bleibt der für diesen Durchlauf des Algorithmus längste LCS zwischen dem AS_{neu} und dem im linguistisch optimierten TM gespeicherten AS-Segment (Abbildung 19).

Die aus dem Suffix entfernten Basiswörter stehen in einem weiteren Durchlauf des Algorithmus wieder für das Matching mit einem anderen Suffix zur Verfügung.

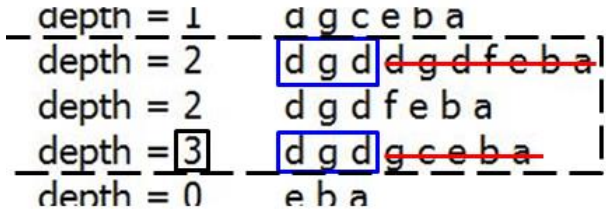

Abbildung 19: Auffinden der LCS im Block: Ist das Suffix länger als der depth-Wert, werden alle nachfolgenden Basiswörter aus dem Suffix entfernt. Der reine LCS bleibt übrig.

4.2.3.3 Bereinigung des GSAs

Nach der Ermittlung des längsten LCS im GSA muss das GSA bereinigt werden, damit der nächstlängste LCS zwischen den beiden miteinander verglichenen Segmenten ermittelt werden kann. Dies geschieht unter folgender Bedingung: Jedes als LCS markierte Basiswort wird in den anderen Suffixen des GSAs als Löschung markiert. Alle Basiswörter, die einem als Löschung markierten Basiswort folgen, werden ebenfalls als Löschung markiert.

Nachdem die als LCS markierten Basiswörter aus den anderen Suffixen im GSA in dem aktuellen Durchlauf des Algorithmus entfernt wurden, wird aus den übrig gebliebenen Suffixen ein neues GSA erstellt. Die Erstellung des neuen GSAs und die Ermittlung des nächstlängsten LCS zwischen den miteinander verglichenen Segmenten erfolgt auf dieselbe Weise wie zuvor beschrieben (siehe Kapitel 4.2.2.2 bis Kapitel 4.2.3.3).

In den nachfolgenden Grafiken (Abbildung 20 bis 23) wird die Bereinigung eines GSAs exemplarisch demonstriert: Im ersten Durchlauf des Algorithmus (Abbildung 20), bei dem der längste LCS zwischen den beiden Segmenten ermittelt wird, werden alle als LCS markierten Basiswörter im GSA entfernt. Daraufhin wird das GSA mit allen übrig gebliebenen Suffixen neu erstellt, wobei ebenso das GLCPA und das DA neu berechnet werden.

Im zweiten Durchlauf des Beispiels (Abbildung 21), bei dem der nächstlängste LCS aufgefunden wird, werden ebenfalls alle als LCS markierten Basiswörter (sowie alle diesen Basiswörtern nachfolgenden Basiswörter)[51] im

[51] Der Algorithmus ist so konzipiert, dass in jedem Durchlauf Basiswörter, die den als LCS markierten Basiswörtern nachfolgen, entfernt werden. Allerdings kann es auch Durchläufe

GSA entfernt. Auch hiernach wird das GSA einschließlich des GLCPAs und des DAs neu erstellt und wiederum der nächstlängste LCS ermittelt etc.

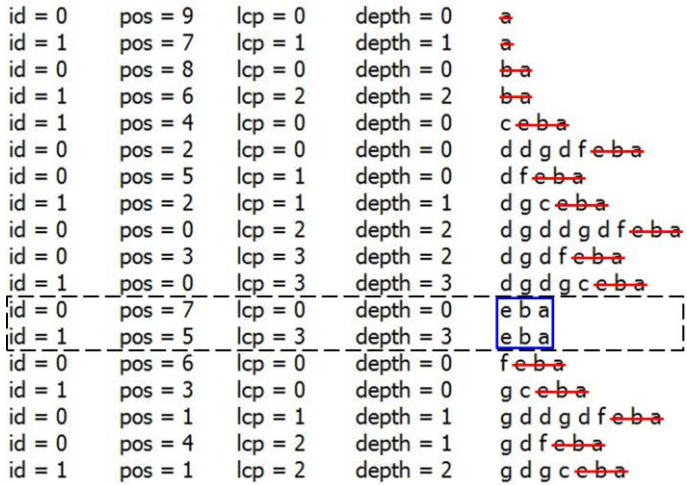

Abbildung 20: Bereinigung des GSAs (Durchlauf 1): Die als LCS markierten Basiswörter werden im GSA entfernt.

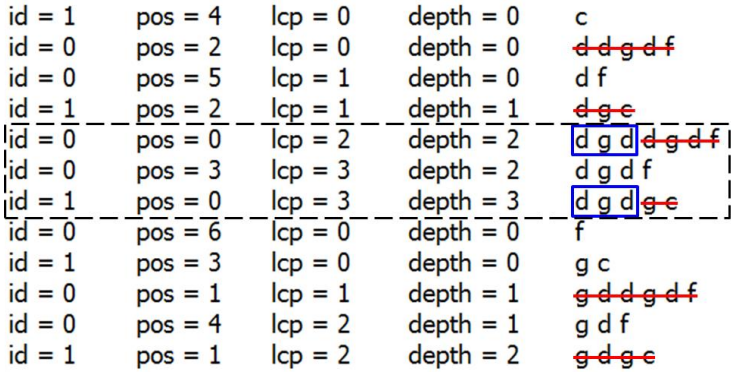

Abbildung 21: Bereinigung des GSAs (Durchlauf 2): Die als LCS markierten Basiswörter werden im GSA entfernt. Alle Basiswörter, die einem als Löschung markierten Basiswort folgen, werden ebenfalls als Löschung markiert.

---

geben, bei denen keine nachfolgenden Basiswörter zu entfernen sind (siehe z. B. Durchlauf 1, bei dem die als LCS markierten Basiswörter am Ende der zu vergleichenden Segmente auftreten).

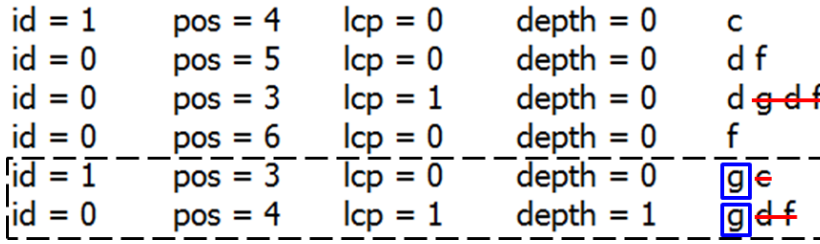

Abbildung 22: Bereinigung des GSAs (Durchlauf 3): Die als LCS markierten Basiswörter werden im GSA entfernt. Alle Basiswörter, die einem als Löschung markierten Basiswort folgen, werden ebenfalls als Löschung markiert.

Dieser Vorgang wiederholt sich solange, bis kein Block mehr gebildet werden kann, d. h., wenn

- nur noch ein Suffix im GSA vorhanden ist oder
- kein Suffix mehr im GSA vorhanden ist oder
- nur noch Suffixe im GSA vorhanden sind, deren depth-Werte = 0 sind (Abbildung 23).

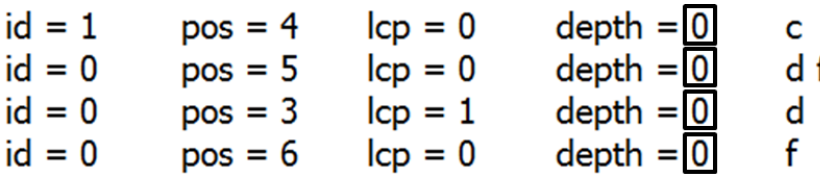

Abbildung 23: Ende des Algorithmus, da keine Blockbildung mehr möglich ist.

Das Ergebnis ist eine Liste mit allen aufgefundenen LCS zwischen dem AS_{neu} und dem im linguistisch optimierten TM gespeicherten AS-Segment (Abbildung 24).

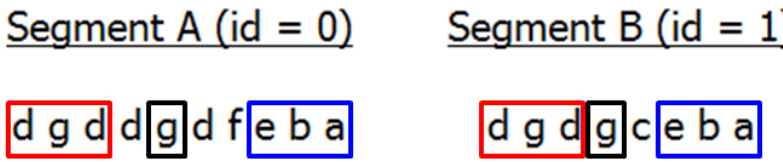

Abbildung 24: Ermittelte LCS unter der Voraussetzung, dass lange LCS sowie minimale Positionsdifferenzen bevorzugt werden.

Bei diesem Algorithmus wird sich Suffix-Arrays aus Gründen der Zeitkomplexität bedient, da sich die Zeitkomplexität zur Berechnung der LCS mithilfe von Suffix-Arrays lediglich auf $O(n)$ beläuft, während die Zeitkomplexität beispielsweise bei der Dynamischen Programmierung $O(nm)$ beträgt.

Die Datenstruktur der GSAs wurde durch die Verfasserin dieser Arbeit insofern erweitert, dass zunächst alle identischen sich wiederholenden LCS aus dem GSA als mögliche Match-Partner identifiziert werden, um anschließend – unter Berücksichtigung der geringsten Positionsdifferenz und der Segmentzugehörigkeit – die bestmögliche Übereinstimmung zwischen zwei Segmenten aufzufinden.

Dabei besteht der Unterschied zu anderen auf der Datenstruktur der GSAs basierenden Algorithmen darin, dass identische sich wiederholende LCS innerhalb desselben Segmentes nicht miteinander gematcht werden, sondern nur identische sich wiederholende LCS zwischen zwei verschiedenen Segmenten. Es wird angenommen, dass zwei Strings umso ähnlicher sind, je geringer die Differenz ihrer Positionen zwischen den Segmenten ist.

4.3 Proximitätsmaß

Nachdem alle LCS zwischen dem AS_{neu} und jedem potenziell matchenden, im linguistisch optimierten TM gespeicherten AS-Segment ermittelt wurden, erfolgt die Berechnung des Ähnlichkeitswertes bzw. des Match-Wertes M zwischen den beiden Segmenten mithilfe des durch die Verfasserin dieser Arbeit selbst entwickelten Proximitätsmaßes. Es wurde ein eigenes Proximitätsmaß erstellt, da keines der in Kapitel 2.2.1.3 vorgestellten Proximitätsmaße die Anforderungen an das Auffinden bedeutungsgleicher Segmente in einem linguistisch optimierten TM vollständig erfüllt. Die drei folgenden Bedingungen muss das Proximitätsmaß für die Zwecke dieser Arbeit erfüllen:

1. *Berücksichtigung der LCS-Struktur*, d. h. der Anzahl und der Länge der ermittelten LCS. Dabei werden sowohl Vertauschungen von Satzfragmenten als auch Wortersetzungen[52] beachtet.
2. *Berücksichtigung der morphosyntaktischen Unterschiede zwischen den identischen Basiswörtern der LCS*, indem die Werte spezifischer morphosyntaktischer Merkmale miteinander verglichen werden.

[52] Unter Wortersetzungen sind in diesem Fall nicht gematchte Wörter zwischen zwei gleichlangen Segmenten zu verstehen.

3. *Berücksichtigung der Anzahl nicht gematchter Basiswörter*, d. h. von zu löschenden oder hinzuzufügenden Basiswörtern zwischen den beiden Segmenten, um einen eventuellen Segmentlängenunterschied in den Match-Wert einzuberechnen.

Aus diesen Anforderungen wurde die nachfolgende Formel zur Berechnung des Match-Wertes erstellt:

$$M = 0{,}7 * {S_{LG}}^{\kappa} + 0{,}3 * S_{SL}$$

Zwei zu vergleichende Segmente besitzen einen Match-Wert M von 1, wenn Sie zu 100 % ähnlich – also identisch – sind. Zwei zu vergleichende Segmente werden als komplett unähnlich bezeichnet, wenn ihr Match-Wert 0 beträgt. Bei dem Proximitätsmaß handelt es sich demnach um ein Ähnlichkeitsmaß: Je höher der Wert M ist, desto ähnlicher sind die beiden miteinander verglichenen Segmente. Die Ähnlichkeit zweier zu vergleichender Segmente wird dabei als S (similarity) bezeichnet.

Die Variable S_{LG} repräsentiert die Ähnlichkeit der gematchten Basiswörter in Form von einem oder mehreren LCS sowie ihren grammatischen Eigenschaften (Bedingung 1, Bedingung 2). S_{LG} setzt sich aus dem Produkt (siehe Kapitel 4.3.3) der Ähnlichkeit der LCS-Struktur (L) und der Ähnlichkeit der grammatischen Struktur (G) der zwischen beiden Segmenten identischen Basiswörter zusammen. Zusätzlich erfolgt eine positive Korrektur dieses Produktes mithilfe des Korrekturfaktors κ (siehe Kapitel 4.3.3). Daraus ergibt sich folgende Umschreibung der oben genannten Formel:

$$M = 0{,}7 * (L * G)^{\kappa} + 0{,}3 * S_{SL}$$

Die Variable S_{SL} repräsentiert die Ähnlichkeit der Segmentlänge, d. h. den Grad des Segmentlängenunterschiedes zwischen den beiden Segmenten (Bedingung 3). Die Gewichtungen von 30 % und 70 % wurden gewählt, da sie einen guten Kompromiss zwischen der Ähnlichkeit der gematchten Basiswörter und dem Segmentlängenunterschied darstellen.

Motiviert durch Blatt (1998: 98) liegt die Berechnung von S_{SL} folgenden Annahmen zugrunde – gemessen am Arbeits- und Identifikationsaufwand, um aus einem gespeicherten AS-Segment das AS_{neu} zu erhalten[53]:

[53] Eigentlich müssten Hinzufügungen und Löschungen in die bzw. aus der ZS-Seite der im linguistisch optimierten TM gespeicherten Übersetzungseinheit erfolgen, um einen ZS_{neu} zu erzeugen. Da jedoch bei der Arbeit mit TMs das ZS_{neu} nicht bekannt ist, muss für dessen Erstellung die AS-Seite herangezogen werden, da sowohl immer das AS_{neu} als auch das im

1. Löschungen nicht gematchter Basiswörter aus dem gespeicherten AS-Segment erfordern einen geringeren Arbeitsaufwand als Hinzufügungen (= Neuübersetzungen) nicht gematchter Basiswörter in das gespeicherte AS-Segment.
2. Eine Löschung nicht gematchter Basiswörter *an einem Stück* aus dem gespeicherten AS-Segment erfordert weniger Arbeits- und Identifikationsaufwand als Löschungen nicht gematchter Basiswörter aus dem gespeicherten AS-Segment *an mehreren Stellen*.
3. Eine Löschung nicht gematchter Basiswörter an einem Stück aus einem gespeicherten AS-Segment kann wie folgt auftreten, wobei die drei Punkte die jeweilige Löschung symbolisieren:

   … LCS
   LCS **…**
   LCS **…** LCS

4. Löschungen nicht gematchter Basiswörter an mehreren Stellen aus einem gespeicherten AS-Segment können wie folgt auftreten, wobei die drei Punkte jeweils eine Löschung symbolisieren:

   … LCS **…**
   LCS **…** LCS **…** LCS **…** etc.
   … LCS **…** LCS **…** LCS etc.

5. Eine Hinzufügung nicht gematchter Basiswörter an einem Stück in ein gespeichertes AS-Segment kann wie folgt auftreten, wobei die Unterstriche die jeweilige Hinzufügung symbolisieren:

   _ LCS
   LCS **_**
   LCS **_** LCS

6. Hinzufügungen nicht gematchter Basiswörter an mehreren Stellen in ein gespeichertes AS-Segment können wie folgt auftreten, wobei die Unterstriche jeweils eine Hinzufügung symbolisieren:

---

linguistisch optimierten TM gespeicherte AS-Segment bekannt ist. Für das eigen erstellte Proximitätsmaß wird somit unterstellt, dass die gespeicherte ZS-Entsprechung und das ZS_{neu} genauso viele Basiswörter enthalten wie das entsprechende gespeicherte AS-Segment und das AS_{neu}.

_ LCS _
LCS _ LCS _ LCS _ etc.
_ LCS _ LCS _ LCS etc.

Die Berechnung von S_{SL} gemäß Kapitel 4.3.4 gilt für folgende Fälle:

LCS … LCS
LCS … LCS … LCS … etc.
… LCS … LCS … LCS etc.
_ LCS
LCS _
LCS _ LCS
_ LCS _
LCS _ LCS _ LCS _ etc.
_ LCS _ LCS _ LCS etc.

Für den Fall, dass das AS_{neu} komplett an einem Stück im gespeicherten AS-Segment enthalten ist und Basiswörter gelöscht werden müssen, d. h., wenn

… LCS
LCS …
… LCS …

tritt eine Ausnahmeregelung für die Berechnung des Match-Wertes in Kraft, bei der der Match-Wert pauschal auf 99 % gesetzt wird:

$$M = 99\ \%$$

Grund für diese Ausnahmeregelung ist, dass angenommen wird, dass die ZS-Entsprechung des im linguistisch optimierten TM gespeicherten AS-Segmentes ebenso über die komplette Übersetzung des AS_{neu} an einem Stück verfügt und somit die Löschung der zusätzlichen Wörter einen minimalen Arbeits- und Identifikationsaufwand für den Übersetzer bedeutet.

Umgekehrt wird für die Fälle, die durch die Berechnung von S_{SL} erfasst werden (siehe oben), ein größerer Arbeits- und Identifikationsaufwand angenommen, der sich folglich auch in der Höhe des Abzuges niederschlagen sollte.

In den nachfolgenden Unterkapiteln wird die Berechnung der Ähnlichkeit der LCS-Struktur, der grammatischen Struktur und der Segmentlänge sowie des Korrekturfaktors κ erläutert und anhand eines fortlaufenden Beispiels veranschaulicht.

4.3.1 Berechnung der Ähnlichkeit der LCS-Struktur (L)

Um die Ähnlichkeit der LCS-Struktur zu berechnen, muss die Anzahl der durch den zuvor durchlaufenen Algorithmus ermittelten LCS bekannt sein, ebenso wie die Anzahl der Basiswörter, die in jedem einzelnen LCS enthalten sind. Des Weiteren ist die Anzahl derjenigen Basiswörter erforderlich, die maximal zwischen dem AS_{neu} und einem im linguistisch optimierten TM gespeicherten AS-Segment gematcht werden könnten (b_{max}). Diese Anzahl entspricht stets dem Minimum der Anzahl an Basiswörtern der beiden Segmente.

Im folgenden Beispiel wurden zwei zu vergleichende Segmente mittels eines morphosyntaktischen Analyseprogramms in ihre Basiswörter zerlegt. Die Buchstaben repräsentieren die Basiswörter, wobei gleiche Buchstaben gleiche Basiswörter bedeuten. Durch den in Kapitel 4.2 beschriebenen Algorithmus konnten zwei LCS ermittelt werden (siehe Markierungen in Tabelle 16). Der eine LCS verfügt über drei Basiswörter, der andere über eins. Die maximale Anzahl matchbarer Basiswörter (b_{max}) beträgt 5.

	Segmente in Form ihrer Basiswörter	**Anzahl LCS**	**Anzahl der im LCS enthaltenen Basiswörter**	**Maximale Anzahl matchbarer Basiswörter (b_{max})**
AS_{neu}	a b c d e f g	2	{3, 1}	5
Im linguistisch optimierten TM gespeichertes AS-Segment	b c d h f			

Tabelle 16: LCS-Struktur zweier zu vergleichender Segmente

Sind alle erforderlichen Werte bekannt, findet eine Partitionierung statt, bei der b_{max} als Summe von Zahlen dargestellt wird. Jeder Darstellungsmöglichkeit wird dabei eine Rangnummer zugewiesen. Zudem wird berücksichtigt, auf wie viele LCS sich die Summe der Zahlen verteilt und wie viele der Basiswörter von b_{max} gematcht und nicht gematcht wurden. Durch diese Partitionierung mit Rangzuweisung ist es möglich, Positionsvertauschungen der LCS sowie Wortersetzungen besser zu bewerten und abzugrenzen. Die Partitionierung erfolgt nach folgenden Gesichtspunkten:

- Der höchste Rang ist derjenige, bei dem alle Basiswörter von b_{max} in nur einem LCS gematcht wurden.
- Der niedrigste Rang (= Rang 1) ist derjenige, bei dem nur ein Basiswort von b_{max} in nur einem LCS gematcht wurde.

- Eine Darstellung erhält einen umso höheren Rang, je mehr Basiswörter in einem LCS gematcht wurden.
- Die Reihenfolge der gematchten Basiswörter, d. h., ob z. B. zuerst ein und dann drei Basiswörter oder umgekehrt gematcht wurden, spielt für die Rangzuweisung keine Rolle.

Da im Falle sehr langer Segmente sehr viele Ränge bei der Partitionierung ermittelt werden müssten und dadurch die Rechenzeit erheblich ansteigen würde, wurde eine Liste vorberechnet und im Quellcode als Funktion hinterlegt. Diese Liste enthält alle Partitionierungsmöglichkeiten mit ihren Rängen für Segmente, die maximal bis zu 350 Basiswörter aufweisen.

Rang	Partition	Anzahl LCS	Anzahl nicht gematchter Basiswörter
18	{5}	1	0
17	{4, 1}	2	0
16	{3, 2}	2	0
15	{3, 1, 1}	3	0
14	{2, 2, 1}	3	0
13	{2, 1, 1, 1}	4	0
12	{1, 1, 1, 1, 1}	5	0
11	{4}	1	1
10	{3, 1}	2	1
9	{2, 2}	2	1
8	{2, 1, 1}	3	1
7	{1, 1, 1, 1}	4	1
6	{3}	1	2
5	{2, 1}	2	2
4	{1, 1, 1}	3	2
3	{2}	1	3
2	{1, 1}	2	3
1	{1}	1	4

Tabelle 17: Beispiel für die Partitionierung der Zahl 5. Insgesamt gibt es 18 Kombinationen, um die Zahl 5 durch eine Summe von Zahlen (resultierend aus der Anzahl und Verteilung gematchter und nicht gematchter Basiswörter) darzustellen.

Aus dieser Liste kann folglich schnell der entsprechende Rang abgelesen werden, ohne eine Partitionierung tatsächlich durchführen zu müssen. Sollte unwahrscheinlicherweise ein Segment mehr als 350 Basiswörter enthalten,

müsste die Liste erweitert werden. Für das Beispiel aus Tabelle 16 gilt die voranstehende Partitionierungstabelle.

Die für die Ermittlung der Ähnlichkeit der LCS-Struktur notwendigen Ränge sind

1. der höchstmögliche Rang, der bei der jeweiligen Partitionierung erzielt werden könnte (r_{max}),
2. der tatsächlich erzielte Rang (r_{match}),
3. der Rang, bei dem die Hälfte der Basiswörter zwischen den beiden Segmenten gematcht würde, verteilt auf nur einen LCS (r_{halb}). Falls die Partition für eine ungerade Zahl vorgenommen wurde, wird dieser Rang durch Abrunden ermittelt[54].

Der Quotient aus dem tatsächlich erzielten Rang und dem höchstmöglichen Rang bildet einen vorläufigen *L*-Wert (L_{vorl}), der zwischen 0 und 1 liegen kann:

$$L_{vorl} = \frac{r_{match}}{r_{max}}$$

Korrektur von L_{vorl}

Da sich die Anzahl an Rängen jedoch mit jedem Basiswort, das zusätzlich partitioniert werden muss, erhöht und somit Partitionierungen von nur sehr kurzen Segmenten in verhältnismäßig höheren *L*-Werten resultieren würden als Partitionierungen von sehr langen Segmenten, muss der vorläufig errechnete *L*-Wert nach oben korrigiert werden.

Dafür kommt der Rang r_{halb} zum Einsatz. Mit ihm kann der zu r_{halb} korrespondierende *L*-Wert L_{halb} analog zur oben aufgeführten Formel errechnet werden:

$$L_{halb} = \frac{r_{halb}}{r_{max}}$$

Damit bei einem im TM vordefinierten Schwellenwert von 70 %, der in der Übersetzungsbranche üblicherweise eingestellt wird, mindestens diejenigen Übersetzungseinheiten angezeigt werden, bei denen die Hälfte der Basiswörter des AS_{neu} mit denen eines im linguistisch optimierten TM gespeicherten AS-Segmentes übereinstimmt, wird L_{halb} auf den Wert 70 % angehoben.

[54] Im Falle von fünf Basiswörtern ist dies Rang 3 (siehe Tabelle 17).
Beispielrechnung: $\left\lfloor \frac{b_{max}}{2} \right\rfloor = \left\lfloor \frac{5}{2} \right\rfloor = 2 \Rightarrow r_{halb} = 3$

Der Korrekturfaktor x zur Anhebung von L_{halb} wird anhand folgender Bedingung ermittelt:

$$L_{halb}{}^{x} = 0{,}7$$

$$\Leftrightarrow x * \log L_{halb} = \log 0{,}7 \Rightarrow x = \frac{\log 0{,}7}{\log L_{halb}} = \frac{\log 0{,}7}{\log \frac{r_{halb}}{r_{max}}}$$

L_{vorl} wird daraufhin mit dem Korrekturfaktor x potenziert.

$$L = L_{vorl}{}^{x} = \left(\frac{r_{match}}{r_{max}}\right)^{x}$$

Das Ergebnis ist der endgültige *L*-Wert für die beiden miteinander verglichenen Segmente.

Werden die soeben erläuterten Formeln auf das Beispiel aus Tabelle 16 bzw. Tabelle 17 angewendet, ergeben sich folgende Werte: Der für das Beispiel höchstmögliche Rang gemäß Tabelle 17 ist Rang 18, da dieser Rang die maximale mögliche Anzahl an matchbaren Basiswörtern in nur einem LCS angibt. Tatsächlich wurde Rang 10 bzw. die Partition {3, 1} erzielt, da zwischen den beiden zu vergleichenden Segmenten zwei LCS ermittelt wurden, die sich einmal aus drei Basiswörtern und einmal aus einem Basiswort zusammensetzen.

Mit dem tatsächlich erzielten Rang $r_{match} = 10$ und dem höchstmöglichen Rang $r_{max} = 18$ lässt sich L_{vorl} bestimmen. Um den endgültigen *L*-Wert berechnen zu können, muss zusätzlich der Korrekturfaktor x ermittelt werden. Dafür ist der Wert von r_{halb} erforderlich. Der Rang r_{halb} entspricht dem Rang in der Partitionierungstabelle, bei dem die Hälfte der Basiswörter zwischen den beiden Segmenten in genau einem LCS gematcht würden. Im Falle von sechs matchbaren Basiswörtern entspräche r_{halb} der Partition {3}, d. h. dem Rang, bei dem drei Basiswörter in genau einem LCS gematcht würden. Im Falle von Tabelle 17, in der jedoch maximal fünf Basiswörter partitioniert werden, entspricht r_{halb} dem Rang, bei dem zwei Basiswörter gematcht würden (d. h. der Partitionierung {2}), da bei der Partitionierung einer ungeraden Anzahl an Basiswörtern r_{halb} durch Abrunden ermittelt wird[55]. Im Falle von Tabelle 17 entspricht r_{halb} folglich Rang 3.

[55] Es können nur ganze Zahlen einer Partition zugeordnet werden, weshalb r_{halb} im Falle einer Partionierung einer ungeraden Anzahl an Basiswörtern durch Abrunden ermittelt wird. Das Heranziehen des abgerundeten anstatt des aufgerundeten Ranges bewirkt dabei, dass der Korrekturfaktor x kleiner und dadurch der endgültige *L*-Wert größer ausfällt.

L_{vorl} beträgt:

$$L_{vorl} = \frac{r_{match}}{r_{max}} = \frac{10}{18}$$

Der Korrekturfaktor x ist:

$$x = \frac{\log 0{,}7}{\log \frac{r_{halb}}{r_{max}}} = \frac{\log 0{,}7}{\log \frac{3}{18}}$$

Der endgültige L-Wert lautet:

$$L = \left(\frac{r_{match}}{r_{max}}\right)^{x} = \left(\frac{10}{18}\right)^{x} \approx 88{,}96\,\%$$

Durch diese Vorgehensweise werden dem Übersetzer mehr brauchbare Matches in den höheren Match-Wert-Bereichen vorgeschlagen, die durch zeichenkettenbasierte TMs mit ihren teilweise zu niedrig angesetzten Match-Werten nicht aufgeführt würden. Jedoch wird der Übersetzer gleichzeitig mit einer höheren Anzahl unbrauchbarer Übersetzungseinheiten konfrontiert, was für ihn im ungünstigsten Fall sogar mehr Arbeitsaufwand bedeuten könnte, als wenn er ein zeichenkettenbasiertes System verwenden würde.

4.3.2 Berechnung der Ähnlichkeit der grammatischen Struktur (G)

Die Ähnlichkeit der grammatischen Struktur wird berechnet, indem die Werte spezifischer Merkmale, die bei der morphosyntaktischen Analyse für jede Wortform der beiden Segmente extrahiert wurden, miteinander verglichen werden. Der Vergleich erfolgt für diejenigen Basiswörter, die bei der Ermittlung der LCS zwischen den beiden Segmenten als identisch markiert wurden.

Zwischen den identischen Basiswörtern werden nachfolgende Merkmale auf Gleichheit ihrer Werte geprüft:

1. In jedem Fall die Wortform/Originalform des Wortes, aus der das Basiswort ermittelt wurde. Falls sich eines der zu vergleichenden Basiswörter am Segmentanfang befindet, wird der Vergleich der Originalform der Wörter ohne Berücksichtigung der Groß- und Kleinschreibung durchgeführt. In allen anderen Fällen wird eine Änderung von Groß- und Kleinschreibung der Originalform des Wortes mitberücksichtigt.

2. Die Wortart des Wortes, aus dem das Basiswort ermittelt wurde, und ggf. die Unterwortart, falls das morphosyntaktische Analyseprogramm Unterwortarten codieren kann.
3. Stimmt die Wortart bzw. Unterwortart überein, werden die Werte folgender wortartenspezifischer Merkmale verglichen: Numerus, Genus, Kasus, Tempus, Verbtyp, Steigerungsform und adjektivische Partizipialform. Werden mehrere Werte für eines dieser Merkmale durch das morphosyntaktische Analyseprogramm ausgegeben, wird jeder dieser Werte in den Vergleich miteinbezogen. Falls das morphosyntaktische Analyseprogramm ebenso semantische Kategorien codieren kann, können auch diese verglichen werden.
4. Stimmt die Wortart bzw. Unterwortart nicht überein, wird der Wortartwechsel untersucht.

Jedem Unterschied werden definierte Kosten zugewiesen:

- Übereinstimmung aller Merkmale inklusive ihrer Werte: Kosten von 0 (= keine Kosten)
- Unterschied in der Originalform des Wortes: Kosten von 1
  Diese geringen Kosten kommen dadurch zustande, dass Unterschiede in der Originalform meistens nur gering sind, z. B. Groß- und Kleinschreibung oder orthografische Varianten.
- Nicht vorhandenes wortartenspezifisches Merkmal bei einem der beiden zu vergleichenden Basiswörter (bei Übereinstimmung der Wortart bzw. Unterwortart)[56]: Kosten von 1 pro nicht vorhandenes Merkmal
- Unterschied in den Werten der wortartenspezifischen Merkmale (bei Übereinstimmung der Wortart bzw. Unterwortart): Kosten von 2 pro unterschiedlichen Wert
- Unterschiede in der Wortart bzw. Unterwortart: Kosten je nach Grad des Wortartwechsels (Tabelle 18)

[56] Diese Regel ist wie folgt begründet: Es kann vorkommen, dass trotz gleicher Wortart dieser Wortart je nach Kontext unterschiedliche wortartenspezifische Merkmale innewohnen. So kann beispielsweise ein Verb als Infinitiv oder als flektierte Wortform vorliegen. Im Falle der flektierten Wortform muss weiterhin u. a. das Tempus in die Analyse mit einfließen, während der Infinitiv zeitlos ist.

Wortartwechsel	**Kosten**
Interpunktion → Interpunktion	5
Adjektiv in attributiver Verwendung ↔ Adjektiv in prädikativer/adverbialer Verwendung	10
Inhaltswort → Inhaltswort	20
Funktionswort → Funktionswort	20
Inhaltswort ↔ Funktionswort	30

Tabelle 18: Kosten für Wortartwechsel

Die einzelnen Kosten, die für jedes Paar an identischen Basiswörtern (nachfolgend *Basiswort-Paar* genannt) identifiziert werden können, werden miteinander addiert, sodass jedem Basiswort-Paar ein finaler Kostenwert zugeschrieben wird. Die größtmöglichen Kosten, die jedem Basiswort-Paar zugewiesen werden können, betragen 31 ($Kosten_{max}$). Dieser Wert ergibt sich aus den Kosten für einen Unterschied in der Originalform des Wortes und einem Wortartwechsel von einem Inhaltswort zu einem Funktionswort (oder umgekehrt).

Die Ähnlichkeit der grammatischen Struktur kann durch ihre Abweichung/Unähnlichkeit D (dissimilarity) ausgedrückt werden. Die Begriffe *ähnlich* und *unähnlich* sind durch die Relation $S = 1 - D$ miteinander verknüpft[57]. Folglich muss die relative Grammatikabweichung von 1 subtrahiert werden:

$$G = 1 - \underbrace{relative\ Grammatikabweichung}_{G_D}$$

Die relative Grammatikabweichung setzt sich dabei zusammen aus der Summe der finalen Kostenwerte pro Basiswort-Paar dividiert durch das Produkt aus der Anzahl an Basiswort-Paaren ($\#Paare$) und dem größtmöglichen Kostenwert pro Basiswort-Paar ($Kosten_{max} = 31$):

$$G_D = \frac{auf\,getretene\ Kosten}{maximal\ mögliche\ Kosten} = \frac{\sum Kosten\ pro\ Paar}{\#Paare * Kosten_{max}}$$

Bei vollständiger morphosyntaktischer Übereinstimmung zwischen allen Basiswort-Paaren entstehen folglich keine Kosten; G_D ist gleich 0 und G gleich 1. Besteht keinerlei morphosyntaktische Übereinstimmung, ist G_D gleich 1 und G gleich 0.

[57] Zwei Segmente, die zu 90 % ähnlich sind, sind sich also zu 10 % unähnlich.

Angenommen, das verglichene Segmentpaar aus Tabelle 16 verfügt über Basiswort-Paare (*b*, *c*, *d* und *f*), die folgende morphosyntaktische Unterschiede aufweisen:

	b	**c**	**d**	**f**
AS_{neu}	Numerus = Singular	Numerus = Singular Steigerung = Positiv	Numerus = Singular	Wortart = Adjektiv
Im linguistisch optimierten TM gespeichertes AS-Segment	Numerus = Plural	Numerus = Plural Steigerung = Superlativ	Numerus = Plural	Wortart = Adverb
Kosten	2	4	2	10

Tabelle 19: Beispiel für die Kostenzuweisung aufgrund von Unterschieden in den morphosyntaktischen Merkmalen zwischen den Basiswort-Paaren

Die Summe der Kosten aller Basiswort-Paare beträgt 18. Der G-Wert, d. h. die Ähnlichkeit der grammatischen Struktur der zwischen den beiden Segmenten identischen Basiswörtern, beträgt für das Beispiel demnach:

$$G = 1 - \left(\frac{2 + 4 + 2 + 10}{4 * 31}\right) \approx 85{,}48\,\%$$

4.3.3 Verrechnung von L und G zur Ermittlung von S_{LG}

Die Verrechnung der Ähnlichkeit der LCS-Struktur (L) mit der Ähnlichkeit der grammatischen Struktur (G) führt zu einer Beschreibung der Gesamtähnlichkeit der gematchten Basiswörter (S_{LG}) zweier Segmente. Diese ist durch das Produkt von L und G gegeben:

$$S_{LG} = L * G$$

Das Produkt wurde gewählt, damit Segmenten, bei denen einer der beiden Faktoren gering ausfällt, dennoch ein relativ schlechter Gesamtähnlichkeitswert S_{LG} zugeschrieben wird.[58] Dies entspricht der Annahme, dass zwei zu vergleichende Segmente umso ähnlicher sind, wenn sowohl ihre LCS-Struktur als auch ihre grammatische Struktur übereinstimmen.

[58] Angenommen, es liegt einerseits eine 100 %-Übereinstimmung in der LCS-Struktur und andererseits eine 0 %-Übereinstimmung in der grammatischen Struktur vor, so beträgt die Gesamtähnlichkeit 0. Dies könnte passieren, falls Homonyme bzw. Homografen gematcht würden.

Korrektur von S_{LG}

Allerdings haben vorläufige Untersuchungen der S_{LG}-Werte mit unterschiedlich langen Segmenten gezeigt, dass der S_{LG}-Wert häufig zu niedrig für das eigene Ähnlichkeitsempfinden (d. h. für das der Verfasserin dieser Arbeit) ausfiel. Dies wurde insbesondere dann so empfunden, wenn alle oder fast alle Basiswörter zwischen den beiden Segmenten gematcht wurden. Daher erfolgt für den Fall, dass alle bzw. fast alle Basiswörter zwischen beiden Segmenten gematcht wurden, eine Verbesserung von S_{LG} durch Einführung des Korrekturfaktors κ, mit dem S_{LG} potenziert wird.

Die Höhe des Korrekturfaktors κ richtet sich nach der Anzahl der nicht gematchten Basiswörter zwischen den beiden Segmenten. Folgende Werte für den Korrekturfaktor κ haben sich dabei als guter Kompromiss ergeben:

Anzahl nicht gematchter Basiswörter	Korrekturfaktor κ
0	0,2
1	0,4
2	0,8
> 2	1

Tabelle 20: Werte des Korrekturfaktors κ

Da die Werte L und G zwischen 0 und 1 liegen, führt ein Korrekturfaktor von $\kappa < 1$ zu einer Erhöhung von S_{LG}. Dies ist der Fall, wenn bis zu zwei Basiswörter nicht gematcht werden. Werden mehr als zwei Basiswörter zwischen den Segmenten nicht gematcht, wird der Korrekturfaktor 1 vergeben, was dazu führt, dass lediglich das Produkt von L und G berechnet wird.

Für das Beispiel aus Tabelle 16 würde folglich der Korrekturfaktor 0,4 vergeben, da vier Basiswörter zwischen den beiden Segmenten gematcht wurden, insgesamt jedoch fünf Basiswörter maximal matchbar gewesen wären.

Die Ähnlichkeit der gematchten Basiswörter beträgt demnach:

$$S_{LG}{}^{\kappa} = (L * G)^{\kappa} \approx (0{,}8896 * 0{,}8548)^{0{,}4} \approx 89{,}62\,\%$$

4.3.4 Berechnung der Ähnlichkeit der Segmentlänge (S_{SL})

Bei der Berechnung der Ähnlichkeit der Segmentlänge müssen zwei Fragestellungen berücksichtigt werden:

1. Müssen Basiswörter aus dem im linguistisch optimierten TM gespeicherten AS-Segment gelöscht oder in dasselbige hinzugefügt werden, um das AS_{neu}, d. h. Segmente mit der gleichen Anzahl an Basiswörtern, zu erhalten?
2. Sind nicht gematchte Basiswörter an einem Stück enthalten oder auf unterschiedliche Stellen in demjenigen Segment verteilt, das das Maximum der Anzahl an Basiswörtern der beiden Segmente enthält?

Ausgehend von diesen Fragestellungen und den Annahmen, die in Kapitel 4.3 zum Arbeits- und Identifikationsaufwand zu löschender oder hinzuzufügender nicht gematchter Basiswörter aus dem bzw. in das im linguistisch optimierten TM gespeicherte AS-Segment getroffen wurden, wurde nachfolgende Formel zur Bestimmung von S_{SL} definiert:

$$S_{SL} = \frac{b_{max}}{SL_{max}}$$

Dabei repräsentiert SL_{max} das Maximum und b_{max} das Minimum der Anzahl an Basiswörtern der beiden Segmente.

Für den Fall, dass das AS_{neu} weniger Basiswörter enthält als das im linguistisch optimierten TM gespeicherte AS-Segment und die nicht gematchten Basiswörter an nur einem Stück im gespeicherten AS-Segment enthalten sind[59], d. h., wenn

LCS … LCS

tritt ein Sonderfall in Kraft. Dabei werden die beiden Segmente so behandelt, als ob sie sich lediglich um ein Basiswort unterscheiden. Dieser Sonderfall liegt in der zuvor getroffenen Annahme begründet, dass die Löschung nicht gematchter Basiswörter an einem Stück einen geringeren Arbeits- und Identifikationsaufwand erfordert als die Löschung nicht gematchter Basiswörter an mehreren Stellen im Segment. Die Formel für den Sonderfall lautet demnach wie folgt:

$$b_{max} = SL_{max} - 1 \Rightarrow S_{SL} = \frac{b_{max}}{SL_{max}} = \frac{SL_{max}-1}{SL_{max}}$$

[59] Beispiel:
AS_{neu}: Den Scherkopf stets reinigen.
Im linguistischen TM gespeichertes AS-Segment: Den Scherkopf, der für die Rasur verwendet wird, stets reinigen.

Da in dem Beispiel aus Tabelle 16 das AS_{neu} mit sieben Basiswörtern länger ist als das im linguistisch optimierten TM gespeicherte AS-Segment (fünf Basiswörter), d. h. eine Hinzufügung von Basiswörtern in das gespeicherte AS-Segment erforderlich ist, und auch die nicht gematchten Basiswörter auf mehrere Stücke im AS_{neu} verteilt sind, muss die allgemeine Formel zur Berechnung von S_{SL} angewendet werden. Aus dieser ergibt sich folgender Wert für die Ähnlichkeit der Segmentlänge:

$$S_{SL} = \frac{5}{7}$$

Nachdem alle Einzelwerte berechnet wurden, können diese in die Match-Formel (siehe Kapitel 4.3) eingetragen werden, um einen finalen Match-Wert für die miteinander verglichenen Segmente zu erhalten. Für das Beispiel aus Tabelle 16 gilt daher:

$$M = 0{,}7 * 0{,}8962 + 0{,}3 * 0{,}7143 \approx 84{,}17\ \%$$

Die beiden Segmente sind sich also zu ca. 84,17 % ähnlich.

5 Realisierung des Modells

In Kapitel 4 wurden bereits der Algorithmus – bei dem die Datenstruktur der GSAs um die Identifikation aller identischen sich wiederholenden LCS im GSA und der bestmöglichen Übereinstimmung zwischen zwei Segmenten unter Berücksichtigung der geringsten Positionsdifferenz und der Segmentzugehörigkeit erweitert wurde – und das Proximitätsmaß beschrieben, die von der Verfasserin dieser Arbeit im Rahmen des iMem-Forschungsprojektes entwickelt wurden. Nachfolgend werden Komponenten erläutert, mit denen der Algorithmus und das Proximitätsmaß durchlaufen werden können. Ebenso werden die iMem-Plug-in-Benutzeroberfläche sowie der Übersetzungsprozess mit dem Plug-in dargestellt und Unterschiede zu anderen linguistisch optimierten TMs aufgeführt.

5.1 Verwendete Komponenten

5.1.1 SDL Trados Studio 2009

In dem iMem-Forschungsprojekt wird SDL Trados Studio 2009 als Vertreter zeichenkettenbasierter, d. h. nicht linguistisch optimierter kommerzieller TM-Systeme, eingesetzt, da die Systeme von SDL Trados zu den meist verbreiteten in der Übersetzungsbranche zählen. Bei SDL Trados Studio 2009 handelt es sich um ein integriertes Übersetzungssystem, das aus unterschiedlichen sprachtechnologischen Werkzeugen besteht. Neben dem TM als Hauptbestandteil umfasst das System die in Kapitel 3.1.2.4 aufgeführten Komponenten.

In SDL Trados Studio 2009 können einem Übersetzungsprojekt ein oder mehrere SDL-TMs zugeteilt werden, die sowohl dateibasiert als auch serverbasiert sein können. Dem Benutzer stehen fünf Optionen[60] zur Verwendung eines jeden zugewiesenen TMs zur Verfügung.

Ferner besteht die Möglichkeit, Einstellungen bezüglich der Suche und weitere Abzüge vorzunehmen, Filter und Feldwerte zu definieren sowie Formate für die automatische Ersetzung von Datumsangaben, Zeitangaben und Maßeinheiten festzusetzen. Da die Einstellmöglichkeiten für die Suche von Übersetzungseinheiten im TM und die Festlegung der Abzüge relevant für die

[60] Es steht zur Auswahl, ein TM zu aktivieren, zu durchsuchen und zu aktualisieren, Abzüge für Übersetzungsergebnisse aus dem entsprechenden TM (maximal möglicher Abzug: 25 %) zu vergeben sowie die Konkordanzsuche einzusetzen. Eine Beschreibung der Optionen findet sich in SDL plc. (2009).

Evaluierung des iMem-TMs sind, werden diese Optionen nachfolgend genauer beschrieben:

Bei der Suche von Übersetzungsergebnissen kann zum einen der Schwellenwert angegeben werden, bis zu dem Matches in der Trefferanzeige aufgeführt werden. Alle Matches, die einen Match-Wert unterhalb dieses Schwellenwertes aufweisen, werden unterdrückt. In SDL Trados Studio 2009 beläuft sich der niedrigste einstellbare Schwellenwert auf 30 %. Zum anderen kann festgelegt werden, wie viele Matches maximal in der Trefferanzeige dargestellt werden können. Die maximal anzeigbare Trefferanzahl in SDL Trados Studio 2009 beträgt 50. Des Weiteren kann der Übersetzer u. a. wählen, ob trotz eines 100 %-Matches nach Fuzzy-Matches gesucht werden soll.

Es stehen fünf verschiedene Arten von Abzügen zur Verfügung. Dabei handelt es sich um Abzüge für fehlende Formatierung, unterschiedliche Formatierung, Ersetzung von Text, mehrere 100 %-Matches und Auto-Lokalisierung. Der Benutzer hat die Möglichkeit, jeden Abzugswert individuell einzustellen, wobei der maximal einstellbare Wert bei 20 % liegt.

Alle oben aufgeführten Einstelloptionen können entweder für einzelne Sprachpaare oder für alle Sprachpaare definiert werden.

Mithilfe des SDL SDK 2.0 können zudem eigene Anwendungen, die in Zusammenhang mit dem TM-System stehen, programmiert und in SDL Trados Studio 2009 über eine API eingebunden werden (siehe SDL plc. 2009–2013: o.S.).

5.1.2 MPRO

Die am IAI entwickelte Software MPRO ist ein Werkzeug zur morphosyntaktischen Analyse von Texten und wird im iMem-Forschungsprojekt für die morphosyntaktische Analyse der AS-Segmente verwendet.

Durch Zugriff auf hinterlegte Wörterbücher wird bei der Analyse mit MPRO jede Wortform innerhalb eines Textes mit entsprechenden linguistischen Informationen versehen. Neben morphosyntaktischen Informationen werden dabei auch semantische Informationen zugewiesen[61] (vgl. Maas 1998: o.S.). Gemäß Maas et al. (2009: 2ff.) durchläuft die MPRO-Analyse vier Module, die nachfolgend lediglich umrissen werden[62]:

[61] Bei diesen Informationen handelt es sich mindestens um die Angabe der Grundform und der Wortart. Je nach Wortart können u. a. Angaben zur Flexion (Kasus, Genus, Numerus, Steigerungsform, Tempus etc.), zur Ableitungs- und Stammformenstruktur, zu Schreibweisen, zum Deklinationstyp attributiver Adjektivformen, zur Kennzeichnung von Partizipialformen sowie zur semantischen Klassifikation ausgegeben werden (vgl. Maas 1998: o.S., IAI o.J.: o.S.).

[62] Für detaillierte Informationen über die Funktionsweise von MPRO sei der interessierte Leser auf Maas et al. (2009) verwiesen.

1. LESEN: Textbereinigung durch Kategorisierung von Tags und Transliteration von Sonderzeichen; Identifikation von Satz- und Wortgrenzen, unbekannter Wörter sowie fester Wendungen mittels spezifischer Regeln; Durchführung einer morphologischen Analyse in Abhängigkeit von der zu analysierenden Sprache durch Konsultation eines Morphologie- und Lemmawörterbuches mit Zuweisung eines Merkmalbündels zu jedem identifizierten Wort
2. LEXNP: Abfrage eines Wörterbuches zur Ermittlung von Mehrworteinheiten (nur verfügbar für die Analyse deutscher Texte)
3. KOMPLETT: Rekonstruktion von Wortteilellipsen (nur verfügbar für die Analyse deutscher Texte)
4. KORRIGIERE: Korrekturvorschläge für falsch geschriebene Wörter

Ein Merkmalbündel ist eine Abfolge von linguistischen Informationen zu jeder Wortform, die der Struktur *Attribut-Operator-Wert* folgt (vgl. Carl/Schmidt-Wigger 1998: 258). Das Merkmalbündel für die Nominalphrase *das mitgelieferte Reinigungsbürstchen* sieht demnach wie folgt aus:

{ori=das,lu=d_art,mlu=d_art,c=w,sc=art,ehead={nb=sg,infl=weak,case=acc;nom,g=n}
,gra=small,ns=Das,ds=d_art,ls=d_art,s=nil}

,{ori=mitgelieferte,lu=mitgeliefert,mlu=mitgeliefert,c=adj,sc~=p;pron;rel;subj;um_zu,
ptc=2,ehead={nb=sg,infl=weak,case=acc;nom,g=n},deg=base,gra=small,ds=mit_$lief
ern,ts=mitgelieferte,t=mitgeliefert,ls=mit_$liefern,s=nil}

,{ori=Reinigungsbürstchen,lu=reinigungsbürstchen,mlu=reinigungsbürstchen,c=noun,eh
ead={nb=sg,infl=weak,case=acc;nom,g=n},g=n,gra=cap,ds=reinigen~ung#bürste~chen,
gs=f#n,ts=reinigungs#bürstchen,t=reinigung#bürstchen,ls=reinigen#bürste,s=instr}.

MPRO verfügt über eine Reihe unterschiedlicher Attribute bzw. Merkmale mit ihren entsprechenden Werten. So gibt es u. a. Merkmale für die Ermittlung der Wortanzahl, der Ableitungsstruktur sowie von morphosyntaktischen und semantischen Eigenschaften, etymologischen Informationen oder Korrekturvorschlägen. Es wird unterschieden zwischen Merkmalen, deren Werte selbst keine Merkmalbündel sind (z. B. c=noun) – wobei je nach Lesart auch mehrere zutreffende Werte aufgeführt werden können (z. B. case=acc;nom) – und Merkmalen, deren Werte sich aus weiteren Merkmalbündeln zusammensetzen (z. B. ehead={nb=sg,infl=weak, case=acc;nom,g=n}) (vgl. Reinke 2004: 354). Die Merkmalbündel des Merkmals *ehead* weisen darauf hin, dass die Bestandteile einer Phrase über eine gemeinsame morphosyntaktische Struktur verfügen. Die für das iMem-Forschungsprojekt relevanten Merkmale und Werte werden in Anhang G aufgelistet. Erklärungen zu weiteren

Merkmalen und Werten finden sich in Reinke (2004: 402f.), Rösener (2005: 209ff.) sowie in Maas et al. (2009: 4ff.).

Eine Besonderheit von MPRO besteht darin, Komposita in ihre Bestandteile zerlegen zu können. Die einzelnen Elemente eines Kompositums werden durch das Symbol # voneinander getrennt. Weitere Symbole sind *$*, *_$*, *$+*, ~ und *~IRREG*, deren jeweilige Bedeutung in der nachfolgenden Tabelle erläutert wird:

Symbol	**Beispiel**	**Originalwort**	**Erläuterung**
#	ls=reinigen#bürste	Reinigungsbürste	Trennung von Kompositumsbestandteilen
$	ls=ver$schicken	verschicken	Nicht abtrennbares Präfix
_$	ls=zu_$hören	zuhören	Abtrennbares Präfix
$+	ls=un$+fassen	unfassbar	Präfix un
~	ds=reinigen~ung	Reinigung	Ableitungsmorphem
~IRREG	ds=fahren~IRREG	Fahrt	Unregelmäßige Ableitung

Tabelle 21: Symbole und ihre Bedeutung, die bei der Codierung von Wortformen durch MPRO ausgegeben werden (siehe auch Reinke 2004: 370, 403).

Die Symbole erscheinen nur bei den Werten der Merkmale, die die morphologische Struktur einer Wortform beschreiben, d. h. bei den Werten der Merkmale *ls*, *ts*, *t* und *ds*. Eine Erläuterung dieser Merkmale wird anhand der Codierung eines Beispielwortes in Tabelle 22 gegeben.

Merkmal	**Wert**	**Erläuterung**
ls	ent$riegeln#taste	Lexikalische Struktur eines Wortes mit Basis der Ableitung
ts	entriegelungs#taste	Oberflächige Zerlegung eines Wortes
t	entriegelung#taste	Lexikalische Einheiten von ts
ds	ent$riegeln~ung#taste	Angaben zu Ableitungen von t

Tabelle 22: Werte der Merkmale *ls*, *ts*, *t* und *ds* für das Wort *Entriegelungstaste*

Aus dieser Gruppe von Merkmalen wird für das iMem-Forschungsprojekt ausschließlich das Merkmal *ls* verwendet, da es die lexikalische Struktur einer beliebigen Zeichenkette und die Basis einer Ableitung angibt. Kompositumsbestandteile werden durch das Symbol # kenntlich gemacht und Informationen zu abtrennbaren und nicht abtrennbaren Präfixen (gekennzeichnet durch die Symbole *_$*, *$* und *$+*) bleiben erhalten, was bei den Merkmalen *ts* und *t* nicht der Fall ist (vgl. Rösener 2005: 212). Informationen zur derivationellen

Struktur, die das Merkmal *ds* liefert (gekennzeichnet durch die Symbole ~ und *~IRREG*), sind hingegen für die Forschungszwecke im iMem-Projekt nicht erforderlich, da die Wahrscheinlichkeit, einen Match zu erhalten, mit jeder weiteren zu vergleichenden Information sinkt. Folglich spielen auch nur die Symbole *$*, *_$*, *$+* und # für die Erstellung eines iMem-TMs eine Rolle.

Trotz der umfangreichen Analyse kann es vorkommen, dass Mehrdeutigkeiten eines Wortes nicht immer aufgelöst werden können. Im Analyseergebnis werden die ambigen Wörter nacheinander aufgelistet, zusammen mit ihren jeweiligen Merkmalbündeln (vgl. Maas et al. 2009: 4). Im nachfolgenden Analyseergebnis des Beispielsegmentes *Reinigen unter Wasser* konnte das Wort *unter* nicht disambiguiert werden. Die unterschiedlichen Lesarten eines Wortes werden durch ein logisches ODER miteinander verknüpft, dargestellt durch das Semikolon am Ende des Merkmalbündels (vgl. Reinke 2004: 354):

{ori=Reinigen,lu=reinigen,mlu=reinigen,c=verb,vtyp=fiv,nb=plu,tns=pres,gra=cap,ds=reinigen,ts=reinigen,t=reinigen,ls=reinigen,s=c0c1},

{ori=unter,lu=unter,mlu=unter,c=w,sc=p,ehead={nb=plu,case=acc,g=n};{nb=sg,case=acc,g=n},gra=small,ds=unter,ls=unter,s=dir};

{ori=unter,lu=unter,mlu=unter,c=w,sc=p,ehead={nb=sg,case=dat,g=n},gra=small,ds=unter,ls=unter,s=loc},

{ori=Wasser,lu=wasser,mlu=wasser,c=noun,ehead={nb=plu,case=acc,g=n};{nb=sg,case=dat,g=n};{nb=sg,case=acc,g=n},g=n,gra=cap,ds=wasser,ts=wasser,t=wasser,ls=wasser,s=mat}.

Ebenso kann ein Merkmalbündel des Merkmals *ehead* über mehrere Werte verfügen, die verschiedene Lesarten repräsentieren. Die Lesarten der Werte sind ebenfalls durch ein logisches ODER miteinander verknüpft (z. B. ehead={nb=sg,infl=weak,case=acc;nom,g=n}).

Laut Maas et al. (2009: 5) verfügt die lexikalische Datenbank für das Deutsche, in der die unterschiedlichen Wörterbücher gewartet werden und Informationen zu Präfixen, Flexions- und Derivationsendungen sowie zu einfachen und komplexen Lexemen (z. B. zu Komposita und Derivaten) enthalten sind, über 736.000 Einträge, worunter etwa 120.000 Einträge fallen, die die morphologischen Eigenschaften von Stammmorphemen beschreiben. Neben diesen Stammmorphemen sind auch zahlreiche Vor- und Nachnamen, Namen von Städten, Ländern und Flüssen sowie Morpheme aus der lateinischen und griechischen Sprache in der lexikalischen Datenbank zu finden (vgl. Maas 1998: o.S., Maas et al. 2009: 5).

MPRO existiert für ein Dutzend von Sprachen, darunter vor allem für Deutsch, Englisch, Französisch und Spanisch. Weitere unterstützte Sprachen sind Bulgarisch, Esperanto, Griechisch, Italienisch, Latein, Niederländisch,

Portugiesisch und Russisch (vgl. Maas 1998: o.S., Maas et al. 2009: 1). Die Einsatzgebiete von MPRO liegen in der Rechtschreib-, Grammatik-, Stil- und Terminologieprüfung, Verschlagwortung, Phonetisierung, Lemmatisierung, Übersetzung und Zusammenfassung von Dokumenten (vgl. Maas 1998: o.S.). Maas et al. (2009: 17) sehen weitere Anwendungsmöglichkeiten in den Bereichen Indexierung, kontrollierte Sprache, Spracherlernung, maschinelle Übersetzung und linguistisch optimiertes Information Retrieval. Ebenso stellen die Ergebnisse der morphosyntaktischen Analyse die Grundlage für das Parsen und Taggen dar (vgl. Maas et al. 2009: 1).

Forschungsprojekte aus den letzten 20 Jahren belegen, dass MPRO zur Lösung von linguistischen Forschungsfragen beitragen kann. So wurde MPRO z. B. in dem Forschungs- und Entwicklungsprojekt MULTILINT eingesetzt, um technische Dokumentation aus dem Automobilbereich morphosyntaktisch zu analysieren. Dieses Projekt hatte zum Ziel „ein linguistisch intelligentes System zur Erstellung und Verwaltung von multilingualer technischer Dokumentation zu entwickeln“ (Haller 1996: 1). Wie in Haller (1996: 5ff.) beschrieben, sollten u. a. die morphologischen Analysen folgende Prozesse verbessern bzw. ermöglichen:

- Kontrolle des Quelltextes (d. h. Rechtschreib-, Syntax-, Stil- und Terminologieprüfung)
- Erstellung eines fremdsprachigen Glossars nach vorheriger Spezifizierung lexikalischer Kategorien und der Angabe, dass Informationen zu Ableitungen erwünscht sind
- Suche nach bestimmten Merkmalen von analysierten Übersetzungseinheiten in TMs
- Maschinelle Übersetzung von kurzen Sätzen oder Satzteilen

So wie das MULTILINT-Projekt befasste sich das EU-Projekt MULTIDOC ebenfalls mit der Autorenunterstützung technischer Dokumentation im Automobilbereich durch sprachtechnologische Mittel (vgl. Haller 2000: 2). Auch hier wurde MPRO für die morphosyntaktische Analyse der technischen Dokumentation verwendet.

In dem Forschungsprojekt EMIS ging es um die Erstellung eines Informationssystems, das die Suche nach deutschen, englischen und französischen Normen aus verschiedenen Fachbereichen des europäischen Medienrechtes erlaubt (vgl. Ripplinger 1998: 507, Ripplinger 2000: 1108f.). MPRO wurde angewendet, um bei der Suche nach Begriffen in den gespeicherten Normen eine höhere Trefferquote zu erzielen. Dazu wurden die Suchbegriffe und die in der Datenbank enthaltenen Dokumente einer morphosyntaktischen Analyse

unterzogen, mit deren Hilfe exakte Treffer (Merkmal *lu*), Basiswörter und einzelne Kompositumsbestandteile (Merkmale *ls* und *t*) nachgeschlagen werden konnten (vgl. Ripplinger 1998: 508f.).

MPRO wird jedoch nicht nur in der Forschung, sondern auch in der Entwicklung kommerzieller Systeme eingesetzt. Das aus einem Forschungsprojekt hervorgegangene System AUTINDEX dient beispielsweise dazu, Schlagwörter aus deutschen und englischen Texten automatisch zu extrahieren (vgl. Haller et al. 2001: 1, IAI 2011: o.S.). Die Verschlagwortung erfolgt durch die auf MPRO beruhende morphosyntaktische Analyse, Identifikation semantischer Klassen und Ermittlung von Wortgruppen (vgl. Haller et al. 2001: 4ff.).

Das System CLAT wiederum unterstützt Autoren dabei, fehlerfreie, konsistente und stilsichere deutsche oder englische technische Dokumentation zu erstellen (vgl. Haller 2007: 73). Mithilfe der Kernkomponente MPRO können Texte auf orthografische und grammatische Richtigkeit überprüft sowie eine Terminologie- und Stilkontrolle durchgeführt werden (vgl. Rösener 2010: 2ff.).

5.1.3 iMem-TM

Das iMem-TM ist eine eigenständige relationale Datenbank, die insgesamt aus drei Relationen[63] besteht. Zwei dieser Relationen dienen der Datenspeicherung, die dritte Relation wird zwecks der Suche im iMem-TM eingesetzt.

In der ersten Relation für die Datenspeicherung (Beispiel siehe Tabelle 23, META_DATA) werden die Metadaten des iMem-TMs verwaltet. Dabei handelt es sich um den Dateinamen des iMem-TMs (Spalte *identifier*) sowie um die Ausgangs- und Zielsprache, in der die Übersetzungseinheiten im iMem-TM gespeichert sind (Spalten *source_language* und *target_language*). Diese Tabelle enthält immer nur einen Eintrag und ist nicht mit den anderen Relationen verknüpft. Sie dient ausschließlich einem verwaltungstechnischen Zweck.

META_DATA

identifier	source_language	target_language
Test.imemtm	DE	EN

Tabelle 23: iMem-TM: Relation mit Metadaten

Um das iMem-TM verwenden zu können, muss es mit Übersetzungseinheiten befüllt werden. Dies kann sowohl vor oder während des Übersetzungsprozesses

[63] Eine Relation ist „eine logisch zusammenhängende Einheit von Informationen“ (Sauer 2002: 33) und wird in Datenbanken in Tabellenform abgebildet.

geschehen. Die AS-Seite der Übersetzungseinheiten, mit denen das iMem-TM befüllt wird, wird einer morphosyntaktischen Analyse unterzogen. Dabei werden die für den weiteren Übersetzungsprozess notwendigen morphosyntaktischen Merkmale mit ihren Werten mittels MPRO extrahiert und im iMem-TM abgespeichert.

TRANSLATION_UNIT

ID	source_ segment	target_ segment	MPRO_result	seriali-zed
1	Die Innenseite des Scherkopfes reinigen.	Clean the inner side of the shaving head.	{ori=Die,lu=d_art,mlu=d_art,c=w,ehead={nb=sg,infl=weak,case=acc;nom,g=f},gra=cap,ds=d_art,ls=d_art} ,{ori=Innenseite,lu=innenseite,mlu=innenseite,c=noun,ehead={nb=sg,infl=weak;strong;null,case=acc;dat;gen;nom,g=f},g=f,nb=sg,case=acc;dat;gen;nom,gra=cap,ds=innen#seite,gs=u#f,ts=innen#seite,t=innen#seite,ls=innen#seite} ,{ori=des,lu=d_art,mlu=d_art,c=w,ehead={nb=sg,infl=weak,case=gen,g=m;n},gra=small,ds=d_art,ls=d_art} ,{ori=Scherkopfes,lu=scherkopf,mlu=scherkopf,c=noun,ehead={nb=sg,infl=weak;strong;null,case=gen,g=m},g=m,nb=sg,case=gen,gra=cap,ds=scheren#kopf,gs=u#m,ts=scher#kopfes,t=scheren#kopf,ls=scheren#kopf} ,{ori=reinigen,lu=reinigen,mlu=reinigen,c=verb,vtyp=inf,gra=small,ds=reinigen,ts=reinigen,t=reinigen,ls=reinigen} ,{ori=.,lu=.,mlu=.,c=w,gra=other,ds=.,ls=.}	[blob]

Tabelle 24: iMem-TM: Relation mit Informationen zu den im iMem-TM gespeicherten Übersetzungseinheiten

Die zweite Relation für die Datenspeicherung (Beispiel siehe Tabelle 24, TRANSLATION_UNIT) beinhaltet demzufolge die AS- und ZS-Segmente der Übersetzungseinheiten des iMem-TMs in ihrer Originalform (Spalten *source_segment* und *target_segment*), die mit einer eindeutigen ID als Primärschlüssel (Spalte *ID*) versehen werden. Des Weiteren verfügt jeder Eintrag über die vollständige MPRO-Analyse seines AS-Segmentes (Spalte *MPRO_ result*). Zudem werden diese strukturierten Daten in einer Sequenz in Binärform, d. h. als sogenannter blob[64], abgebildet (Spalte *serialized*).

Der blob enthält eine gefilterte MPRO-Analyse, d. h. nur die Merkmale mit ihren Werten, die für den Vergleich eines AS_{neu} mit einem AS_{iMem}

[64] Akronym für *binary large object*; bezeichnet „Binärdaten von maximal *n* Bytes Länge“ (Krypczyk 2011: 69).

notwendig sind. Die gefilterten Merkmale sind gemäß folgender Segment-Datenstruktur angeordnet:

- SEGMENT: AS- und ZS-Segment einer Übersetzungseinheit sowie die vollständige MPRO-Analyse des AS-Segmentes. Ein Segment besteht aus n Wörtern (WORD).
- WORD: Wort des AS-Segmentes in seiner Originalform sowie seine Wortposition. Ein Wort kann aus n Basiswörtern (BASE) bestehen. Ebenso können jedem Wort n ehead-Strukturen (EHEAD) sowie n MPRO-Merkmale (FEATURE) zugeordnet werden.
- BASE: Basiswort eines AS-Originalwortes mit seiner Position im Segment.
- EHEAD: ehead-Struktur eines AS-Originalwortes. Einer ehead-Struktur können n MPRO-Merkmale (FEATURE) zugeordnet werden.
- FEATURE: wortartenspezifisches Merkmal mit genau einem Wert (ohne Disjunktionen).

Durch die Abbildung der Daten als blob kann ein schnelles Auslesen aller notwendigen Daten für den Vergleich eines AS_{neu} mit einem AS_{iMem} erfolgen, da diese Daten an nur einer Stelle in der Datenbank komprimiert gespeichert sind; ein zeitaufwendiges Nachschlagen in andernfalls zusätzlich verknüpften Relationen entfällt[65] (vgl. Kofler et al. 2011: 424).

Die Relation für die Suche (Beispiel siehe Tabelle 25, SEARCH) wird verwendet, um schnell die AS-Segmente in der Datenbank auffinden zu können, die potenzielle Match-Partner des AS_{neu} sind. Dazu werden in dieser Relation die Basiswörter eines jeden AS_{iMem} abgelegt (Spalte *source_base*), wobei jedoch nur Basiswörter von eindeutigen Inhaltswörtern (d. h. deren Wortart ein Substantiv, Verb, Adjektiv oder Adverb ist) gespeichert werden. Die Basiswörter der Stoppwörter *brauchen, dürfen, können, mögen, müssen, sollen, wollen, haben, sein, werden, tun, machen*[66] und Interpunktionszeichen werden nicht gespeichert. Den Basiswörtern der gespeicherten AS-Segmente wird eine ID zugewiesen (Spalte *TRANSLATION_UNIT_ID*), die der Fremdschlüssel zu der Relation *TRANSLATION_UNIT* ist. Wurde ein Eintrag als potenzieller Match-Partner eines AS_{neu} in der Relation *SEARCH* aufgefunden,

[65] Zu Beginn der Arbeit wurden die Daten in vielen einzelnen Tabellen gespeichert. Aus Gründen der Abfragezeit der Datenbank wurde diese Konzeption jedoch wieder verworfen. Die Daten werden stattdessen in nur einer Tabelle in serialisierter Form als blob abgebildet. Die Relationen eines Segmentes sind im blob erhalten.

[66] Das heißt dieselben Stoppwörter, die auch bei der Vorfilterung auf das AS_{neu} angewendet werden (siehe Kapitel 4.2.1, Arbeitsschritt 2).

kann mittels des Fremdschlüssels der dazugehörige Eintrag in der Relation *TRANSLATION_UNIT* nachgeschlagen und der blob ausgelesen werden.

SEARCH

source_base	TRANSLATION_UNIT_ID
innen seite scheren kopf reinigen	1

Tabelle 25: iMem-TM: Relation für eine schnelle Suche

Die Spalte *MPRO_result* der Tabelle *TRANSLATION_UNIT* wird mit einem selbst erstellten Parser durchlaufen, um die Relationen *SEGMENT*, *WORD*, *BASE*, *EHEAD* und *FEATURE* zu extrahieren und in der Spalte *serialized* zu speichern. Mit dem Parser werden die Symbole *_$* und *#* des Merkmals *ls* eliminiert, sodass die Basiswörter einzelner Kompositumsbestandteile und abtrennbare Präfixe als einzelne, durch Leerzeichen getrennte Werte (BASE) behandelt werden können (z. B. ls=reinigen#bürste → BASE=reinigen, BASE=bürste; ls=mit_$liefern → BASE=mit, BASE= liefern). Somit können die einzelnen Elemente von Komposita und Verbklammern für die Ermittlung der LCS mitberücksichtigt werden.

Die Symbole *$* und *$+* bleiben hingegen erhalten, da nicht abtrennbare Präfixe sowie das Präfix *un* immer unmittelbar vor dem Basiswort stehen – entgegen abtrennbaren Präfixen, die auch an anderer Stelle als vor dem Basiswort im Segment auftreten können. Somit wird ein Basiswort inklusive seines nicht abtrennbaren Präfixes bzw. des Präfixes *un* als eine Einheit bei der Ermittlung der LCS behandelt. Die Symbole *$* und *$+* werden im iMem-TM mit abgespeichert (z. B. ls=ent$riegeln → BASE=ent$riegeln; ls=un$+brauchen → BASE= un$+brauchen).

Werden für ein wortartenspezifisches Merkmal mehrere Werte analysiert, so werden diese aufgetrennt (z. B. case=acc;nom → FEATURE case=acc, FEATURE case=nom). Ebenso werden ehead-Strukturen in ihre einzelnen Merkmale aufgespalten (z. B. ehead={nb=sg, infl=weak, case=acc;nom,g=n} → EHEAD {FEATURE nb=sg, FEATURE infl=weak, FEATURE case=acc, FEATURE case=nom, FEATURE g=n}).

In dem nachfolgenden Schema, auch Entity-Relationship-Modell[67] genannt, werden die Datenbankstruktur und die im blob gespeicherte Segment-Datenstruktur veranschaulicht.

[67] Mit einem Entity-Relationship-Modell kann „die grundlegende Tabellen- und Beziehungsstruktur einer Datenbank entworfen und abgebildet [werden]“ (Krypczyk 2011: 70).

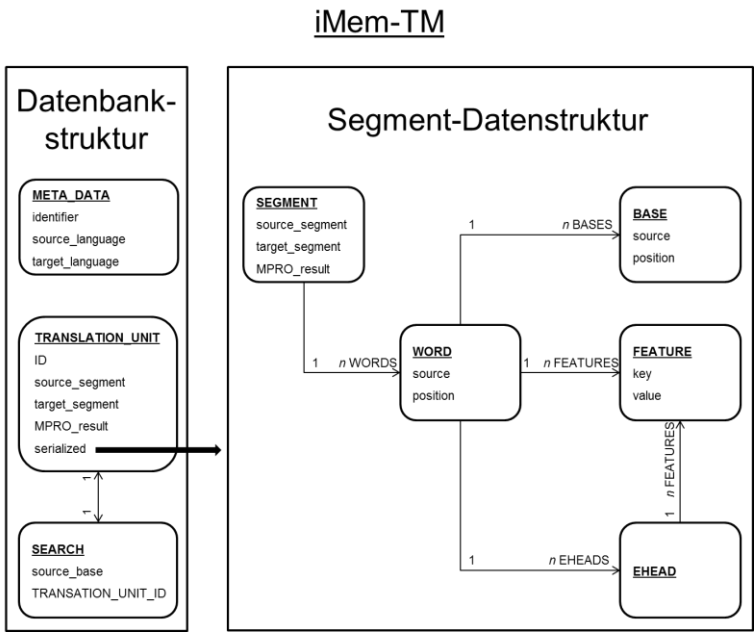

Abbildung 25: Entity-Relationship-Modell des iMem-TMs

Im iMem-TM werden keine Dubletten gespeichert, d. h., ein und dieselbe Übersetzungseinheit kann nicht mehrfach abgelegt werden. Gibt es zu einem AS_{iMem} jedoch mehrere ZS-Entsprechungen, wird jede dieser Übersetzungseinheiten im iMem-TM abgespeichert.

Da die Datenbank an das kommerzielle TM-System über ein Plug-in angebunden wird, muss sie auf demselben Betriebssystem lauffähig sein wie das kommerzielle TM-System. Daher wurde das iMem-TM für Windows konzipiert. Um schnelle Antwortzeiten des Plug-ins zu gewährleisten, sind mindestens 512 MB RAM notwendig. Des Weiteren wird die Java-Version *Java SE 7u51 32Bit* für das iMem-TM verwendet. Die 32Bit-Version ist notwendig, da die MPRO-DLL nur als 32Bit-Version vorliegt.

5.2 Praktische Anwendung des iMem-TMs

5.2.1 Benutzeroberfläche

Das iMem-TM ist in Form eines Plug-ins mit SDL Trados Studio 2009 verknüpft. Der Benutzer kann jedoch selbst entscheiden, ob er das Plug-in

aktivieren[68] möchte oder nicht. Ist das Plug-in aktiviert, besteht die Möglichkeit, ein oder mehrere iMem-TMs in den Übersetzungsprozess einzubinden. Es sind drei Übersetzungsszenarien möglich:

1. Die Übersetzung kann weiterhin konventionell erfolgen, d. h. ausschließlich mit einem oder mehreren SDL-TMs.
2. Die Übersetzung kann ausschließlich linguistisch optimiert erfolgen, indem nur ein oder mehrere iMem-TMs aktiviert werden.
3. Die Übersetzung kann gemischt erfolgen, d. h., zu einem oder mehreren SDL-TMs können ein oder mehrere iMem-TMs hinzugeschaltet werden. In der Trefferanzeige werden sowohl Matches aus dem SDL-TM als auch aus dem iMem-TM angezeigt.

Über die Benutzeroberfläche von SDL Trados Studio 2009 kann der Übersetzer festlegen, in welchem Umfang das jeweilige dem Übersetzungsprojekt zugewiesene iMem-TM verwendet werden soll (TM aktivieren, durchsuchen, aktualisieren, Abzugsvergabe, Einsatz der Konkordanzsuche; siehe auch Kapitel 5.1.1).

Ein iMem-TM kann einem Übersetzungsprojekt zugewiesen werden, indem entweder ein bestehendes iMem-TM ausgewählt oder ein neues iMem-TM erstellt wird[69]. Diese Optionen können im entsprechenden iMem-Dialogfenster ausgewählt werden (Abbildung 26).

Aufgrund des prototypischen Charakters des Plug-ins ist die Benutzeroberfläche rudimentär gehalten. Die Felder *Source-Lang* und *Target-Lang* geben die Ausgangs- und Zielsprache des ausgewählten bzw. neu zu generierenden iMem-TMs an. Da im iMem-Forschungsprojekt der Algorithmus und das Proximitätsmaß lediglich für die Sprache Deutsch konzipiert wurden, muss im Dialogfenster als Ausgangssprache Deutsch gewählt werden. Andere Ausgangssprachen werden von diesem Prototypen noch nicht unterstützt. Als Zielsprache kann hingegen jede Sprache gewählt werden, die mit dem kommerziellen TM-System bearbeitet werden kann. Sollen mehrere iMem-TMs in den Übersetzungsprozess eingebunden werden, muss jedes iMem-TM einzeln ausgewählt bzw. neu erstellt werden.

---

[68] In SDL Trados Studio 2009 über das Menü *Extras → Plug-ins…*

[69] In SDL Trados Studio 2009 z. B. über das Menü *Projekt → Projekteinstellungen… → Sprachpaare →* [Sprachpaar auswählen] *→ Translation Memorys und automatisierte Übersetzung → Hinzu → iMem Linguistic Translation Memory*

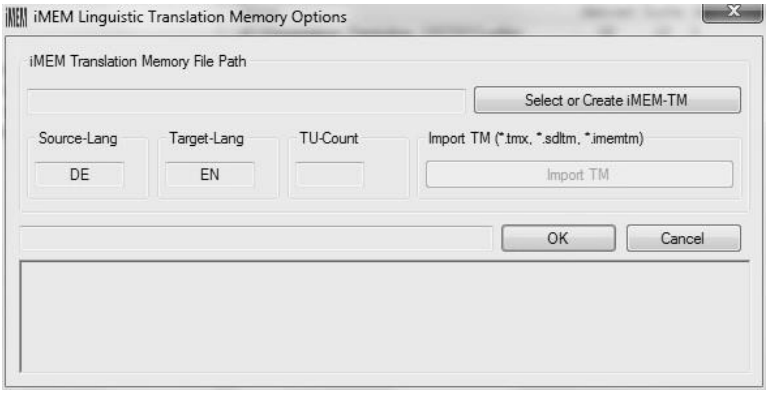

Abbildung 26: Dialogfenster zum Auswählen oder Erstellen eines iMem-TMs

Zum Auswählen eines existierenden iMem-TMs muss die Schaltfläche *Select or Create iMEM-TM* angeklickt und das iMem-TM aus demjenigen Ordner ausgesucht werden, in dem es auf dem lokalen Rechner abgelegt wurde. Das Feld *TU-Count* zeigt an, wie viele Übersetzungseinheiten im selektierten iMem-TM enthalten sind. Über die Statuszeile wird der Benutzer benachrichtigt, wenn das iMem-TM fertig geladen ist (Abbildung 27). Zu einem bereits existierenden iMem-TM können nachträglich weitere Übersetzungseinheiten über die Schaltfläche *Import TM* hinzugefügt werden. Durch Drücken der *OK*-Taste wird die Auswahl bestätigt. Über die Taste *Cancel* kann der Vorgang abgebrochen werden.

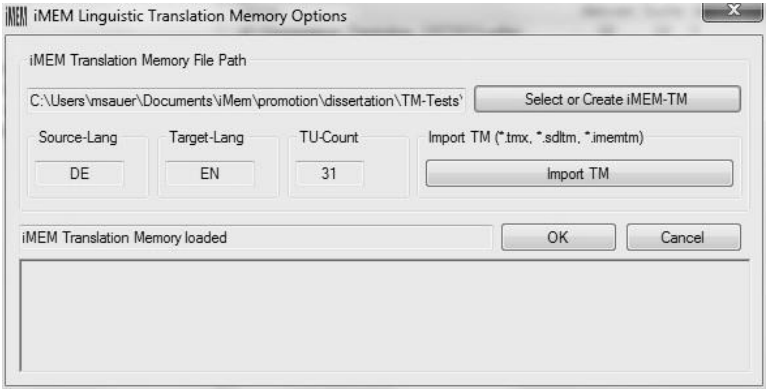

Abbildung 27: Ausgewähltes iMem-TM

Soll hingegen ein neues iMem-TM erzeugt werden, dient die Schaltfläche *Select or Create iMEM-TM* dazu, das zu generierende iMem-TM mit einem Namen zu versehen und den Speicherort auf dem lokalen Rechner festzulegen. Daraufhin wird ein iMem-TM mit der Dateiendung *.imemtm* angelegt. Da das neue iMem-TM noch keine Übersetzungseinheiten enthält, wird in dem Feld *TU-Count* der Wert 0 angezeigt. Auch hier gibt die Statuszeile an, wenn das neu erstellte iMem-TM fertig geladen ist (Abbildung 28). Ab diesem Zeitpunkt können Übersetzungseinheiten in das leere iMem-TM über die Schaltfläche *Import TM* importiert werden.

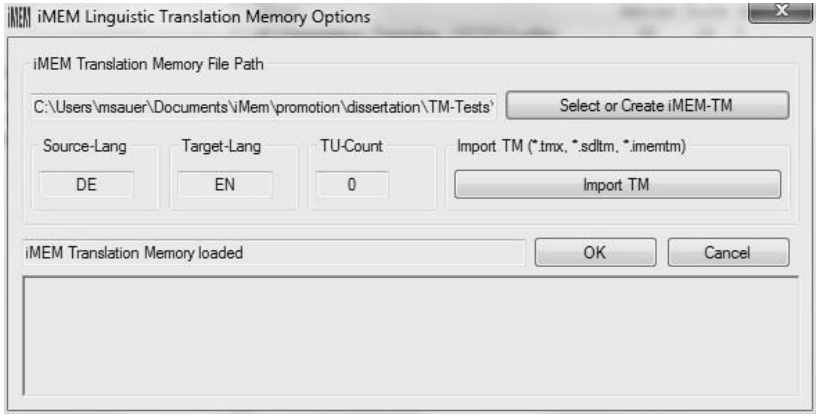

Abbildung 28: Neu erstelltes, leeres iMem-TM

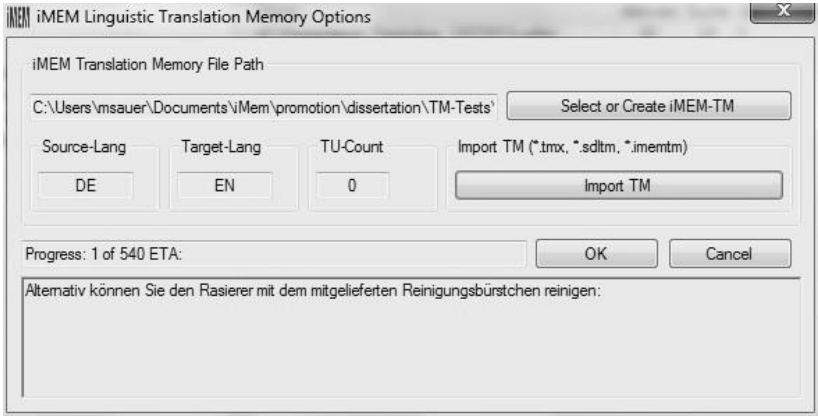

Abbildung 29: Import einer tmx-Datei in ein neues, leeres iMem-TM

Es können nur tmx-Dateien, SDL-TMs oder andere iMem-TMs importiert werden. Falls eine tmx-Datei oder ein SDL-TM für den Import ausgewählt wurde, wird jede einzelne Übersetzungseinheit aus der zu importierenden Datei mithilfe von MPRO analysiert und die für die Anwendung des Algorithmus und des Proximitätsmaßes (siehe Kapitel 4.2 und 4.3) benötigten morphosyntaktischen Merkmale mit ihren Werten extrahiert und in das neue iMem-TM geschrieben. In der Statuszeile wird der Fortschritt angezeigt, während in dem Kommentarfeld das Segment wiedergegeben wird, das momentan analysiert wird (Abbildung 29).

Soll jedoch ein anderes iMem-TM importiert werden, werden die Übersetzungseinheiten des zu importierenden iMem-TMs mit ihren bereits durch MPRO analysierten und extrahierten morphosyntaktischen Merkmalen und Werten dem neuen iMem-TM hinzugefügt.

In der Statuszeile und dem Feld *TU-Count* wird letztlich dargestellt, wie viele Übersetzungseinheiten hinzugefügt wurden. Nach dem Import kann durch Drücken der Schaltfläche *OK* das Dialogfenster geschlossen werden. Das iMem-TM steht fortan zur Übersetzung des Dokumentes zur Verfügung. Matches aus dem iMem-TM werden dabei über die Trefferanzeige des SDL-TMs dargeboten.

Findet ein gemischtes Übersetzungsszenario statt, d. h., wird während der Übersetzung sowohl auf mindestens ein SDL-TM als auch auf mindestens ein iMem-TM zugegriffen, werden in der Trefferanzeige des SDL-TMs die Matches beider TM-Typen aufgeführt. Mithilfe der Statuszeile der Trefferanzeige kann differenziert werden, aus welchem TM der Treffer stammt: Matches aus dem iMem-TM sind durch das iMem-Logo und den Dateipfad des iMem-TMs gekennzeichnet. Handelt es sich hingegen um einen Match aus einem SDL-TM, wird lediglich der Dateiname des SDL-TMs in der Statuszeile angegeben. Des Weiteren können für Matches aus dem SDL-TM Anzeigeoptionen eingestellt werden, z. B., indem festgelegt wird, wie zu löschender oder einzufügender Text markiert werden soll. In Abbildung 30 wird die Unterscheidung zwischen den verschiedenen TMs verdeutlicht: Match Nummer 1 stammt aus dem iMem-TM, da für Matches aus dem iMem-TM keine Anzeigeoptionen einstellbar sind, wohl aber in der Statuszeile das iMem-Logo zu sehen ist. Matches 2 und 3 wurden im SDL-TM gefunden, was durch die Textmarkierung kenntlich gemacht ist.

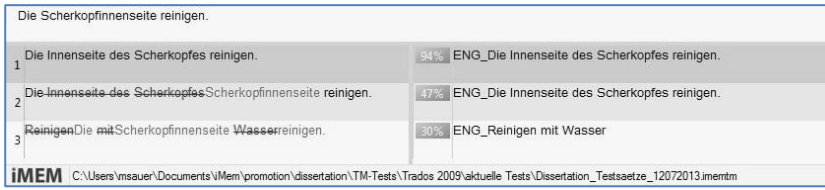

Abbildung 30: Trefferanzeige in SDL Trados Studio 2009 mit Unterscheidung der TMs

5.2.2 Übersetzungsprozess

Bei der Übersetzung eines AS_{neu} durch das iMem-TM wird das AS_{neu} zunächst mittels MPRO analysiert und das Analyseergebnis mit dem selbst erstellten Parser bereinigt, mit dem ebenso das iMem-TM durchlaufen wurde (siehe Kapitel 5.1.3). Auf diese Weise wird die gleiche Repräsentation des AS_{neu} wie die der AS_{iMem} erzeugt, wodurch ein Vergleich der Segmente möglich ist.

5.2.2.1 Vorfilterung

Nach der morphosyntaktischen Analyse des AS_{neu} und der Bereinigung des Analyseergebnisses kann die Vorfilterung wie in Kapitel 4.2.1 beschrieben erfolgen. Dabei spielen die folgenden MPRO-Merkmale eine bedeutsame Rolle:

- Merkmal *ls*: Mit ihm werden die Basiswörter codiert. Da der Vergleich eines AS_{neu} mit einem AS_{iMem} auf Grundlage der Basiswörter erfolgt, müssen auch für die Vorfilterung die Basiswörter herangezogen werden.
- Merkmal *c*: Mit ihm wird die Wortart einer Wortform codiert. Es handelt sich um ein Inhaltswort, wenn das Merkmal *c* den Wert *noun* (= Substantiv), *verb* (= Verb), *adj* (= Adjektiv) oder *adv* (= Adverb) annimmt.

Damit ein AS_{iMem} als potenzieller Match-Partner infrage kommt, müssen drei weitere Anforderungen[70] erfüllt sein:

1. Mindestens 50 % der eindeutigen Inhaltswörter des AS_{neu} müssen in einem AS_{iMem} enthalten sein. Mit dem Wert von 50 % wird eine Mindestähnlichkeit zwischen den beiden Segmenten erzielt. Alle AS_{iMem}, die weniger als

[70] Die Anforderungen beziehen sich auf die Arbeitsschritte 3, 4 und 6 zur Vorfilterung (siehe Kapitel 4.2.1).

50 % der eindeutigen Inhaltswörter des AS_{neu} besitzen, werden als potenziell unähnlich erachtet.
2. Die Segmentlänge des AS_{iMem} darf nicht länger oder kürzer als 400 % im Vergleich zum AS_{neu} sein. Der Wert von 400 % wurde gewählt, da größere Segmentlängenunterschiede in einer größeren Unähnlichkeit der beiden Segmente resultieren würden – insbesondere wenn das AS_{neu} sehr kurz ist.
3. Die sortierte Liste enthält die maximal 50 besten AS_{iMem} hinsichtlich der Anzahl an mit dem AS_{neu} übereinstimmenden Inhaltswörtern. Alle nachfolgenden Einträge werden verworfen. Die Erstellung der GSAs, die Ermittlung der LCS und die Anwendung des Proximitätsmaßes werden demnach höchstens 50 Mal für ein AS_{neu} durchgeführt. Auf diese Weise wird ein guter Kompromiss zwischen akzeptabler Rechenzeit und Anzahl der Übersetzungsergebnisse erreicht.

5.2.2.2 Erstellung der GSAs und Ermittlung der LCS

Die Erstellung der GSAs und die Ermittlung der LCS zwischen dem AS_{neu} und jedem der herausgefilterten AS_{iMem} wird wie in Kapitel 4.2.2 und 4.2.3 beschrieben durchgeführt. Dazu werden die Werte des MPRO-Merkmals *ls* herangezogen. Auf diese Weise besteht eine größere Wahrscheinlichkeit, eine höhere Anzahl an LCS zwischen den beiden Segmenten zu finden als bei der Verwendung der Wörter in ihrer Originalform. Werden bei der MPRO-Analyse eines AS_{neu} bzw. eines AS_{iMem} mehrere Lesarten einer Wortform codiert, so wird der Algorithmus für beide zu vergleichenden Segmenten lediglich für die erste Lesart dieser Wortform durchgeführt.

In der nachfolgenden Tabelle sind zwei miteinander gematchte Segmente aufgeführt. Die jeweiligen LCS sind gleichartig unterstrichen. In diesem Beispiel-Segmentpaar konnten drei LCS ermittelt werden, da das Symbol #, das die einzelnen Kompositumsbestandteile voneinander trennt, eliminiert wurde. Würde hingegen die Originalform der Wörter zur Ermittlung der LCS herangezogen, könnte die Kompositazerlegung nicht berücksichtigt und lediglich zwei LCS identifiziert werden (*Die* und *reinigen.*).

	Segmente in Originalform	**Segmente in Form ihrer Basiswörter (Merkmal *ls*, ohne #-Symbol)**
AS_{neu}	Die Scherkopfinnenseite reinigen.	d_art scheren kopf innen seite reinigen .
AS_{iMem}	Die Innenseite des Scherkopfes reinigen.	d_art innen seite d_art scheren kopf reinigen .

Tabelle 26: Ermittlung der LCS zwischen zwei Segmenten mithilfe des Merkmals *ls*

Enthält ein Wort ein Präfix, muss unterschieden werden, ob es sich entweder um ein abtrennbares Präfix (durch *_$* gekennzeichnet) oder um ein nicht abtrennbares Präfix bzw. das Präfix *un* (durch *$* bzw. *$+* gekennzeichnet) handelt. Nicht abtrennbare Präfixe sowie das Präfix *un* werden für die Ermittlung der LCS mit dem Basiswort als eine Einheit betrachtet, da diese Präfixe an keiner anderen Stelle als vor diesem Basiswort im Segment vorkommen können.

Dagegen werden abtrennbare Präfixe als eigenständige Einheiten beim Vergleich zweier Segmente betrachtet, da abtrennbare Präfixe sowohl am Basiswort als auch – im Falle von Verbklammern – an anderer Stelle im Segment auftreten können.

Ein Ignorieren der Präfixe hätte hingegen zur Folge, dass bei bloßer Betrachtung der Basiswörter eventuell semantisch unterschiedliche Wörter als identisch angesehen würden (z. B. entriegeln, verriegeln → Basiswort: riegeln, obwohl ls=ent$riegeln und ls=ver$riegeln konträre Bedeutungen haben).

5.2.2.3 Proximitätsmaß

Das Proximitätsmaß wird wie in Kapitel 4.3 erläutert angewendet. Bei der Berechnung des G-Wertes werden stets die MPRO-Merkmale *ori* (= Originalform des Wortes), *c* (= Wortart) und – für den Fall, dass es sich bei der Wortart um ein Funktionswort handelt – zusätzlich das Merkmal *sc* (= Unterwortart) zwischen den einzelnen gematchten Basiswörtern des AS_{neu} und des jeweiligen AS_{iMem} miteinander verglichen. Besteht eine Übereinstimmung des Merkmals *c* (und ggf. *sc*), werden weitere wortartenspezifische MPRO-Merkmale zwischen den gematchten Basiswörtern auf Gleichheit geprüft. Tabelle 27 gibt Aufschluss darüber, welche weiteren Merkmale für welche Wortart miteinander verglichen werden müssen.

Bei den Wortarten *verb*, *adj*, *adv*, *fromto* und *z* kann es vorkommen, dass bei der MPRO-Analyse unterschiedliche Merkmale ermittelt werden (in Tabelle 27 durch *oder* gekennzeichnet). Welche Merkmale bestimmt werden, hängt davon ab, wie das analysierte Wort im Segment eingebunden ist. In Anhang H werden Beispiele für die einzelnen Möglichkeiten aufgelistet.

Werden zwei Segmente miteinander verglichen, besteht folglich die Möglichkeit, dass in dem einen Segment bestimmte MPRO-Merkmale ermittelt werden, in dem anderen Segment jedoch nicht, obwohl die Basiswörter (und deren Wortart) identisch sind. Daher werden verschiedene Arten von Kosten – wie in Kapitel 4.3.2 aufgeführt – unterschieden.

Enthält eine Lesart einer Wortform eine ehead-Struktur mit mehreren Merkmalbündeln, wird für alle Werte der in den Merkmalbündeln enthaltenen

Merkmale *nb*, *case* und *g* die Ähnlichkeit der grammatischen Struktur (G-Wert) berechnet. Mit jedem dieser G-Werte wird ein finaler Match-Wert für das Segmentpaar ermittelt. Der beste finale Match-Wert wird letztlich dem Übersetzer angezeigt.

Übereinstimmende Wortart	**Weitere zu vergleichende Merkmale**
c=noun	nb, g, case
c=verb	nb, tns, vtyp *oder* nb, vtyp *oder* vtyp, deg *oder* vtyp
c=adj (Adjektiv in attributiver Verwendung)	nb, g, case, deg, ptc *oder* nb, g, case, deg
c=adv (Adjektiv in prädikativer/adverbialer Verwendung)	deg, ptc *oder* deg
c=fromto	nb, g, case *oder* –
c=z	nb, g, case *oder* –
c=vpref	–
c= sgml	–
c=w, sc=art; card; conj; dem; interr1; interr2; noun; p; pers; poss; post; pron; quant; rel; relposs	nb, g, case
c=w, sc=refl	case
c=w, sc=adv; cit; clause; comma; compar; coord; enum; itj; leer; part; pred; punct; siehe; slash; subj; um_zu; unknown; webadresse; wh_adv; zu_inf	–

Tabelle 27: Weitere zu vergleichende Merkmale in Abhängigkeit zur übereinstimmenden Wortart zwischen AS_{neu} und AS_{iMem}

5.2.2.4 Anzeige und Verarbeitung der Übersetzungsergebnisse

Nachdem mithilfe des iMem-Proximitätsmaßes die Match-Werte zwischen dem AS_{neu} und eines jeden herausgefilterten AS_{iMem} berechnet wurden, können zwei Szenarien eintreten: Entweder handelt es sich bei den verglichenen Segmenten um einen Match (Exact- bzw. Fuzzy-Match) oder um einen No-Match.

Im ersten Fall wird jedes gematchte AS_{iMem} mit seiner ZS-Entsprechung (ZS_{iMem}) in der Trefferanzeige aufgeführt. Der Übersetzer kann daraufhin das geeignetste ZS_{iMem} in die ZS-Seite des Übersetzungseditors von SDL Trados Studio 2009 einfügen und entscheiden, ob er es so belassen oder bearbeiten möchte. Wird das ZS_{iMem} modifiziert, kann die dadurch neu entstandene Übersetzungseinheit in das iMem-TM gespeichert werden. Bei diesem Vorgang

wird automatisch eine morphosyntaktische Analyse angestoßen, sodass die neue Übersetzungseinheit unmittelbar mit ihren linguistischen Merkmalen in der Datenbank abgelegt wird.

No-Matches werden hingegen nicht in der Trefferanzeige dargestellt. Kann sonst kein Match gefunden werden, muss der Übersetzer das AS_{neu} selbst übersetzen. Die dadurch neu entstandene Übersetzungseinheit kann in das iMem-TM inklusive ihrer linguistischen Merkmale – aufgrund der MPRO-Analyse beim Speichervorgang – übernommen werden.

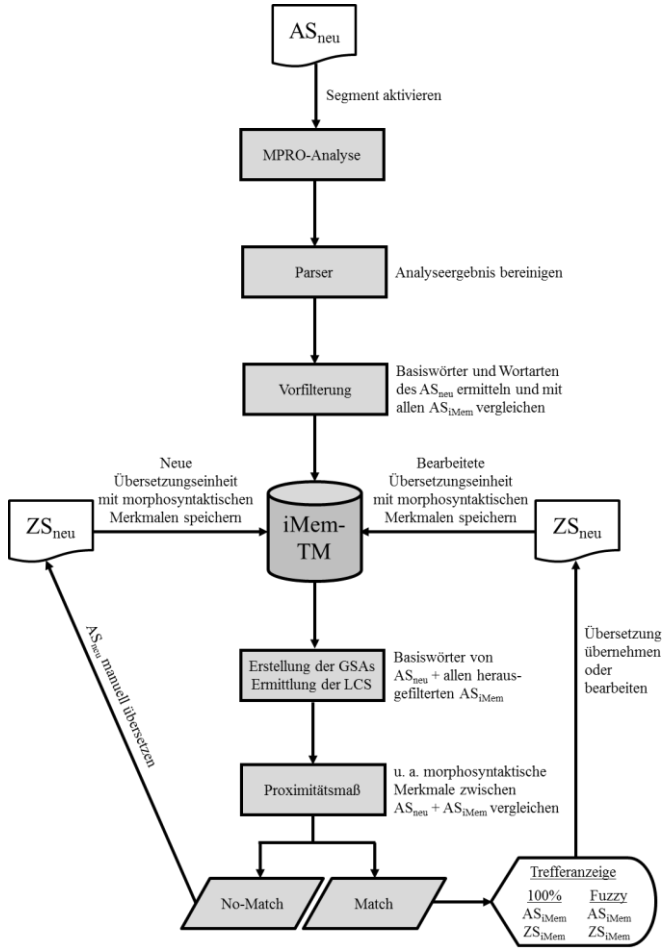

Abbildung 31: Übersetzungsprozess mit dem iMem-TM

In Abbildung 31 wird der komplette Übersetzungsprozess mit dem iMem-TM nochmals veranschaulicht – beginnend bei der Aktivierung des AS_{neu} im Übersetzungseditor, hin zur MPRO-Analyse und Bereinigung durch den Parser, weiter zur Vorfilterung, Erstellung der GSAs und Ermittlung der LCS, Anwendung des Proximitätsmaßes und schließlich zur Anzeige der Matches bzw. manuellen Übersetzung im Falle eines No-Matches. Der illustrierte Ablauf bildet dabei eine ausschließlich linguistisch optimierte Übersetzung ab.

Wie in Kapitel 5.2.1 erwähnt, kann der Übersetzungsprozess jedoch auch gemischt, d. h. mit SDL-TMs und iMem-TMs, stattfinden. In diesem Fall wird parallel zum oben beschriebenen Vorgehen eine konventionelle Abfrage aller hinzugeschalteten SDL-TMs durchgeführt; der Algorithmus von SDL Trados Studio 2009 bleibt also unverändert. In der Trefferanzeige werden sowohl Matches aus allen aktivierten iMem-TMs als auch aus allen aktivierten SDL-TMs angegeben. Der Übersetzer kann somit ebenfalls einen Match aus einem SDL-TM für die Übersetzung des AS_{neu} wählen.

Auch bei einem gemischten Szenario können neue Übersetzungseinheiten im Übersetzungseditor entweder durch eine manuelle Übersetzung durch den Humanübersetzer oder durch Bearbeitung der aus dem iMem-TM oder SDL-TM stammenden ZS-Entsprechungen erstellt werden. Soll die neue Übersetzungseinheit für den weiteren Übersetzungsprozess zur Verfügung stehen, wird sie sowohl im iMem-TM, wie oben beschrieben, als auch im SDL-TM, ohne zwischengeschaltete linguistische Analyse, gespeichert.

5.3 Unterschiede zu anderen linguistisch optimierten TMs

Sowohl das iMem-TM als auch alle anderen in Kapitel 3.2.2 und 3.3.2 erwähnten linguistisch optimierten TMs benötigen computerlinguistische Analyseverfahren, um die zu vergleichenden Segmente linguistisch zu analysieren.

Dennoch bestehen Unterschiede zwischen dem iMem-TM und anderen linguistisch optimierten TMs: So segmentieren die meisten linguistisch optimierten TMs die zu vergleichenden Segmente in linguistisch motivierte (meist syntaktische) Subsegmente. Falls eine Generalisierung der Segmente vorgenommen wird, geschieht dies meistens, indem die Grundform der Wörter gebildet wird. Weitere extrahierte linguistische Informationen beschränken sich auf die Wortart oder semantische Informationen. Lediglich die TELA-Struktur (siehe Kapitel 3.3.2) ermöglicht den Vergleich umfangreicherer Informationen, die jedoch neben den soeben erwähnten linguistischen Informationen eher nicht linguistischer Natur sind.

Die TELA-Struktur ist zudem ein Vertreter linguistisch optimierter TMs, die die kompletten zu vergleichenden Segmente in verschiedenen Repräsentationen darstellen (als Folge von Grundformen, als Folge von Wortarten etc.) und den Vergleich der Segmente mit jeder dieser Repräsentationen durchführen.

Des Weiteren verfügen andere linguistisch optimierte TMs häufig über zusätzliche Nachschlagewerke (z. B. morphologische Wörterbücher, Phrasenlexika, Wissensdatenbanken oder im Falle von Similis sogar eine linguistisch optimierte Terminologiedatenbank), mit denen der Vergleich der Segmente ermöglicht bzw. erweitert wird. Eine Kombination mit MÜ-Komponenten wird zudem in manchen Systemen eingesetzt, um No-Matches zu übersetzen.

Im Vergleich zu anderen linguistisch optimierten TMs ist das iMem-TM schlanker gehalten: Es werden weder Nachschlagewerke noch MÜ-Systeme hinzugeschaltet. Stattdessen wird ausschließlich auf ein einziges morphosyntaktisches Analyseprogramm zurückgegriffen. Dadurch müssen jedoch die extrahierten linguistischen Informationen umfangreicher sein: Neben der Wortart bzw. Unterwortart werden Informationen zu Kasus, Genus, Numerus, Tempus, Verbtyp, Steigerungsform und adjektivischer Partizipialform ermittelt sowie ggf. ein Wortartwechsel untersucht. Des Weiteren wird für eine Generalisierung der Segmente nicht die Grundform der Wörter, sondern deren Basiswörter ermittelt. Die im iMem-TM gespeicherten AS-Segmente werden dabei im Gegensatz zu anderen linguistisch optimierten TMs nicht als Folge von linguistischen Merkmalen in der Datenbank gespeichert, sondern in verschiedenen Relationen in der Datenbank abgelegt und mit der Übersetzungseinheit verknüpft.

Semantische Informationen werden derzeit noch nicht aus MPRO extrahiert. Im Falle einer Weiterentwicklung des iMem-TMs wäre dieser Schritt jedoch denkbar, sofern MPRO bis dahin ausreichend semantische Informationen ausgeben kann. Ebenso werden im Gegensatz zu anderen linguistisch optimierten TMs keine linguistisch motivierten Subsegmente/ Phrasen erzeugt. Durch die Anwendung der Datenstruktur der GSAs – die von keinem anderen linguistisch optimierten kommerziellen TM eingesetzt wird – werden lediglich nicht linguistisch motivierte Teilzeichenketten (d. h. die LCS zwischen dem AS_{neu} und dem AS_{iMem}) ermittelt. Diese LCS werden dem Übersetzer jedoch nicht als einzelne Subsegmente dargeboten, sondern lediglich für die Berechnung des Match-Wertes zwischen dem kompletten AS_{neu} und dem kompletten AS_{iMem} verwendet.

Bezüglich der Match-Wert-Berechnung können ebenfalls Unterschiede zwischen dem iMem-TM und anderen linguistisch optimierten TMs festgestellt werden. So gibt z. B. Similis in der Trefferanzeige auch Subsegmente aus und berechnet für diese Subsegmente einen eigenen Match-Wert. Das

iMem-TM findet hingegen lediglich vollständige Übersetzungseinheiten und berechnet auch den Match-Wert nur für das komplette AS_{iMem}.

In den meisten Systemen wird zur Berechnung der Match-Werte die Edit Distance als Proximitätsmaß angewendet. Obwohl das iMem-Proximitätsmaß wie die Edit Distance Kosten für Unterschiede zwischen den beiden zu vergleichenden Segmenten zuteilt, bezieht das iMem-Proximitätsmaß mehr Informationen in die Berechnung des Match-Wertes ein: Während sich die Edit Distance auf einfache Hinzufügungen, Löschungen und Ersetzungen von Tokens beschränkt, werden beim iMem-Proximitätsmaß zusätzlich linguistische Unterschiede berücksichtigt.

6 Evaluierung des iMem-TMs

Um zu testen, ob die Übersetzung mit dem iMem-TM effektiv und effizient[71] ist, ist es notwendig, das entwickelte System zu evaluieren. Dazu muss eine Testreihe erstellt werden, die laut Manning et al. (2008: 140) standardmäßig drei Merkmale erfüllen muss. In Bezug auf TMs lauten diese wie folgt:

1. Das TM muss über eine ausreichend große Anzahl an Übersetzungseinheiten für den Vergleich mit dem AT_{neu} verfügen.
2. Der AT_{neu} muss über eine ausreichend große Anzahl an Segmenten verfügen[72].
3. Es muss ein Goldstandard erstellt werden, d. h., jedes Segmentpaar bestehend aus einem AS_{neu} und einem im TM gespeicherten AS-Segment muss dahin gehend bewertet werden, ob das im TM gespeicherte AS-Segment mit seiner ZS-Entsprechung relevant oder nicht relevant für die Übersetzung eines AS_{neu} ist.

6.1 Verwendete Korpora und neu zu übersetzende Dokumente

Es werden zwei unterschiedliche Korpora für die Evaluierung des iMem-TMs eingesetzt: Ein bilinguales Textpaar (nachfolgend *Korpus A* genannt) wurde im iMem-Forschungsprojekt von der Verfasserin dieser Arbeit selbst erstellt und besteht aus 535 deutschen und englischen Segmenten, die diversen Bedienungsanleitungen zu Bart- und Haarschneidern, Epiliergeräten, Lady Shavern, Präzisionshaarschneidern und Rasierern der Marke Braun entstammen (vgl. Procter & Gamble Manufacturing GmbH 2011: o.S.). Diese Bedienungsanleitungen wurden gewählt, da sie aufgrund des hohen technischen Grades der Texte für die Übersetzung mit TMs gut geeignet sind und viele ähnliche Sätze aufweisen. Aus Korpus A wurden ein iMem-TM sowie ein SDL-TM mit der Sprachkombination Deutsch-Englisch erstellt.

---

[71] Ein Information-Retrieval-System ist effektiv, wenn die Kosten und der Nutzen des Systems in gutem Verhältnis zueinanderstehen, d. h., wenn die richtigen Dinge ausgeführt werden. Ein Information-Retrieval-System ist hingegen effizient, wenn Ressourcen, beispielsweise Speicherplatz und Laufzeit, sparsam genutzt werden, d. h., wenn die Dinge richtig ausgeführt werden (vgl. van Slype 1979: 51ff., Henrich 2008: 61).

[72] Eine Anzahl von „50 information needs“ (Manning et al. 2008: 140) ist ausreichend. Der Begriff *information needs* wird in dieser Arbeit den Segmenten des neu zu übersetzenden Dokumentes gleichgesetzt.

Um möglichst viele sprachliche Phänomene abdecken und das Verhalten des Systems hinsichtlich einzelner Unterschiede zum AT_{neu} untersuchen zu können (z. B. Änderung des Numerus oder Kasus, Auslassung eines Präfixes, Vertauschung von Haupt- und Nebensatz, Zerlegung eines Kompositums), wurden 49 Segmente manuell angepasst. Damit der Bezug zu einem realistischen Übersetzungsszenario jedoch nicht verloren geht, wurden die restlichen 486 Segmente unverändert aus den Bedienungsanleitungen übernommen. Somit verfügt Korpus A ebenso über komplexe Segmente, die mehr als nur einen sprachlichen Unterschied beim Vergleich mit den AS_{neu} aufweisen. Von diesen 486 nicht modifizierten Segmenten sind des Weiteren 374 enthalten, die bewusst keine inhaltliche oder formale Ähnlichkeit mit irgendeinem Segment des AT_{neu} besitzen.

Der neu zu übersetzende Text (*AT_{neu} A*) enthält 100 deutsche Segmente, die aus denselben Bedienungsanleitungen herrühren wie diejenigen des Korpus A. Auch im Falle des AT_{neu} A wurden – aus denselben Gründen wie bei Korpus A – Segmente modifiziert (47 Segmente) und original beibehalten (53 Segmente). Korpus A sowie der AT_{neu} A wurden so konzipiert, dass ein Vergleich ausschließlich Fuzzy- oder No-Matches, jedoch keine 100 %-Matches liefert. In der nachfolgenden Tabelle wird eine Übersicht über die Anzahl modifizierter und nicht modifizierter Segmente gegeben. Eine detaillierte Auflistung der modifizierten und original beibehaltenen Segmente des Korpus A und AT_{neu} A findet sich zudem in Anhang C bis F.

	AT_{neu} A	**Korpus A**
Anzahl Segmentpaare	**100**	**535**
Anzahl modifizierter AS-Segmente	**47**	**49**
Anzahl original beibehaltener AS-Segmente	**53**	**486**

Tabelle 28: Anzahl modifizierter und nicht modifizierter AS-Segmente im AT_{neu} A und Korpus A

Da Korpus A mit nur 535 Übersetzungseinheiten sehr klein ist, in der Realität jedoch auch TMs in einer Größenordnung von mehreren Tausend Übersetzungseinheiten nicht unüblich sind, wurde ein zweites Testkorpus (*Korpus B*) erstellt. Bei Korpus B handelt es sich um das im Jahr 2013 veröffentlichte DGT-TM (Vol_2012), das den sogenannten acquis communautaire[73] umfasst. Zwar sind Rechtstexte weniger für die Übersetzung mit TMs geeignet, jedoch

[73] Der acquis communautaire ein gemeinschaftlicher Besitzstand der Länder der Europäischen Union, in dem das gesamte geltende EU-Recht dokumentiert ist (vgl. Europäisches Parlament o.J.: o.S.).

sind nur wenige TMs dieser Größenordnung frei verfügbar. So besitzt z. B. die 2013 veröffentlichte Version des DGT-TMs u. a. 472.081 deutsche und 538.949 englische Segmente und ist auf der Internetseite des Joint Research Centre in sechs Teilen kostenlos herunterladbar (vgl. European Commission 2014: o.S.). Für die Erstellung von Korpus B wurden die ersten fünf Teile des DGT-TMs mit der Sprachrichtung Deutsch-Englisch herangezogen, sodass das Testkorpus 407.783 Übersetzungseinheiten enthält, von denen 305.324 Übersetzungseinheiten eindeutig sind. Keine der Übersetzungseinheiten wurde modifiziert. Aus Korpus B wurden ein iMem- und ein SDL-TM erzeugt.

Das zu Korpus B gehörige neu zu übersetzende Dokument (AT_{neu} *B*) entspricht dem sechsten Teil des 2013 veröffentlichten DGT-TMs, aus dem die AS-Segmente der Übersetzungseinheiten extrahiert wurden. AT_{neu} B besteht somit aus 64.298 deutschen Segmenten, die dasselbe Fachgebiet wie die Übersetzungseinheiten in Korpus B behandeln. Die Segmente des AT_{neu} B wurden ebenfalls original beibehalten.

6.2 Relevanz

Bei der Messung der Effektivität eines Systems spielt die sogenannte Relevanz eine tragende Rolle. Stock (2007) definiert den Begriff *relevant* wie folgt:

> „Es geht nicht um das Finden von ‚irgendwelchen' Informationen, sondern nur um das Aufspüren von zutreffendem Wissen, das dem Nutzer bei seinem Informationsbedarf hilft." (Stock 2007: 4)

Wird diese Definition auf die Evaluierung von TMs übertragen, gilt eine Übersetzungseinheit als relevant, wenn der *Inhalt* ihrer AS-Seite identisch mit oder ähnlich dem *Inhalt* des AS_{neu} ist und demnach ein möglichst geringer Bearbeitungsaufwand der ZS-Seite der Übersetzungseinheit erforderlich ist, um eine Übersetzung des AS_{neu} zu erstellen (vgl. Reinke 2004: 154f.).

Ob eine Übersetzungseinheit relevant ist oder nicht, muss durch eine zuvor durchgeführte Bewertung der Datensätze bestimmt werden, was sowohl einen großen zeitlichen Aufwand als auch immer die Gefahr von subjektiven Einschätzungen der bewertenden Person mit sich bringt (vgl. Manning et al. 2008: 143). Für Korpus A und AT_{neu} A wurde eine solche Relevanzbestimmung vorgenommen, bei der jedes Segment des AT_{neu} A jedem AS-Segment des Korpus A gegenübergestellt und bewertet wurde: Von den insgesamt

53.500 möglichen Relevanzmarkierungen[74] wurde 315 Mal die Bewertung *relevant* durch drei Juroren vergeben. Die 315 positiven Bewertungen verteilen sich dabei auf insgesamt 161 Segmente des Korpus A, was bedeutet, dass ein Segment aus Korpus A auch zu mehreren Segmenten des AT_{neu} A als relevant eingestuft werden konnte. Jedem Segment aus AT_{neu} A wurde wiederum mindestens ein relevantes Segment aus Korpus A zugewiesen. Eine Korpus A-AT_{neu} A-Segmentkombination wurde im Endergebnis als relevant angesehen, wenn mindestens zwei der drei Juroren dieser Segmentkombination die Bewertung *relevant* zugeteilt hatten. Bewertete nur einer der drei Juroren eine Korpus A-AT_{neu} A-Segmentkombination als relevant, wurde sie im Endergebnis als *nicht relevant* eingestuft. Es wurde kein Ranking der als relevant befundenen Segmente aus Korpus A vorgenommen.

Insgesamt wurden 53.185 Korpus A-AT_{neu} A-Segmentkombinationen als nicht relevant markiert. 374 Segmente des Korpus A enthalten dabei keine einzige Bewertung mit dem Wert *relevant*. Die nachfolgende Grafik verbildlicht die oben erwähnten Bewertungen der Juroren nochmals. Die linke Grafik gibt die Anzahl der als relevant markierten Korpus A-AT_{neu} A-Segmentkombinationen an, während die rechte Grafik die Anzahl an nicht relevanten Markierungen wiedergibt. Die Schnittmengen der Kreise sollen die identischen Bewertungen zwischen den verschiedenen Juroren verdeutlichen.

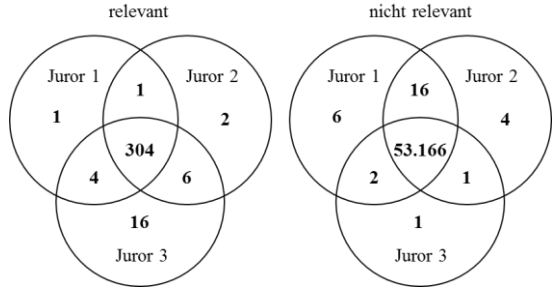

Abbildung 32: Anzahl relevant markierter (links) und nicht relevant markierter (rechts) Korpus A-AT_{neu} A-Segmentkombinationen

Die Relevanzbestimmung wurde nach bestimmten Kriterien durchgeführt. Als relevant eingestuft wurden diejenigen Segmentpaare, die linguistische Unterschiede aufwiesen, durch die jedoch eine Inhaltsgleichheit bzw. -ähnlichkeit zwischen AS_{neu} und AS_{TM} bzw. AS_{iMem} schnell zu erkennen und/oder

[74] Herrührend aus 535 Segmenten des Korpus A in der Bewertung mit jedem der 100 Segmente des AT_{neu} A.

ein geringer Modifikationsaufwand der ZS_{TM} bzw. ZS_{iMem} zur Erstellung der ZS_{neu} zu erwarten war. Einige dieser linguistischen Phänomene werden mit passenden Beispiel-Segmentpaaren in Anhang I aufgelistet.

Für Korpus B wurde hingegen auf eine Relevanzbestimmung verzichtet, da der zeitliche Aufwand aufgrund des Umfangs von Korpus B und AT_{neu} B zu groß für den Zeitrahmen des iMem-Forschungsprojektes gewesen wäre.

6.3 Evaluierungsverfahren im Information Retrieval

In dem 1996 erschienenen EAGLES[75]-Bericht werden unterschiedliche Kriterien zur Evaluierung von TMs aufgeführt, beispielsweise die Analyse der Datenbankkapazität, der Antwortzeit, der Anzahl aufgefundener Matches oder der Qualität der Segmentierung (vgl. EAGLES 1996: o.S.). Da TMs zu Systemen aus dem Bereich des Information Retrievals gezählt werden, können demnach auch Verfahren aus dem Information Retrieval zur Evaluierung eines TMs herangezogen werden.

So haben bereits Whyman und Somers (1999: 1269ff.) TMs mithilfe der Berechnung bestimmter Maße, die aus dem Information Retrieval herrühren, evaluiert. Die Berechnung erfolgte auf Grundlage verschiedener Schwellenwerte für die Trefferanzeige zwischen 0 % und 99 %; 100 %-Matches blieben unberücksichtigt. Auf diese Weise sollen die Retrieval-Leistung der Systeme und die Nützlichkeit der aufgefundenen Matches beurteilt werden.

Des Weiteren können Evaluierungsmethoden aus dem Fachbereich der MÜ für die Evaluierung von TMs hilfreich sein. So beschreibt insbesondere van Slype (1979) verschiedene Evaluierungsmethoden für MÜ-Systeme, die auf TMs übertragen werden können. Im Rahmen einer Makroevaluierung zählt zu diesen Evaluierungsmethoden z. B. das Kriterium der sogenannten Fidelity, mit dem der Grad der inhaltlichen Übereinstimmung zwischen dem Quelltext und seiner Übersetzung z. B. mittels einer Bewertungsskala gemessen wird (vgl. van Slype 1979: 72ff.)[76]. Mit dem Kriterium der sogenannten Usability wird hingegen die Brauchbarkeit der generierten Übersetzungen bewertet (vgl. van Slype 1979: 79).

Ferner kann die in van Slype (1979) beschriebene diagnostische Evaluierung als Teil einer Mikroevaluierung auf TMs angewendet werden. Dabei geht es darum, Fehler in Übersetzungssystemen aufzufinden, zu analysieren und Verbesserungsvorschläge zu geben (vgl. van Slype 1979: 14).

---

[75] Akronym für *Expert Advisory Group on Language Engineering Standards*.

[76] Weitere Beschreibungen dieses Kriteriums finden sich in Lehrberger und Bourbeau (1988: 208f.) sowie in Arnold et al. (1994: 171ff.).

Trotz oder wegen dieser Vielfalt an Evaluierungsmethoden ist es – wie Lehrberger und Bourbeau (1988: 192) betonen – jedoch unmöglich, feste Evaluierungskriterien zu benennen, die auf alle Arten von Übersetzungssystemen anwendbar sind. Die Wahl des Evaluierungsverfahrens hängt nämlich stark vom Übersetzungssystem sowie vom Bedarf des Anwenders ab.

Für die Evaluierung des iMem-TMs werden zum einen Verfahren aus dem Information Retrieval eingesetzt[77]. In den beiden nachfolgenden Unterkapiteln (Kapitel 6.3.1 und 6.3.2) wird dabei auf diejenigen Maße Bezug genommen, die für die Evaluierung des iMem-TMs sinnvoll erscheinen. Zum anderen werden Evaluierungsmethoden aus dem Bereich der MÜ auf das iMem-TM übertragen: Mithilfe eines Online-Fragebogens wird die Fidelity, d. h. die inhaltliche Übereinstimmung zwischen einer Reihe von AS_{neu} und von für relevant befundenen AS_{TM} bzw. AS_{iMem} mithilfe einer Bewertungsskala durch Testpersonen gemessen (siehe Kapitel 6.5.3). Die Brauchbarkeit aufgefundener ZS_{TM} bzw. ZS_{iMem} wird hingegen durch die Bewertung des Nachbearbeitungsaufwandes angegeben, der erforderlich ist, um aus dem ZS_{TM} bzw. ZS_{iMem} die Übersetzung des AS_{neu} zu erstellen (siehe Kapitel 6.5.4). Dies geschieht ebenfalls anhand einer Bewertungsskala, die von sechs Juroren auszufüllen war. In einem weiteren Schritt werden die Gründe für eventuell schlechte Bewertungen diagnostiziert und Optimierungsvorschläge diskutiert (siehe Kapitel 7). In Kapitel 6.5.5 und 6.5.6 wird des Weiteren Bezug auf im EAGLES-Bericht genannte Möglichkeiten zur Effizienzmessung genommen.

6.3.1 Maße für unsortierte Retrieval-Ergebnisse

6.3.1.1 Precision und Recall

Die Berechnung von *Precision* (Präzision, Genauigkeit) und *Recall* (Abdeckung, Vollständigkeit) sind Standardverfahren im Information Retrieval zur Messung der Effektivität eines Systems (vgl. Manning et al. 2008: 142). In Bezug auf die Evaluierung von TMs steht der Begriff *Precision* für den Anteil an aufgefundenen relevanten Übersetzungseinheiten im Verhältnis zu allen aufgefundenen Übersetzungseinheiten. Der Recall-Wert gibt hingegen den Anteil an aufgefundenen relevanten Übersetzungseinheiten im Verhältnis zu allen relevanten Übersetzungseinheiten an (vgl. Henrich 2008: 63f.). Die beiden Maße lassen sich wie folgt berechnen (nach Reinke 2004: 153):

[77] Ein Überblick über verschiedene Evaluierungsmaße findet sich in Manning et al. (2008: 142ff.).

$$\text{Precision} = \frac{\text{Anzahl relevanter ÜEs, die aufgefunden wurden}}{\text{Anzahl aufgefundener ÜEs}}$$

$$\text{Recall} = \frac{\text{Anzahl relevanter ÜEs, die aufgefunden wurden}}{\text{Anzahl aller im TM gespeicherten relevanten ÜEs}}$$

Dieser Sachverhalt kann mithilfe der nachfolgenden Konfusionsmatrix verdeutlicht werden (nach Manning et al. 2008: 143, Carstensen et al. 2010: 586):

	ÜE relevant	**ÜE nicht relevant**
ÜE aufgefunden	richtig-positiv (rp)	falsch-positiv (fp)
ÜE nicht aufgefunden	falsch-negativ (fn)	richtig-negativ (rn)

Tabelle 29: Konfusionsmatrix

Die Precision ist folglich der Quotient aus richtig-positiven Übersetzungseinheiten und der Summe aus richtig-positiven und falsch-positiven Übersetzungseinheiten:

$$Precision = \frac{rp}{(rp + fp)}$$

Beim Recall handelt es sich hingegen um den Quotienten aus richtig-positiven Übersetzungseinheiten und der Summe aus richtig-positiven und falsch-negativen Übersetzungseinheiten:

$$Recall = \frac{rp}{(rp + fn)}$$

Die Werte für Precision und Recall liegen zwischen 0 und 1 bzw. 0 % und 100 % (vgl. Manning et al. 2008: 144). Ein hoher Precision-Wert sagt dabei aus, dass viele der aufgefundenen Übersetzungseinheiten relevant sind. Ist der Precision-Wert hingegen niedrig, deutet dies darauf hin, dass viele der aufgefundenen Übersetzungseinheiten für die Übersetzung eines AS_{neu} unbrauchbar sind. Ein hoher Recall-Wert bedeutet, dass viele der im TM gespeicherten relevanten Übersetzungseinheiten aufgefunden wurden, wohingegen ein niedriger Recall-Wert darauf schließen lässt, dass viele der im TM gespeicherten relevanten Übersetzungseinheiten beim Retrieval übersehen wurden (vgl. Carstensen et al. 2010: 585f.).

Ein optimales System würde bei jeder Anfrage sowohl einen Precision- als auch einen Recall-Wert von 1 ausgeben. In der Realität weisen die beiden Maße jedoch ein gegenläufiges Verhalten auf: Eine Erhöhung der Precision geht einher mit der Verringerung des Recalls, und eine Erhöhung des Recalls führt zu einer Verringerung der Precision (vgl. Stock 2007: 63).

Bei der Berechnung der Precision- und Recall-Werte können Spezialfälle auftreten, nämlich dann, wenn entweder keine relevanten Elemente in einer Dokumentensammlung vorhanden sind oder keine Elemente aufgefunden wurden, was die Division durch 0 zur Folge hätte. So ist es auch bei der Evaluierung des iMem-TMs in wenigen Fällen vorgekommen, dass keine Übersetzungseinheit, d. h. eine leere Antwortmenge, ausgegeben wurde, obwohl zu jedem AS_{neu} mindestens ein relevantes Segment im TM gespeichert war. In diesem Fall beträgt der Recall 0 und die Precision wird ebenfalls auf 0 gesetzt.

Die Berechnung des Recalls kann zudem ein Problem darstellen, da dafür die Anzahl aller in der Datenbank gespeicherten relevanten Elemente bekannt sein muss und bei sehr großen Datensätzen die Relevanzbestimmung häufig nicht oder nicht exakt möglich ist (vgl. Henrich 2008: 69). Da jedoch die Anzahl an Segmenten in Korpus A und AT_{neu} A klein genug ist, konnte eine vollständige Relevanzbestimmung für dieses Korpus durchgeführt werden, sodass die Berechnung des Recalls für jede Anfrage möglich war.

6.3.1.2 Fallout

Beim *Fallout* handelt es sich um die Ausfallquote beim Retrieval. Hinsichtlich TMs wird mit diesem Maß ermittelt, wie hoch der Anteil an aufgefundenen nicht relevanten Übersetzungseinheiten im Verhältnis zu allen im TM gespeicherten nicht relevanten Übersetzungseinheiten ist (vgl. Schmitt 2006: 49f.). Mathematisch kann dieser Sachverhalt bezüglich TMs wie folgt dargestellt werden:

$$\text{Fallout} = \frac{\text{Anzahl nicht relevanter ÜEs, die aufgefunden wurden}}{\text{Anzahl aller im TM gespeicherten nicht relevanten ÜEs}}$$

Unter Zuhilfenahme der Konfusionsmatrix lässt sich der Fallout berechnen durch Division der falsch-positiven Übersetzungseinheiten durch die Summe aus falsch-positiven und richtig-negativen Übersetzungseinheiten:

$$Fallout = \frac{fp}{(fp + rn)}$$

Der Fallout misst demnach, wie gut ein System nicht relevante Übersetzungseinheiten unterdrücken kann. Auch im Falle des Fallouts werden Werte zwischen 0 und 1 bzw. 0 % und 100 % ausgegeben, wobei – entgegen den Maßen *Precision* und *Recall* – ein System umso besser ist, je kleiner der Fallout-Wert ausfällt (vgl. Schmitt 2006: 50). Wird keine Übersetzungseinheit zu einer Anfrage aufgefunden, obwohl sich nicht relevante Übersetzungseinheiten im TM befinden, beträgt der Fallout-Wert 0.

Da die Anzahl an nicht relevanten Elementen einer Dokumentensammlung oft unbekannt ist, wird dieses Evaluierungsmaß seltener angewendet. Dieser Umstand trifft auf Korpus A jedoch nicht zu: Da aufgrund der überschaubaren Anzahl an Segmenten des Korpus A und des AT_{neu} A eine Relevanzbestimmung vorgenommen werden konnte, war es möglich, auch die Ausfallquote für jedes einzelne Segment des AT_{neu} A sowie die durchschnittliche Ausfallquote über alle Anfragen (hier *Mean Fallout* genannt) für beide Systeme zu ermitteln.

6.3.2 Maße für sortierte Retrieval-Ergebnisse

Während die Maße *Precision*, *Recall* und *Fallout* für die Evaluierung unsortierter Retrieval-Ergebnisse verwendet werden können, dienen u. a. (interpolierte) Precision-Recall-Kurven, Precision-Histogramme sowie die Maße *Average Precision*, *Mean Average Precision (MAP)*, *Precision at k*, *R-Precision* und *Mean R-Precision* dazu, die Sortierung der Ergebnisse zu berücksichtigen, wie es beispielsweise bei der Trefferanzeige in einem TM-System der Fall ist. Nachfolgend werden die oben genannten Maße und Verfahren zur Effektivitätsmessung für sortierte Retrieval-Ergebnisse erläutert.

6.3.2.1 Precision-Recall-Kurve

Mithilfe der Precision-Recall-Kurve können die beiden Maße *Precision* und *Recall* in Abhängigkeit voneinander betrachtet werden, d. h., die Kurve gibt wieder, wie hoch die Precision an einem spezifischen Recall-Level *j* zu einer Anfrage *i* ist.

Eine Precision-Recall-Kurve entsteht, indem die Precision- und Recall-Werte, die bei der schrittweisen Abarbeitung der Trefferliste nach relevanten und nicht relevanten Übersetzungseinheiten ermittelt werden, in einem Diagramm abgetragen werden. Auf der X-Achse wird der Recall standardmäßig in elf Einheiten von 0 bis 1 (bzw. 0 % bis 100 %) gleichmäßig unterteilt (vgl. Baeza-Yates/Ribeiro-Neto 1999: 76), während die Precision auf der Y-Achse eingetragen wird. Typischerweise hat die Precision-Recall-Kurve eine gezackte Form, die dadurch zustande kommt, dass für relevante Treffer sowohl

die Precision als auch der Recall ansteigen, für nicht relevante Treffer die Precision jedoch sinkt und der Recall gleich bleibt (vgl. Manning et al. 2008: 145).

6.3.2.2 Interpolierte Precision-Recall-Kurve

Es kann vorkommen, dass die Recall-Levels der einzelnen Anfragen nicht immer den Standard-Recall-Levels entsprechen. Daher ist es notwendig, eine Interpolation der Precision-Recall-Kurve vorzunehmen. Die Interpolation eines Precision-Wertes an einem bestimmten Standard-Recall-Level r erfolgt, indem stets der höhere Precision-Wert zwischen den Recall-Levels j und $(j + 1)$ herangezogen wird (vgl. Baeza-Yates/Ribeiro-Neto 1999: 78).

Eine solche Kurve wird für jede Anfrage an das zu evaluierende System erstellt. Werden mehrere Anfragen an das System gestellt, kann durch Berechnung des arithmetischen Mittels der interpolierten Precision-Werte an jedem Recall-Level eine durchschnittliche interpolierte Precision-Recall-Kurve generiert werden (vgl. Manning et al. 2008: 146). Auf diese Weise lassen sich die Retrieval-Algorithmen unterschiedlicher Systeme vergleichen. Dabei ist dasjenige System am effektivsten, dessen Kurve sich am nächsten an der rechten oberen Ecke des Diagramms befindet.

6.3.2.3 Average Precision

Die *Average Precision* liefert die durchschnittliche Precision über alle aufgefundenen relevanten Übersetzungseinheiten einer Anfrage. Auch hierbei gilt, dass das System für eine spezifische Anfrage um so genauer ist, je größer der Average-Precision-Wert ausfällt.

Für die Berechnung werden alle Precision-Werte jeder aufgefundenen relevanten Übersetzungseinheit zu einer Anfrage addiert und durch die Anzahl aller im TM gespeicherten relevanten Übersetzungseinheiten dividiert. Wird eine relevante Übersetzungseinheit nicht aufgefunden, wird dieser Übersetzungseinheit automatisch der Precision-Wert 0 zugewiesen (vgl. Büttcher et al. 2010: 408).

6.3.2.4 Mean Average Precision (MAP)

Mit der *Mean Average Precision* (*MAP*) kann die durchschnittliche Precision nicht nur für eine einzelne Anfrage, sondern für das komplette System ermittelt werden. Dazu wird das arithmetische Mittel aller Average-Precision-Werte der einzelnen Anfragen gebildet. Dieses Maß dient dazu, die Genauigkeit eines Systems in einer einzigen Zahl auszudrücken.

Die Average-Precision-Werte der einzelnen Anfragen innerhalb eines Systems können aufgrund der unterschiedlichen Anzahl und Rangfolge aufgefundener relevanter Übersetzungseinheiten pro Anfrage stark voneinander abweichen. Da jedoch jede Anfrage gleichgewichtet in die MAP-Wert-Berechnung eingeht, müssen die Anzahl und die Art der Anfragen groß und verschiedenartig genug sein, um einen MAP-Wert zu erhalten, der die Effektivität des zu evaluierenden Systems adäquat repräsentiert (vgl. Manning et al. 2008: 147f.).

6.3.2.5 Precision at *k*

Üblicherweise sind Übersetzer nur an den ersten paar Übersetzungseinheiten der Trefferliste interessiert, da sie selten Zeit haben, alle angezeigten Matches auf ihre Brauchbarkeit hin zu überprüfen. Ein Maß, das diese Vorgehensweise berücksichtigt, ist die *Precision at k* (auch *P@k* abgekürzt). Mithilfe der Precision at k wird die Precision nach k Einträgen in der Ergebnisliste gemessen. Sie wird analog zur Precision ermittelt, jedoch werden lediglich die ersten k aufgefundenen Übersetzungseinheiten für die Berechnung herangezogen:

$$\text{Precision at } k = \frac{\text{Anzahl relevanter ÜEs unter den ersten } k \text{ aufgefundenen ÜEs}}{\text{Anzahl aufgefundener ÜEs bis zur } k\text{-ten Position}}$$

Je höher der Precision-Wert an der *k*-ten Position ist, desto mehr relevante Übersetzungseinheiten befinden sich in den oberen Rängen der Ergebnisliste. Der P@k-Wert spiegelt jedoch nicht die Anordnung der relevanten Übersetzungseinheiten innerhalb der ersten k Positionen wider (vgl. Büttcher et al. 2010: 408).

Obwohl dieses Evaluierungsmaß den Vorteil bietet, dass die Gesamtanzahl aller im TM gespeicherten Übersetzungseinheiten für die Berechnung des P@k-Wertes nicht bekannt sein muss, gilt es als das instabilste Maß zur Effektivitätsmessung eines Systems (vgl. Manning et al. 2008: 148). Durch die willkürliche Wahl des Parameters k (z. B. P@1, P@5, P@10 oder P@20) und die Nichtbeachtung der Gesamtzahl aller gespeicherten relevanten Übersetzungseinheiten kann der P@k-Wert ein verzerrtes Ergebnis hervorbringen: Befinden sich nur wenig relevante Übersetzungseinheiten im TM, ist die Trefferliste jedoch sehr lang, kann der P@k-Wert trotzdem relativ klein ausfallen, obwohl alle relevanten Übersetzungseinheiten aufgefunden wurden. Ebenso ist eine Mittelwertbildung nicht immer geeignet, da die Gesamtanzahl relevanter Übersetzungseinheiten pro Anfrage für einen aussagekräftigen Mittelwert essenziell ist (vgl. Manning et al. 2008: 148).

6.3.2.6 *R*-Precision

Zur Lösung dieses Problems wurde das Maß *R-Precision* eingeführt. Es gibt den Precision-Wert an der R-ten Position der nach Match-Wert absteigend sortieren Ergebnisliste an, wobei R für die Anzahl aller relevanten im TM gespeicherten Übersetzungseinheiten zu einer Anfrage steht (vgl. Baeza-Yates/ Ribeiro-Neto 1999: 80). Das heißt, dass der Parameter k des Maßes *Precision at k* auf R gesetzt wird. Werden alle relevanten Übersetzungseinheiten gefunden und belegen diese zudem die obersten Plätze der Sortierung, liegt der R-Precision-Wert bei 1, wodurch ein gutes System ausgezeichnet ist. Die R-Precision gibt somit einen Hinweis darauf, wie viele der relevanten Übersetzungseinheiten sich auf den vorderen Positionen der Ergebnisliste befinden.

Der Recall-Wert an der R-ten Position entspricht stets dem Precision-Wert an dieser Stelle. Dabei handelt es sich auch um den sogenannten Break-even-Point, mit dem untersucht werden kann, an welcher Position der Sortierung der Precision-Wert identisch mit dem Recall-Wert ist (vgl. Manning et al. 2008: 148).

Des Weiteren kann auch der Durchschnitt der R-Precision-Werte über alle Anfragen (*Mean R-Precision* genannt) ermittelt werden, wodurch das Verhalten des Algorithmus über das gesamte System beobachtet werden kann.

6.3.2.7 Precision-Histogramme

Zum schnellen bildlichen Vergleich des Verhaltens zweier Retrieval-Algorithmen eignen sich Precision-Histogramme. Zur Erstellung eines solchen Histogramms muss zunächst für jeden Algorithmus der R-Precision-Wert zu jeder Anfrage berechnet werden. Der R-Precision-Wert der i-ten Anfrage des zweiten Algorithmus wird anschließend vom R-Precision-Wert der i-ten Anfrage des ersten Algorithmus subtrahiert. Ist der Wert der Subtraktion $RP_{A/B}(i)$ gleich 0, ist das Verhalten beider Algorithmen für die i-te Anfrage identisch. Ein positiver Wert deutet darauf hin, dass die Retrieval-Leistung des ersten Algorithmus besser ist. Ein negativer Wert indiziert demgegenüber eine bessere Retrieval-Leistung des zweiten Algorithmus (vgl. Baeza-Yates/ Ribeiro-Neto 1999: 80).

Die zu jeder Anfrage errechneten Differenzwerte werden daraufhin in einem Precision-Histogramm abgetragen. Die Länge der Balken weist dabei auf das Ausmaß hin, in dem der eine Algorithmus dem anderen bei einer spezifischen Anfrage überlegen ist.

6.4 Ziele der Evaluierung

Mit der Erstellung der zwei Testkorpora bzw. iMem- und SDL-TMs und der dazugehörigen neu zu übersetzenden Dokumente werden zwei unterschiedliche Evaluierungsziele verfolgt: Basierend auf Korpus A und AT_{neu} A werden sowohl für das iMem-TM als auch für das SDL-TM die Maße *Average Precision*, *Mean Average Precision*, *Precision at k*, *R-Precision*, *Mean R-Precision*, *Fallout* und *Mean Fallout* berechnet sowie interpolierte Precision-Recall-Kurven und ein Precision-Histogramm erstellt, um zu ermitteln, wie effektiv das iMem-TM gegenüber dem zeichenkettenbasierten TM-System SDL Trados Studio 2009 im Falle eines Schwellenwertes für die Trefferanzeige von 30 % und 70 % ist.

Des Weiteren wird untersucht, in welchen Match-Wert-Bereichen und bei welchen linguistischen Phänomenen das iMem-TM bessere, schlechtere oder identische Ergebnisse im Vergleich zum SDL-TM liefert. Ein auf ausgewählten Segmenten des Korpus A und AT_{neu} A beruhender und an Testpersonen verteilter Online-Fragebogen soll zudem Aufschluss darüber geben, welche Match-Werte (entweder die des iMem-TMs oder des SDL-TMs) dem menschlichen Ähnlichkeitsempfinden näher kommen. Die für den Fragebogen ausgewählten Segmente finden sich in Anhang J wieder. Dabei wird trotz der unterschiedlichen Algorithmen beider Systeme angenommen, dass die Systeme gleich arbeiten und ihre Match-Werte sowie die Schwellenwerte für die Trefferanzeige miteinander vergleichbar sind.

Wissend über diese Problematik wird zudem mittels Korpus A und AT_{neu} A der erforderliche Nachbearbeitungsaufwand der besten Matches beider Systeme für ausgewählte ZS_{TM} bzw. ZS_{iMem} zur Erstellung der Übersetzung eines AS_{neu} bewertet. Dadurch wird die Brauchbarkeit der ausgegebenen ZS_{TM} bzw. ZS_{iMem} unabhängig von ihren ermittelten Match-Werten beurteilt.

Korpus B und AT_{neu} B hingegen werden dazu verwendet, die Effizienz des iMem-TMs zu untersuchen. Hierbei wird die Antwortzeit im Falle eines iMem-TMs von mehreren Tausend Datensätzen dokumentiert und mit derjenigen eines iMem-TMs mit nur sehr wenigen Übersetzungseinheiten (Korpus A) verglichen, um festzustellen, ob die Menge an Datensätzen eine Auswirkung auf die Antwortzeit hat und ob die gemessene Antwortzeit akzeptabel für die Übersetzungspraxis ist. Ebenso wird der Speicherplatzbedarf der insgesamt vier TMs (iMem-TM des Korpus A und B, SDL-TM des Korpus A und B) gemessen und miteinander verglichen. Analog zur Untersuchung der Antwortzeit zielt auch diese Untersuchung darauf ab, zu ermitteln, ob der gemessene Speicherplatzbedarf für die Übersetzungspraxis annehmbar ist.

6.5 Durchführung und Ergebnisse der Evaluierung

6.5.1 Effektivitätsmessung: statistische Auswertung unter Anwendung der Evaluierungsmaße

Das iMem-TM des Korpus A wurde zusammen mit dem SDL-TM des Korpus A in ein Übersetzungsprojekt in SDL Trados Studio 2009 eingebunden, dem der AT_{neu} A zugewiesen wurde. Folglich wurde ein gemischtes Übersetzungsszenario simuliert. Es wurden keinerlei Abzüge in den Projekteinstellungen definiert und der Wert für die maximale Anzahl anzeigbarer Matches wurde auf 50 gesetzt. Des Weiteren wurde die Option gewählt, dass im Falle eines aufgefundenen 100 %-Matches trotzdem nach Fuzzy-Matches gesucht werden soll. Für diesen Teil der Evaluierung wurden interpolierte Precision-Recall-Kurven und Precision-Histogramme erstellt sowie die Maße *Average Precision*, *Mean Average Precision*, *Precision at k*, *R-Precision*, *Mean R-Precision*, *Fallout* und *Mean Fallout* berechnet. Das Maximum der *R*-Precision-Werte über alle 100 Anfragen betrug 11.

Motiviert durch Whyman und Somers (1999) wurden alle Werte für zwei verschiedene Schwellenwerte für die Trefferanzeige des iMem-TMs und des SDL-TMs ermittelt: einerseits für einen vordefinierten Schwellenwert von 30 % und andererseits von 70 %. Der Schwellenwert von 70 % wurde gewählt, um ein realitätsnahes Evaluierungsergebnis aufzuzeigen, denn in der Übersetzungspraxis wird üblicherweise ein Schwellenwert für die Trefferanzeige von 70 % eingestellt. Der Schwellenwert von 30 % wurde hingegen gewählt, da dieser der niedrigste einzustellende Match-Wert in SDL Trados Studio 2009 ist und demzufolge das umfassendste Retrieval-Ergebnis für die zu vergleichenden TMs ermittelt werden kann – einschließlich relevanter Übersetzungseinheiten unterhalb der 70 %-Grenze.

Zwischen dem AT_{neu} A und dem iMem- bzw. SDL-TM wurden alle Treffer dokumentiert und mithilfe der zuvor erstellten Relevanzbestimmung abgeglichen, wodurch die Anzahl aufgefundener relevanter, aufgefundener nicht relevanter sowie insgesamt aufgefundener Übersetzungseinheiten pro angefragten AS_{neu} ermittelt werden konnte. Auf diese Weise konnten interpolierte Precision-Recall-Kurven und Precision-Histogramme generiert sowie die anderen oben erwähnten Maße für jedes der beiden TMs berechnet werden.

6.5.1.1 Auswertung für einen Schwellenwert von 30 %

Nach der Sichtung der aufgefundenen Übersetzungseinheiten zu jeder Anfrage ergibt sich, dass bei einem vordefinierten Schwellenwert von 30 % das iMem-TM zwar insgesamt mehr Übersetzungseinheiten über alle 100 Anfragen

ausgibt als das SDL-TM (485 ÜEs vs. 417 ÜEs), jedoch sind im Vergleich zum SDL-TM weniger dieser aufgefundenen Übersetzungseinheiten relevant (214 ÜEs vs. 218 ÜEs). Insgesamt hätten über alle Anfragen 313 relevante Übersetzungseinheiten durch beide TMs aufgefunden werden können. Ebenso liefert das iMem-TM insgesamt über alle Anfragen mehr nicht relevante Übersetzungseinheiten als das SDL-TM (271 ÜEs vs. 200 ÜEs).

Die längste Trefferliste des iMem-TMs zu einer Anfrage verfügt über 33 Übersetzungseinheiten, wovon 29 Übersetzungseinheiten nicht relevant sind, was gleichzeitig die Anzahl der meisten aufgefundenen nicht relevanten Übersetzungseinheiten angibt. Trotz dieser hohen Ausfallquote für eine spezifische Anfrage liegt die Anzahl an Anfragen, die keinerlei Fallout aufweisen bei 55. Die höchste zu einer Anfrage ermittelte Anzahl an relevanten Übersetzungseinheiten liegt bei 6, während die niedrigste Anzahl an aufgefundenen relevanten Übersetzungseinheiten im Falle von zwei AS_{neu} 0 beträgt.

	SDL-TM	**iMem-TM**
Anzahl gestellter Anfragen	100	
Gesamtanzahl aller gespeicherten ÜEs in Korpus A	535	
Gesamtanzahl aufgefundener ÜEs über alle Anfragen	417	485
Gesamtanzahl aufgefundener relevanter ÜEs über alle Anfragen	218	214
Gesamtanzahl relevanter ÜEs, die über alle Anfragen hätten aufgefunden werden können	313	
Gesamtanzahl aufgefundener nicht relevanter ÜEs über alle Anfragen	200	271
Höchste Anzahl an ÜEs, die zu einer Anfrage aufgefunden wurden	17	33
Höchste Anzahl an relevanten ÜEs, die zu einer Anfrage aufgefunden wurden	8	6
Höchste Anzahl an nicht relevanten ÜEs, die zu einer Anfrage aufgefunden wurden	11	29
Niedrigste Anzahl an relevanten ÜEs, die zu einer Anfrage aufgefunden wurden	1	0
Anzahl an Anfragen mit den wenigsten aufgefundenen relevanten ÜEs	41	2
Anzahl an Anfragen ohne jeglichen Fallout	42	55
Anzahl an Anfragen, bei denen alle aufgefundenen relevanten ÜEs zu den ersten k ÜEs zählen	85	91

Tabelle 30: Übersicht über die Anzahl an Anfragen bzw. aufgefundenen Übersetzungseinheiten durch das iMem- und SDL-TM bei einem Schwellenwert von 30 %

Das SDL-TM liefert hingegen für fast alle der oben beschriebenen Untersuchungen ein besseres Ergebnis: Die längste Liste zu einem AS_{neu} enthält

17 Übersetzungseinheiten, die auch eine der Trefferlisten mit den meisten relevanten Übersetzungseinheiten (8 ÜEs) ist. Im Gegensatz zum iMem-TM wurde zudem zu jeder Anfrage mindestens eine relevante Übersetzungseinheit gefunden, worunter 41 Anfragen mit nur einer einzigen aufgefundenen relevanten Übersetzungseinheit fallen. Die höchste Anzahl an aufgefundenen nicht relevanten Übersetzungseinheiten beträgt 11, während die Anzahl an Anfragen, bei denen keinerlei Fallout verzeichnet wurde, bei 42 liegt.

Nicht nur im letzteren Fall liefert das iMem-TM einen besseren Wert als das SDL-TM. Auch bei der Anzahl an Anfragen, bei denen alle aufgefundenen relevanten Übersetzungseinheiten zu den ersten *k* Übersetzungseinheiten zählen, übersteigt das iMem-TM mit 91 Anfragen gegenüber 85 Anfragen das SDL-TM.

Diese Statistik vermittelt zunächst den Eindruck, dass die Arbeit mit dem iMem-TM bei einem niedrigen vordefinierten Schwellenwert stets weniger effektiv ist als mit dem SDL-TM. Beim Vergleich der durchschnittlichen interpolierten Precision-Recall-Kurven beider Systeme wird jedoch deutlich, dass diese Annahme lediglich für Recall-Levels unter 60 % zutrifft, da die Kurve für das SDL-TM bis zu diesem Recall-Level über derjenigen des iMem-TMs verläuft (Abbildung 33). Für Recall-Levels über 70 % fällt jedoch die Precision-Recall-Kurve des iMem-TMs weniger steil ab als die des SDL-TMs. Demnach ist im Falle des Korpus A und des AT_{neu} A das iMem-TM im hohen Recall-Level-Bereich effektiver als das SDL-TM.

Nachfolgend werden die durchschnittlichen interpolierten Precision-Recall-Kurven beider Systeme gezeigt. In Tabelle 31 werden zudem die über allen 100 Anfragen gemittelten interpolierten Precision-Werte für das iMem- und SDL-TM zu jedem Standard-Recall-Level aufgeführt.

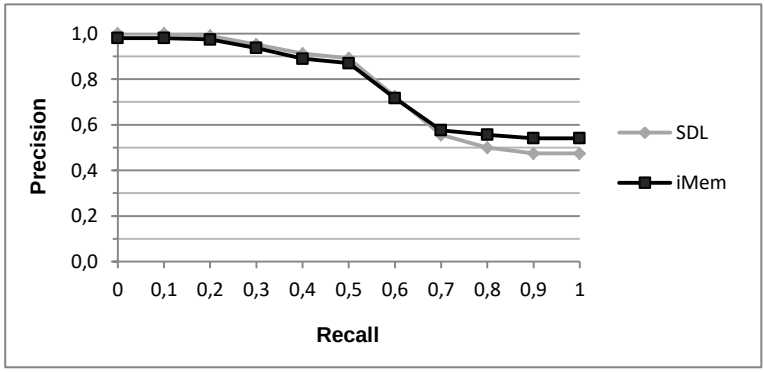

Abbildung 33: Durchschnittliche interpolierte Precision-Recall-Kurven des iMem- und SDL-TMs bei einem Schwellenwert von 30 %

Recall-Level	Interpolierter Precision-Mittelwert SDL	Interpolierter Precision-Mittelwert iMem
0	1,00	0,98
0,1	1,00	0,98
0,2	0,99	0,97
0,3	0,95	0,94
0,4	0,91	0,89
0,5	0,89	0,87
0,6	0,72	0,72
0,7	0,56	0,58
0,8	0,50	0,57
0,9	0,47	0,54
1	0,47	0,54

Tabelle 31: Über 100 Anfragen gemittelte interpolierte Precision-Werte (gerundet) des iMem- und SDL-TMs zu jedem Standard-Recall-Level bei einem Schwellenwert von 30 %

Im Falle der Average Precision konnte beobachtet werden, dass das iMem-TM in 23 zu 18 Anfragen bessere Werte als das SDL-TM liefert. In über der Hälfte der Anfragen (59 Anfragen) konnte zudem ein identischer Average-Precision-Wert verzeichnet werden. Ebenfalls liegt der MAP-Wert des iMem-TMs mit 77,78 % zu 76,56 % über dem des SDL-TMs (Tabelle 32). Wie in Abbildung 34 ersichtlich, wurde auch öfter ein Average-Precision-Wert von 1 beim iMem-TM nachgewiesen als beim SDL-TM (53 vs. 46 Anfragen). Allerdings wurde auch bei zwei Anfragen des iMem-TMs ein Average-Precision-Wert von 0 errechnet. Bei diesen Anfragen handelt es sich zum einen um eine Anfrage, bei der nur nicht relevante Übersetzungseinheiten gefunden wurden, und zum anderen um eine Anfrage, bei der weder relevante noch nicht relevante Übersetzungseinheiten ausgegeben wurden.

	SDL-TM	iMem-TM
Höherer Average-Precision-Wert	18 Anfragen	23 Anfragen
Identischer Average-Precision-Wert	59 Anfragen	
Mean Average Precision (MAP)	76,56 %	77,78 %

Tabelle 32: Average-Precision- und Mean-Average-Precision-Werte (gerundet) für das iMem- und SDL-TM bei einem Schwellenwert von 30 %

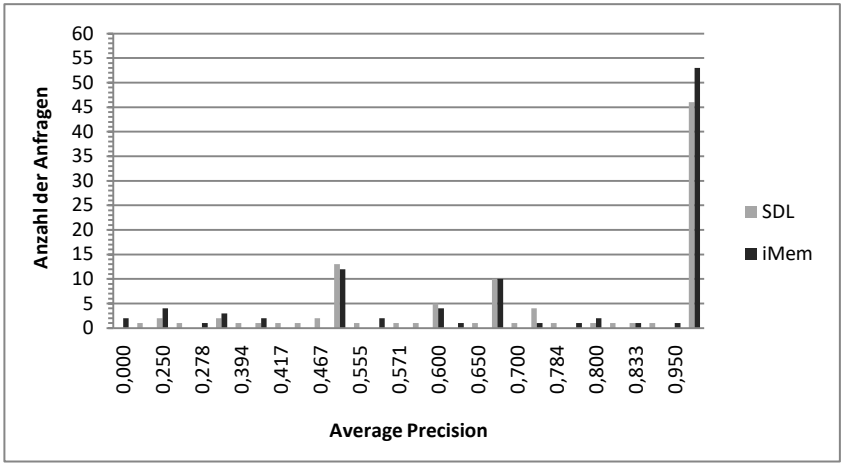

Abbildung 34: Anzahl an Anfragen, bei denen ein spezifischer Average-Precision-Wert (gerundet) des iMem- und SDL-TMs ermittelt wurde bei einem Schwellenwert von 30 %.

Der zuletzt beschriebene Sachverhalt wird nochmals durch Abbildung 35 verdeutlicht, in der die Häufigkeit spezifischer P@1-Werte dargestellt wird. Bei einem Schwellenwert von 30 % wurde immer mindestens eine relevante Übersetzungseinheit durch das SDL-TM aufgefunden. Das iMem-TM hingegen gab bei einer Anfrage nur eine nicht relevante Übersetzungseinheit (P@1 = 0) und bei einer anderen eine leere Antwortmenge (P@1 = Ø) aus.

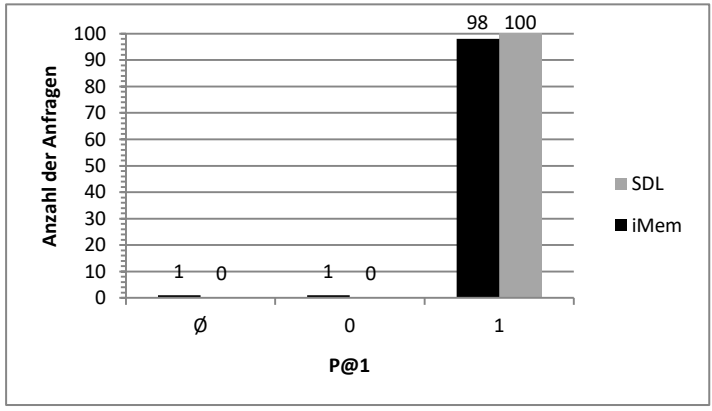

Abbildung 35: Anzahl an Anfragen, bei denen ein spezifischer P@1-Wert des iMem- und SDL-TMs ermittelt wurde bei einem Schwellenwert von 30 %.

Bei Betrachtung der ersten beiden Treffer der Ergebnislisten (P@2) kann zwar häufiger ein P@2-Wert von 1 für das iMem-TM beobachtet werden als für das SDL-TM, allerdings existiert auch bei P@2 häufiger eine leere Antwortmenge für das iMem-TM als für das SDL-TM (Abbildung 36). Analog verhalten sich die Systeme für P@3 (Abbildung 37).

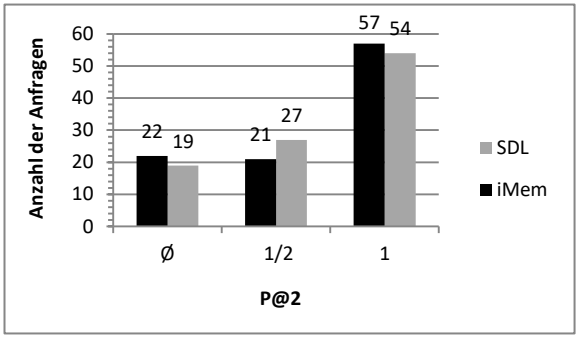

Abbildung 36: Anzahl an Anfragen, bei denen ein spezifischer P@2-Wert des iMem- und SDL-TMs ermittelt wurde bei einem Schwellenwert von 30 %.

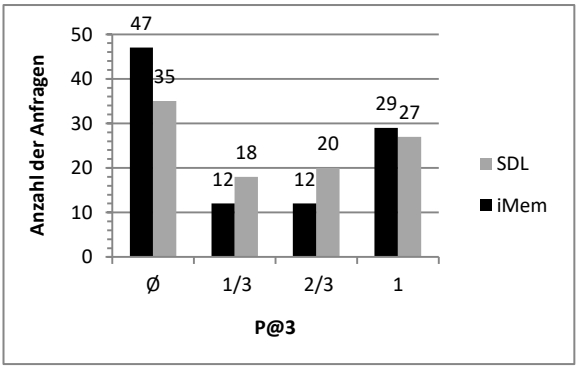

Abbildung 37: Anzahl an Anfragen, bei denen ein spezifischer P@3-Wert des iMem- und SDL-TMs ermittelt wurde bei einem Schwellenwert von 30 %.

Bei der Berechnung der *R*-Precision-Werte beider Systeme konnte festgestellt werden, dass bei fast zwei Dritteln der Anfragen ein identischer *R*-Precision-Wert vorliegt, während in 22 Fällen das iMem-TM einen höheren Wert aufweist als das SDL-TM mit 16 Anfragen (Tabelle 33). Folglich liegt auch die durchschnittliche *R*-Precision des iMem-TMs über derjenigen des SDL-TMs (76,47 % vs. 75 %).

	SDL-TM	iMem-TM
Höherer *R*-Precision-Wert	16 Anfragen	22 Anfragen
Identischer *R*-Precision-Wert	62 Anfragen	
Mean *R*-Precision	75 %	76,47 %

Tabelle 33: R-Precision- und Mean-R-Precision-Werte (gerundet) für das iMem- und SDL-TM bei einem Schwellenwert von 30 %

Das Precision-Histogramm (Abbildung 38), mit dem die Differenzen zwischen den R-Precision-Werten des iMem-TMs und denen des SDL-TMs zu jeder Anfrage für einen Schwellenwert von 30 % ermittelt wurden, liefert den bildlichen Vergleich beider Algorithmen. Die 22 Balken oberhalb der X-Achse bedeuten, dass in diesen Fällen der Algorithmus des iMem-TMs eine bessere Leistung erbringt. Die 16 Balken unterhalb der X-Achse weisen dagegen auf eine bessere Leistung des Algorithmus des SDL-TMs hin. Die Länge der Balken gibt dabei einen Hinweis darauf, wie viele relevante Übersetzungseinheiten durch den jeweiligen Algorithmus zusätzlich bis zur R-ten Position der sortierten Ergebnisliste aufgefunden wurden. Folglich kann in den 62 Fällen, in denen im Precision-Histogramm kein Balken dargestellt ist, eine identische Leistung beider Algorithmen angenommen werden. Die größte Differenz ($RP_{iMem/SDL}(42) = -1$) wurde bei Anfrage 42 ermittelt, die dadurch zustande kommt, dass beim iMem-TM keine Übersetzungseinheit ausgegeben wurde, während das SDL-TM die einzige für relevant befundene Übersetzungseinheit liefert, wenn auch nur mit einem Match-Wert von 45 %.

Obwohl mit 41 zu 28 Anfragen das iMem-TM niedrigere Fallout-Werte liefert als das SDL-TM und zudem in fast einem Drittel der Anfragen ein identischer Fallout-Wert beider Systeme nachgewiesen werden kann, liegt der durchschnittliche, über allen 100 Anfragen ermittelte Fallout-Wert des iMem-TMs mit 0,51 % über demjenigen des SDL-TMs, der lediglich 0,38 % beträgt (Tabelle 34). Dies liegt daran, dass beim iMem-TM die Anfragen, die einen Fallout aufweisen, mehr nicht relevante Übersetzungseinheiten liefern als beim SDL-TM.

	SDL-TM	iMem-TM
Niedrigerer Fallout-Wert	28 Anfragen	41 Anfragen
Identischer Fallout-Wert	31 Anfragen	
Mean Fallout	0,38 %	0,51 %

Tabelle 34: Fallout- und Mean-Fallout-Werte (gerundet) für das iMem- und SDL-TM bei einem Schwellenwert von 30 %

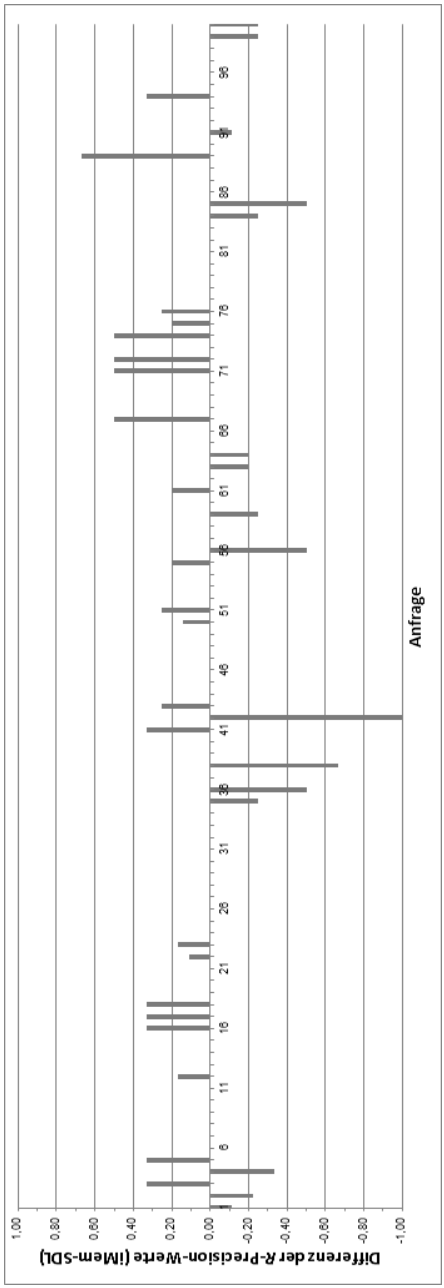

Abbildung 38: Precision-Histogramm: Differenzen der *R*-Precision-Werte (gerundet) zwischen dem iMem- und SDL-TM zu jeder Anfrage bei einem Schwellenwert von 30 %. Die Balken oberhalb der X-Achse zeigen eine bessere Leistung des iMem-Algorithmus. Die Balken unterhalb der X-Achse bedeuten, dass der SDL-Algorithmus leistungsfähiger für eine spezifische Anfrage ist.

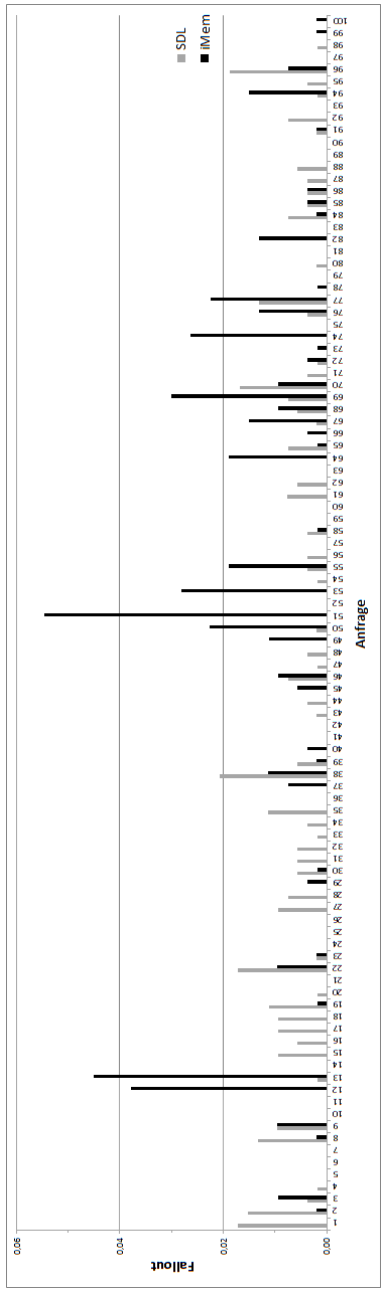

Abbildung 39: Fallout-Werte (gerundet) des iMem- und SDL-TMs pro Anfrage bei einem Schwellenwert von 30 %

In Abbildung 39, in der zu jeder Anfrage die Fallout-Werte des iMem- und SDL-TMs vergleichend gegenübergestellt sind, wird dieser Sachverhalt verdeutlicht. Vor allem bei den Anfragen 12, 13 und 51 können große Unterschiede im Fallout beider Systeme beobachtet werden. Während das SDL-TM bei diesen drei Anfragen maximal drei Übersetzungseinheiten ausgibt, liefert das iMem-TM 22, 26 und 33 Übersetzungseinheiten, darunter dementsprechend viele nicht relevante. Demnach ist das SDL-TM bei einem Schwellenwert von 30 % besser dazu geeignet, nicht relevante Übersetzungseinheiten zu unterdrücken als das iMem-TM.

6.5.1.2 Auswertung für einen Schwellenwert von 70 %

Bei Betrachtung der Evaluierungsergebnisse für einen vordefinierten Schwellenwert von 70 % wird ersichtlich, dass das iMem-TM öfter bessere Ergebnisse liefert als bei der Evaluierung mit einem Schwellenwert von 30 %.

Zwar werden durch das iMem-TM bei einem Schwellenwert von 70 % ebenfalls insgesamt mehr Übersetzungseinheiten über allen 100 Anfragen aufgefunden als durch das SDL-TM (236 vs. 115 ÜEs), jedoch übersteigt das iMem-TM – im Gegensatz zur Evaluierung mit einem Schwellenwert von 30 % – dieses Mal auch die Gesamtanzahl an aufgefundenen relevanten Übersetzungseinheiten im Vergleich zum SDL-TM (204 vs. 115 ÜEs). Die Gesamtanzahl an aufgefundenen nicht relevanten Übersetzungseinheiten liegt allerdings mit 32 beim iMem-TM über derjenigen des SDL-TMs, die 0 beträgt.

Das SDL-TM liefert folglich bei allen Anfragen keinen Fallout, während das iMem-TM nur bei 79 Anfragen keinen Fallout ausgibt. Diese Zahlen schließen jedoch auch Anfragen ein, die eine leere Antwortmenge liefern: Beim iMem-TM kam lediglich bei zwei Anfragen eine leere Antwortmenge zustande, während dies beim SDL-TM bei 22 Anfragen der Fall war. Die höchste zu einer Anfrage ermittelte Anzahl an nicht relevanten Übersetzungseinheiten beläuft sich beim iMem-TM maximal auf vier. Aus dem Vergleich dieser Fallout-bezogenen Werte mit denjenigen der Evaluierung mit einem Schwellenwert von 30 % lässt sich schlussfolgern, dass sich bei beiden Systemen mehr nicht relevante Übersetzungseinheiten im unteren Match-Wert-Bereich befinden.

Die längste Trefferliste des iMem-TMs zählt zehn Übersetzungseinheiten, die ebenso mit sechs Übersetzungseinheiten die meisten relevanten birgt. Im Vergleich dazu ist die längste Trefferliste des SDL-TMs mit fünf Übersetzungseinheiten nur halb so lang, jedoch sind alle dieser aufgefundenen Übersetzungseinheiten auch relevant.

Bei der Untersuchung, bei wie vielen Anfragen alle aufgefundenen relevanten Übersetzungseinheiten zu den ersten *k* Übersetzungseinheiten zählen, liefert das iMem-TM mit 97 Anfragen zu 78 Anfragen ebenfalls einen besseren Wert als das SDL-TM, was u. a. damit zusammenhängt, dass häufig eine leere Antwortmenge durch das SDL-TM eintrat. Die oben beschriebenen Evaluierungsergebnisse werden in der nachfolgenden Tabelle nochmals vergleichend gegenübergestellt.

	SDL-TM	**iMem-TM**
Anzahl gestellter Anfragen	100	
Gesamtanzahl aller gespeicherten ÜEs in Korpus A	535	
Gesamtanzahl aufgefundener ÜEs über alle Anfragen	115	236
Gesamtanzahl aufgefundener relevanter ÜEs über alle Anfragen	115	204
Gesamtanzahl relevanter ÜEs, die über alle Anfragen hätten aufgefunden werden können	313	
Gesamtanzahl aufgefundener nicht relevanter ÜEs über alle Anfragen	0	32
Höchste Anzahl an ÜEs, die zu einer Anfrage aufgefunden wurden	5	10
Höchste Anzahl an relevanten ÜEs, die zu einer Anfrage aufgefunden wurden	5	6
Höchste Anzahl an nicht relevanten ÜEs, die zu einer Anfrage aufgefunden wurden	0	4
Niedrigste Anzahl an relevanten ÜEs, die zu einer Anfrage aufgefunden wurden	0	0
Anzahl an Anfragen mit den wenigsten aufgefundenen relevanten ÜEs	22	2
Anzahl an Anfragen ohne jeglichen Fallout	100	79
Anzahl an Anfragen, bei denen alle aufgefundenen relevanten ÜEs zu den ersten *k* ÜEs zählen	78	97

Tabelle 35: Übersicht über die Anzahl an Anfragen bzw. aufgefundenen Übersetzungseinheiten durch das iMem- und SDL-TM bei einem Schwellenwert von 70 %

Im nachfolgenden Diagramm werden die durchschnittlichen interpolierten Precision-Recall-Kurven beider Systeme dargestellt. In Tabelle 36 werden die genauen dazugehörigen Werte zu jedem Standard-Recall-Level aufgelistet.

Der Vergleich der durchschnittlichen interpolierten Precision-Recall-Kurven beider Systeme für einen Schwellenwert von 70 % belegt durch den deutlichen Abstand beider Kurven zueinander (Abbildung 40), dass das iMem-TM bei höheren vordefinierten Schwellenwerten bei allen Recall-Levels präzisere Übersetzungseinheiten ausgibt als das SDL-TM.

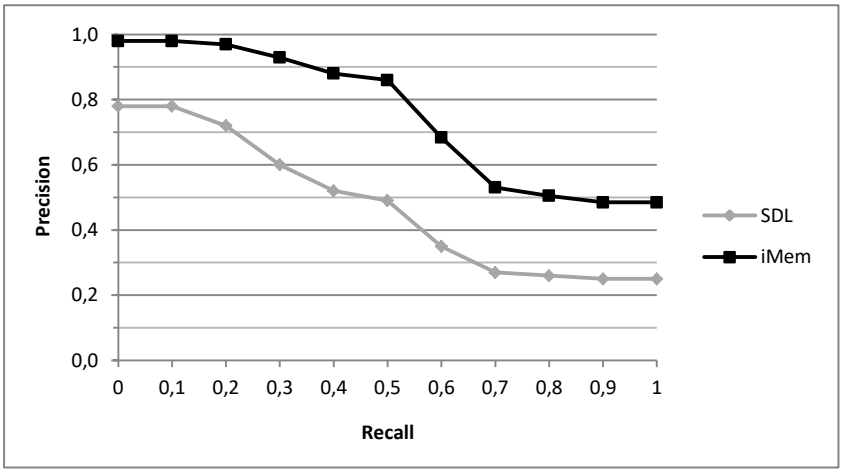

Abbildung 40: Durchschnittliche interpolierte Precision-Recall-Kurven des iMem- und SDL-TMs bei einem Schwellenwert von 70 %

Recall-Level	Interpolierter Precision-Mittelwert SDL	Interpolierter Precision-Mittelwert iMem
0	0,78	0,98
0,1	0,78	0,98
0,2	0,72	0,97
0,3	0,60	0,93
0,4	0,52	0,88
0,5	0,49	0,86
0,6	0,35	0,68
0,7	0,27	0,53
0,8	0,26	0,51
0,9	0,25	0,48
1	0,25	0,48

Tabelle 36: Über 100 Anfragen gemittelte interpolierte Precision-Werte (gerundet) des iMem- und SDL-TMs zu jedem Standard-Recall-Level bei einem Schwellenwert von 70 %

Auch bezüglich der Average Precision kann ein deutlicher Unterschied zwischen beiden Systemen festgestellt werden: Das iMem-TM lieferte bei über der Hälfte der Anfragen einen höheren Average-Precision-Wert als das SDL-TM und auch der MAP-Wert des iMem-TMs liegt mit 75 % zu 45,83 % eindeutig über dem des SDL-TMs (Tabelle 37).

Ebenso konnte bei fast doppelt so vielen Anfragen ein Average-Precision-Wert von 1 durch das iMem-TM ermittelt werden als durch das SDL-TM (Abbildung 41). Der hohe Ausschlag beim Average-Precision-Wert von 0 beim SDL-TM kommt durch die 22 Anfragen zustande, bei denen eine leere Antwortmenge geliefert wurde.

	SDL-TM	**iMem-TM**
Höherer Average-Precision-Wert	0 Anfragen	52 Anfragen
Identischer Average-Precision-Wert	48 Anfragen	
Mean Average Precision (MAP)	45,83 %	75 %

Tabelle 37: Average-Precision- und Mean-Average-Precision-Werte (gerundet) für das iMem- und SDL-TM bei einem Schwellenwert von 70 %

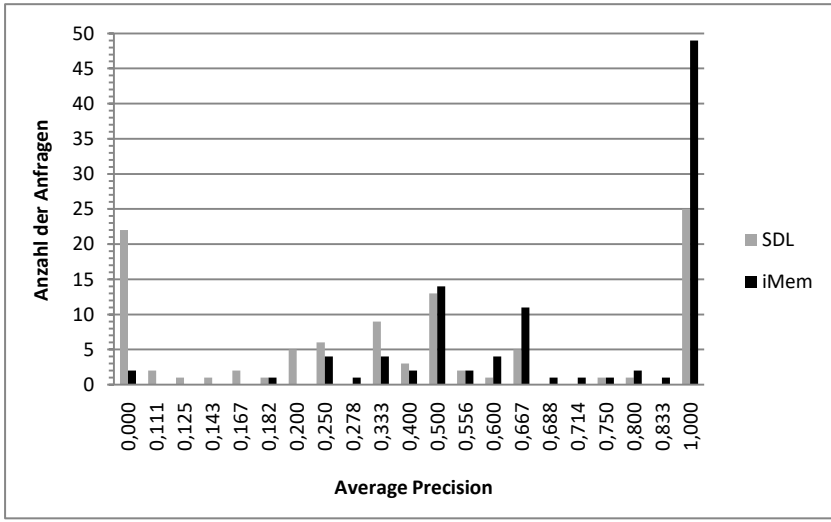

Abbildung 41: Anzahl an Anfragen, bei denen ein spezifischer Average-Precision-Wert (gerundet) des iMem- und SDL-TMs ermittelt wurde bei einem Schwellenwert von 70 %.

Auch in den Abbildungen 42 bis 44, in denen die Häufigkeit von P@1-, P@2- und P@3-Werten beider Systeme miteinander verglichen werden, wird nochmals deutlich, dass das iMem-TM stets häufiger einen P@k-Wert von 1 liefert als das SDL-TM. Entgegen der Untersuchung mit einem Schwellenwert von 30 % konnte zudem stets seltener eine leere Antwortmenge durch das iMem-TM vermerkt werden.

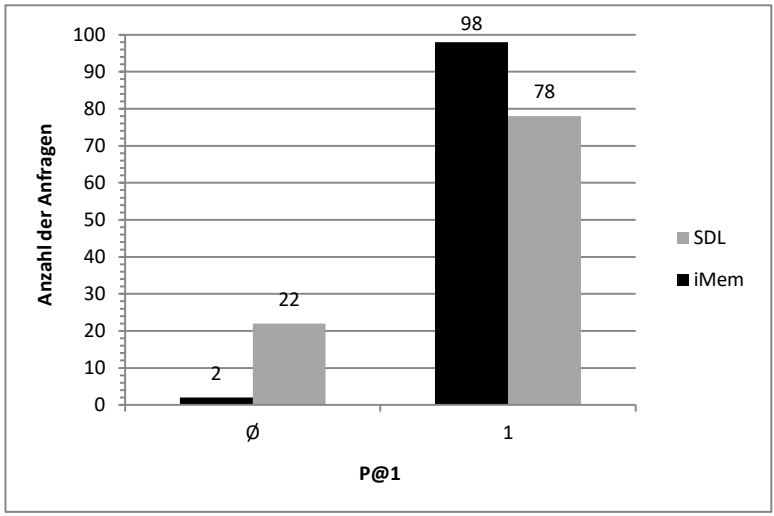

Abbildung 42: Anzahl an Anfragen, bei denen ein spezifischer P@1-Wert des iMem- und SDL-TMs ermittelt wurde bei einem Schwellenwert von 70 %.

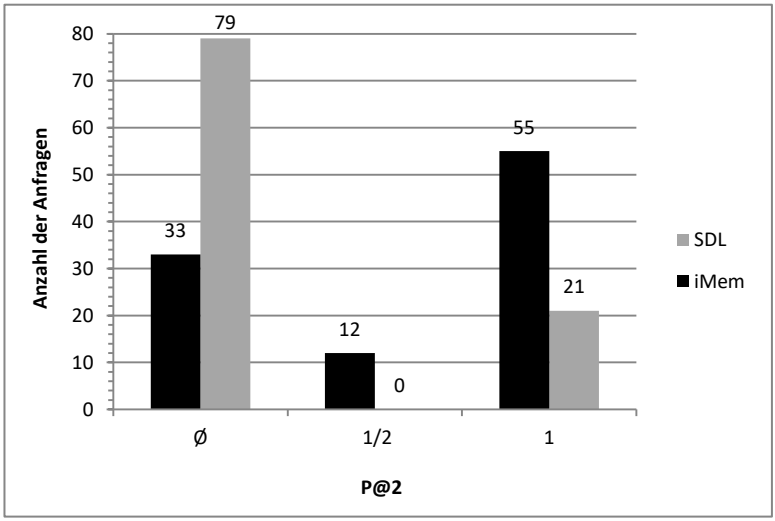

Abbildung 43: Anzahl an Anfragen, bei denen ein spezifischer P@2-Wert des iMem- und SDL-TMs ermittelt wurde bei einem Schwellenwert von 70 %.

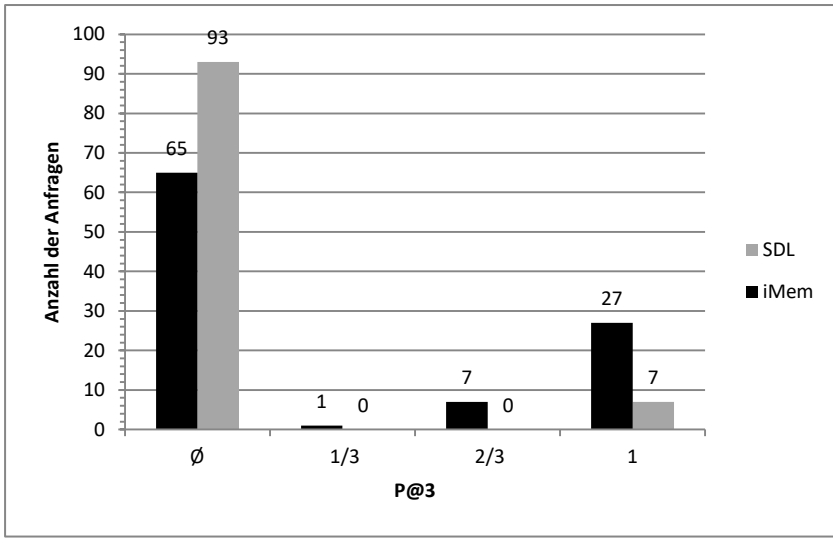

Abbildung 44: Anzahl an Anfragen, bei denen ein spezifischer P@3-Wert des iMem- und SDL-TMs ermittelt wurde bei einem Schwellenwert von 70 %.

Analog zur Evaluierung mit einem Schwellenwert von 30 % liegen die *R*-Precision-Werte des iMem-TMs bei einem Schwellenwert von 70 % über denen des SDL-TMs (Tabelle 38): Bei 52 Anfragen wird ein höherer *R*-Precision-Wert durch das iMem-TM gegenüber dem SDL-TM ausgegeben und die durchschnittliche *R*-Precision des iMem-TMs liegt ebenfalls mit 75,31 % deutlich über derjenigen des SDL-TMs mit nur 45,83 %. Bei knapp der Hälfte der Anfragen sind die *R*-Precision-Werte beider Systeme identisch, woraus im Umkehrschluss folgt, dass das SDL-TM kein Segment mit einem höheren *R*-Precision-Wert als das iMem-TM vorweisen kann.

	SDL-TM	**iMem-TM**
Höherer *R*-Precision-Wert	0 Anfragen	52 Anfragen
Identischer *R*-Precision-Wert	48 Anfragen	
Mean *R*-Precision	45,83 %	75,31 %

Tabelle 38: *R*-Precision- und Mean-*R*-Precision-Werte (gerundet) für das iMem- und SDL-TM bei einem Schwellenwert von 70 %

Das Precision-Histogramm (Abbildung 45) verdeutlicht diesen Sachverhalt: Es sind 52 Balken oberhalb der X-Achse sichtbar, was bedeutet, dass der

Algorithmus des iMem-TMs in diesen Fällen eine bessere Leistung erbringt als derjenige des SDL-TMs. Bei 11 dieser Anfragen wird zudem die höchste Differenz von $RP_{iMem/SDL}(i) = 1$ erzielt. Der Grund für diese hohe Differenz ist der Gleiche wie bei der Evaluierung für einen Schwellenwert von 30 %: Das SDL-TM gibt bei keiner dieser 11 Anfragen eine Übersetzungseinheit aus, während im Gegenzug alle relevanten Übersetzungseinheiten durch das iMem-TM gefunden werden. Die 48 balkenlosen Anfragen im Precision-Histogramm weisen darauf hin, dass in diesen Fällen die Algorithmen beider Systeme gleich effektiv sind.

Das SDL-TM gibt bei einem Schwellenwert von 70 % keine einzige nicht relevante Übersetzungseinheit aus. Der durchschnittliche Fallout-Wert des SDL-TMs liegt demnach bei 0 %, wobei jedoch die leeren Antwortmengen bei 22 Anfragen nicht zu vergessen sind. Dagegen konnte in 21 Fällen beim iMem-TM durchaus ein Fallout ermittelt werden, woraus sich ein geringer durchschnittlicher Fallout-Wert von 0,06 % ergibt. Dieser liegt jedoch deutlich unter dem Fallout-Wert, der bei der Auswertung für einen Schwellenwert von 30 % bestimmt wurde. Folglich kann auch bei einem Schwellenwert von 70 % das SDL-TM nicht relevante Übersetzungseinheiten besser unterdrücken als das iMem-TM. Der Fallout ist somit das einzige Maß, bei dem das iMem-TM stets schlechtere Werte ausgibt als das SDL-TM. Die Werte werden in Tabelle 39 sowie in Abbildung 46 nochmals vergleichend aufgeführt.

	SDL-TM	**iMem-TM**
Niedrigerer Fallout-Wert	21 Anfragen	0 Anfragen
Identischer Fallout-Wert	79 Anfragen	
Mean Fallout	0 %	0,06 %

Tabelle 39: Fallout- und Mean-Fallout-Werte (gerundet) für das iMem- und SDL-TM bei einem Schwellenwert von 70 %

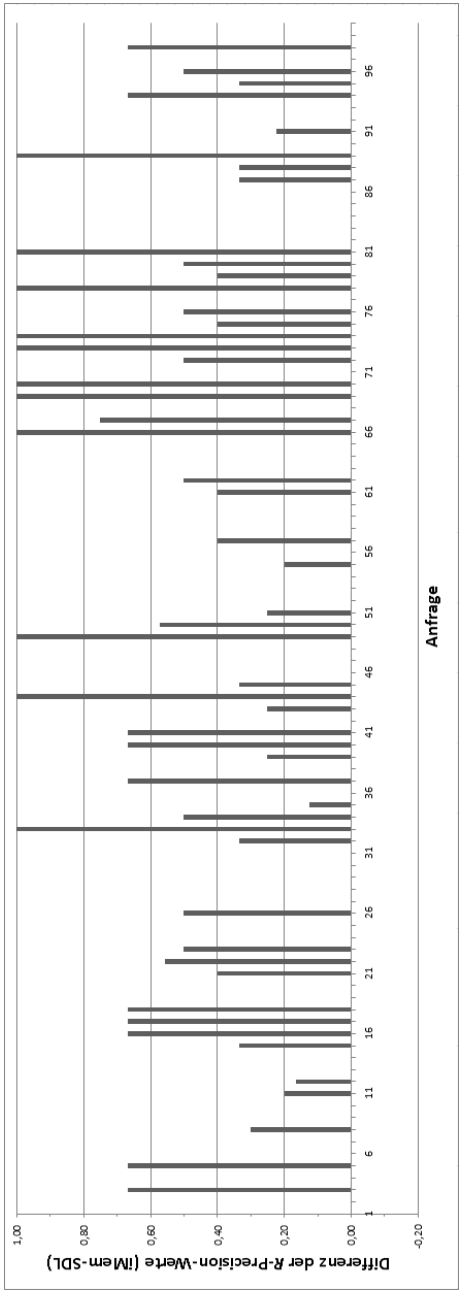

Abbildung 45: Precision-Histogramm: Differenzen der *R*-Precision-Werte (gerundet) zwischen dem iMem- und SDL-TM zu jeder Anfrage bei einem Schwellenwert von 70 %. Die Balken oberhalb der X-Achse zeigen eine bessere Leistung des iMem-Algorithmus.

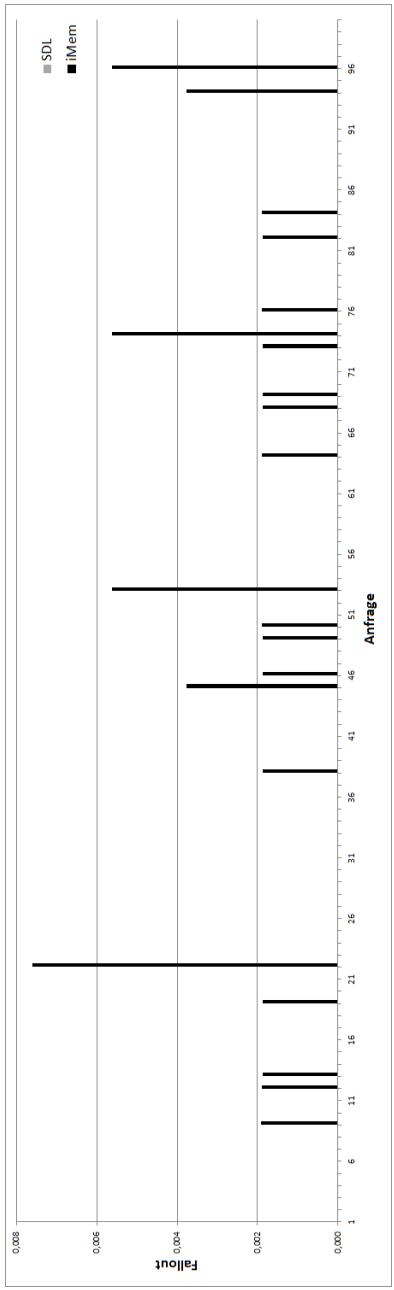

Abbildung 46: Fallout-Werte (gerundet) des iMem- und SDL-TMs pro Anfrage bei einem Schwellenwert von 70 %

6.5.2 Effektivitätsmessung: statistische Auswertung für spezifische Match-Wert-Bereiche und Identifikation linguistischer Phänomene

Neben der Berechnung und Auswertung der einzelnen Maße zur Messung der Effektivität beider Systeme wurde eine Statistik über das Vorkommnis relevanter und nicht relevanter Übersetzungseinheiten in spezifischen, selbst definierten Fuzzy-Match-Bereichen erstellt. Dazu wurde der Match-Wert jeder aufgefundenen relevanten und nicht relevanten Übersetzungseinheit vermerkt und dem entsprechenden Bereich zugeordnet.

Im weiteren Verlauf dieser Untersuchung wurden die Match-Wert-Bereiche von 70 %–74 %, 75 %–84 %, 85 %–94 % und 95 %–99 % genauer betrachtet und dahin gehend untersucht, in welchen Match-Wert-Bereichen welche linguistischen Unterschiede zwischen den AS_{neu} und den aufgefundenen relevanten AS_{TM} bzw. AS_{iMem} häufig auftreten. Im Zuge dessen wurden die Match-Werte der relevanten gespeicherten AS-Segmente, die durch das SDL- und das iMem-TM ermittelt wurden, gegenübergestellt.

6.5.2.1 Vorkommnis relevanter und nicht relevanter Übersetzungseinheiten in spezifischen Match-Wert-Bereichen

Bei dieser Untersuchung wird nochmals deutlich, dass das iMem-TM in den Match-Wert-Bereichen zwischen 70 % und 99 % mehr relevante Übersetzungseinheiten ausgibt als das SDL-TM (Tabelle 40). Während sich die Zahlen in den Bereichen zwischen 70 % und 94 % zwischen den beiden Systemen nur geringfügig unterscheiden, fällt vor allem der große Unterschied im Bereich von 95 %–99 % ins Auge (97 vs. 23 relevante Übersetzungseinheiten). Ebenso fällt auf, dass in den Match-Wert-Bereichen zwischen 30 % und 69 % weitaus weniger relevante Übersetzungseinheiten durch das iMem-TM gefunden wurden als durch das SDL-TM (10 vs. 103 Übersetzungseinheiten). Dabei ist anzumerken, dass die durch das SDL-TM im unteren Match-Wert-Bereich (unter 70 %) aufgefundenen Übersetzungseinheiten meistens im oberen Match-Wert-Bereich (über 70 %) des iMem-TMs wiederzufinden sind.

Letzterer Sachverhalt ist darin begründet, dass das iMem-Proximitätsmaß so konzipiert ist, dass in den meisten Fällen die Übersetzungseinheiten, die eine Ähnlichkeit mit dem AS_{neu} aufweisen, mittels des Korrekturfaktors x, der zur Berechnung des L-Wertes eingesetzt wird (siehe Kapitel 4.3.1), mindestens einen Match-Wert von 70 % erreichen. Dadurch wird gewährleistet, dass dem Übersetzer in einem realen Übersetzungsprozess, bei dem der Schwellenwert meist auf 70 % eingestellt wird, ähnliche Übersetzungseinheiten dargeboten werden.

Match-Wert-Bereiche	SDL-TM Anzahl aufgefundener relevanter ÜEs	iMem-TM Anzahl aufgefundener relevanter ÜEs
95 %–99 %	23	97
85 %–94 %	52	53
75 %–84 %	30	37
70 %–74 %	10	17
60 %–69 %	24	10
50 %–59 %	20	0
40 %–49 %	30	0
30 %–39 %	29	0
Gesamt (30 %–99 %)	218	214
Gesamt (30 %–69 %)	103	10
Gesamt (70 %–99 %)	115	204

Tabelle 40: Anzahl der durch das iMem- und SDL-TM aufgefundenen relevanten Übersetzungseinheiten, eingeteilt in verschiedene Match-Wert-Bereiche

Die Statistik der aufgefundenen nicht relevanten Übersetzungseinheiten (Tabelle 41) verhält sich dazu konträr: Während beim SDL-TM in den Match-Wert-Bereichen zwischen 70 %–99 % keinerlei Fallout verzeichnet werden konnte, werden durch das iMem-TM immerhin noch 32 nicht relevante Übersetzungseinheiten in den Match-Wert-Bereichen zwischen 70 % und 84 % aufgefunden.

Erst ab einem Match-Wert unter 70 % treten mit jedem tiefer liegenden Match-Wert-Bereich stetig mehr nicht relevante Übersetzungseinheiten durch das SDL-TM auf; der höchste Wert (150 nicht relevante Übersetzungseinheiten) wird im niedrigsten Match-Wert-Bereich registriert. Die Werte des iMem-TMs nehmen einen anderen Verlauf: Während die Werte in den Match-Wert-Bereichen von 70 %–74 % und 40 %–49 % sowie von 75 %–84 % und 30 %–39 % relativ gering sind und sich auch nur geringfügig voneinander unterscheiden, werden in den mittleren Match-Wert-Bereichen von 60 %–69 % und 50 %–59 % die höchsten Werte an nicht relevanten Übersetzungseinheiten erfasst.

Match-Wert-Bereiche	SDL-TM Anzahl aufgefundener nicht relevanter ÜEs	iMem-TM Anzahl aufgefundener nicht relevanter ÜEs
95 %–99 %	0	0
85 %–94 %	0	0
75 %–84 %	0	10
70 %–74 %	0	22
60 %–69 %	1	105
50 %–59 %	9	92
40 %–49 %	40	27
30 %–39 %	150	15
Gesamt (30 %–99 %)	200	271
Gesamt (30 %–69 %)	200	239
Gesamt (70 %–99 %)	0	32

Tabelle 41: Anzahl der durch das iMem- und SDL-TM aufgefundenen nicht relevanten Übersetzungseinheiten, eingeteilt in verschiedene Match-Wert-Bereiche

6.5.2.2 Identifikation linguistischer Phänomene in spezifischen Match-Wert-Bereichen

Beim Vergleich der AS_{neu} mit den aufgefundenen relevanten gespeicherten AS-Segmenten konnte festgestellt werden, dass bestimmte linguistische Phänomene spezifisch für bestimmte Match-Wert-Bereiche sind.

In den nachfolgenden Tabellen (Tabelle 42 bis 46) sind die AS_{neu} mit den sowohl durch das SDL-TM als auch durch das iMem-TM aufgefundenen relevanten gespeicherten Übersetzungseinheiten mit ihren jeweiligen Match-Werten für die Match-Wert-Bereiche von 95 %–99 %, 85 %–94 %, 75 %–84 % sowie 70 %–74 % aufgeführt. Die Zuordnung der Match-Werte zu den jeweiligen Match-Wert-Bereichen erfolgt dabei gemäß den Match-Werten des iMem-TMs. Es wurden lediglich Tabellen für diese vier Match-Wert-Bereiche generiert, da einerseits das iMem-TM in den unteren Match-Wert-Bereichen kaum relevante Übersetzungseinheiten ausgibt (Tabelle 40) und demnach keine große Vergleichsmöglichkeit zwischen den beiden Systemen

mehr besteht und da andererseits in der Übersetzungspraxis selten ein Schwellenwert unter 70 % eingestellt wird.

In den Tabellen wird unterschieden, ob alle oder nicht alle Basiswörter des AS_{neu} im AS_{TM} bzw. AS_{iMem} enthalten sind und ob die Segmentpaare über einen LCS oder mehr als einen LCS verfügen. Dazu werden Beispiele[78] für linguistische Phänomene aufgeführt, die häufig in den jeweiligen Match-Wert-Bereichen identifiziert werden konnten. Bei der Identifikation der linguistischen Phänomene wird stets untersucht, welche Änderungen vorgenommen werden müssten, um aus dem AS_{TM} bzw. AS_{iMem} das AS_{neu} herzustellen.

Zu den linguistischen Phänomenen werden Beispiele von AS_{neu}-AS_{TM}-Segmentpaaren bzw. AS_{neu}-AS_{iMem}-Segmentpaaren gegeben, die in den spezifischen Match-Wert-Bereichen aufgefunden wurden. Die Segmentpaare sind gemäß der Match-Werte des iMem-TMs absteigend sortiert. Dabei werden nur Segmentpaare aufgeführt, bei denen ein markanter Unterschied zwischen den Match-Werten beider Systeme besteht. Es werden auch nicht alle bei der Evaluierung aufgefundenen Segmentpaare aufgelistet, da sich die linguistischen Phänomene teilweise wiederholen. Die Unterstreichungen in den Segmentpaaren weisen auf die linguistischen Unterschiede zwischen den Segmenten hin[79]. Die Tabellen dienen folglich neben der Auflistung der linguistischen Phänomene in den spezifischen Match-Wert-Bereichen dazu, aufzuzeigen, wie sehr sich die Match-Werte der beiden Systeme unterscheiden.

Im obersten Match-Wert-Bereich (95 %–99 %; Tabelle 42) treten nur geringfügige Unterschiede zwischen den zu vergleichenden Segmenten auf. Häufig handelt es sich dabei um ausschließlich morphologische oder lexikalische Unterschiede und somit um Unterschiede einfacher Art. Ebenso finden sich in diesem Match-Wert-Bereich Segmentpaare wieder, bei denen die Ausnahmeregelung des iMem-Proximitätsmaßes greift: Der Match-Wert wird pauschal auf 99 % gesetzt, wenn das AS_{neu} komplett an einem Stück im AS_{iMem} auftritt.

---

[78] Bei der Auflistung der Beispiele wird kein Anspruch auf Vollständigkeit erhoben.

[79] Im Falle mehrerer, verschiedenartiger Unterstreichungen innerhalb eines Segmentpaares beziehen sich jeweils gleiche Unterstreichungen auf die zu vergleichenden Unterschiede. Existiert lediglich in einem Segment (also entweder nur im AS_{TM}/AS_{iMem} oder im AS_{neu}) eine Unterstreichung, weist dies auf eine Löschung bzw. Hinzufügung hin.

Linguistische Phänomene im Match-Wert-Bereich von 95 %–99 %					
Form der Basiswörter	**Anzahl LCS**	**Linguistische Phänomene (Beispiele)**	**Segmentpaare (Beispiele)**	**Match-Wert iMem-TM**	**Match-Wert SDL-TM**
alle Basiswörter des AS_{neu} sind im AS_{TM} bzw. AS_{iMem} enthalten	1	• Änderung des Numerus • Orthografische Varianten • Ersetzung eines Substantivs durch einen substantivierten Infinitiv • Änderung von Basis zu Diminutivum • AS_{neu} komplett im AS_{TM}/AS_{iMem} enthalten • Ersetzung oder Löschung eines Interpunktions-, Nummerierungs- oder Aufzählungszeichens	AS_{TM}/AS_{iMem}: Wenn der Akku leer ist, können Sie das Gerät auch direkt über das Spezialkabel vom Netz betreiben. AS_{neu}: Wenn die Akkus leer sind, können Sie das Gerät auch direkt über das Spezialkabel vom Netz betreiben.	99 %	91 %
			AS_{TM}/AS_{iMem}: Einige praktische Tipps AS_{neu}: Einige praktische Tips	99 %	90 %
			AS_{TM}/AS_{iMem}: Nach der Epilation empfehlen wir die Verwendung einer Feuchtigkeitscreme. AS_{neu}: Nach dem Epilieren empfehlen wir die Verwendung einer Feuchtigkeitscreme.	99 %	85 %
			AS_{TM}/AS_{iMem}: Klingenblock gründlich mit der Bürste reinigen. AS_{neu}: Klingenblock gründlich mit dem Bürstchen reinigen.	99 %	78 %
			AS_{TM}/AS_{iMem}: Halten Sie den Distanzkamm flach auf dem Haar, parallel zur Kopfhaut und führen Sie den Haarschneider langsam gegen die Haarwuchsrichtung. AS_{neu}: Führen Sie den Haarschneider langsam gegen die Haarwuchsrichtung.	99 %	37 %
			AS_{TM}/AS_{iMem}: Reinigen mit Wasser AS_{neu}: Reinigung mit Wasser	96 %	90 %
alle Basiswörter des AS_{neu} sind im AS_{TM} bzw. AS_{iMem} enthalten	> 1	• Vertauschungen von Phrasen • Vertauschung von Haupt- und Nebensatz • Ersetzung einer mehrteiligen Verbphrase durch eine einteilige Verbphrase • Löschung von Phrasen oder Teilsätzen an einem Stück	AS_{TM}/AS_{iMem}: Vergewissern Sie sich, dass das Batteriefach trocken und sauber ist, bevor Sie die Batteriefach-Abdeckung wieder schließen. AS_{neu}: Vergewissern Sie sich, dass die Batteriefach-Abdeckung trocken und sauber ist, bevor Sie das Batteriefach wieder schließen.	99 %	88 %
			AS_{TM}/AS_{iMem}: Bitte lesen Sie vor Gebrauch des Gerätes die Gebrauchsanweisung sorgfältig durch und bewahren Sie sie auf. AS_{neu}: Bitte lesen Sie die Gebrauchsanweisung vor Gebrauch des Gerätes sorgfältig durch und bewahren Sie sie auf.	99 %	75 %

Linguistische Phänomene im Match-Wert-Bereich von 95 %–99 %					
Form der Basiswörter	**Anzahl LCS**	**Linguistische Phänomene (Beispiele)**	**Segmentpaare (Beispiele)**	**Match-Wert iMem-TM**	**Match-Wert SDL-TM**
		• Hinzufügung kurzer Phrasen • Aktiv-Passiv-Konverse • Änderung eines Imperativsatzes in einen imperativischen Infinitiv	AS_{TM}/AS_{iMem}: Halten Sie das Gerät senkrecht zur Haut und führen Sie es ohne Druck mit der Schalterseite gegen den Haarwuchs, um eine optimale Epilation zu gewährleisten. AS_{neu}: Um eine optimale Epilation zu gewährleisten, halten Sie das Gerät senkrecht zur Haut und führen Sie es ohne Druck mit der Schalterseite gegen den Haarwuchs.	99 %	81 %
			AS_{TM}/AS_{iMem}: Nach der Epilation empfehlen wir die Verwendung einer Feuchtigkeitscreme. AS_{neu}: Wir empfehlen die Verwendung einer Feuchtigkeitscreme nach der Epilation.	99 %	32 %
			AS_{TM}/AS_{iMem}: Sie sollten ihn abnehmen und säubern, wenn sich der Distanzkamm mit Haaren zusetzt. AS_{neu}: Wenn sich der Distanzkamm mit Haaren zusetzt, sollten Sie ihn abnehmen und säubern.	99 %	30 %
			AS_{TM}/AS_{iMem}: Halten Sie die Haut gestrafft (B) und führen Sie das Gerät langsam gegen die Haarwuchsrichtung. AS_{neu}: Straffen Sie die Haut (B) und führen Sie das Gerät langsam gegen die Haarwuchsrichtung.	98 %	91 %
			AS_{TM}/AS_{iMem}: Die Entsorgung kann über den Braun Kundendienst oder lokal verfügbare Rückgabe- und Sammelstellen erfolgen. AS_{neu}: Die Entsorgung kann über den Braun Kundendienst erfolgen.	98 %	61 %
			AS_{TM}/AS_{iMem}: Um die Haut zu entspannen, empfehlen wir die Verwendung einer Feuchtigkeitscreme nach der Epilation. AS_{neu}: Nach dem Epilieren empfehlen wir die Verwendung einer Feuchtigkeitscreme.	97 %	51 %
			AS_{TM}/AS_{iMem}: Scherfolie und Klingenblock sind Präzisionsteile, die mit der Zeit verschleißen. AS_{neu}: Die Scherfolie und der Klingenblock sind Präzisionsteile, die mit der Zeit verschleißen.	96 %	87 %

Linguistische Phänomene im Match-Wert-Bereich von 95 %–99 %					
Form der Basiswörter	**Anzahl LCS**	**Linguistische Phänomene (Beispiele)**	**Segmentpaare (Beispiele)**	**Match-Wert iMem-TM**	**Match-Wert SDL-TM**
			AS_{TM}/AS_{iMem}: Das bewährte Silk·épil Epiliersystem entfernt das Haar an der Wurzel und die Haut bleibt wochenlang glatt. AS_{neu}: Mit dem bewährten Silk·épil Epiliersystem wird das Haar an der Wurzel entfernt und die Haut bleibt wochenlang glatt.	96 %	80 %
			AS_{TM}/AS_{iMem}: Prüfen Sie, ob die auf dem Transformator angegebene Spannung mit Ihrer Netzspannung übereinstimmt. AS_{neu}: Prüfen Sie, ob die Spannungsangabe mit Ihrer Netzspannung übereinstimmt.	96 %	71 %
			AS_{TM}/AS_{iMem}: Wickeln Sie das Netzkabel nicht um das Gerät. AS_{neu}: Netzkabel nicht um das Gerät wickeln.	96 %	69 %
nicht alle Basiswörter des AS_{neu} sind im AS_{TM} bzw. AS_{iMem} enthalten	> 1	• Wortersetzung • Kompositazerlegung • Kompositabildung	AS_{TM}/AS_{iMem}: Scherkopf und Klingenblock separat unter fließendem Wasser reinigen. AS_{neu}: Scherkopf und Klingenblock separat unter fließendes Wasser halten.	97 %	88 %
			AS_{TM}/AS_{iMem}: Mit der Bürste den Klingenblock und die Scherkopf-Innenseite reinigen. AS_{neu}: Mit der Bürste den Klingenblock und die Innenseite des Scherkopfes reinigen.	97 %	79 %
			AS_{TM}/AS_{iMem}: Mit der Bürste die Innenseite des Scherkopfes reinigen. AS_{neu}: Mit der Bürste die Scherkopf-Innenseite reinigen.	96 %	70 %

Tabelle 42: Linguistische Unterschiede zwischen ausgewählten AS_{neu} und aufgefundenen relevanten AS_{iMem} bzw. AS_{TM} im Match-Wert-Bereich von 95 %–99 % mit den dazugehörigen Match-Werten des iMem- und SDL-TMs

Linguistische Phänomene im Match-Wert-Bereich von 85 %–94 %					
Form der Basiswörter	**Anzahl LCS**	**Linguistische Phänomene (Beispiele)**	**Segmentpaare (Beispiele)**	**Match-Wert iMem-TM**	**Match-Wert SDL-TM**
alle Basiswörter des AS_{neu} sind im AS_{TM} bzw. AS_{iMem} enthalten	> 1	• Ersetzung eines Funktionsverbgefüges durch ein Verb • Löschung von Phrasen an mehreren Stellen	AS_{TM}/AS_{iMem}: Wir empfehlen die Verwendung von Rasierschaum oder -gel, wenn Sie es beim Duschen verwenden. AS_{neu}: Wir empfehlen die Verwendung von Rasierschaum oder -gel, wenn Sie duschen.	94 %	82 %
			AS_{TM}/AS_{iMem}: Klingenblock (2) gründlich mit dem Bürstchen (7) reinigen (i). AS_{neu}: Klingenblock gründlich mit dem Bürstchen reinigen.	91 %	69 %
nicht alle Basiswörter des AS_{neu} sind im AS_{TM} bzw. AS_{iMem} enthalten	1	• Hinzufügung einer kurzen Phrase • Hinzufügung eines Interpunktions-, Nummerierungs- oder Aufzählungszeichens	AS_{TM}/AS_{iMem}: Im Handel oder beim Braun Kundendienst erhältlich. AS_{neu}: Es ist im Handel oder beim Braun Kundendienst erhältlich.	94 %	81 %
			AS_{TM}/AS_{iMem}: Rasierer ausschalten. AS_{neu}: • Rasierer ausschalten.	93 %	96 %
nicht alle Basiswörter des AS_{neu} sind im AS_{TM} bzw. AS_{iMem} enthalten	> 1	• Ersetzung einer (Adjektiv-)Phrase durch einen Relativsatz oder eine andere Phrase mit weiteren Unterschieden, z. B. Löschung von Wörtern oder Teilsätzen, Änderung des Numerus	AS_{TM}/AS_{iMem}: Da die Haut in diesem Bereich nach der Epilation besonders empfindlich ist, sollten Sie keine hautreizenden Substanzen, wie z.B. alkoholhaltige Deodorants, verwenden. AS_{neu}: Da die Haut in diesem Bereich nach der Epilation besonders empfindlich ist, sollten Sie keine Substanzen, die Hautreizungen hervorrufen können, wie z.B. alkoholhaltige Deodorants, verwenden.	94 %	85 %
		• Änderung eines Imperativsatzes in einen Indikativsatz	AS_{TM}/AS_{iMem}: Wickeln Sie das Netzkabel nicht um das Gerät. AS_{neu}: Das Netzkabel darf nicht um das Gerät gewickelt werden.	94 %	48 %
		• Wortersetzung in einem sehr kurzen Satz	AS_{TM}/AS_{iMem}: Reinigen mit Wasser AS_{neu}: Reinigen unter Wasser	93 %	77 %
		• Verwendung synonymer Benennungen mit weiteren Unterschieden, z. B. Aktiv-Passiv-Konverse oder Vertauschung von Haupt- und Nebensatz	AS_{TM}/AS_{iMem}: Epilation im Achselbereich und in der Bikinizone AS_{neu}: Epilation von Achselbereich und Bikinizone	90 %	64 %
		• Hinzufügung langer Teilsätze an einem Stück	AS_{TM}/AS_{iMem}: Dennoch sollten Sie im Interesse der Rohstoff-Rückgewinnung das Gerät am Ende seiner Lebensdauer nicht mit dem Hausmüll entsorgen. AS_{neu}: Dieses Gerät darf am Ende seiner Lebensdauer nicht mit dem Hausmüll entsorgt werden.	90 %	48 %

Linguistische Phänomene im Match-Wert-Bereich von 85 %–94 %					
Form der Basiswörter	**Anzahl LCS**	**Linguistische Phänomene (Beispiele)**	**Segmentpaare (Beispiele)**	**Match-Wert iMem-TM**	**Match-Wert SDL-TM**
			AS_{TM}/AS_{iMem}: Im Handel oder bei Braun Kundendienststellen ist das Braun Clean&Charge erhältlich. AS_{neu}: Es ist im Handel oder beim Braun Kundendienst erhältlich.	90 %	38 %
			AS_{TM}/AS_{iMem}: Gerät ist von der Anschlussleitung zu trennen, bevor Sie es unter Wasser reinigen. AS_{neu}: Das handgehaltene Teil ist von der Anschlussleitung zu trennen, bevor es im Wasser gereinigt wird.	87 %	61 %
			AS_{TM}/AS_{iMem}: Bei allen Formen der Epilation an der Wurzel kann es zu kleinen Hautverletzungen und zum Einwachsen von Haaren kommen, abhängig auch von Ihrem jeweiligen Haut- und Haartyp. AS_{neu}: Bei allen Formen der Epilation, bei denen die Haare an den Wurzeln entfernt werden, kann es zu kleinen Hautverletzungen und zum Einwachsen von Haaren kommen.	87 %	41 %
			AS_{TM}/AS_{iMem}: Das Gerät nicht längere Zeit Temperaturen über 50 °C aussetzen. AS_{neu}: Das Gerät nicht längere Zeit direktem Sonnenlicht aussetzen.	86 %	76 %
			AS_{TM}/AS_{iMem}: Um eine optimale Epilation zu gewährleisten, halten Sie das Gerät senkrecht zur Haut. AS_{neu}: Um eine optimale Epilation zu gewährleisten, halten Sie das Gerät senkrecht zur Haut und führen Sie es ohne Druck mit der Schalterseite gegen den Haarwuchs.	85 %	55 %
			AS_{TM}/AS_{iMem}: Dieses Akku-/Netzgerät lässt sich jedoch auch direkt über das Spezialkabel vom Netz betreiben, falls die Akkus leer sind. AS_{neu}: Wenn die Akkus leer sind, können Sie das Gerät auch direkt über das Spezialkabel vom Netz betreiben.	85 %	30 %

Tabelle 43: Linguistische Unterschiede zwischen ausgewählten AS_{neu} und aufgefundenen relevanten AS_{iMem} bzw. AS_{TM} im Match-Wert-Bereich von 85 %–94 % mit den dazugehörigen Match-Werten des iMem- und SDL-TMs

Bereits im nächsttieferen Match-Wert-Bereich (85 %–94 %; Tabelle 43) fällt auf, dass die linguistischen Phänomene zwischen dem AS_{neu} und den gespeicherten AS-Segmenten immer komplexer werden (weniger morphologische und lexikalische; mehr syntaktische oder gemischte Unterschiede). Dennoch wiederholen sich einige linguistische Phänomene, die im oberen Match-Wert-Bereich identifiziert werden konnten, auch in den unteren Match-Wert-Bereichen. Die niedrigeren Match-Werte kommen dadurch zustande, dass z. B. mehr LCS vorhanden sind oder mehr Wörter neu übersetzt werden müssen.

Besonders gut ist an den in Tabelle 44 aufgeführten Segmentpaaren zu erkennen, wie stark sich beim SDL-TM eine geringe Veränderung in der Zeichenabfolge auf den Match-Wert auswirkt (69 % vs. 48 %), obwohl es zwischen den beiden Segmentpaaren keinen Bedeutungsunterschied gibt. Die Match-Werte des iMem-TMs für diese beiden Segmentpaare liegen hingegen sehr nah beieinander (96 % vs. 94 %). Zudem divergieren die Match-Werte nicht nur stark innerhalb des SDL-TMs, sondern ebenso zwischen den beiden Systemen.

Segmentpaare	**Match-Wert iMem-TM**	**Match-Wert SDL-TM**
AS_{TM}/AS_{iMem}: Wickeln Sie das Netzkabel nicht um das Gerät. AS_{neu}: Netzkabel nicht um das Gerät wickeln.	96 %	69 %
AS_{TM}/AS_{iMem}: Wickeln Sie das Netzkabel nicht um das Gerät. AS_{neu}: Das Netzkabel darf nicht um das Gerät gewickelt werden.	94 %	48 %

Tabelle 44: Vergleich der Match-Werte des iMem- und SDL-TMs für ausgewählte Segmentpaare: Ein geringer Unterschied in der Zeichenfolge hat große Auswirkungen auf den Match-Wert des SDL-TMs.

Im Match-Wert-Bereich von 75 %–84 % wiederholen sich die zuvor genannten linguistischen Phänomene, wobei auffällt, dass in jedem Segmentpaar stets mehr als ein Unterschied besteht. Das zeigt wiederum, dass mit sinkendem Match-Wert die Unterschiede immer komplexer werden. Besonders häufig sind Hinzufügungen und Löschungen von Teilsätzen oder Phrasen, ggf. mit Kongruenz anderer Wörter, Ersetzungen von Phrasen, Vertauschungen der Positionen von Phrasen sowie Paraphrasen zu beobachten. Diese Änderungen haben stets zur Folge, dass nicht alle Basiswörter des AS_{neu} mit denen des AS_{TM} bzw. AS_{iMem} identisch sind und zudem mehr als ein LCS vorhanden ist. Auffällig ist dabei, dass trotz der festgestellten Bedeutungsgleichheit bzw. -ähnlichkeit der ausgewählten Segmentpaare keiner der Match-Werte des SDL-TMs über 70 % liegt; die Match-Werte des iMem-TMs erreichen hingegen alle mindestens einen Wert von 75 % (Tabelle 45).

Segmentpaare (75 %–84 %)		**Match-Wert iMem-TM**	**Match-Wert SDL-TM**
AS_{TM}/AS_{iMem}:	Das bewegliche Schersystem sorgt dabei automatisch für eine optimale Anpassung der Doppel-Scherfolie und des Integral-Schneiders an die Gesichtsform.	84 %	40 %
AS_{neu}:	• Der bewegliche Schwingkopf und die flexiblen Scherfolien sorgen automatisch für eine optimale Anpassung an die Gesichtsform.		
AS_{TM}/AS_{iMem}:	Schäden durch unsachgemäßen Gebrauch (Knickstellen an der Scherfolie, Bruch), normaler Verschleiß (z.B. Schersystem) sowie Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen.	83 %	67 %
AS_{neu}:	Schäden, die auf unsachgemäßen Gebrauch zurückzuführen sind, normaler Verschleiß und Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen.		
AS_{TM}/AS_{iMem}:	Halten Sie die Haut gestrafft (B) und führen Sie das Gerät langsam gegen die Haarwuchsrichtung.	83 %	43 %
AS_{neu}:	Führen Sie den Haarschneider langsam gegen die Haarwuchsrichtung.		
AS_{TM}/AS_{iMem}:	Dann reicht die Ladung noch für ca. 2 bis 3 Rasuren.	82 %	58 %
AS_{neu}:	Die verbleibende Ladung reicht noch für 2–3 Rasuren.		
AS_{TM}/AS_{iMem}:	Durch regelmäßiges Reinigen verbessern Sie die Rasierleistung Ihres Rasierers.	81 %	60 %
AS_{neu}:	Durch regelmäßiges Reinigen erhalten Sie eine optimale Rasierleistung.		
AS_{TM}/AS_{iMem}:	Setzen Sie das Rasiersystem (3) auf die gestraffte Haut und rasieren Sie sanft gegen die Haarwuchsrichtung.	81 %	46 %
AS_{neu}:	Straffen Sie die Haut (B) und führen Sie das Gerät langsam gegen die Haarwuchsrichtung.		
AS_{TM}/AS_{iMem}:	Im Garantiefall senden Sie bitte das vollständige Gerät mit der ausgefüllten Garantiekarte einem unserer autorisierten Servicehändler oder an eine Braun Kundendienststelle.	80 %	52 %
AS_{neu}:	Im Garantiefall senden Sie das Gerät mit Kaufbeleg bitte an einen autorisierten Braun Kundendienstpartner.		
AS_{TM}/AS_{iMem}:	Lesen Sie bitte vor der ersten Anwendung die Gebrauchsanweisung vollständig und sorgfältig durch.	79 %	38 %
AS_{neu}:	Bitte lesen Sie die Gebrauchsanweisung vor Gebrauch des Gerätes sorgfältig durch und bewahren Sie sie auf.		

Segmentpaare (75 %–84 %)		**Match-Wert iMem-TM**	**Match-Wert SDL-TM**
AS_{TM}/AS_{iMem}:	Um ein optimales Ergebnis zu erzielen, können Sie die Haut mit einer Hand glatt ziehen.	77 %	53 %
AS_{neu}:	Beste Ergebnisse erzielen Sie, wenn Sie die Haut mit der anderen Hand straff ziehen.		
AS_{TM}/AS_{iMem}:	Die ideale Umgebungstemperatur für das Laden liegt zwischen 15 °C und 35 °C.	76 %	45 %
AS_{neu}:	Günstige Umgebungstemperatur beim Laden: 15 °C bis 35 °C.		
AS_{TM}/AS_{iMem}:	Der bewegliche Scherfolienrahmen passt sich automatisch der Gesichtsform an und sorgt für eine gründliche und sanfte Rasur.	76 %	30 %
AS_{neu}:	• Der bewegliche Schwingkopf und die flexiblen Scherfolien sorgen automatisch für eine optimale Anpassung an die Gesichtsform.		

Tabelle 45: Linguistische Unterschiede zwischen ausgewählten AS_{neu} und aufgefundenen relevanten AS_{iMem} bzw. AS_{TM} im Match-Wert-Bereich von 75 %–84 % mit den dazugehörigen Match-Werten des iMem- und SDL-TMs

Segmentpaare (70 %–74 %)		**Match-Wert iMem-TM**	**Match-Wert SDL-TM**
AS_{TM}/AS_{iMem}:	Aus hygienischen Gründen möchten wir Sie bitten, Ihren BodycruZer nicht mit anderen Personen zu teilen.	74 %	44 %
AS_{neu}:	Das Gerät ist aus hygienischen Gründen nicht zum gemeinsamen Gebrauch mit anderen Personen gedacht.		
AS_{TM}/AS_{iMem}:	Sollte es jedoch wegen der leeren Akku-Einheit beim Einschalten nicht laufen, reicht 1 Minute Ladezeit (Schalterstellung «0») für den Gebrauch direkt am Stromnetz.	74 %	35 %
AS_{neu}:	(Sollte der Bartschneider nach dem Einschalten nicht sofort laufen, ca. 1 Minute bei Schalterstellung «off» laden.)		
AS_{TM}/AS_{iMem}:	Da die Haut in diesem Bereich nach der Epilation besonders empfindlich ist, sollten Sie keine hautreizenden Substanzen, wie z.B. alkoholhaltige Deodorants, verwenden.	71 %	45 %
AS_{neu}:	Vermeiden Sie jedoch unmittelbar nach der Haarentfernung die Verwendung von Substanzen, die Hautreizungen hervorrufen können, wie z.B. alkoholhaltige Deodorants.		

Tabelle 46: Linguistische Unterschiede zwischen ausgewählten AS_{neu} und aufgefundenen relevanten AS_{iMem} bzw. AS_{TM} im Match-Wert-Bereich von 70 %–74 % mit den dazugehörigen Match-Werten des iMem- und SDL-TMs

Auch im Match-Wert-Bereich von 70 %–74 % treten dieselben linguistischen Phänomene auf, wobei sich die Anzahl an Paraphrasen jedoch häuft. Die Match-Werte des SDL-TMs liegen meist unter 50 %, während diejenigen des iMem-TMs bei einem Schwellenwert von 70 % immerhin noch in der Trefferanzeige dargeboten würden (Tabelle 46).

Von 183 Segmentpaaren, die sowohl beim SDL-TM als auch beim iMem-TM aufgefunden werden konnten, lieferte das iMem-TM in 173 Fällen (94,54 %) höhere Match-Werte als das SDL-TM. Bei nur drei Segmentpaaren wurde ein identischer Match-Wert verzeichnet.

	iMem-TM	**SDL-TM**
Höherer Match-Wert	173	7
Identischer Match-Wert	3	

Tabelle 47: Anzahl an aufgefundenen Übersetzungseinheiten, die im Vergleich zum jeweils anderen TM einen höheren oder identischen Match-Wert liefern.

Des Weiteren konnte bei 85 von 100 Anfragen festgestellt werden, dass bei beiden Systemen das gleiche AS_{TM} bzw. AS_{iMem} an der ersten Position in der Trefferanzeige stand.

Zusammenfassend kann für die statistische Auswertung für spezifische Match-Wert-Bereiche und die Identifikation linguistischer Phänomene festgehalten werden, dass die Anzahl und Komplexität linguistischer Unterschiede in niedrigeren Match-Wert-Bereichen stetig zunimmt. Unterschiede in kurzen Segmenten wirken sich dabei in beiden Systemen stärker auf den Match-Wert aus als in langen Segmenten. Die oben aufgeführten Tabellen zeigen, dass das iMem-TM in den meisten Fällen einen (erheblich) höheren Match-Wert liefert als das SDL-TM. Ob die Match-Werte des iMem-TMs aber tatsächlich dem menschlichen Ähnlichkeitsempfinden näher kommen, wurde in einer Online-Umfrage untersucht, deren Auswertung im folgenden Abschnitt aufgeführt ist.

6.5.3 Effektivitätsmessung: Vergleich mit dem menschlichen Ähnlichkeitsempfinden

6.5.3.1 Aufbau des Fragebogens

Um zu testen, ob die durch das iMem-TM errechneten Match-Werte mit dem menschlichen Ähnlichkeitsempfinden korrespondieren, wurde ein Online-Fragebogen aufgesetzt, in dem 50 ausgewählte Segmente des AT_{neu} A jeweils einem AS-Segment einer relevanten Übersetzungseinheit des Korpus A gegenübergestellt wurden. Voraussetzung dafür war, dass die Übersetzungseinheiten

sowohl durch das iMem-TM als auch durch das SDL-TM aufgefunden wurden. Die Match-Werte beider TMs wurden für jedes ausgewählte AS_{neu}-AS_{TM}- bzw. AS_{neu}-AS_{iMem}-Paar dokumentiert.

Bei der Auswahl der AS_{neu} und AS_{TM} bzw. AS_{iMem} wurde darauf geachtet, dass so viele verschiedene linguistische Phänomene wie möglich abgedeckt wurden. So gibt es AS_{neu}-AS_{TM}-/AS_{neu}-AS_{iMem}-Paare, zwischen denen ausschließlich morphologische Unterschiede bestehen (z. B. Unterschiede im Numerus). Andere Segmentpaare beinhalten hingegen lexikalische Unterschiede (z. B. Wortersetzungen), syntaktische Unterschiede (z. B. Vertauschungen von Phrasen, Hinzufügungen/Löschungen eines oder mehrerer Wörter), bis hin zu komplexen Unterschieden (z. B. Verwendung von Paraphrasen, gleichzeitiges Vorkommnis von lexikalischen und syntaktischen Unterschieden). Ein weiteres Auswahlkriterium war, dass die Match-Werte beider TMs deutlich voneinander abweichen (z. B. 48 % vs. 90 %) oder zumindest eine relativ große Differenz in Relation zum linguistischen Phänomen aufweisen mussten (z. B. 90 % vs. 99 % im Falle einer orthografischen Variante).

Zu jedem dargebotenen AS_{neu}-AS_{TM}-/AS_{neu}-AS_{iMem}-Paar wurden die Match-Werte beider TMs aufgeführt. Es wurde nicht angegeben, welcher Match-Wert durch welches TM ermittelt wurde. Jeder Match-Wert wurde mit einer Bewertungsskala versehen, die aus den fünf Bewertungsstufen *viel zu niedrig*, *zu niedrig*, *genau richtig*, *zu hoch* und *viel zu hoch* bestand. Für jeden Match-Wert konnte immer nur eine Bewertungsstufe ausgewählt werden.

Die Aufgabe der befragten Personen bestand darin, anzugeben, wie hoch die *Bedeutungs*gleichheit bzw. *Bedeutungs*ähnlichkeit zwischen dem AS_{neu} und dem AS_{TM} bzw. AS_{iMem} ist, d. h., anzugeben, inwieweit die beiden Segmente das Gleiche/Ähnliches *aussagen* – gemessen am subjektiven Ähnlichkeitsempfinden. Beide Match-Werte mussten demnach bewertet werden. So war es z. B. möglich, dass der Match-Wert aus dem SDL-TM als zu niedrig erachtet wurde, während der Match-Wert aus dem iMem-TM zu hoch erschien.

Zudem stand zu jedem AS_{neu}-AS_{TM}-/AS_{neu}-AS_{iMem}-Paar ein freies Textfeld zur Verfügung, in dem die befragten Personen Bemerkungen, Begründungen der Bewertungen, den subjektiven Ähnlichkeitswert etc. angeben konnten.

Der Fragebogen umfasst 51 Fragen, wobei es sich bei der ersten Frage um die Berufsgruppenzugehörigkeit handelt, d. h., ob die Testpersonen im Übersetzungsbereich tätig sind oder nicht (Abbildung 47). Die restlichen 50 Fragen führen jeweils ein AS_{neu}-AS_{TM}-/AS_{neu}-AS_{iMem}-Paar mit den Match-Werten und den Bewertungsskalen sowie das freie Textfeld auf. Der Fragebogen mit den Befragungsergebnissen inklusive etwaiger subjektiver Ähnlichkeitswerte oder sonstiger Bemerkungen und Informationen zur Berufsgruppenzugehörigkeit sind in Anhang J aufgeführt. Bei Satz A (Abbildung 48) handelt es sich

immer um das AS_{neu}, während Satz B stets das im TM gespeicherte Segment ist. Der erste Match-Wert stammt immer aus dem SDL-TM, der zweite immer aus dem iMem-TM. Keine der gemachten Angaben wurde personalisiert.

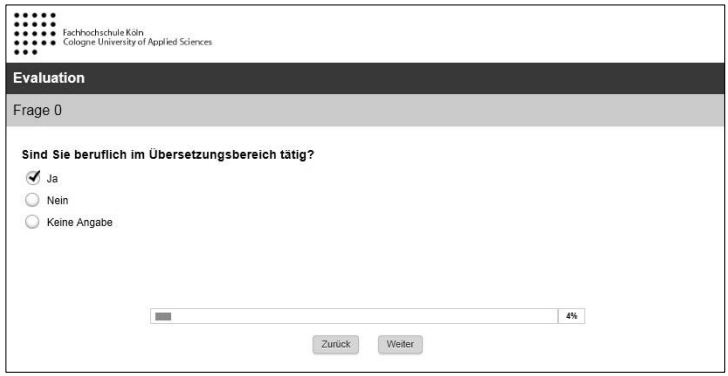

Abbildung 47: Frage zur Berufsgruppenzugehörigkeit

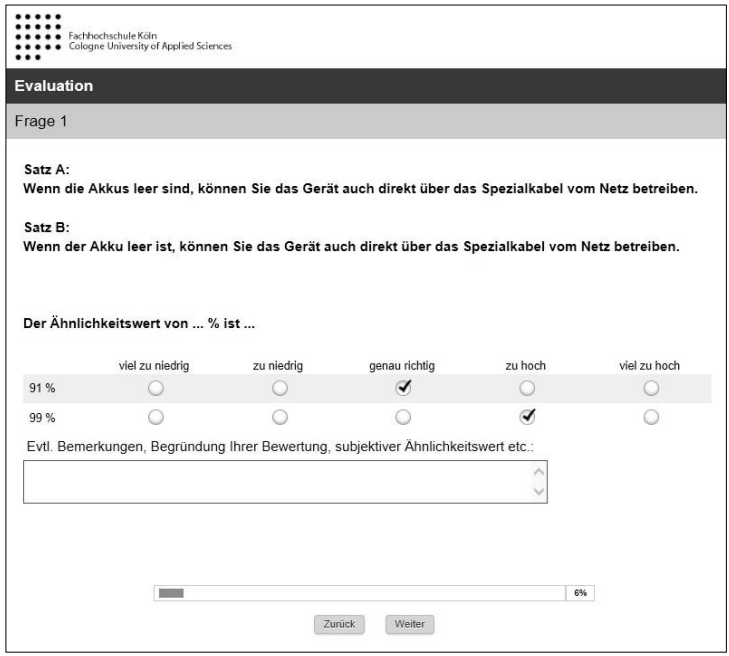

Abbildung 48: Aufbau des Online-Fragebogens am Beispiel einer der insgesamt 51 Fragen

6.5.3.2 Beschreibung der Evaluierungsteilnehmer

Die befragten Personen stammen aus unterschiedlichen beruflichen Bereichen. Einerseits wurde der Online-Fragebogen insgesamt 779 Übersetzungsagenturen, Sprachdienstleistern, freiberuflichen Übersetzern, Dolmetschern, Lektoren und technischen Redakteuren sowie Studenten, Doktoranden und Lehrkräften aus übersetzungsrelevanten Studiengängen zugänglich gemacht. Dieses Klientel wurde gewählt, da sie in den meisten Fällen mit der TM-Technologie vertraut ist und der Berufszweig ein gutes linguistisches Gespür voraussetzt. Da jedoch die Gefahr bestand, dass gerade das Wissen über die TM-Technologie einen Einfluss auf die Bewertung der Match-Werte haben könnte, d. h., dass eher Unterschiede in den Zeichenketten als in der Bedeutung bewertet wurden, wurde der Fragebogen zudem an 61 Testpersonen geschickt, denen TM-Systeme vermutlich bislang unbekannt waren. Diese Testpersonen entstammen z. B. der Berufsgruppe der Pädagogen, Rechtsanwälte, Kaufleute etc. Die erste Frage im Fragebogen (*Zugehörigkeit der Berufsgruppe*) dient folglich dem Zweck, herauszufinden, ob ein vermeintliches Wissen über die TM-Technologie einen Einfluss auf die Bewertung der Ähnlichkeit zwischen AS_{neu} und AS_{TM} bzw. AS_{iMem} hat.

Die nachfolgende statistische Auswertung aller vollständig ausgefüllten Fragebogen gibt an, in wie vielen Fällen und bei welchen linguistischen Phänomenen das iMem-TM mit dem menschlichen Ähnlichkeitsempfinden korrespondiert oder divergiert. Diese Analyse soll demnach dazu dienen, auszuloten, ob und bei welchen linguistischen Unterschieden das iMem-Proximitätsmaß angepasst werden muss oder nicht.

6.5.3.3 Statistische Auswertung des Fragebogens

Die Laufzeit des Fragebogens betrug zwei Monate. In dieser Zeit wurde der Fragebogen von 209 der insgesamt 840 befragten Personen ausgefüllt. Jeder der 209 Fragebogen wurde anschließend auf Plausibilität und Vollständigkeit[80] geprüft, sodass letztlich 94 vollständig ausgefüllte Fragebogen in die Auswertung einfließen konnten.

Beim Auswerten aller gemachten Angaben pro Bewertungsstufe und System kann festgestellt werden, dass die Match-Werte des SDL-TMs in den

[80] Bei der Prüfung auf Plausibilität und Vollständigkeit wurde untersucht, ob alle Fragen des Fragebogens vollständig ausgefüllt wurden und ob die angeklickten Antworten sich nicht gegenseitig ausschlossen (Negativbeispiel: 91 % genau richtig und gleichzeitig 99 % zu niedrig). Es wurden letztlich nur diejenigen Fragebogen für die weitere Auswertung berücksichtigt, die sowohl vollständig als auch plausibel waren.

meisten Fällen als zu niedrig und die Match-Werte des iMem-TMs in den meisten Fällen als genau richtig bewertet wurden.

Die Anzahl an Antworten für die Match-Werte des SDL-TMs ist für die Bewertungsstufen *viel zu niedrig* und *genau richtig* ungefähr gleich groß und macht einen Gesamtanteil von über 50 % aller Antworten aus. Dennoch kann über allen Bewertungsstufen hinweg für das SDL-TM eine Tendenz der Beurteilungen in Richtung *zu niedrig* und *viel zu niedrig* beobachtet werden.

Dagegen verhalten sich die Antworthäufigkeiten für die Match-Werte des iMem-TMs konträr: Am seltensten wurden die Bewertungsstufen *zu niedrig* und *viel zu niedrig* angeklickt, während die Bewertungsstufen *zu hoch* und *viel zu hoch* insgesamt einen Anteil von über 50 % aller Beantwortungen ausmachen. Die Tendenz verläuft demnach entgegengesetzt zu derjenigen des SDL-TMs, nämlich in Richtung *zu hoch* und *viel zu hoch*. In Tabelle 48 werden die detaillierten Antworthäufigkeiten aufgeführt. In Abbildung 49 werden zudem die unterschiedlichen Tendenzen beider Systeme erkennbar.

	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
SDL (Anzahl Antworten)	1199	1853	1177	445	26
iMem (Anzahl Antworten)	22	445	1704	1503	1026
SDL (Ø-Häufigkeit)	25,51 %	39,43 %	25,04 %	9,47 %	0,55 %
iMem (Ø-Häufigkeit)	0,47 %	9,47 %	36,26 %	32,00 %	21,83 %

Tabelle 48: Antworthäufigkeiten pro Bewertungsstufe (gerundete Prozentwerte)

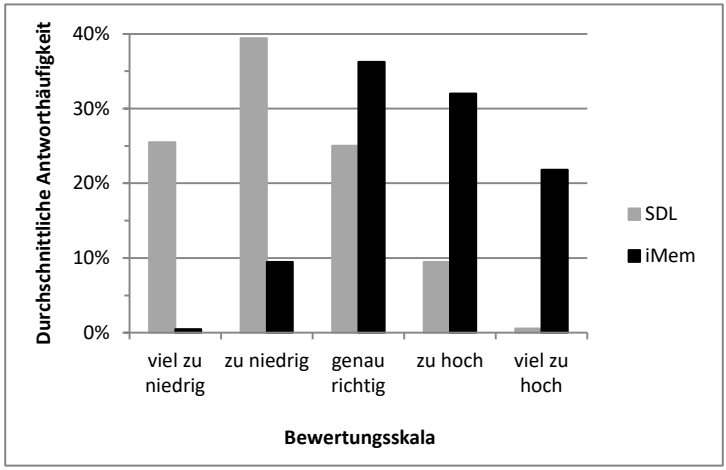

Abbildung 49: Durchschnittliche Antworthäufigkeiten pro Bewertungsstufe

Die Tendenz, die bereits in Tabelle 48 und Abbildung 49 beobachtet werden konnte, spiegelt sich auch wider, wenn pro Bewertungsstufe ausgezählt wird, bei wie vielen Fragen häufiger für eines der beiden Systeme die jeweilige Bewertungsstufe angeklickt wurde. Hierbei fällt auf, dass die Anzahl für die Bewertungsstufen *zu niedrig* und *viel zu niedrig* für das SDL-TM identisch ist mit der Anzahl für die Bewertungsstufen *zu hoch* und *viel zu hoch* für das iMem-TM. Für die Bewertungsstufe *genau richtig* unterscheiden sich die Werte: Es wurde häufiger zugunsten des iMem-TMs entschieden als zugunsten des SDL-TMs (Tabelle 49).

	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
SDL (Anzahl an Fragen, bei denen Bewertungsstufe häufiger gewählt wurde)	50	49	20	1	0
iMem (Anzahl an Fragen, bei denen Bewertungsstufe häufiger gewählt wurde)	0	1	30	49	50

Tabelle 49: Anzahl an Fragen, bei denen häufiger die entsprechende Bewertungsstufe für ein System im Vergleich zum anderen System gewählt wurde.

Die Angaben in Tabelle 49 können präzisiert werden, indem untersucht wird, welche Bewertungsstufe bei welchen linguistischen Unterschieden zwischen AS_{neu} und AS_{TM} am häufigsten gewählt wurde (Tabelle 50 bis 53). Dabei kann vermerkt werden, dass bei ausschließlich morphologischen und bei komplexen Unterschieden das iMem-TM öfter einen höheren *genau richtig*-Wert aufweist als das SDL-TM. Nur bei den lexikalischen Unterschieden unterliegt das iMem-TM dem SDL-TM in der Bewertungsstufe *genau richtig*. Bei den syntaktischen Unterschieden halten sich beide Systeme bei den *genau richtig*-Bewertungen die Waage.

Bei der Betrachtung der anderen Bewertungsstufen zeichnet sich die gleiche Tendenz wie zuvor beschrieben ab: Unabhängig von der Art des linguistischen Unterschiedes können immer mehr Fragen gezählt werden, bei denen das SDL-TM eine höhere *zu niedrig*- oder *viel zu niedrig*-Bewertung erhalten hat als das iMem-TM. Die Match-Werte des iMem-TMs hingegen wurden für alle Arten an linguistischen Unterschieden öfter als *zu hoch* oder *viel zu hoch* empfunden als diejenigen des SDL-TMs.

Gesamtanzahl der Fragen mit rein morphologischen Unterschieden zwischen AS_{neu} und AS_{TM}/AS_{iMem}: 6					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
SDL	6	6	1	0	0
iMem	0	0	5	6	6

Tabelle 50: Anzahl an Fragen mit ausschließlich morphologischen Unterschieden, bei denen häufiger die entsprechende Bewertungsstufe für ein System im Vergleich zum anderen System gewählt wurde.

Gesamtanzahl der Fragen mit rein lexikalischen Unterschieden zwischen AS_{neu} und AS_{TM}/AS_{iMem}: 2					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
SDL	2	2	2	0	0
iMem	0	0	0	2	2

Tabelle 51: Anzahl an Fragen mit ausschließlich lexikalischen Unterschieden, bei denen häufiger die entsprechende Bewertungsstufe für ein System im Vergleich zum anderen System gewählt wurde.

Gesamtanzahl der Fragen mit rein syntaktischen Unterschieden zwischen AS_{neu} und AS_{TM}/AS_{iMem}: 12					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
SDL	12	12	6	0	0
iMem	0	0	6	12	12

Tabelle 52: Anzahl an Fragen mit ausschließlich syntaktischen Unterschieden, bei denen häufiger die entsprechende Bewertungsstufe für ein System im Vergleich zum anderen System gewählt wurde.

Gesamtanzahl der Fragen mit komplexen Unterschieden zwischen AS_{neu} und AS_{TM}/AS_{iMem}: 30					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
SDL	30	29	11	1	0
iMem	0	1	19	29	30

Tabelle 53: Anzahl an Fragen mit komplexen Unterschieden, bei denen häufiger die entsprechende Bewertungsstufe für ein System im Vergleich zum anderen System gewählt wurde.

Zur Auslotung, ob das iMem-Proximitätsmaß das menschliche Ähnlichkeitsempfinden widerspiegelt, ist vor allem die Anzahl an *genau richtig*-Bewertungen pro Frage für den jeweiligen Match-Wert des iMem-TMs interessant:

Bei etwa der Hälfte aller Fragen kann eine höhere Anzahl an *genau richtig*-Bewertungen für das iMem-TM im Vergleich zu den anderen Bewertungsstufen vermerkt werden.

Dabei muss jedoch beachtet werden, dass auch Fragen existieren, bei denen sich – unabhängig vom linguistischen Unterschied – die Anzahl der *genau richtig*- und *zu hoch*- bzw. *viel zu hoch*-Bewertungen nur minimal unterscheidet, was sich auch an den subjektiven Ähnlichkeitswerten feststellen lässt, die zu manchen Fragen genannt wurden. So wurde beispielsweise bei Frage 2 (siehe Anhang J) der Match-Wert des iMem-TMs in 30 Fällen als *genau richtig* und in 33 Fällen als *zu hoch* erachtet. Weitere solche Beispiele finden sich in Anhang J, Fragen 4, 9, 10, 14, 20, 35, 43 und 50. Bei Frage 48 ist die Anzahl an *genau richtig*- und *zu hoch*-Bewertungen sogar identisch. Bei Frage 31 wurde hingegen öfter *zu niedrig* ausgewählt als *genau richtig*.

Nachfolgend werden linguistische Unterschiede aufgeführt, bei denen ein höherer *genau richtig*-Wert im Vergleich zu den anderen Bewertungsstufen durch das iMem-TM erreicht wurde. Die Unterschiede traten teilweise isoliert, teilweise kombiniert auf. Es werden die Änderungen beschrieben, die notwendig sind, um aus einem AS_{iMem} das entsprechende AS_{neu} herzustellen:

- Kompositabildung und -zerlegung
- Orthografische Variante
- Änderung der Diathese
- Änderung von Nominal- in Verbalstil
- Ersetzung eines Substantivs durch einen substantivierten Infinitiv
- Änderung eines Imperativsatzes in einen imperativischen Infinitiv oder Indikativsatz
- Ersetzung einer Präpositionalphrase oder eines Adjektivs durch einen Relativsatz
- Ersetzung eines Verbs durch ein Funktionsverbgefüge
- Phrasenvertauschung
- Löschung mehrerer Wörter oder kurzer Phrasen
- Hinzufügungen mehrerer Wörter, Phrasen oder Teilsätze

Dagegen konnten bei folgenden linguistischen Unterschieden – teilweise auch nur knapp – niedrigere *genau richtig*-Werte im Vergleich zu den restlichen Bewertungsstufen durch das iMem-TM festgestellt werden. Auch diese Unterschiede traten sowohl ausschließlich als auch in Kombination miteinander auf:

- Änderung des Numerus
- Änderung von Basis zu Diminutivum
- Wortersetzungen in sehr kurzen Segmenten
- Ersetzung einer mehrteiligen Verbphrase durch eine einteilige Verbphrase
- Hinzufügung einer Phrase in einem kurzen Segment
- Hinzufügung eines Teilsatzes in einem langen Segment
- Löschung eines längeren Teilsatzes
- Löschung eines längeren Teilsatzes; das AS_{neu} ist komplett im AS_{iMem} enthalten
- Phrasenvertauschung, Vertauschung von Haupt- und Nebensatz
- Paraphrasen

Es fällt auf, dass insbesondere AS_{neu}-AS_{iMem}-Paare mit großem Segmentlängenunterschied sowie mit einer großen Anzahl an verschiedenen linguistischen Unterschieden, wie es z. B. bei Paraphrasen der Fall ist, von einem niedrigen *genau richtig*-Wert betroffen sind.

Die Berufsgruppenzugehörigkeit – und damit einhergehend das vermeintliche Wissen der befragten Personen über die Funktionsweise von TMs – hat dabei nur einen geringen Einfluss auf die Bewertung. In Tabelle 54 und 55 werden die Antworthäufigkeiten pro Bewertungsstufe und System hinsichtlich der Berufsgruppenzugehörigkeit aufgeschlüsselt.

Wenn in beiden Tabellen die Antworthäufigkeiten für die Kategorie *Keine Angabe*[81] außer Acht gelassen wird, unterscheiden sich die Antworthäufigkeiten im Falle des SDL-TMs maximal um 3,67 % (Tabelle 54, Bewertungsstufe *viel zu niedrig*). Bei den Bewertungsstufen *genau richtig* und *zu hoch* für das iMem-TM kann hingegen ein größerer Unterschied von 6,33 % bzw. 7,54 % zwischen den Antworthäufigkeiten vermerkt werden.

SDL-TM	**viel zu niedrig**	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
Im Übersetzungsbereich tätig	26,13 %	39,64 %	24,16 %	9,54 %	0,52 %
Nicht im Übersetzungsbereich tätig	29,80 %	38,70 %	21,20 %	9,50 %	0,80 %
Keine Angabe	16,00 %	39,54 %	35,08 %	9,08 %	0,31 %

Tabelle 54: Antworthäufigkeit (gerundet) pro Bewertungsstufe für das SDL-TM gemäß Berufsgruppenzugehörigkeit

[81] Die Kategorie *Keine Angabe* ist lediglich der Vollständigkeit halber mit aufgeführt. Übersprungene Fragen wurden als *Keine Angabe* gezählt.

iMem-TM	viel zu niedrig	zu niedrig	genau richtig	zu hoch	viel zu hoch
Im Übersetzungsbereich tätig	0,66 %	10,00 %	35,77 %	31,64 %	21,93 %
Nicht im Übersetzungsbereich tätig	0,10 %	10,30 %	42,10 %	24,10 %	23,40 %
Keine Angabe	0,15 %	5,69 %	29,54 %	45,69 %	18,92 %

Tabelle 55: Antworthäufigkeit (gerundet) pro Bewertungsstufe für das iMem-TM gemäß Berufsgruppenzugehörigkeit

Aus dem Vergleich beider Tabellen lässt sich schließen, dass sowohl im Übersetzungsbereich tätige als auch nicht im Übersetzungsbereich tätige Personen eher dazu tendieren, linguistisch optimierte Match-Werte als genau richtig zu bewerten, während nicht linguistisch optimierte Match-Werte unabhängig der Berufsgruppenzugehörigkeit eher als zu niedrig erachtet werden. Anmerkungen zu bestimmten Fragen deuten jedoch darauf hin, dass bei der Bewertung der Segmentpaare manchmal der Inhalt vernachlässigt und stattdessen vielmehr die Funktionsweise von TMs berücksichtigt wurde (siehe Anhang J, z. B. Frage 3, 5 und 50).

6.5.4 Effektivitätsmessung: Untersuchung des Nachbearbeitungsaufwandes

Die bisher durchgeführten Effektivitätsmessungen haben starken Bezug auf die Match-Werte beider Systeme genommen. Da diese Match-Werte jedoch aufgrund der unterschiedlichen Algorithmen relativ sind, wird nachfolgend untersucht, wie groß der Nachbearbeitungsaufwand der besten Matches beider Systeme für ausgewählte ZS_{TM} bzw. ZS_{iMem} zur Erstellung der Übersetzung eines AS_{neu} ist. Auf diese Weise werden nicht die Match-Werte, sondern die ausgegebenen Übersetzungseinheiten bewertet.

Dazu wird wie beim Vergleich mit dem menschlichen Ähnlichkeitsempfinden (siehe Kapitel 6.5.3) eine Skala zur Bewertung des Nachbearbeitungsaufwandes eingesetzt. Motiviert durch Castilho Monteiro de Sousa et al. (2011), die den Post-Editing-Aufwand der Übersetzungen von Untertiteln messen, die durch MÜ-Systeme generiert sowie durch ein TM-System aufgefunden wurden, verfügt diese Skala über die folgenden fünf Bewertungsstufen, mit denen eingeschätzt werden soll, ob und in welchem Maße eine Nachbearbeitung eines ZS_{TM} bzw. ZS_{iMem} notwendig ist, um eine korrekte Übersetzung des AS_{neu} zu erhalten:

Bewertungsstufe	Beschreibung
1	Kein Nachbearbeitungsaufwand notwendig
2	Geringe Nachbearbeitung notwendig
3	Viel Nachbearbeitung notwendig, aber weniger Aufwand als komplette Neuübersetzung
4	Komplette Neuübersetzung erfordert weniger Aufwand als Nachbearbeitung
5	Komplette Neuübersetzung notwendig, da kein Match aufgefunden werden konnte

Tabelle 56: Skala zur Bewertung des Nachbearbeitungsaufwandes

Die Beurteilung des Nachbearbeitungsaufwandes wurde mit AT_{neu} A und Korpus A für jedes System durchgeführt, wobei nur die jeweils besten Matches eines jeden Systems für die Bewertung herangezogen wurden. Auch bei dieser Untersuchung wurden nur Fuzzy-Matches bzw. No-Matches berücksichtigt; 100 %-Matches wurden außer Acht gelassen.

Da bei 85 der 100 Segmente aus AT_{neu} A bei beiden Systemen dieselbe Übersetzungseinheit an erster Position in der Trefferliste ausgegeben wurde (siehe Kapitel 6.5.2.2, Seite 174), wurde der Nachbearbeitungsaufwand lediglich für diejenigen 15 AS_{neu} beurteilt, zu denen von beiden Systemen unterschiedliche Übersetzungseinheiten als erste Treffer dargeboten wurden. Alle 15 Übersetzungseinheiten des SDL-TMs waren dabei zuvor als relevant eingestuft worden, während im Falle des iMem-TMs lediglich 13 Übersetzungseinheiten als relevant und eine als nicht relevant bewertet wurden. Zu einer Anfrage wurde sogar keine Übersetzungseinheit durch das iMem-TM ausgegeben. In Anhang K sind alle 15 AS_{neu} mit den durch das SDL-TM und iMem-TM aufgefundenen Übersetzungseinheiten mit der Anzahl der durch die Juroren gewählten Bewertungsstufen detailliert aufgeführt.

Die Einschätzung des Nachbearbeitungsaufwandes erfolgte durch sechs unabhängige Juroren, die mit der TM-Technologie vertraut sind und sowohl die deutsche als auch die englische Sprache beherrschen. Vier der Juroren sind Diplom-Übersetzer mit mehrjähriger Praxiserfahrung. Einer dieser Übersetzer verfügt über eine zusätzliche Qualifikation zum Terminologen und Sprachtechnologen, während ein anderer zudem Lehrbeauftragter für die Bereiche IT-Fachübersetzungen Englisch-Deutsch, Übersetzungstechnologie und Softwarelokalisierung an der Technischen Hochschule Köln ist. Die restlichen zwei Juroren sind ausgebildete Terminologen und Sprachtechnologen, von denen einer über eine mehrjährige Berufserfahrung als Softwarelokalisierer verfügt. Der andere ist als freiberuflicher Terminologe und Dozent für

Übersetzer für das Fach Trados an der Übersetzer- und Dolmetscherschule Köln tätig.

Den Juroren wurden die 15 AS_{neu} mit den durch beide Systeme aufgefundenen Übersetzungseinheiten (d. h. AS_{TM} mit ZS_{TM} und AS_{iMem} mit ZS_{iMem}) ohne Angabe des jeweiligen Match-Wertes präsentiert. Es wurde nicht angegeben, welche Übersetzungseinheit aus welchem System stammt. Anhand der Bewertungsskala sollten die Juroren den gemäß ihrem subjektiven Empfinden angemessenen Nachbearbeitungsaufwand einschätzen[82], wobei für jeden ZS_{TM} bzw. ZS_{iMem} immer nur eine Bewertungsstufe ausgewählt werden durfte. In Tabelle 57 wird angegeben, wie oft welche Bewertungsstufe für jedes System über die 15 AS_{neu} im Vergleich mit den aufgefundenen Übersetzungseinheiten von den sechs Juroren ausgewählt wurde.

	1	**2**	**3**	**4**	**5**
SDL (Anzahl Stimmen)	26	48	12	4	0
iMem (Anzahl Stimmen)	30	34	10	7	9
SDL (Ø-Häufigkeit)	28,89 %	53,33 %	13,33 %	4,44 %	0 %
iMem (Ø-Häufigkeit)	33,33 %	37,78 %	11,11 %	7,78 %	10,00 %

Tabelle 57: Antworthäufigkeiten der sechs Juroren pro Bewertungsstufe (gerundete Prozentwerte) über die 15 AS_{neu} im Vergleich mit den durch das iMem- und SDL-TM aufgefundenen Übersetzungseinheiten

Die Tabelle macht deutlich, dass bei den durch das iMem-TM aufgefundenen Übersetzungseinheiten öfter kein Nachbearbeitungsaufwand erforderlich ist als bei den durch das SDL-TM aufgefundenen Übersetzungseinheiten. Dagegen werden die Übersetzungseinheiten des SDL-TMs am häufigsten als Übersetzungseinheiten eingestuft, die nur eines geringen Nachbearbeitungsaufwandes bedürfen. Die höchste Anzahl an Bewertungen für das iMem-TM ist ebenso für die Bewertungsstufe 2 zu verzeichnen, auch wenn diese fast ein Drittel weniger Stimmen erhielt als für das SDL-TM. Auch bei der Bewertungsstufe 3 wurde öfter zugunsten des SDL-TMs als zugunsten des iMem-TMs entschieden. Die Bewertungsstufen 4 und 5, bei denen eine komplette Neuübersetzung weniger Nachbearbeitungsaufwand bedeutet als eine Überarbeitung der gespeicherten Übersetzung, wurden jeweils öfter beim

[82] Gemäß Moorkens et al. (2015) ist die subjektive Einschätzung kein verlässliches Maß zur Messung des Nachbearbeitungsaufwandes. Vielmehr sollte beispielsweise die tatsächlich aufgewendete Zeit für die Nachbearbeitung gemessen werden. Aufgrund der dazu fehlenden technischen Möglichkeiten im Rahmen des iMem-Forschungsprojektes war die Messung der Nachbearbeitungszeit jedoch nicht durchführbar, sodass auf die subjektive Einschätzung zurückgegriffen werden musste.

iMem-TM gewählt. Mit der Bewertungsstufe 5 wurden sowohl leere Antwortmengen als auch die Fälle bewertet, bei denen zwar eine Übersetzungseinheit ausgegeben, diese jedoch nicht als relevant eingestuft wurde. In Tabelle 58 wird weiterführend verdeutlicht, bei wie vielen der 15 Fragen die jeweilige Bewertungsstufe öfter zugunsten eines der beiden Systeme im Vergleich zum anderen System ausgewählt wurde bzw. bei wie vielen Fragen gleich oft oder auch gar nicht die jeweilige Bewertungsstufe selektiert wurde.

Bewertungsstufe	Anzahl der Fragen, bei denen die jeweilige Bewertungsstufe öfter beim SDL-TM gewählt wurde	Anzahl der Fragen, bei denen die jeweilige Bewertungsstufe öfter beim iMem-TM gewählt wurde	Anzahl der Fragen, bei denen die Bewertungsstufe gleich oft gewählt wurde	Anzahl der Fragen, bei denen die Bewertungsstufe nicht gewählt wurde
1	4	7	0	4
2	10	4	1	0
3	4	3	1	7
4	1	2	0	12
5	0	3	0	12

Tabelle 58: Anzahl an Fragen, bei denen die jeweilige Bewertungsstufe öfter bei einem der beiden Systeme bzw. gleich oft oder nicht gewählt wurde.

Die in dieser Tabelle aufgeführten Ergebnisse verhalten sich analog zu denen aus Tabelle 57, von denen nachfolgend nur die markanten Ergebnisse erläutert werden: Beim iMem-TM wurden mehr Fragen der Bewertungsstufe 1 zugeschrieben als beim SDL-TM. Im Gegensatz dazu wurde bei zwei Dritteln aller Fragen die Bewertungsstufe 2 öfter zugunsten des SDL-TMs als zugunsten des iMem-TMs gewählt. Eine Neuübersetzung (Bewertungsstufe 4 und 5) erschien den Juroren bei mehr Fragen für das iMem-TM notwendig.

Zusammenfassend kann festgehalten werden, dass, obwohl die durch das iMem-TM ausgegebenen Übersetzungseinheiten im Vergleich zu denen des SDL-TMs häufiger keiner Nachbearbeitung bedürfen, jedoch auch häufiger Übersetzungseinheiten beim iMem-TM als beim SDL-TM auftraten, bei denen eine komplette Neuübersetzung als effektiver erachtet wurde als eine Überarbeitung der gespeicherten Übersetzung.

6.5.5 Effizienzmessung: Untersuchung der Antwortzeit

Ziel dieser Effizienzmessung ist es, zu untersuchen, wie lange ein sehr großes iMem-TM im Vergleich zu einem sehr kleinen iMem-TM für das Auffinden potenzieller Matches zu einem AS_{neu} und das Berechnen des dazugehörigen

Match-Wertes benötigt. Dazu wurde der AT_{neu} B mit einem zuvor erstellten iMem-TM des Korpus B verglichen. Selbiges galt für den AT_{neu} A und das iMem-TM des Korpus A.

Die Untersuchung der Antwortzeit wurde in zwei Teile unterteilt: Zum einen wurde die Zeit für die Datenbankabfrage gemessen. Unter der *Datenbankabfragezeit* ist diejenige Zeit zu verstehen, die für die Vorfilterung (siehe Kapitel 5.2.2.1) durch das iMem-TM benötigt wird, um die maximal 50 besten potenziellen Match-Partner für das AS_{neu} zu bestimmen. Zum anderen wurde die *LCS-Zeit* ausgelesen, die die Zeit angibt, die für die Ermittlung der LCS zwischen dem AS_{neu} und jedem identifizierten potenziellen Match-Partner (siehe Kapitel 5.2.2.2) erforderlich ist. Ebenso wurde die *Match-Zeit* ermittelt, die für die Berechnung des Match-Wertes (siehe Kapitel 5.2.2.3) zwischen dem AS_{neu} und jedem der identifizierten potenziell matchenden AS_{iMem} benötigt wird.

Mittels der Java-Zeitfunktion *System.nanoTime* (siehe Oracle 1993, 2016: o.S.) wurde die Antwortzeit in Nanosekunden ermittelt. Diese Zeitfunktion bietet den Vorteil, auch Antwortzeiten unter einer Millisekunde messen zu können. Die Messungen wurden auf einem Rechner mit 2,66 GHz unter Windows 7, 32Bit sowie mit der Java-Version *Java SE 7u51 32Bit* durchgeführt.

Zu Beginn der Messung der Datenbankabfragezeit sowie der LCS- und Match-Zeit wurde die Startzeit jeweils in Nanosekunden festgehalten. Am Ende der Messung wurde erneut die Zeit erfasst. Die Differenz von Start- und Endezeit ergibt die jeweilige Antwortzeit für die Datenbankabfrage, die Ermittlung der LCS und die Berechnung des Match-Wertes.

Um mögliche Messfehler durch die Zeitfunktion zu vernachlässigen, wurde die Datenbankabfragezeit sowie die LCS- und Match-Zeit pro nachgefragten AS_{neu} hundert Mal erfasst, indem die Startzeit zu Beginn des ersten Durchlaufes und die Endezeit nach dem hundertsten Durchlauf festgehalten wurden. Die daraus resultierenden Zeiten wurden anschließend über die 100 Durchläufe gemittelt. Es wurden 100 Durchläufe pro AS_{neu} durchgeführt, da sich diese Anzahl als eine gute Balance zwischen Genauigkeit und Schnelligkeit der Messung anbot.

Des Weiteren wird während der Ermittlung der maximal 50 besten potenziellen Match-Partner pro AS_{neu} und der Match-Wert-Berechnung – insbesondere beim Nachschlagen der ersten AS_{neu} im iMem-TM – der Programmcode durch die Java Virtual Machine dynamisch optimiert. Dieser Optimierungsvorgang hat zur Folge, dass die Messung der eigentlichen Antwortzeiten – vor allem bei den ersten nachgefragten AS_{neu} – stark verfälscht wird; es werden viel zu hohe Antwortzeiten im Vergleich zu den gemessenen Antwortzeiten der restlichen AS_{neu} ausgegeben. Im Laufe der Messung normalisieren

sich die Antwortzeiten jedoch, sodass die Messergebnisse der ersten 50 AS_{neu} verworfen wurden.

Letztlich wurden die Messergebnisse der restlichen AS_{neu} über die Anzahl der AS_{neu} (exklusive der ersten 50 AS_{neu}) gemittelt. Folgende durchschnittlichen Antwortzeiten konnten schließlich für beide iMem-TMs ausgelesen werden:

	iMem-TM (Korpus A)	**iMem-TM (Korpus B)**
Anzahl eindeutiger Segmente	535	305.324
Ø Anzahl Wörter in Originalform/Segment	15,34	20,51
Ø Anzahl Basiswörter/Segment	17,59	23,76
Datenbankabfragezeit	2,11 ms	422,83 ms
LCS-Zeit	0,04 ms	0,09 ms
Match-Zeit	0,16 ms	0,17 ms
Gesamtdauer	2,31 ms	423,09 ms

Tabelle 59: Antwortzeiten (gerundet) für das iMem-TM des Korpus A und Korpus B

Wie in Tabelle 59 ersichtlich, ist die Datenbankabfragezeit für das iMem-TM des Korpus B zweihundert Mal und die LCS-Zeit etwa zwei Mal länger als die entsprechenden Zeiten des iMem-TMs des Korpus A. Die Match-Zeit ist hingegen unwesentlich länger.

Die größere Datenbankabfrage-, LCS- und Match-Zeit bei Korpus B im Vergleich zu Korpus A kommt dabei zum einen aufgrund der größeren Anzahl an AS_{iMem} zustande und zum anderen dadurch, dass Korpus B mehr deutlich längere Segmente als Korpus A enthält. Längere Segmente enthalten mehr Inhaltswörter, die bei der Vorfilterung festgestellt und verglichen werden müssen. Ebenso müssen ggf. mehr LCS ermittelt und mehr MPRO-Merkmale für die Identifikation morphosyntaktischer Unterschiede untersucht werden, als es bei kurzen Segmenten der Fall ist.

Wie in Nielsen (1993: 135) erläutert, hat der Anwender bei einer Antwortzeit bis maximal 0,1 Sekunden den Eindruck einer unmittelbaren Reaktion des Systems. Bei einer Antwortzeit zwischen 0,1 und maximal 1,0 Sekunden fällt die Antwortverzögerung zwar auf, jedoch bleibt die Aufmerksamkeit des Nutzers weiterhin erhalten. Erst Antwortzeiten von mehr als 10 Sekunden führen dazu, dass sich der Anwender in der Zwischenzeit anderen Aktivitäten widmet.

Beim Vergleich dieser Einstufung der Antwortzeiten mit den in Tabelle 59 gelisteten kann festgestellt werden, dass das iMem-TM im Falle von

Korpus A mit 0,00231 Sekunden sehr schnell reagiert. Die Gesamtantwortzeit von unter einer halben Sekunde pro nachgefragtem AS_{neu} im Falle von Korpus B ist ebenfalls akzeptabel für den Einsatz des iMem-TMs in der Übersetzungspraxis.

6.5.6 Effizienzmessung: Untersuchung des Speicherplatzbedarfes

Für die Untersuchung des Speicherplatzbedarfes des iMem-TMs wurden insgesamt vier TMs erstellt, deren Dateigrößen miteinander verglichen wurden. Dabei handelt es sich um das SDL-TM des Korpus A und Korpus B sowie um das iMem-TM des Korpus A und Korpus B.

In Tabelle 60 ist zu erkennen, dass beide iMem-TMs größer im Vergleich zum entsprechenden SDL-TM sind. Es fällt auf, dass der Unterschied im Speicherplatzbedarf zwischen einem iMem-TM und einem SDL-TM umso größer wird, je mehr Datensätze in den TMs enthalten sind.

	Speicherplatz
SDL-TM (Korpus A)	1,2 MB
SDL-TM (Korpus B)	360,2 MB
iMem-TM (Korpus A)	3,2 MB
iMem-TM (Korpus B)	2,46 GB

Tabelle 60: Speicherplatzbedarf für das iMem- und SDL-TM des Korpus A und Korpus B

Der Grund für den Unterschied im Speicherplatzbedarf ist, dass in einem iMem-TM – im Gegensatz zu einem konventionellen zeichenkettenbasierten TM – zusätzlich Inhaltswörter zum Zwecke der Vorfilterung sowie MPRO-Werte zur Berechnung des Match-Wertes gespeichert werden müssen. Diese Zusatzinformationen, die unerlässlich für den iMem-Algorithmus sind, bedürfen dieses zusätzlichen Speicherplatzes. Allerdings stellt ein Speicherplatzbedarf für ein TM von 2,46 GB oder mehr mit der heutigen Computer-Technologie kein großes Problem in der Übersetzungspraxis dar.

7 Diskussion

7.1 Zusammenfassung

Im Rahmen dieser Arbeit wurde gezeigt, wie die Retrieval-Leistung eines kommerziellen TM-Systems mittels Integration morphosyntaktischer Analyseverfahren verbessert werden kann. Dazu wurde zunächst der Begriff *Ähnlichkeit* aus menschlicher und informatischer Perspektive erläutert. Anschließend wurden Grundlagen der TM-Technologie beschrieben, wobei u. a. auf in der Literatur zu findende Ansätze zur linguistischen Optimierung kommerzieller und in Forschungsprojekten entwickelter TMs eingegangen wurde. Die darauffolgende Beschreibung des eigenen entwickelten Modells eines linguistisch optimierten TMs umfasste die Darlegung folgender Aspekte:

- Identifikation der für den Vergleich eines AS_{neu} mit einem AS_{iMem} notwendigen morphosyntaktischen Merkmale
- Bestimmung von Arbeitsschritten zur Vorfilterung für das iMem-TM
- Erarbeitung eines Algorithmus zur Ermittlung der LCS zwischen einem AS_{neu} und einem AS_{iMem}
- Erstellung und Auslotung eines Proximitätsmaßes zur Berechnung und Ausgabe von dem menschlichen Ähnlichkeitsempfinden entsprechenden Match-Werten

Das konzipierte Modell wurde in Form eines Plug-ins in ein bestehendes kommerzielles TM-System eingebunden. Das Datenmodell wurde mithilfe einer relationalen Datenbank für die Sprachkombination Deutsch-Englisch gespeichert. Schließlich erfolgte die Evaluierung des iMem-TMs, indem zunächst die dafür notwendigen Evaluierungsmaße und erstellten Korpora vorgestellt und unterschiedliche Effektivitäts- und Effizienzmessungen des neu entwickelten Systems durchgeführt wurden.

7.2 Diskussion der Evaluierungsergebnisse

Die Retrieval-Leistung kommerzieller TM-Systeme wurde durch den Einsatz morphosyntaktischer Analyseverfahren verbessert. Linguistische Unterschiede zwischen einem AS_{neu} und einem AS_{iMem} (Unterschiede im Numerus, Kompositazerlegung bzw. -bildung, Phrasenvertauschungen etc.) konnten abgebildet werden.

Durch die statistische Auswertung unter Anwendung ausgewählter Evaluierungsmaße sowie die Untersuchung des Vorkommnisses relevanter und nicht relevanter Übersetzungseinheiten in spezifischen Match-Wert-Bereichen konnte festgestellt werden, dass das iMem-TM stets bessere Precision-Werte (Average Precision, Precision at k und R-Precision inbegriffen) in den hohen Match-Wert-Bereichen (ab 70 % bis 99 %) ausgibt. Ebenso konnte in den hohen Match-Wert-Bereichen ein höherer Recall, d. h. insgesamt mehr relevante Übersetzungseinheiten, im Vergleich zum SDL-TM registriert werden. Allerdings ist auch der Fallout des iMem-TMs, d. h. die Anzahl an durch das iMem-TM aufgefundenen nicht relevanten Übersetzungseinheiten, meistens höher als beim SDL-TM.

Die besseren Precision-Werte in den hohen Match-Wert-Bereichen kommen dadurch zustande, dass für die Berechnung der Ähnlichkeit der LCS-Struktur (L) diese mithilfe des Korrekturfaktors x angehoben wird, unter der Bedingung, dass ${L_{halb}}^{x} = 0{,}7$ (siehe Kapitel 4.3.1). Der Grund für den hohen Recall und gleichzeitig hohen Fallout durch das iMem-TM ist, dass beim Vergleich des AS_{neu} mit dem AS_{iMem} die Basiswörter der Segmente miteinander verglichen werden. Dadurch wird veranlasst, dass mehr Wörter als identisch angesehen werden, als es beim bloßen Zeichenkettenvergleich der Fall ist. Die Verwendung der einfachen Wortformen würde hingegen den Nachteil mit sich bringen, dass beispielsweise ein Kompositum als Ganzes und nicht die einzelnen Kompositumsbestandteile für die Ermittlung der LCS zur Verfügung stände. Eine Kompositazerlegung als einziger Unterschied zwischen dem AS_{neu} und dem AS_{iMem} würde daher mit zu hohen Kosten gewertet, obwohl die Segmente semantisch gleich sind.

Da sich der Algorithmus jedoch auf das Auffinden der LCS auf lexikalischer Ebene beschränkt, können ebenso grammatisch unkorrekte bzw. nicht linguistisch motivierte Einheiten in die Ähnlichkeitsberechnung einbezogen werden. Aus diesem Grund ist eine Erweiterung des Algorithmus denkbar, der das Ermitteln ähnlicher Satzfragmente auf syntaktischer Ebene ermöglicht. Dazu ist ein Syntax-Parser erforderlich, der die AS_{neu} und AS_{iMem} in syntaktische Einheiten zerlegt. Ob und inwieweit sich das in dieser Arbeit angewendete Modell der GSAs auch auf Syntaxebene anwenden lässt, ist noch zu erforschen.

Die Tatsache, dass die Bedeutungsgleichheit/-ähnlichkeit zwischen einem AS_{neu} und einem AS_{iMem} auf Grundlage morphosyntaktischer Mittel ausgedrückt und berechnet wird, stellt die größte Schwierigkeit bei der Erstellung eines linguistisch optimierten TMs dar. So kann es vorkommen, dass Segmentpaare, die für den Menschen nur eine geringe Ähnlichkeit aufweisen, durch das System zu hoch bewertet werden, da es lediglich auf die

morphosyntaktischen Eigenschaften der zu vergleichenden Wörter zurückgreifen kann; semantische Eigenschaften bleiben in diesem Prototypen vorerst noch unberücksichtigt.

Obwohl MPRO über ein Merkmal verfügt, das semantische Zugehörigkeiten berücksichtigt (Merkmal *s*), war dieses Merkmal zur Zeit der Entwicklung des iMem-TMs noch zu unausgereift (in den meisten Fällen wurde der Wert *nil* ausgegeben), um einen tatsächlichen Nutzen daraus ziehen zu können. Falls das MPRO-Merkmal *s* zukünftig vervollständigt werden sollte, könnten AS_{neu} und AS_{iMem} zusätzlich auf semantischer Ebene verglichen werden. Bei der Berechnung der Ähnlichkeit der grammatischen Struktur (G) müssten folglich auch semantische Unterschiede berücksichtigt und entsprechende Kosten definiert werden. Ob und inwieweit sich eine Erweiterung von MPRO realisieren lässt, ist jedoch zum jetzigen Zeitpunkt unklar.

Aus den Ergebnissen der Untersuchung des Vorkommnisses relevanter und nicht relevanter Übersetzungseinheiten in spezifischen Match-Wert-Bereichen (siehe Tabelle 40 und 41) lässt sich die Frage ableiten, ob sich der in dieser Arbeit beschriebene Aufwand zur linguistischen Optimierung eines zeichenkettenbasierten kommerziellen TMs lohnt, wenn mittels der linearen Regressionsgleichung[83] der gleiche Effekt durch das SDL-TM erzeugt werden kann wie durch das iMem-TM (die Anzahl an relevanten Matches in den oberen Match-Wert-Bereichen des SDL-TMs nimmt zu, während die Anzahl an nicht relevanten Matches sinkt).

Dennoch kann der Einsatz linguistischen Wissens als nutzbringend erachtet werden, da mit ihm linguistische Phänomene, wie z. B. zerlegte Komposita, beim Auffinden potenziell matchender AS_{iMem} mitberücksichtigt werden. Dies kann dazu führen, dass andere, eventuell relevantere Matches an erster Stelle der Trefferliste stehen als beim Einsatz des nicht linguistisch optimierten TMs. Dieser Effekt könnte eventuell bei einer Weiterentwicklung des iMem-TMs durch das Hinzuschalten weiterer linguistischer Komponenten (wie z. B. der semantischen oder syntaktischen Ebene, sofern das morphosyntaktische Analyseprogramm ausgereift genug ist) verstärkt werden. Denn es muss berücksichtigt werden, dass es sich beim iMem-TM zurzeit um einen Prototypen handelt und nicht alle denkbaren linguistischen Verbesserungen, wie die Verwendung des Merkmals *s* oder eines Syntax-Parsers, umgesetzt

[83] So lassen sich beispielsweise mit der Gleichung $f(x) = 0{,}5x + 50$ die Ergebnisse des SDL-TMs des derzeitigen Match-Wert-Bereiches von 40 %–100 % auf den Match-Wert-Bereich von 70 %–100 % abbilden, wobei die relative Anordnung der Matches beibehalten bleibt. Die Variable x in der Gleichung steht dabei für einen Match-Wert des SDL-TMs.

werden konnten. Für den Übersetzer würde eine Weiterentwicklung wiederum eine mögliche Zeitersparnis bedeuten.

Zudem muss beachtet werden, dass die Evaluierung nur für ein sehr kleines Korpus durchgeführt wurde und die Ergebnisse ausschließlich für dieses Korpus gelten. Ob das iMem-TM bei anderen Korpora analoge Ergebnisse liefert, müsste in weiteren Tests bzw. einer Signifikanzanalyse bewiesen werden.

Die Identifikation linguistischer Unterschiede in spezifischen Match-Wert-Bereichen hat gezeigt, dass die Match-Werte des iMem-TMs umso kleiner sind, je komplexer die Unterschiede zwischen dem AS_{neu} und dem AS_{iMem} werden. Beim Vergleich der durch die beiden TMs berechneten Match-Werte mit dem menschlichen Ähnlichkeitsempfinden für ausgewählte Segmentpaare zwischen Korpus A und AT_{neu} A, für dessen Zweck ein Online-Fragebogen erstellt und verteilt wurde, konnte festgestellt werden, dass der Großteil der Match-Werte des iMem-TMs die Bewertungsstufe *genau richtig* zugeordnet bekam. Dennoch konnte auch die Tendenz beobachtet werden, dass die Match-Werte des iMem-TMs häufig als zu hoch oder viel zu hoch und nur selten als zu niedrig oder viel zu niedrig erachtet wurden.

Die Untersuchung, welche linguistischen Unterschiede mit einem zu hoch oder viel zu hoch befundenen Match-Wert des iMem-TMs einhergingen, ergab, dass insbesondere syntaktische und komplexe Unterschiede zu diesem Ergebnis führen.

Dass sehr viele der relevanten Matches beim iMem-TM einen Match-Wert über 70 % aufweisen, liegt auch in diesem Fall an der Anhebung von L_{halb} auf 70 % mittels des Korrekturfaktors x.

Zum anderen kommen die zu hohen Match-Werte durch die Bedingung zustande, dass Löschungen an einem Stück aus einem AS_{iMem} geringere Kosten zugewiesen werden als Löschungen an mehreren Stellen aus dem AS_{iMem} oder Hinzufügungen in das AS_{iMem}, da ein geringerer Arbeitsaufwand für den Übersetzer im Falle von Löschungen an einem Stück angenommen wird. Diese Annahme scheint jedoch oft nicht auf Zustimmung bei den befragten Personen zu treffen: Die Auswertung des Online-Fragebogens hat gezeigt, dass Segmentlängenunterschiede einen großen Einfluss auf die Bewertung des Match-Wertes haben. Je größer der Segmentlängenunterschied zwischen dem AS_{neu} und dem AS_{TM} bzw. AS_{iMem} war, desto seltener wurde die Bewertungsstufe *genau richtig* ausgewählt. Ein Beispiel dafür ist Frage 42 des Online-Fragebogens (siehe Anhang J), bei der es sich um die Frage mit der größten Diskrepanz zwischen der Bewertungsstufe *genau richtig* (0 Antworten) und *viel zu hoch* (81 Antworten) handelt. Bei dieser Frage existiert ein großer

Segmentlängenunterschied zwischen dem AS_{neu} und dem AS_{TM} bzw. AS_{iMem}, gleichzeitig ist jedoch das AS_{neu} komplett im AS_{TM}/AS_{iMem} enthalten.

Eine nachträgliche Korrektur des iMem-Proximitätsmaßes ist somit notwendig, damit die Match-Werte dem menschlichen Ähnlichkeitsempfinden näher kommen. Derzeit verfügt das iMem-Proximitätsmaß über eine komplexe Struktur: die Gewichtung von 70 % für die Ähnlichkeit der LCS-Struktur (L) sowie der grammatischen Struktur (G) und von 30 % für die Ähnlichkeit der Segmentlänge (S_{SL}); das Produkt von L und G; die Korrekturfaktoren x und κ sowie Kosten für einen Wortartwechsel und für Unterschiede in den morphosyntaktischen Merkmalen.

Die einzelnen Parameter wurden gezielt gewählt. So wurden Vertauschungen von LCS als relevanter eingestuft als ein Unterschied in der Segmentlänge, weswegen die Ähnlichkeit der LCS- und der grammatischen Struktur stärker gewichtet wurde als die Ähnlichkeit der Segmentlänge. Das Produkt aus L und G wird ermittelt – anstatt beispielsweise das arithmetische Mittel –, um identifizierten LCS, die grammatisch nicht übereinstimmen, eine geringere Ähnlichkeit zuzuweisen. Da L sensitiv auf die Segmentlänge reagiert (je kürzer die Segmente sind, desto weniger Ränge werden partitioniert (siehe Tabelle 17), weswegen kurze Segmente einen höheren vorläufigen L-Wert zugewiesen bekommen), wurde der Korrekturfaktor x zur Erhöhung des L-Wertes eingeführt. Mit dem Korrekturfaktor κ werden hingegen Segmente begünstigt, die sich in weniger als zwei Basiswörtern unterscheiden. Die Kosten für einen Wortartwechsel wurden gemäß dem Grad der Bedeutungsverschiebung gewählt: Ein Wechsel in der Interpunktion bewirkt beispielsweise eine geringere Bedeutungsänderung als ein Wechsel von einem Inhaltswort zu einem Funktionswort. Aus diesem Grund werden für den ersten Fall geringere Kosten definiert als für den letzteren. Gleiches gilt für die Kosten für Unterschiede in den morphosyntaktischen Merkmalen: Größere Unterschiede werden mit höheren Kosten bedacht.

Die jeweiligen Parameter und ihre zugewiesenen Werte wurden von der Verfasserin dieser Arbeit nach ausgiebigen Untersuchungen festgelegt. Die daraus resultierenden Match-Werte liegen in einem Bereich zwischen 0 % und 100 %. Eine Änderung der Parameter oder Werte ist möglich, was andere Match-Werte als die zurzeit errechneten zur Folge hätte.

Bei einer Korrektur des iMem-Proximitätsmaßes könnten beispielsweise Segmentlängenunterschiede mehr berücksichtigt werden, indem S_{SL} stärker gewichtet wird als L und G. Ebenso muss die Ausnahmeregelung, bei der pauschal nur 1 % Abzug erteilt wird, wenn das AS_{neu} komplett im AS_{iMem} enthalten ist, überdacht werden. Hier sollte ebenso vielmehr der tatsächliche Segmentlängenunterschied in die Ermittlung des Match-Wertes einfließen.

Die Identifikation des Teilsatzes im AS_{iMem} und damit einhergehend die Identifikation der wiederverwertbaren Teile in der ZS-Entsprechung der Übersetzungseinheit machen einen anscheinend größeren Arbeitsaufwand aus als zunächst angenommen.

Eine weitere Überlegung könnte sein, L_{halb} auf einen niedrigeren Wert als 70 % anzuheben, damit alle durch das iMem-Proximitätsmaß berechneten Match-Werte niedriger ausfallen. Dies hätte jedoch zur Folge, dass weniger relevante AS_{iMem} in den oberen Match-Wert-Bereichen angezeigt würden. Relevante AS_{iMem}, die mit dem jetzigen iMem-Proximitätsmaß Match-Werte von knapp 70 % aufweisen, würden bei einer Verringerung des Korrekturfaktors x voraussichtlich einen Match-Wert unter 70 % zugewiesen bekommen und somit bei einem vordefinierten Schwellenwert für die Trefferanzeige von 70 % nicht mehr angezeigt werden.

Da die Parameter aufeinander aufbauen, liegt eine Datenabhängigkeit vor. Eine Sprachenabhängigkeit besteht hingegen lediglich bei der Definition der Kosten für einen Wortartwechsel und der Kosten für die Unterschiede in den morphosyntaktischen Merkmalen und somit für die Ermittlung des G-Wertes. Sollte das iMem-TM für eine andere Ausgangssprache als Deutsch verwendet werden, müssten zunächst die morphosyntaktische Struktur (d. h. mögliche Wortartwechsel und potenziell analysierbare Merkmale mit den dazugehörigen Werten pro Wortart) der neuen Ausgangssprache identifiziert und eventuell neue Kosten definiert werden. Auf diese Weise ließe sich das iMem-Proximitätsmaß auch auf andere Sprachen übertragen. Da MPRO im iMem-Forschungsprojekt als morphosyntaktisches Analyseprogramm eingesetzt wurde, wurden die aus diesem Programm resultierenden Merkmale und Werte für die Ermittlung des G-Wertes herangezogen. Anstelle von MPRO könnte jedoch auch ein anderes morphosyntaktisches Analyseprogramm herangezogen werden. Auch in diesem Fall müsste zunächst identifiziert werden, welche Merkmale und Werte ausgegeben werden, um daraufhin eventuell die Parameter für die Ermittlung des G-Wertes anpassen zu können. Alle anderen in der Match-Formel verwendeten Parameter sind – ausgenommen der Tatsache, dass die Berechnung des Proximitätsmaßes auf den Basiswörtern beruht, die zunächst analysiert werden müssen – sprachenunabhängig, wodurch zum Teil eine Allgemeingültigkeit erreicht werden kann.

Bei der Bewertung des Nachbearbeitungsaufwandes der besten Matches beider Systeme für ausgewählte ZS_{TM} bzw. ZS_{iMem} zur Erstellung der Übersetzung eines AS_{neu} durch unabhängige Juroren konnte eine häufige Auswahl der Bewertungsstufen 1 (*Kein Nachbearbeitungsaufwand notwendig*), 4 (*Komplette Neuübersetzung erfordert weniger Aufwand als Nachbearbeitung*) und 5 (*Komplette Neuübersetzung notwendig, da kein Match aufgefunden werden*

konnte) für das iMem-TM registriert werden (siehe Kapitel 6.5.4). Diese Ergebnisse wurden dadurch erzielt, dass bei der Vorfilterung der potenziellen Match-Partner des AS_{neu} die Basiswörter miteinander verglichen werden. Dadurch werden einerseits mehr relevante Übersetzungseinheiten aufgefunden als beim bloßen Zeichenkettenvergleich, was wiederum erklärt, dass beim iMem-TM öfter die Bewertungsstufe 1 ausgewählt wurde als beim SDL-TM. Andererseits bewirkt die Vorfilterung jedoch auch, dass gleichzeitig mehr nicht relevante Übersetzungseinheiten durch das iMem-TM ermittelt werden und die ZS_{iMem} somit nicht oder nur in geringem Maße für die Übersetzung des AS_{neu} brauchbar sind. Eine Überarbeitung der Arbeitsschritte zur Vorfilterung wäre folglich notwendig, um den Anteil nicht relevanter AS_{iMem}, die als potenzielle Match-Partner infrage kommen, zu reduzieren.

Problematisch bei dieser Art der Evaluierung ist jedoch, dass die Einschätzung des Nachbearbeitungsaufwandes auf Basis der zielsprachlichen Entsprechungen der Übersetzungseinheiten, das Durchlaufen des iMem-Algorithmus allerdings auf Basis der ausgangssprachlichen Segmente der Übersetzungseinheiten erfolgt. Es kann vorkommen, dass die zielsprachlichen Segmente den Inhalt des ausgangssprachlichen gespeicherten Segmentes nicht vollständig wiedergeben oder dass referentielle Mehrdeutigkeiten (siehe Kapitel 3.1.2.3.1) eine vermeintliche inhaltliche Übereinstimmung zwischen AS_{neu} und dem gespeicherten zielsprachlichen Segment hervorrufen (siehe beispielsweise Anhang K, Frage 14, AS_{neu} im Vergleich zum ZS_{TM}). Ebenso ist die Einschätzung der Juroren stets subjektiv, weswegen die Meinungen, ob viel, wenig oder gar keine Nachbearbeitung des zielsprachlichen Segmentes zur Erstellung des ZS_{neu} erforderlich ist, je nach Juror variiert. Daher wäre es in einer weiterführenden Forschung sinnvoll, auch objektive Kriterien, wie beispielsweise die aufgewendete Nachbearbeitungszeit, zur Messung des Nachbearbeitungsaufwandes heranzuziehen, um ein verlässlicheres Ergebnis zu erzielen. Dazu sollte ebenso eine größere Anzahl an Juroren befragt sowie mehr als 15 Segmentpaare untersucht werden, um ein repräsentativeres Ergebnis hinsichtlich des Nachbearbeitungsaufwandes liefern zu können.

Bezüglich der Untersuchung der Antwortzeit und des Speicherplatzbedarfes sind die Messergebnisse für die Übersetzungspraxis annehmbar (siehe Kapitel 6.5.5 und 6.5.6). Jedoch sollten auch hierbei – ähnlich zur statistischen Auswertung – Tests mit weiteren Korpora durchgeführt werden, um die Antwortzeit und den Speicherplatzbedarf zu präzisieren.

8 Ausblick

Die Erkenntnisse dieser Arbeit können als Anstoß zu weiteren Forschungsbemühungen im Bereich der TM-Technologie angesehen werden.

So könnten weitere Optimierungen auf linguistischer Basis konzipiert werden. Als zusätzliches zu vergleichendes Merkmal könnten semantische Informationen zu den einzelnen Wörtern herangezogen werden. Dazu müsste entweder MPRO dahin gehend optimiert werden, dass semantische Zugehörigkeiten nutzbringend analysiert werden oder Strategien für die semantische Suche (siehe Kapitel 2.2.2) in das iMem-TM implementiert werden, bei denen beispielsweise Übersetzungseinheiten mit Ontologien annotiert werden, die folglich die Grundlage für die Suche im TM darstellen. Ebenso könnte ein Syntax-Parser hinzugeschaltet werden, mit dem linguistisch motivierte Phrasen erkannt werden. Des Weiteren könnten Einträge in Terminologiedatenbanken mithilfe linguistischer Verfahren analysiert und die linguistisch optimierten Ergebnisse über die Terminologieerkennungskomponente des TMs während des Übersetzungsprozesses angezeigt werden. Die Ergebnisse der linguistischen Analyse der im TM gespeicherten AS-Segmente könnten zudem für eine verbesserte Konkordanzsuche verwendet werden.

Wie bereits in Kapitel 7 erörtert, lässt ebenso die Überarbeitung des Proximitätsmaßes Spielraum für weitere Forschungsbemühungen. Dabei wäre es denkbar, Placeables sowie Abzüge für Unterschiede in der Formatierung von Wörtern zusätzlich in die Berechnung des Match-Wertes mit einfließen zu lassen.

Darüber hinaus könnte das iMem-Proximitätsmaß auf weitere Sprachen, die durch MPRO analysierbar sind, ausgeweitet werden. Dazu müssten die für den Vergleich zweier Wörter benötigten morphosyntaktischen Merkmale für jede Sprache definiert und entsprechende Kosten im Falle von Unterschieden in den Werten der Merkmale zugeteilt werden. Die Erweiterung des iMem-Proximitätsmaßes auf weitere analysierbare Sprachen würde wiederum zur Folge haben, dass die Benutzeroberfläche des iMem-Plug-ins zum Auswählen und Erstellen eines iMem-TMs überarbeitet werden müsste, damit auch die zusätzlichen Sprachen als Ausgangssprache auswählbar sind.

Zuletzt könnte die Kennzeichnung eines linguistisch optimierten Matches in der Trefferanzeige des kommerziellen TMs eindeutiger gestaltet werden. Denn die derzeitige Lösung, bei der im Falle eines gemischten Übersetzungsszenarios immer alle gefundenen Matches aus beiden TMs (d. h. aus dem SDL-TM sowie aus dem iMem-TM) in der Trefferanzeige aufgeführt werden, sofern die Match-Werte beider Systeme für dieselbe Übersetzungseinheit

mindestens den voreingestellten Schwellenwert für die Trefferanzeige erreichen, könnte einen mit der Funktionsweise des iMem-Plug-ins nicht vertrauten Anwender irritieren. Demnach könnten beispielsweise linguistisch optimierte Übersetzungseinheiten andersfarbig markiert werden. Auch eine Markierung von Phrasen im ZS_{iMem}, die den Phrasen im AS_{iMem} entsprechen, ist vorstellbar. Für die Zuordnung der Phrasen könnten statistische Verfahren zum Einsatz kommen.

In weiteren Tests mithilfe zusätzlicher Korpora und Fragebogen oder Befragungen anderer Art müsste das modifizierte iMem-TM anschließend evaluiert werden. Ein Folgeprojekt könnte dafür den entsprechenden Rahmen bieten.

Für weitverbreitete Sprachen wie Deutsch, Englisch, Französisch oder Spanisch wäre eine Realisierung eines linguistisch optimierten TMs denkbar, da für diese Sprachen morphosyntaktische Analyseprogramme bzw. genügend Material zur Konzeption eines solchen Programms existieren. Dies kann von wenig verbreiteten Sprachen hingegen nicht behauptet werden. Inwieweit sich die TM-Hersteller Gedanken über eine linguistische Optimierung gemacht haben, bleibt unklar. Das iMem-TM könnte jedoch den TM-Herstellern einen Denkanstoß dahin gehend geben, dass eine linguistische Optimierung nicht immer fest in einem System verankert sein muss, sondern auch in Form eines Plug-ins flexibel, d. h. je nach zu übersetzender Sprache, hinzugeschaltet werden kann. Für weitverbreitete Sprachen würde dies bedeuten, dass das Plug-in aktiviert und eine linguistische Unterstützung des Übersetzungsprozesses durchgeführt werden kann. Bei der Übersetzung wenig verbreiteter Sprachen könnte das Plug-in hingegen deaktiviert werden und der konventionelle zeichenkettenbasierte Übersetzungsprozess erfolgen.

Anhang A

Modifizierte, original beibehaltene und eigenständig hinzugefügte AS_{neu} zur Demonstration des Matchings zeichenkettenbasierter und linguistisch optimierter TMs

	AS_{neu}	**Status**	**Original-Segment**
1	Alternativ können Sie den Rasierer mit der mitgelieferten Reinigungsbürste reinigen:	modifiziert	Alternativ können Sie den Rasierer mit der gelieferten Bürste reinigen:
2	Günstigste Umgebungstemperatur beim Laden:	original beibehalten	
3	Dieses Gerät darf nicht in der Nähe von mit Wasser gefüllten Behältern, z.B. Badewanne, Dusche, Waschbecken, verwendet werden.	modifiziert	Nicht in der Nähe von mit Wasser gefüllten Behältern, z.B. Badewanne, Dusche, Waschbecken, verwenden.
4	Einige praktische Tips	original beibehalten	
5	Reinigung mit Wasser	modifiziert	Reinigen mit Wasser
6	Vergewissern Sie sich, dass das Batteriefach trocken und sauber ist, bevor Sie die Batteriefach-Abdeckung wieder schließen.	original beibehalten	
7	Scherkopf und Klingenblock separat unter fließendem Wasser reinigen.	original beibehalten	
8	Die Scherfolie darf nicht mit der Bürste gereinigt werden, da dieses zu Beschädigungen führen könnte.	original beibehalten	
9	Reinigen unter Wasser	original beibehalten	
10	Die Bürsten werden gereinigt.	hinzugefügt	
11	Das Reinigen der Bürste.	hinzugefügt	
12	Ein Reinigungsbürstchen wird mitgeliefert.	hinzugefügt	
13	Nach der Epilation empfehlen wir die Verwendung einer Feuchtigkeitscreme.	modifiziert	Um die Haut zu entspannen, empfehlen wir die Verwendung einer Feuchtigkeitscreme nach der Epilation.
14	Halten Sie die Haut gestrafft.	modifiziert	Halten Sie die Haut gestrafft (B) und führen Sie das Gerät langsam gegen die Haarwuchsrichtung.
15	Die Reinigung der Bürste.	hinzugefügt	

	AS_{neu}	Status	Original-Segment
16	Die Scherkopf-Innenseite reinigen.	modifiziert	Mit der Bürste den Klingenblock und die Scherkopf-Innenseite reinigen.
17	Wir empfehlen die Verwendung einer Feuchtigkeitscreme nach der Epilation.	modifiziert	Um die Haut zu entspannen, empfehlen wir die Verwendung einer Feuchtigkeitscreme nach der Epilation.
18	Wenn sich der Distanzkamm mit Haaren zusetzt, sollten Sie ihn abnehmen und säubern.	original beibehalten	
19	Die Frau ist nett.	hinzugefügt	
20	Die Maus ist klein.	hinzugefügt	

Anhang B

Modifizierte, original beibehaltene und eigenständig hinzugefügte AS_{TM} zur Demonstration des Matchings zeichenkettenbasierter und linguistisch optimierter TMs

	AS_{TM}	**Status**	**Original-Segment**
1	Alternativ können Sie den Rasierer mit der mitgelieferten Reinigungsbürste reinigen.	modifiziert	Alternativ können Sie den Rasierer mit der mitgelieferten Bürste reinigen:
2	Günstige Umgebungstemperatur beim Laden:	modifiziert	Günstigste Umgebungstemperatur beim Laden:
3	Dieses Gerät darf nicht in der Nähe von mit Wasser gefüllten Behältern, z.B. Dusche, Badewanne, Waschbecken, verwendet werden.	modifiziert	Nicht in der Nähe von mit Wasser gefüllten Behältern, z.B. Badewanne, Dusche, Waschbecken, verwenden.
4	Einige praktische Tips	original beibehalten	
5	Reinigen mit Wasser	original beibehalten	
6	Vergewissern Sie sich, dass das Batteriefach trocken und sauber ist, bevor Sie die Batteriefach-Abdeckung wieder schließen.	original beibehalten	
7	Scherkopf und Klingenblock separat unter fließendes Wasser halten.	modifiziert	Scherkopf und Klingenblock separat unter fließendem Wasser reinigen.
8	Die Scherfolie darf nicht mit der Bürste gereinigt werden, da dieses zu Beschädigungen führen könnte.	modifiziert	Die Scherfolie darf nicht mit der Bürste gereinigt werden, da es die Scherfolie beschädigen könnte.
9	Die Bürste wird gereinigt.	hinzugefügt	
10	Eine Reinigungsbürste wird mitgeliefert.	hinzugefügt	
11	Nach der Epilation empfehlen wir die Verwendung einer Feuchtigkeitscreme.	modifiziert	Um die Haut zu entspannen, empfehlen wir die Verwendung einer Feuchtigkeitscreme nach der Epilation.
12	Straffen sie die Haut.	modifiziert	Halten Sie die Haut gestrafft (B) und führen Sie das Gerät langsam gegen die Haarwuchsrichtung.
13	Die Reinigung der Bürste.	hinzugefügt	
14	Die Innenseite des Scherkopfes reinigen.	modifiziert	Mit der Bürste den Klingenblock und die Innenseite des Scherkopfes reinigen.

	AS_{TM}	**Status**	**Original-Segment**
15	Nach der Epilation empfehlen wir, eine Feuchtigkeitscreme zu verwenden.	modifiziert	Um die Haut zu entspannen, empfehlen wir die Verwendung einer Feuchtigkeitscreme nach der Epilation.
16	Sie sollten ihn abnehmen und säubern, wenn sich der Distanzkamm mit Haaren zusetzt.	modifiziert	Wenn sich der Distanzkamm mit Haaren zusetzt, sollten Sie ihn abnehmen und säubern.
17	Die Frau, die an mir vorbei gegangen ist, ist nett.	hinzugefügt	
18	Die Laus ist klein.	hinzugefügt	
19	Die Mäuse sind klein.	hinzugefügt	

Anhang C

Modifikationen der Segmente des AT_{neu} A:

Original-Segment	Wie modifiziert?	Original-Segment auch im AT_{neu} A enthalten?
Alternativ können Sie den Rasierer mit der gelieferten Bürste reinigen:	Alternativ können Sie den Rasierer mit der mitgelieferten Reinigungsbürste reinigen:	ja
Wenn die Akku-Einheit leer ist, kann der Haarschneider auch direkt über Netzanschluss betrieben werden.	Wenn die Akku-Einheit leer ist, kann die Rasur auch direkt über Netzanschluss erfolgen.	nein
Günstigste Umgebungstemperatur beim Laden: 15 °C bis 35 °C.	Günstige Umgebungstemperatur beim Laden: 15 °C bis 35 °C.	nein
Schäden durch unsachgemäßen Gebrauch (Knickstellen an der Scherfolie, Bruch), normaler Verschleiß (z.B. Schersystem) sowie Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen.	Schäden durch unsachgemäßen Gebrauch, normaler Verschleiß sowie Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen.	nein
Reinigen mit Wasser	Reinigung mit Wasser	nein
Eine gründliche Reinigung des Epilierkopfes vor jeder Anwendung reduziert weitestgehend dieses Infektionsrisiko.	Eine gründliche Reinigung des Epilierkopfes vor jeder Anwendung reduziert weitestgehend diese Infektionsgefahr.	nein
Um die Haut zu entspannen, empfehlen wir die Verwendung einer Feuchtigkeitscreme nach der Epilation.	Nach dem Epilieren empfehlen wir die Verwendung einer Feuchtigkeitscreme. Wir empfehlen die Verwendung einer Feuchtigkeitscreme nach der Epilation. Um die Haut zu entspannen, empfehlen wir, eine Feuchtigkeitscreme nach der Epilation zu verwenden. Um die Haut zu entspannen, empfehlen wir, nach der Epilation eine Feuchtigkeitscreme zu verwenden.	nein
(Sollte der Haarschneider nach dem Einschalten nicht sofort laufen, ca. 1 Minute bei Schalterstellung «off» laden.)	(Sollte der Bartschneider nach dem Einschalten nicht sofort laufen, ca. 1 Minute bei Schalterstellung «off» laden.)	nein

Original-Segment	**Wie modifiziert?**	**Original-Segment auch im AT_{neu} A enthalten?**
Gemäß nationaler oder lokaler Bestimmungen geben Sie leere Batterien zur umweltgerechten Entsorgung beim Handel oder entsprechenden Sammelstellen ab.	Gemäß nationaler oder lokaler Bestimmungen geben Sie die leeren Batterien zur umweltgerechten Entsorgung beim Handel oder entsprechenden Sammelstellen ab. Gemäß nationaler oder lokaler Bestimmungen geben Sie leere Batterien zur umweltgerechten Entsorgung beim Handel oder bei entsprechenden Sammelstellen ab.	nein
Halten Sie die Haut gestrafft (B) und führen Sie das Gerät langsam gegen die Haarwuchsrichtung.	Straffen Sie die Haut (B) und führen Sie das Gerät langsam gegen die Haarwuchsrichtung.	nein
Scherkopf und Klingenblock separat unter fließendem Wasser reinigen.	Scherkopf und Klingenblock separat unter fließendes Wasser halten.	nein
Da die Haut in diesem Bereich nach der Epilation besonders empfindlich ist, sollten Sie keine hautreizenden Substanzen, wie z.B. alkoholhaltige Deodorants, verwenden.	Da die Haut in diesem Bereich nach der Epilation besonders empfindlich ist, sollten Sie die Verwendung von hautreizenden Substanzen, wie z.B. alkoholhaltige Deodorants, vermeiden. Da die Haut in diesem Bereich nach der Epilation besonders empfindlich ist, sollten Sie keine Substanzen, die Hautreizungen hervorrufen können, wie z.B. alkoholhaltige Deodorants, verwenden.	nein
Die Scherfolie darf nicht mit der Bürste gereinigt werden, da dieses zu Beschädigungen führen könnte.	Die Scherfolie nicht mit der Bürste reinigen, da dieses zu Beschädigungen führen könnte. Die Scherfolie darf nicht mit der Bürste gereinigt werden, da es die Scherfolie beschädigen könnte.	nein
Leere Batterien sofort aus dem Rasierer entfernen.	Leere Batterien bitte sofort aus dem Rasierer entfernen.	nein
Prüfen Sie vor Inbetriebnahme, ob die auf dem Transformator angegebene Spannung mit Ihrer Netzspannung übereinstimmt.	Prüfen Sie, ob die Spannungsangabe mit Ihrer Netzspannung übereinstimmt.	nein
Das sind normale Reaktionen, die auch rasch wieder abklingen.	Für die Haarentfernung an der Wurzel sind das normale Reaktionen, die auch rasch wieder abklingen.	nein

Original-Segment	Wie modifiziert?	Original-Segment auch im AT_{neu} A enthalten?
Bitte lesen Sie vor Gebrauch des Gerätes die Gebrauchsanweisung sorgfältig durch und bewahren Sie sie auf.	Bitte lesen Sie die Gebrauchsanweisung vor Gebrauch des Gerätes sorgfältig durch und bewahren Sie sie auf.	nein
Grundsätzlich raten wir aber, das Gerät von Kindern fern zu halten.	Gerät von Kindern fernhalten.	ja
Um eine optimale Epilation zu gewährleisten, halten Sie das Gerät senkrecht (90°) zur Haut und führen Sie es ohne Druck mit der Schalterseite gegen den Haarwuchs.	Um eine optimale Epilation zu gewährleisten, halten Sie das Gerät senkrecht zur Haut und führen Sie es ohne Druck mit der Schalterseite gegen den Haarwuchs.	nein
Die Garantie kann in allen Ländern in Anspruch genommen werden, in denen dieses Braun Gerät von uns autorisiert verkauft wird.	Sie kann in allen Ländern in Anspruch genommen werden, in denen dieses Braun Gerät von uns autorisiert verkauft wird.	nein
Schieben Sie den Distanzkamm (1) auf das Gerät, bis er einrastet.	Schieben Sie den Distanzkamm auf das Gerät, bis er einrastet.	nein
Gerät von der Anschlussleitung trennen, bevor Sie es unter Wasser reinigen.	Trennen Sie das Gerät von der Anschlussleitung, bevor Sie es unter Wasser reinigen.	nein
Das Gerät ist aus hygienischen Gründen nicht zum gemeinsamen Gebrauch mit Dritten gedacht.	Das Gerät ist aus hygienischen Gründen nicht zum gemeinsamen Gebrauch mit anderen Personen gedacht.	nein
Epilation im Achselbereich und in der Bikinizone	Epilation von Achselbereich und Bikinizone	nein
Scherfolie und Klingenblock sind Präzisionsteile, die mit der Zeit verschleißen.	Die Scherfolie und der Klingenblock sind Präzisionsteile, die mit der Zeit verschleißen.	nein
Die Entsorgung kann über den Braun Kundendienst oder lokal verfügbare Rückgabe- und Sammelstellen erfolgen.	Die Entsorgung kann über den Braun Kundendienst erfolgen.	nein
• Nach jedem Gebrauch das Schneidsystem reinigen und ölen.	• Nach jedem Gebrauch das Schneidsystem ölen.	ja
Netzkabel nicht um das Gerät wickeln.	Das Netzkabel darf nicht um das Gerät gewickelt werden.	ja
Wir empfehlen die Verwendung von Rasierschaum oder -gel, wenn Sie es beim Duschen verwenden.	Wir empfehlen die Verwendung von Rasierschaum oder -gel, wenn Sie das Rasiersystem beim Duschen verwenden. Wir empfehlen die Verwendung von Rasierschaum oder -gel, wenn Sie duschen.	nein

Original-Segment	Wie modifiziert?	Original-Segment auch im AT_{neu} A enthalten?
Öffnen Sie das Gehäuse wie auf Seite 68 dargestellt; nehmen Sie die Akku-Einheit heraus und geben Sie sie gemäß nationaler oder lokaler Bestimmungen beim Handel oder entsprechenden Sammelstellen ab.	Öffnen Sie das Gehäuse wie auf Seite 68 dargestellt, nehmen Sie die Akku-Einheit heraus und geben Sie die Akku-Einheit zum Schutz der Umwelt gemäß nationaler oder lokaler Bestimmungen beim Handel oder bei entsprechenden Sammelstellen ab. Öffnen Sie das Gehäuse wie auf Seite 68 dargestellt; nehmen Sie die Akku-Einheit heraus, geben Sie sie zum Schutz der Umwelt gemäß nationaler oder lokaler Bestimmungen beim Handel oder bei entsprechenden Sammelstellen ab.	nein
Führen Sie langsame und kontrollierte Bewegungen aus, zwingen Sie den Haarschneider nicht schneller durch das Haar, als das Gerät schneiden kann.	Zwingen Sie den Haarschneider nicht schneller durch das Haar, als das Gerät schneiden kann.	nein
Den Haarschneider bei Schalterstellung «off» (6) mindestens 8 Stunden am Netz aufladen.	Den Haarschneider bei Schalterstellung «6» mindestens 8 Stunden am Netz aufladen.	nein
Voll geladen kann der Haarschneider je nach Haarstärke ca. 30 Minuten ohne Netzanschluss betrieben werden.	Voll geladen kann der Haarschneider je nach Bartstärke ca. 30 Minuten ohne Netzanschluss betrieben werden.	nein
Verwenden Sie den Rasier-Aufsatz nicht mit beschädigter Scherfolie oder defektem Spezialkabel.	Verwenden Sie den Rasier-Aufsatz nicht mit beschädigter Scherfolie oder defektem Kabel.	nein
Um die Kapazität der Akku-Einheit zu erhalten, sollte das Gerät ca. alle 6 Monate durch regulären Gebrauch entladen werden.	Um die maximale Kapazität der Akku-Einheit zu erhalten, sollte das Gerät ca. alle 6 Monate durch regulären Gebrauch entladen werden.	nein
Gelegentlich Seife verwenden (Flüssigseife auf natürlicher Basis ohne Scheuermittel)	Gelegentlich Flüssigseife (ohne Scheuermittel) verwenden.	nein
Klingenblock (2) gründlich mit dem Bürstchen (7) reinigen (i).	Klingenblock gründlich mit dem Bürstchen reinigen.	nein
Nicht in der Nähe von mit Wasser gefüllten Behältern, z.B. Badewanne, Dusche, Waschbecken, verwenden.	Dieses Gerät darf nicht in der Nähe von mit Wasser gefüllten Behältern, z.B. Dusche, Badewanne, Waschbecken, verwendet werden.	nein
Am besten epilieren Sie beim ersten Mal am Abend, damit eventuelle Hautrötungen über Nacht abklingen können.	Am besten epilieren Sie beim ersten Mal am Abend, so daß eventuelle Hautrötungen über Nacht abklingen können.	nein

Anhang D

Original beibehaltene Segmente des AT_{neu} A:

Wenn die Akkus leer sind, können Sie das Gerät auch direkt über das Spezialkabel vom Netz betreiben.
Unsere Produkte wurden hergestellt, um höchste Ansprüche an Qualität, Funktionalität und Design zu erfüllen.
Einige praktische Tips
Mit der Bürste den Klingenblock und die Innenseite des Scherkopfes reinigen.
Mit der Bürste die Scherkopf-Innenseite reinigen.
Schäden, die auf unsachgemäßen Gebrauch zurückzuführen sind, normaler Verschleiß und Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen.
Schäden durch unsachgemäßen Gebrauch, normaler Verschleiß und Verbrauch sowie Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen.
Reinigen unter Wasser
Eine gründliche Reinigung und Desinfektion des Epilierkopfes vor jeder Anwendung reduziert weitestgehend diese Infektionsgefahr.
Vergewissern Sie sich, dass die Batteriefach-Abdeckung trocken und sauber ist, bevor Sie das Batteriefach wieder schließen.
Falls diese Reaktionen nach 36 Stunden noch anhalten, sollten Sie Ihren Arzt um Rat fragen.
Vermeiden Sie jedoch unmittelbar nach der Haarentfernung die Verwendung von Substanzen, die Hautreizungen hervorrufen können, wie z.B. alkoholhaltige Deodorants.
Mit dem bewährten Silk·épil Epiliersystem wird das Haar an der Wurzel entfernt und die Haut bleibt wochenlang glatt.
Das bewährte Silk·épil Epiliersystem entfernt das Haar an der Wurzel und hält die Haut wochenlang glatt.
Wir wünschen Ihnen mit Ihrem Braun Rasierer viel Freude.
Danach den Rasierer wieder voll aufladen.
Danach 2 Stunden aufladen.
Lesen Sie bitte vor der ersten Anwendung die Gebrauchsanweisung sorgfältig und vollständig durch.
Es ist im Handel oder beim Braun Kundendienst erhältlich.
Durch regelmäßiges Reinigen erhalten Sie eine optimale Rasierleistung.
Durch regelmäßiges Reinigen verbessern Sie die Rasierleistung.
Jede durch Haarentfernung entstandene Kleinstverletzung birgt die Gefahr der Entzündung durch das Eindringen von Bakterien, unter anderem durch das Gleiten des Gerätes über die Haut.
Bei allen Formen der Epilation, bei denen die Haare an den Wurzeln entfernt werden, kann es zu kleinen Hautverletzungen und zum Einwachsen von Haaren kommen.
Schaum gut abspülen und den Rasierer noch einige Sekunden laufen lassen.

Ihr Rasierer ist mit einem Spezialkabel mit integriertem Netzteil für Schutzkleinspannung ausgestattet.
Führen Sie den Haarschneider langsam gegen die Haarwuchsrichtung.
Da die nachwachsenden Härchen zart und weich sind, entstehen keine Stoppeln mehr.
• Rasierer ausschalten.
Vor dem Epilieren sollten Sie den entsprechenden Bereich gründlich reinigen, um Rückstände zu entfernen (z.B. Deodorant) und dann mit einem Handtuch trockentupfen.
Die verbleibende Ladung reicht noch für 2–3 Rasuren.
Nach jedem Gebrauch Netzstecker ziehen und den Epilierkopf reinigen.
Rasierer einschalten und den Scherkopf unter fließendes, warmes Wasser halten.
• Der bewegliche Schwingkopf und die flexiblen Scherfolien sorgen automatisch für eine optimale Anpassung an die Gesichtsform.
Das Gerät nicht längere Zeit direktem Sonnenlicht aussetzen.
Die Unterseite des Scherkopfes leicht ausklopfen (nicht auf die Metallseite klopfen).
Die Unterseite des Scherkopfes leicht ausklopfen.
Der Distanzkamm kann nur bei ausgeschaltetem Bartschneider abgenommen werden.
Wenn sich der Distanzkamm mit Haaren zusetzt, sollten Sie ihn abnehmen und säubern.
Beste Ergebnisse erzielen Sie, wenn Sie die Haut mit der anderen Hand straff ziehen.
Das handgehaltene Teil ist von der Anschlussleitung zu trennen, bevor es im Wasser gereinigt wird.
Es dürfen keine Teile ausgetauscht oder Veränderungen vorgenommen werden, da sonst Stromschlaggefahr besteht.
So halten Sie Ihren Bartschneider in Bestform
Dieses Gerät darf am Ende seiner Lebensdauer nicht mit dem Hausmüll entsorgt werden.
Wenn Sie den bodycruZer regelmäßig mit Wasser reinigen, sollten Sie wöchentlich einen Tropfen Leichtmaschinenöl auf dem Langhaarschneider verteilen.
Drücken Sie die Entriegelungstaste (3), um das Schneidsystem (2) zu öffnen.
Bei Eingriffen durch nicht von uns autorisierte Braun Kundendienstpartner sowie bei Verwendung anderer als Original Braun Ersatzteile erlischt die Garantie.
Im Garantiefall senden Sie das Gerät mit Kaufbeleg bitte an einen autorisierten Braun Kundendienstpartner.
Diese Batterien gewährleisten eine Rasierleistung von ca. 60 Minuten.
Die grüne Ladekontroll-Leuchte (5a) zeigt durch Blinken an, dass das Gerät geladen wird.

Anhang E

Modifikationen der AS-Segmente des Korpus A:

Original-Segment	Wie modifiziert?	Original-Segment auch in Korpus A enthalten?
Alternativ können Sie den Rasierer mit der mitgelieferten Bürste reinigen:	Alternativ können Sie den Rasierer mit dem mitgelieferten Reinigungsbürstchen reinigen: Alternativ können Sie den Rasierer mit der mitgelieferten Reinigungsbürste reinigen. Alternativ können Sie den Rasierer mit dem gelieferten Reinigungsbürstchen reinigen. Alternativ können Sie den Rasierer mit der mitgelieferten Bürste reinigen.	ja
Schäden durch unsachgemäßen Gebrauch (Knickstellen an der Scherfolie, Bruch), normaler Verschleiß (z.B. Schersystem) sowie Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen.	Schäden durch unsachgemäßen Gebrauch, normaler Verschleiß sowie Mängel, die den Wert des Gerätes nur unerheblich beeinflussen. Schäden durch unsachgemäßen Gebrauch, normaler Verschleiß, Verbrauch sowie Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen. Schäden durch unsachgemäßen Gebrauch, normaler Verschleiß sowie Verbrauch und Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen. Schäden durch unsachgemäßen Gebrauch, normaler Verschleiß und Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen.	ja
Mit der Bürste den Klingenblock und die Scherkopf-Innenseite reinigen.	Mit der Bürste die Innenseite des Scherkopfes reinigen.	ja
Um die Haut zu entspannen, empfehlen wir die Verwendung einer Feuchtigkeitscreme nach der Epilation.	Nach der Epilation empfehlen wir die Verwendung einer Feuchtigkeitscreme.	ja

Original-Segment	Wie modifiziert?	Original-Segment auch in Korpus A enthalten?
Eine gründliche Reinigung des Epilierkopfes vor jeder Anwendung reduziert weitestgehend dieses Infektionsrisiko.	Die gründliche Reinigung des Epilierkopfes vor jeder Anwendung reduziert weitestgehend diese Infektionsgefahr.	ja
Gemäß nationaler oder lokaler Bestimmungen geben Sie leere Batterien zur umweltgerechten Entsorgung beim Handel oder entsprechenden Sammelstellen ab.	Geben Sie die Akku-Einheit zum Schutz der Umwelt gemäß nationaler oder lokaler Bestimmungen beim Handel oder bei entsprechenden Sammelstellen ab.	ja
Leere Batterien sofort aus dem Rasierer entfernen.	Leere Batterien bitte unmittelbar aus dem Rasierer entfernen.	ja
Mit dem bewährten Silk·épil Epiliersystem wird das Haar an der Wurzel entfernt und die Haut bleibt wochenlang glatt.	Das bewährte Silk·épil Epiliersystem entfernt das Haar an der Wurzel und die Haut bleibt wochenlang glatt.	nein
Danach wieder voll aufladen.	Dann den Rasierer wieder voll aufladen. Den Rasierer wieder voll aufladen.	ja
Prüfen Sie vor Inbetriebnahme, ob die auf dem Transformator angegebene Spannung mit Ihrer Netzspannung übereinstimmt.	Prüfen Sie, ob die auf dem Transformator angegebene Spannung mit Ihrer Netzspannung übereinstimmt. Prüfen Sie vor Inbetriebnahme, ob die Spannungsangabe mit Ihrer Netzspannung übereinstimmt.	ja
Bitte lesen Sie vor Gebrauch des Gerätes die Gebrauchsanweisung sorgfältig durch und bewahren Sie sie auf.	Lesen Sie bitte vor der ersten Anwendung die Gebrauchsanweisung vollständig und sorgfältig durch.	ja
Um eine optimale Epilation zu gewährleisten, halten Sie das Gerät senkrecht (90°) zur Haut.	Um eine optimale Epilation zu gewährleisten, halten Sie das Gerät senkrecht zur Haut.	nein
Schieben Sie den Distanzkamm (1) auf das Gerät, bis er einrastet.	Schieben Sie den Distanzkamm auf das Gerät.	nein
Gerät von der Anschlussleitung trennen, bevor Sie es unter Wasser reinigen.	Gerät ist von der Anschlussleitung zu trennen, bevor Sie es unter Wasser reinigen.	ja
Durch regelmäßiges Reinigen erhalten Sie eine optimale Rasierleistung.	Reinigen Sie das Gerät regelmäßig, um seine Rasierleistung zu erhalten.	nein
Bei allen Formen der Epilation, bei denen die Haare an den Wurzeln entfernt werden, kann es zu kleinen Hautverletzungen und zum Einwachsen von Haaren kommen.	Bei allen Formen der Epilation an der Wurzel kann es zu kleinen Hautverletzungen und zum Einwachsen von Haaren kommen.	nein
Klingenblock (2) gründlich mit dem Bürstchen (7) reinigen (i).	Klingenblock gründlich mit der Bürste reinigen.	ja

Original-Segment	Wie modifiziert?	Original-Segment auch in Korpus A enthalten?
Epilation im Achselbereich und in der Bikinizone	Epilation von Achselbereich und von Bikinizone 4 Epilation von Achselbereich und Bikinizone	ja
Schaum gut abspülen und den Rasierer noch einige Sekunden laufen lassen.	Den Schaum gut abspülen und den Rasierer noch einige Sekunden laufen lassen.	nein
Führen Sie den Haarschneider langsam gegen die Haarwuchsrichtung.	Führen Sie den Haarschneider vorsichtig gegen die Haarwuchsrichtung.	nein
Vor dem Epilieren sollten Sie den entsprechenden Be-reich gründlich reinigen, um Rückstände zu entfernen (z.B. Deodorant) und dann mit einem Handtuch trockentupfen.	Vor dem Epilieren sollten Sie den entsprechenden Bereich gründlich reinigen, um Rückstände zu entfernen (z.B. Deodorant), und dann mit einem Handtuch trockentupfen.	nein
Die verbleibende Ladung reicht dann noch für 2–3 Rasuren.	Die Ladung reicht noch für 2–3 Rasuren. Die restliche Ladung reicht dann noch für 2–3 Rasuren.	ja
Nach jedem Gebrauch Netzstecker ziehen und den Epilierkopf reinigen.	Nach jedem Gebrauch den Epilierkopf reinigen.	nein
Rasierer einschalten und den Scherkopf unter fließendes, warmes Wasser halten.	Rasierer einschalten und den Scherkopf unter fließendes und warmes Wasser halten. Rasierer einschalten und den Scherkopf unter warmes, fließendes Wasser halten.	nein
• Der bewegliche Schwingkopf und die flexiblen Scherfolien sorgen automatisch für eine optimale Anpassung an die Gesichtsform.	– Der bewegliche Schwingkopf und die flexiblen Scherfolien sorgen automatisch für eine optimale Anpassung an die Gesichtsform. • Der bewegliche Schwingkopf und die flexiblen Scherfolien sorgen automatisch für eine optimale Anpassung an die Gesichtsform	nein
• Nach jedem Gebrauch das Schneidsystem reinigen und ölen.	• Nach jedem Gebrauch das Schneidsystem ölen und reinigen. • Nach jedem Gebrauch das Schneidsystem reinigen.	nein
Grundsätzlich raten wir aber, das Gerät von Kindern fern zu halten.	Das Gerät von Kindern fern zu halten raten wir aber grundsätzlich.	nein
Wenn sich der Distanzkamm mit Haaren zusetzt, sollten Sie ihn abnehmen und säubern.	Sie sollten ihn abnehmen und säubern, wenn sich der Distanzkamm mit Haaren zusetzt.	nein

Original-Segment	Wie modifiziert?	Original-Segment auch in Korpus A enthalten?
Um zu vermeiden, dass sich der Distanzkamm bei längerer Anwendung mit Haaren zusetzt, sollte er zwischendurch ausgeschüttelt und gereinigt werden.	Um zu vermeiden, dass sich der Distanzkamm mit Haaren zusetzt, sollten Sie ihn abnehmen und säubern.	ja
Um eine optimale Epilation zu gewährleisten, halten Sie das Gerät senkrecht (90°) zur Haut und führen Sie es ohne Druck mit der Schalterseite gegen den Haarwuchs.	Halten Sie das Gerät senkrecht zur Haut und führen Sie es ohne Druck mit der Schalterseite gegen den Haarwuchs, um eine optimale Epilation zu gewährleisten.	nein
Öffnen Sie das Gehäuse wie auf Seite 68 dargestellt; nehmen Sie die Akku-Einheit heraus und geben Sie sie gemäß nationaler oder lokaler Bestimmungen beim Handel oder entsprechenden Sammelstellen ab.	Öffnen Sie das Gehäuse wie auf Seite 68 dargestellt, nehmen Sie die Akku-Einheit heraus; geben Sie sie zum Schutz der Umwelt gemäß nationaler oder lokaler Bestimmungen beim Handel oder bei entsprechenden Sammelstellen ab.	nein
Bei allen Formen der Haarentfernung an der Wurzel kann es zum Einwachsen von Haaren oder zu Hautreizungen (z.B. Brennen, Rötungen, Jucken) kommen, abhängig auch von Ihrem jeweiligen Haut- und Haartyp.	Bei allen Formen der Epilation an der Wurzel kann es zu kleinen Hautverletzungen und zum Einwachsen von Haaren kommen, abhängig auch von Ihrem jeweiligen Haut- und Haartyp.	nein
Beste Ergebnisse erzielen Sie, wenn Sie die Haut mit der anderen Hand straff ziehen.	Ziehen Sie die Haut mit der anderen Hand straff.	nein
Den Haarschneider bei Schalterstellung «off» (6) mindestens 8 Stunden am Netz aufladen.	Den Haarschneider bei Schalterstellung «off» mindestens 8 Stunden am Netz aufladen.	nein
Nicht in der Nähe von mit Wasser gefüllten Behältern, z.B. Badewanne, Dusche, Waschbecken, verwenden.	Dieses Gerät nicht in der Nähe von mit Wasser gefüllten Behältern, z.B. Badewanne, Dusche, Waschbecken, verwenden.	ja
Die grüne Ladekontroll-Leuchte (5a) zeigt durch Blinken an, dass das Gerät geladen wird.	Die Kontroll-Leuchte (5a) zeigt durch Blinken an, dass das Gerät geladen wird.	nein

Anhang F

Original beibehaltene AS-Segmente des Korpus A:

Wenn der Akku leer ist, können Sie das Gerät auch direkt über das Spezialkabel vom Netz betreiben.
Günstigste Umgebungstemperatur beim Laden: 15 °C bis 35 °C.
Unsere Produkte werden hergestellt, um höchste Ansprüche an Qualität, Funktionalität und Design zu erfüllen.
Einige praktische Tipps
Reinigen mit Wasser
Vergewissern Sie sich, dass das Batteriefach trocken und sauber ist, bevor Sie die Batteriefach-Abdeckung wieder schließen.
Falls diese Hautreaktionen nach 36 Stunden noch anhalten, sollten Sie Ihren Arzt um Rat fragen.
Halten Sie die Haut gestrafft (B) und führen Sie das Gerät langsam gegen die Haarwuchsrichtung.
Scherkopf und Klingenblock separat unter fließendem Wasser reinigen.
Da die Haut in diesem Bereich nach der Epilation besonders empfindlich ist, sollten Sie keine hautreizenden Substanzen, wie z.B. alkoholhaltige Deodorants, verwenden.
Die Scherfolie darf nicht mit der Bürste gereinigt werden, da dieses zu Beschädigungen führen könnte.
Wir wünschen Ihnen mit Ihrem neuen Braun Rasierer viel Freude.
Das sind normale Reaktionen, die auch rasch wieder abklingen.
Die Garantie kann in allen Ländern in Anspruch genommen werden, in denen dieses Braun Gerät von uns autorisiert verkauft wird.
Im Handel oder beim Braun Kundendienst erhältlich.
Regelmäßiges Reinigen verbessert die Rasierleistung.
Es kann vorkommen, dass sich die Haut durch das Eindringen von Bakterien entzündet (z.B. wenn das Gerät über die Haut gleitet).
Aus hygienischen Gründen sollten Sie das Gerät nicht gemeinsam mit anderen Personen benutzen.
Scherfolie und Klingenblock sind Präzisionsteile, die mit der Zeit verschleißen.
Ihr Rasierer ist mit einem Spezialkabel mit integriertem Netzteil für Sicherheitskleinspannung ausgestattet.
Die Entsorgung kann über den Braun Kundendienst oder lokal verfügbare Rückgabe- und Sammelstellen erfolgen.
Da die nachwachsenden Härchen zart und weich sind, entstehen auch keine Stoppeln mehr.
Rasierer ausschalten.
Wenn die Akku-Einheit leer ist, kann der Haarschneider auch direkt über Netzanschluss betrieben werden.
Durch regelmäßiges Reinigen verbessern Sie die Rasierleistung Ihres Rasierers.
Das Gerät nicht längere Zeit Temperaturen über 50 °C aussetzen.
Die Unterseite des Scherkopfes leicht ausklopfen (Folie befindet sich oben).
Der Distanzkamm kann nur bei Schnittstufe «1» abgenommen werden.
Wickeln Sie das Netzkabel nicht um das Gerät.

Wir empfehlen die Verwendung von Rasierschaum oder -gel, wenn Sie es beim Duschen verwenden.
Halten Sie das Gerät von Kindern fern.
Führen Sie langsame und kontrollierte Bewegungen aus, zwingen Sie den Haarschneider nicht schneller durch das Haar, als das Gerät schneiden kann.
Es dürfen weder Teile ausgetauscht noch Veränderungen vorgenommen werden, da sonst Stromschlaggefahr besteht.
Voll geladen kann der Haarschneider je nach Haarstärke ca. 30 Minuten ohne Netzanschluss betrieben werden.
So halten Sie Ihren Haarschneider in Bestform
Verwenden Sie den Rasier-Aufsatz nicht mit beschädigter Scherfolie oder defektem Spezialkabel.
Aus Umweltschutzgründen darf das Gerät am Ende seiner Lebensdauer nicht mit dem Hausmüll entsorgt werden.
Um die Kapazität der Akku-Einheit zu erhalten, sollte das Gerät ca. alle 6 Monate durch regulären Gebrauch entladen werden.
Gelegentlich Flüssigseife verwenden.
Am besten epilieren Sie beim ersten Mal am Abend, damit eventuelle Hautrötungen über Nacht abklingen können.
Zwingen Sie den Haar- und Bartschneider nicht schneller durch das Haar, als das Gerät schneiden kann.
(Sollte der Haarschneider nach dem Einschalten nicht sofort laufen, ca. 1 Minute bei Schalterstellung «off» laden.)
Gelegentlich Seife verwenden (Flüssigseife auf natürlicher Basis ohne Scheuermittel)
Mit der Bürste den Klingenblock und den inneren Bereich des Rasierkopfes reinigen.
Wir empfehlen, die Klingen der Scherköpfe zweimal pro Jahr oder nach Reinigen unter Wasser mit einem Tropfen Leichtmaschinenöl zu ölen.
Drücken Sie die Entriegelungstaste (3), um den Kontaktbügel (2) zu öffnen.
Bei Eingriffen nicht von uns autorisierter Stellen oder bei Verwendung anderer als original Braun Ersatzteile erlischt die Garantie.
Im Garantiefall senden Sie bitte das vollständige Gerät mit der ausgefüllten Garantiekarte einem unserer autorisierten Servicehändler oder an eine Braun Kundendienststelle.
Eine Volladung reicht ca. 60 Minuten.
Führen Sie das Gerät langsam über Ihre Haut, ohne fest aufzudrücken.
Die Kontrolllampe zeigt an, dass der Rasierer mit Spannung versorgt wird.
Setzen Sie ihn auf den normalen Scherkopf, bis er einrastet.
Seife auf natürlicher Basis (ohne Scheuermittel) darf verwendet werden.
Benutzen Sie das Gerät nie, wenn der Scherkopf beschädigt oder defekt ist.
Gebrauchte Batterien sollten nicht im Haushaltsabfall entsorgt werden.
Bitte entsorgen Sie sie in Batterie-Sammelbehältern, die überall dort zu finden sind wo Batterien verkauft werden, oder bringen Sie sie zu Ihrem Fachhändler zurück.
Setzen Sie den Scherkopf dort an, wo Sie Haare entfernen wollen und fahren Sie langsam gegen die Wuchsrichtung des Haares.
Um ein optimales Ergebnis zu erzielen, können Sie die Haut mit einer Hand glatt ziehen.
Diese Garantie ist in Ländern gültig, in denen dieses Produkt offiziell verkauft wird.
Aus hygienischen Gründen möchten wir Sie bitten, Ihren BodycruZer nicht mit anderen Personen zu teilen.

Für diese Funktion kann Gillette Rasierschaum oder -gel verwendet werden.
Setzen Sie ihn auf den Langhaarschneider wie dargestellt, bis er hörbar einrastet.
Setzen Sie ihn auf den Langhaarschneider wie dargestellt, bis er hörbar einrastet.
Setzen Sie das Rasiersystem (3) auf die gestraffte Haut und rasieren Sie sanft gegen die Haarwuchsrichtung.
Schiebeschalter ausfahren und die Entriegelungstaste (4) drücken, um das benutzte Rasiersystem auszuwerfen.
Sie können nach der Haarentfernung etwas Creme oder Körperlotion auftragen.
Wenn Sie den Scherkopf unter Wasser reinigen, sollten die Scherteile nach jeder Reinigung geschmiert werden.
Verteilen Sie etwas Leichtmaschinenöl oder Vaseline auf der Scherfolie und dem Langhaarschneider.
Zum Lösen der Scherfolie drücken Sie den blauen Kunststoffrahmen nach unten (F).
Die ideale Umgebungstemperatur für das Laden liegt zwischen 15 °C und 35 °C.
Je nach Haartyp und Anwendungsgewohnheiten können Sie das Gerät bis zu 40 Minuten kabellos verwenden.
Dieses Akku-/Netzgerät lässt sich jedoch auch direkt über das Spezialkabel vom Netz betreiben, falls die Akkus leer sind.
Um die maximale Kapazität der Akkus zu erhalten, sollte der Rasierer alle sechs Monate durch Rasieren vollständig entladen werden.
Sollte es jedoch wegen der leeren Akku- Einheit beim Einschalten nicht laufen, reicht 1 Minute Ladezeit (Schalterstellung «0») für den Gebrauch direkt am Stromnetz.
Um die optimale Leistung und Lebensdauer der Akku-Einheit zu erhalten, sollte dieser Lade-/Entladevorgang alle sechs Monate wiederholt werden.
Vollständig geladen können Sie sich ca. 45 Minuten schnurlos rasieren.
Der ideale Temperaturbereich zum Wiederaufladen liegt zwischen 15 °C und 35 °C.
Unmittelbar nach der Benutzung sollte man jedoch mit Parfum oder Deodorant vorsichtig sein.
So halten Sie Ihren Lady Braun style in Top-Form
Der Scherkopf Ihres Lady Braun style ist ein Präzisionsteil.
Dennoch sollten Sie im Interesse der Rohstoff-Rückgewinnung das Gerät am Ende seiner Lebensdauer nicht mit dem Hausmüll entsorgen.
Sollte jedoch der Bartschneider wegen leeren Akkus beim Einschalten nicht laufen, reicht 1 Minute Ladezeit (Schalterstellung «Aus» = «0»), um dann direkt (nach dem Wiedereinschalten des Gerätes) aus dem Netz den Bart zu schneiden.
Die Kontroll-Leuchte leuchtet zu Beginn des Ladens und blinkt mit zunehmender Ladung in immer längeren Abständen.
Alle folgenden Ladezeiten betragen ebenfalls 2 Stunden.
Bei längerer Anwendung Distanz-Kamm zwischendurch mit der Bürste reinigen, um ein Zusetzen mit Haaren zu vermeiden (d).
Im Garantiefall geben Sie bitte das vollständige Gerät mit der ausgefüllten Garantiekarte einem unserer autorisierten Service-Händler oder senden Sie beides an die nächst- gelegene Braun Kundendienststelle.
Halten Sie den Distanzkamm flach auf dem Haar, parallel zur Kopfhaut und führen Sie den Haarschneider langsam gegen die Haarwuchsrichtung.
Nehmen Sie die abgelaufenen Batterien sofort aus dem Gerät.
Die Haut mit der freien Hand straffen, so dass sich die Haare aufstellen.
Das Schwenkelement passt sich automatisch der Hautkontur an.

Zubehörteile (Rasierfolie, Klingenblock) sind beim Händler oder Braun Kundendienst erhältlich.
Der Schwingkopf ist beweglich und paßt sich so selbsttätig innerhalb des Schwenkwinkels den Gesichtsformen an.
Das bewegliche Schersystem sorgt dabei automatisch für eine optimale Anpassung der Doppel-Scherfolie und des Integral-Schneiders an die Gesichtsform.
Dann reicht die Ladung noch für ca. 2 bis 3 Rasuren.
Der bewegliche Scherfolienrahmen passt sich automatisch der Gesichtsform an und sorgt für eine gründliche und sanfte Rasur.
Im Handel oder bei Braun Kundendienststellen ist das Braun Clean&Charge erhältlich.
Für den Betrieb dieses Gerätes benötigen Sie eine 1,5 V Alkali-Mangan-Batterie (Typ Micro, LR03, AM4, AAA, z.B. Duracell).
Setzen Sie die Batterie polrichtig ein.
Wenn Sie das Gerät längere Zeit nicht benutzen, nehmen Sie die Batterie heraus (Auslaufgefahr)
Braun SilkFinish ist mit zwei extra schmalen Scherköpfen unterschiedlicher Länge ausgestattet.
Beide sind einzigartig geformt, um präzise und sicher jedes unerwünschte Härchen an Körper und Gesicht (Augenbrauen, Kinn, Oberlippe etc.) zu entfernen.
Der normale Scherkopf (2) eignet sich hervorragend zum Entfernen und Trimmen von Haaren.
Mit dem Präzisionsscherkopf (6) können Sie gezielt einzelne Härchen entfernen.
• Schutzkappe (1) abziehen.
• Ein-/Aus-Schalter (3) nach oben schieben, um das Gerät einzuschalten.
Um den normalen Scherkopf (2) gegen den Präzisionskopf (6) auszutauschen, drehen Sie ihn um 90° gegen den Uhrzeigersinn und heben ihn ab.
Dann den Präzisionsscherkopf aufsetzen und um 90° im Uhrzeigersinn drehen (b).
Mit Ihrem Braun SilkFinish erhalten Sie einen Trimmerkamm (5).
Er ist ideal zum Kürzen der Augenbrauen oder zum Trimmen von Haaren an anderen Körperstellen auf gleiche Länge:
Achten Sie darauf, dass die Kammseitenfläche flach über die Haut geführt wird.
Dieses Gerät entspricht der EU-Richtlinie EMV 89/336/EWG.
Benutzen Sie das Gerät nicht bei empfindlicher Haut und Hautirritationen.
Dieses Gerät funktioniert mit einer AAA 1,5 V Batterie.
Entfernen Sie die Abdeckung des Batteriefachs (4), indem Sie diese herunterziehen (a).
Legen Sie die Batterie mit (+) und (–) Pol in die angegebene Richtung ein.
Es besteht die Gefahr, dass die Batterie ausläuft.
Ihr Exact Series Precision Trimmer ist hervorragend geeignet für eine gründliche und sichere Haarentfernung am Körper und im Gesicht (z.B. Nacken, Rücken, Augenbrauen).
Zum Gerät gehören zwei Trimmaufsätze, um eine gleichmäßige Länge zu erzielen (z.B. beim Formen der Augenbrauen):
Wir empfehlen Ihnen, mit dem langen Trimmaufsatz zu beginnen.
2 Jahre Garantie gelten für Material- und Herstellungsfehler (ausgenommen der Batterie).
Änderungen vorbehalten.
Der aufladbare Braun BodycruZer ist eine Kombination aus Nassrasierer und Trimmer.
Sie können damit alle Körperpartien unterhalb der Halslinie sicher, sanft, schnell und glatt rasieren oder trimmen – nass oder trocken.
Der BodycruZer ist elektrisch sicher und kann ohne Bedenken im Bad, in der Badewanne und unter der Dusche benutzt werden.

Benutzen Sie den BodycruZer nicht zum Entfernen von Bart- oder Kopfhaar.
Für das präzise Trimmen und Stylen benutzen Sie nur den Langhaarschneider (2) (Abb.a).
Für ein optimales Trimmergebnis benutzen Sie alle Trimming-Aufsätze nur bei trockenem Haar (weder Rasierschaum noch -gel benutzen).
Entfernen Sie angesammelte Haare regelmäßig von den Trimming-Aufsätzen.
Für das Trimmen in empfindlichen Körperzonen sollten Sie für eine bessere Hautschonung den «sensitive» Aufsatz (1a) wählen (Abb.b).
Trimmen der Haarlänge mit dem «medium» oder «long» Aufsatz
Mit den Trimming Aufsätzen (1b/1c) können Sie die Haare auf zwei unterschiedliche Längen kürzen («medium» = 3 mm oder «long» = 5 mm).
Beginnen Sie mit dem Trimming-Aufsatz «long» (= 5 mm) um Übung zu bekommen und führen Sie den Kamm gegen die Haarwuchsrichtung (Abb.c).
Trimmen und rasieren in einem Schritt für eine glatte Rasur in Langhaar-Zonen mit dem Langhaarschneider (2) und Rasiersystem (3)
Die Kombirasur eignet sich bestens für größere Flächen wie Brust/Rücken.
Drücken Sie die «lock»-Taste und schieben Sie den Schalter auf Stellung «trim&shave» (Abb.d).
Der Langhaarschneider (2) richtet zunächst die längeren Haare auf und kürzt sie.
Schalten Sie das Gerät ein.
Dann folgt das Rasiersystem (3) und sorgt für eine gründliche Rasur.
Bei der Anwendung in sensiblen Zonen empfehlen wir, mit dem Sensitive-Aufsatz (1a) vorzutrimmen und dann mit dem Rasiersystem (3) glatt zu rasieren.
Drücken Sie die «lock»-Taste und schieben Sie den Schalter (5) so weit es geht nach oben auf Stellung «shave» (Abb.e).
Der BodycruZer wird mit einer Schutzhülle ausgeliefert, die auch als Gerätehalter in der Dusche verwendet werden kann.
Bei Verwendung als Schutzhülle setzen Sie den BodycruZer mit dem Langhaarschneider nach innen ein.
Bei Verwendung als Gerätehalter setzen Sie den BodycruZer mit dem Langhaarschneider nach außen ein (9).
Das Wasser gut abschütteln und den BodycruZer trocknen lassen.
Wechseln Sie das Rasiersystem (3), sobald sich der orangefarbene Streifen verfärbt.
Für das optimale Rasurergebnis verwenden Sie M3 POWER Klingen.
Alle MACH3 Klingen passen auf den BodycruZer.
Zur Aufnahme eines neuen Rasiersystems klicken Sie den BodycruZer direkt in eine im Organizer befindliche Ersatzklinge und entnehmen sie wie in (g) dargestellt.
Falls der BodycruZer auf das Rasiersystem gefallen sein sollte, bitte aus Sicherheitsgründen das Rasiersystem austauschen.
Dieses Gerät enthält Akkus.
Elektrische Angaben siehe Bedruckung auf dem Ladeteil.
Als Hersteller übernehmen wir für dieses Gerät – nach Wahl des Käufers zusätzlich zu den gesetzlichen Gewährleistungsansprüchen gegen den Verkäufer – eine Garantie von 2 Jahren ab Kaufdatum.
Innerhalb dieser Garantiezeit beseitigen wir nach unserer Wahl durch Reparatur oder Austausch des Gerätes unentgeltlich alle Mängel, die auf Material- oder Herstellungsfehlern beruhen.
Die Anschrift für Deutschland können Sie kostenlos unter 00800/27 28 64 63 erfragen.
Mit Braun Silk&Soft haben Sie die perfekte Wahl für eine gründliche und zugleich schonende Rasur der Beine sowie des Achsel- und Bikini-Bereichs getroffen.

Die EasyGlide-Fläche erleichtert das Gleiten des Scherkopfs über die Haut und verringert so Hautreizungen.
Der zusätzliche OptiShave-Aufsatz ermöglicht eine besonders gründliche und schonende Rasur der Beine.
Er sorgt für perfekte Gründlichkeit und einen optimalen Haltewinkel, bei dem Scherfolie und Langhaarschneider gleichzeitig die Haut berühren.
Der bewegliche Langhaarschneider passt sich der Hautoberfläche an – er richtet die längeren Haare auf und schneidet sie ab.
Dann folgt die flexible Scherfolie und entfernt alle noch verbliebenen Härchen.
Wenn Sie längere Zeit nicht rasiert haben, nehmen Sie den OptiShave-Aufsatz ab, um längere Haare schneller vorkürzen zu können (B).
Auch für die Rasur des Achselbereichs sollte der OptiShave-Aufsatz abgenommen werden, um auch an schwer zu erreichenden Stellen alle Härchen zu erfassen.
Um Haare auf eine einheitliche Länge (ca. 4 mm) zu kürzen (z.B. Bikini-Bereich), stellen Sie den Langhaarschneider ebenfalls fest und setzen dann den OptiTrim-Aufsatz 2 auf den Scherkopf (C2).
Die neue Scherfolie wird von innen in den Scherkopf eingesetzt.
Um den Klingenblock abzunehmen, drücken und drehen Sie ihn um 90° (G1).
Beim Aufsetzen des neuen Klingenblocks, wieder drücken und um 90° drehen (G2).
Das Spezialkabel sorgt für eine automatische Spannungsanpassung zwischen 100 und 240 Volt.
Zum Trimmen exakter Linien und Konturen, wie z.B. der Bikini-Linie, stellen Sie den Langhaarschneider 3d fest, indem Sie den «trim/shave»-Schalter 5 auf die Position «trim» schieben (C1).
Der praktische Gerätehalter 10 eignet sich hervorragend als Hängehaken beim Aufladen oder auch für das Trocknen der Scherteile.
Wenn Sie das Gerät mit aufgesetztem OptiTrim-Aufsatz und mit der Schalterseite nach innen in den Gerätehalter setzen, rastet es ein und ist für den Transport besonders kompakt.
Prüfen Sie gelegentlich das Netzkabel, ob es Schadstellen aufweist oder ob der Stecker im Gerät locker sitzt.
Sollte dies der Fall sein, ersetzen Sie es aus Sicherheitsgründen sofort durch ein neues Netzkabel.
Dieses Gerät enthält eine Nickel-Hydrid-Akku-Einheit, die frei von umweltbelastenden Schwermetallen ist.
Sollten Sie jedoch die Entsorgung der Akku-Einheit selbst vornehmen wollen, nehmen Sie die Akku-Einheit wie auf Seite 43 dargestellt heraus.
Die komfortable two-way Technik Ihres Lady Braun style macht Sie beim Gebrauch unabhängig vom Stromnetz mit Hilfe der eingebauten aufladbaren Akku-Einheit und ermöglicht auch jederzeit den Betrieb direkt mit dem Netzkabel.
Ihr Lady Braun style passt sich jeder Wechselspannung von 100-240 Volt automatisch an.
Sie können ihn also ohne Umschaltung überall auf der Welt benutzen.
In einigen Ländern werden Sie allerdings einen Adapterstecker benötigen, der in die ortsüblichen Steckdosen passt.
Bei Netzbetrieb (also direktem Netzkabelanschluss an das Stromnetz) ist Ihr Gerät sofort funktionsbereit.
Ihr Lady Braun style ist mit einer Schalter-Sperre ausgestattet.
Sie verhindert ein versehentliches Einschalten.
Der Geräteschalter schaltet sowohl das Scherfolien-System als auch die beiden Langhaarschneider-Systeme ein:

Ein = Schalter nach oben schieben
Aus = Schalter nach unten schieben
Nehmen Sie die Schutzkappe vor Gebrauch ab:
Einfach in der Mitte anfassen und abziehen.
Nach dem Gebrauch sollten Sie die Schutzkappe immer wieder aufsetzen, um die Scherfolie vor Beschädigung zu schützen.
Achten Sie bitte darauf, dass Ihre Haut vor der Haarentfernung trocken ist.
Die Haarentfernung mit dem kombinierten Scherfolien-/Langhaarschneider-System Ihres Lady Braun style ist schnell, gründlich, aber auch sanft und schonend:
Am besten halten Sie den Scherkopf so gegen Ihre Haut, dass etwa die Hälfte der Scherfolie und einer der Langhaarschneider auf der Haut aufliegen.
Den Lady Braun style können Sie auf beiden Langhaarschneider-Seiten verwenden.
Durch die symmetrische Form des Scherkopfes können Sie ihn also mit der einen Seite vorwärts und mit der anderen zurück führen.
Durch die abgerundeten Seiten des Scherkopfes kommen Sie auch an die «schwierigen» Stellen bequem heran – z.B. an den Knien, Knöcheln, Unterarmen und an die Bikini-Linie.
Sie können also das Gerät in jeder beliebigen Richtung benutzen.
Ihre Haare wachsen nicht schneller nach, wenn sie mit dem Lady Braun style entfernt werden.
Nach Gebrauch und nach dem Laden sollten Sie den Lady Braun style im praktischen Beutel aufbewahren.
Darin ist er platzsparend verstaut und gleichzeitig gegen Beschädigung geschützt (a).
Damit könnten Sie sich verletzen.
Sollte die Scherfolie oder das Kabel einmal beschädigt sein, wird das Gerät bei Ihrem Fachhändler oder beim Braun Kundendienst (siehe Anhang) wieder instand gesetzt.
Nehmen Sie dazu bitte den Rasierkopf ab und tragen Sie – wie in Abb.(d) gezeigt – eine winzige Menge Vaseline auf.
Dieses Gerät entspricht dem EMV-Gesetz (EG-Richtlinie (89/336/EWG) sowie der Niederspannungsrichtlinie (73/23 EWG).
Das Gerät kann ohne Spannungsumschaltung an allen internationalen Wechselspannungsnetzen betrieben werden.
Schnelladung für eine Rasur:
Diese Akku-Einheit ist frei von giftigen Schwermetallen.
Nach beendeter Ladung das Gerät vom Netz trennen (Netzstecker ziehen).
Das Gerät soll nicht mehrere Wochen am Netz angeschlossen sein.
Gelegentliches Überladen von mehreren Tagen schadet jedoch nicht.
Distanz-Kamm auf den ausgeschalteten Bartschneider setzen (Schalter in Stellung «0») und auf die gewünschte Bartlänge einstellen.
Die Bartlänge kann leicht mit einer Hand eingestellt werden (a).
Es stehen 6 Schnittstufen (2, 5, 8,11,14, 18 mm) zur Verfügung; Schalterstellung «6» für die längste Bartlänge, Schalterstellung «1 » für die kürzeste Bartlänge.
Es empfiehlt sich, mit Schalterstellung «6» zu beginnen und dann stufenweise bis zur gewünschten Bartlänge zu kürzen.
Distanz-Kamm mit der abgeschrägten Seite an den Bart ansetzen (b).
Der Distanz-Kamm vermeidet ein Verkanten und übernimmt die Funktion eines Kammes.
Für stärkeres Kürzen bzw. Ausdünnen des Bartes empfiehlt es sich, gegen den Strich zu schneiden.

Memory-Funktion (Speicherung der eingestellten Schnittstufe)
Ihr Bartschneider ist mit einer Memory-Funktion ausgestattet, die verhindert, daß Sie Ihren Bart kürzer als beabsichtigt stutzen, weil Sie irrtümlich eine zu kurze Schnittstufe gewählt haben.
Wenn Sie die für Sie richtige Schnittstufe zwischen 1 und 6 gefunden haben, stellen Sie den «Memory»-Schieber auf diese Schnittstufe ein (a).
Beim Einschalten mit dem Schalter läßt sich das Gerät nicht einschalten, bevor Sie die mit dem «Memory»-Schieber eingestellte Schnittstufe erreicht haben.
Eine längere Schnittstufe können Sie einschalten, ohne den «Memory»-Schieber zu verändern.
Die Memory-Funktion bleibt erhalten, selbst wenn Sie den Distanz-Kamm abnehmen und wieder aufsetzen.
Das Schneidsystem dient zum Konturenschneiden, Entfernen einzelner langer Haare und Formen von Vollbärten (c).
Für diese Anwendungen wird der Bartschneider ohne Distanz-Kamm benutzt und auf die mit dem «Memory»-Schieber fixierte Schnittstufe eingestellt.
Kaufen Sie nur original Braun Teile, insbesondere wenn Sie das Netzkabel erneuern.
Nur so haben Sie die Garantie für eine stets sichere und optimale Rasur.
Die Garantie tritt nur in Kraft, wenn das Kaufdatum durch Stempel und Unterschrift des Händlers auf der Garantiekarte und der Registrierkarte bestätigt ist.
• Nur für den Hausgebrauch.
Bei abgenommenem Distanzkamm darf das Schneidsystem nicht in die Haut gedrückt werden.
Aufladen des Haarschneiders (nur HC 50)
Die Person, deren Haar Sie schneiden wollen, sollte so vor Ihnen sitzen, dass Sie mit Ihren Augen etwa auf Scheitelhöhe sind.
Das Haar sollte sauber, gut gekämmt, entwirrt und trocken sein.
Sie sollten den Haarschneider fest, aber nicht angespannt in der Hand halten.
Wenn Sie das Haareschneiden noch nicht gewöhnt sind, beginnen Sie mit einer längeren Einstellung der Schnittstufe und reduzieren Sie sie später.
So schneiden Sie das Haar nicht versehentlich kürzer als gewünscht.
Schütteln oder blasen Sie zwischendurch die Haare vom Gerät.
Um Ihren Fortschritt während des Schneidens zu kontrollieren, sollten Sie das Haar hin und wieder in die Richtung der gewünschten Frisur kämmen.
Der Schnittstufen-Schalter (5) muss dazu auf Stufe «1» stehen.
Den Schnittstufen-Schalter stellen Sie auf die gewünschte Schnittstufe (4), indem Sie ihn gedrückt halten und nach oben schieben.
Um mit dem Schneiden zu beginnen, schieben Sie den Ein-/Ausschalter (6) auf «on».
Konturenschneiden und Schneiden ohne Distanzkamm
Gerät ausschließlich parallel zur Haut führen (A, Haare schneiden).
Nur zum Konturenschneiden das Schneidsystem genau senkrecht aufsetzen (B, Konturenschneiden).
Das Gerät auf keinen Fall kippen, solange es die Haut berührt (C).
Vermeiden Sie, bzw. die Person, deren Haare geschnitten werden, ruckartige Bewegungen.
Schneiden mit dem Friseurkamm
Um längeres Haar zu schneiden, halten Sie es mit Hilfe des Friseurkamms (10) hoch und schneiden Sie es mit der Friseurschere (11) oder mit dem Haarschneider ab.
Zum Schutz des Kabels sollte es nach Gebrauch vom Gerät getrennt werden.
Akkupflege (nur HC 50)

Der Haarschneider sollte nicht permanent geladen werden.
Umweltschutz (nur HC 50)
Das Gerät trocken halten.
Kleiner Distanzkamm (schneidet 3 Geschwindigkeitseinstellung Barthaar und kurzes Haar)
Großer Distanzkamm (schneidet Kopfhaar)
100 – 240 V 2 / 50 oder 60 Hz (automatische Anpassung)
Ein/Aus-Schalter mit Einstellung der Schnittstufen
Memory-Schalter (für 6 Schnittstufen)
Bei den drei ersten Ladevorgängen der Akku-Einheit:
Nach Vollladung der Akku-Einheit leuchtet die Kontroll-Leuchte deutlich schwächer.
So verwenden Sie Ihren Bartschneider
Normale Geschwindigkeit für Trimmen und Konturenschneiden
Hohe Geschwindigkeit für starken Bartwuchs oder schwierige Gesichtspartien (z.B. Kinn) und Haareschneiden
Es stehen 6 Schnittstufen zur Verfügung: ca. 1, 3, 6, 9, 12 und 16 mm.
Wenn Sie mit der Bartlängenabstufung vertraut sind, können Sie den Memory-Schalter auf Ihre bevorzugte Schnittstufe einstellen.
Fahren Sie so fort wie unter (A) beschrieben, aber benutzen Sie den größeren Distanzkamm (Schnittlängen ca. 10 / 12 / 15 /18 / 21 / 25 mm).
Wenn Sie das Haar kürzer als 10 mm schneiden wollen, empfehlen wir Ihnen, zuerst mit dem großen Distanzkamm auf 10 mm abzuschneiden und dann den kleinen Distanzkamm für die gewünschte Länge zu nehmen.
Der Konturenschneider dient zum präzisen Schneiden der Bartkonturen, zum Formen von Teilbärten und auch zum Entfernen einzelner langer Haare.
Zum Konturenschneiden muss der Distanzkamm abgenommen werden.
Kippen Sie das Doppel-Schneidsystem, so dass der schmale Konturenschneider in der oberen Position einrastet.
Dieser Haar- und Bartschneider funktioniert mit drei 1,5 Volt Batterien AA, MN 1500, LR 6.
Für die beste Leistung benutzen Sie Alkaline Mangan Batterien (z.B. von Duracell).
Einsetzen der Batterien
Lassen Sie nie die Batterien in dem Haar- und Bartschneider, wenn dieser über eine längere Zeit nicht benutzt wird (Gefahr des Auslaufens).
Setzen Sie die Batterien so wie gekennzeichnet ein (A) und schließen Sie das Batterienfach, indem Sie es einrasten lassen.
Um breitere Konturen zu schneiden, kippen Sie den breiten Bartschneider (3a) nach oben.
Silk·épil SuperSoft Plus ist mit einem hochpräzisen Epilationssystem ausgestattet, das speziell zur schnellen und langanhaltenden Haarentfernung entwickelt wurde.
Zusätzlich sorgt der Relax-System-Aufsatz für einen sanften Epiliervorgang.
Das Epiliergefühl wird durch ein angenehmes Kribbeln überlagert.
Falls Sie Zweifel haben, ob Sie dieses Gerät benutzen sollen, fragen Sie bitte Ihren Arzt.
In folgenden Fällen sollten Sie das Gerät nur nach ärztlichem Rat anwenden:
– bei Ekzemen, Wunden, entzündeten Hautreaktionen wie Follikulitiden («Eiterknötchen») und Krampfadern
– im Bereich von Muttermalen
– bei Schwächung der Abwehrkräfte Ihrer Haut, die auftreten kann bei Diabetes, Schwangerschaft, bei Vorliegen des Raynaud Syndroms

– bei Blutern oder bei Immunschwäche.
• Silk·épil wurde für die Haarentfernung an den Beinen entwickelt, kann aber auch an allen empfindlichen Körperzonen wie Unterarm, Achselbereich oder Bikini-Linie angewendet werden.
• Das laufende Gerät sollte nicht mit anderen Hautpartien (z.B. Wimpern, Kopfhaaren usw.), Kleidern und Schnüren in Kontakt kommen, um jede Verletzungsgefahr, ein Blockieren oder ein Beschädigen des Gerätes zu vermeiden.
Benutzen Sie nur den mitgelieferten 12 V Transformator, Typ PI-41-77 V-3.
Epilierkopf mit Pinzettenwalze
Buchse für den Verbindungsstecker
12 V Transformator mit Netzstecker
Gerätebeschreibung (s. Seite 4)
Relax-System-Aufsatz (nicht bei Modell EE1020)
Dermatologisch kontrollierte Anwendungstests haben gezeigt, daß der Epilierkopf auch im Achselbereich und an der Bikini-Linie eingesetzt werden kann.
Allerdings ist in diesen Bereichen mit erhöhtem Schmerzempfinden zu rechnen, das sich aber mit jeder weiteren Anwendung verringert.
Für diese spezielle Anwendung möchten wir Ihnen folgende Hinweise geben:
So vermeiden Sie Hautirritationen.
Die Härchen sollten nicht länger als 5 mm sein.
Ihre Haut muß trocken und fettfrei sein.
• Verbindungsstecker in die Buchse stecken, Transformator-Stecker ans Netz anschließen.
0 = aus
1 = für die behutsame Epilation
2 = für die schnelle Epilation
Im Achselbereich muß das Gerät dazu in verschiedene Richtungen geführt werden.
Für beste Epilierergebnisse sollten die Rädchen des Relax-System-Aufsatzes immer die Haut berühren.
An knochigen Stellen empfehlen wir, eine niedrigere Schaltstufe zu wählen.
Bei der Anwendung an den Kniekehlen muß das Bein immer gestreckt sein; bei der Anwendung im Achselbereich sollten Sie den Arm nach oben strecken.
Dabei kann die Pinzettenwalze von Hand weitergedreht werden.
Bei Erstanwendung oder wenn längere Zeit nicht epiliert wurde, empfehlen wir, längere Haare zunächst zu rasieren.
Nach 1-2 Wochen sind die nachwachsenden, kürzeren Haare leichter zu epilieren.
Nach dem Baden oder Duschen geht die Haarentfernung leichter.
Haut und Haare müssen jedoch trocken sein.
Um dem vorzubeugen, empfehlen wir die regelmäßige Verwendung eines Massageschwammes (z.B. nach dem Duschen) oder eines Körper-Peelings.
Somit wird die obere Hautschicht entfernt, und feine Haare können an die Oberfläche gelangen.
Im Folgenden möchten wir Sie mit dem Gerät vertraut machen und Ihnen einige nützliche Informationen zur Epilation geben.
Braun Silk·épil EverSoft wurde entwickelt, um die Entfernung unerwünschter Härchen so gründlich, behutsam und leicht wie möglich zu machen.
Der hochpräzise Epilierkopf 1 sorgt durch seine einzigartige Anordnung von Pinzetten und die integrierte Einfädelgeometrie für eine schnelle und effiziente Haarentfernung.
Benutzen Sie nur das mitgelieferte Steckernetzteil.

Die Epilation ist leichter und angenehmer, wenn die Haare die optimale Länge von 2 bis 5 mm haben.
Sind die Haare länger, empfehlen wir, auf diese Länge vorzukürzen.
1.Zum Einschalten Schalter 3 auf Stufe «2» schieben («2» = normale Geschwindigkeit, «1» = reduzierte Geschwindigkeit).
Epilieren Sie von unten nach oben.
Bitte beachten Sie, dass diese Bereiche besonders schmerzempfindlich sind.
Bei wiederholter Anwendung wird das Schmerzempfinden nachlassen.
Der Epilierkopf sorgt mit seiner einzigartigen Anordnung von 40 Pinzetten und SoftLift Tips für noch mehr Gründlichkeit.
Das Relax-System (1a) stimuliert die Haut vor und nach der Epilation und minimiert so den Zupfschmerz.
Zwei zusätzliche Aufsätze werden mitgeliefert:
Der Effizienz-Aufsatz (1b) ermöglicht eine besonders schnelle Epilation, da das Schwenkelement für optimalen Hautkontakt und die ideale Anwendungsposition sorgt.
Der Einsteiger-Aufsatz (1c) (auch für empfindliche Haut) hat einen verkleinerten Epilationsbereich, so dass weniger Härchen gleichzeitig gezupft werden.
Dadurch wird Ihnen der Einstieg in die Epilation leichter und angenehmer gemacht.
Die Achsel-Epilierkappe (8) wurde speziell für die Epilation in empfindlichen Körperzonen wie Achselbereich und Bikinizone entwickelt.
Der Präzisionskopf (9) sorgt für eine präzise Haarentfernung im Gesicht und anderen empfindlichen Körperbereichen.
Sein spitz zulaufendes Design ermöglicht eine bessere Sicht auf die Hautpartie, die Sie epilieren möchten.
Der Rasieraufsatz (10) dient zur schnellen und gründlichen Rasur von Achsel- und Bikinizone.
Verwenden Sie den Epilierkopf (2) nie ohne Aufsatz (1).
Wenn Sie bisher noch kein Epiliergerät verwendet haben, oder wenn Sie längere Zeit nicht epiliert haben, kann es eine kurze Zeit dauern, bis sich Ihre Haut an die Epilation gewöhnt hat.
Der zunächst stärker empfundene Zupfschmerz wird bei wiederholter Anwendung deutlich geringer, denn die Zahl der zu entfernenden Haare nimmt ab und die Haut gewöhnt sich an die Epilation.
• Sobald die rote Restkapazitätsleuchte (5b) aufleuchtet, sollten Sie das Gerät über das Spezialkabel ans Netz anschließen um es wieder aufzuladen.
Aus Sicherheitsgründen kann es bei niedrigem Ladezustand vorkommen, dass sich das Gerät ausschaltet und beide Ladeleuchten (grün/rot) blinken.
Stellen Sie stets sicher, dass der Epilierkopf (2) sauber und mit einem Aufsatz (1) versehen ist.
Drücken Sie eine der Freigabetasten (4a) und drehen Sie den Schalter im Uhrzeigersinn auf Stufe 2 (empfohlene Stufe).
Für verringerte Geschwindigkeit wählen Sie die Stufe 1.
Die «smartlight»-Funktion sorgt für ideale Lichtverhältnisse:
Solange das Gerät eingeschaltet ist, leuchtet das Licht und hilft Ihnen, auch feinste Härchen zu entdecken und gründlich zu entfernen.
Achten Sie darauf, dass die Epilierwalze zwischen den Massagerollen des Relax-Systems (1a) immer Hautkontakt hat.
Da die Haare nicht immer in eine einheitliche Richtung wachsen, kann es hilfreich sein, das Gerät in verschiedenen Richtungen über die Haut zu führen, um ein optimales Ergebnis zu erzielen.
Die Massage-Rollen stimulieren und entspannen die Haut mit ihren pulsierenden Bewegungen.

Das macht die Epilation angenehmer.
Wenn Sie bereits mit dem Epilieren vertraut sind und eine schnellere, effizientere Methode bevorzugen, tauschen Sie das Relax-System (1a) gegen den Effizienz-Aufsatz (1b).
Speziell für diese Anwendung wurde die Achsel-Epilierkappe (8) entwickelt.
Mit dem Präzisionskopf (9) entfernt Ihr Silk·épil einfach alle ungewünschten Haare rund um Mund, Kinn oder von anderen empfindlichen Körperstellen.
Für die Reinigung der Pinzettenwalze empfehlen wir, die Bürste mit Alkohol zu benetzen.
Vorkürzen der Haare für die Epilation
Wenn Sie die Haare auf die ideale Länge für die Epilation vorkürzen wollen, setzen Sie den OptiTrim-Kamm (a) auf den Rasierkopf.
Den trim/shave Schalter auf Position «trim» schieben.
Halten Sie das Gerät mit dem OptiTrim-Kamm so, dass die Kammfläche immer flach auf der Haut aufliegt.
Durch unterschiedliche Wuchsrichtungen lassen sich manche Haare nur schwer schneiden.
Führen Sie das Gerät in diesem Fall leicht schräg oder quer zum Haarwuchs.
Beide Teile separat trocknen lassen, bevor der Scherkopf wieder aufgesetzt werden kann.
Der Epilierkopf 2 ist mit 40 Pinzetten ausgestattet, die einzigartig angeordnet sind, um mehr Haare mit einem einzigen Zug zu entfernen – für noch mehr Gründlichkeit.
Die Massage-Rollen des 4-fach Relax-Systems A sollen immer in Kontakt mit der Haut bleiben, damit die pulsierenden Bewegungen die Haut stimulieren und entspannen können und so die Epilation angenehmer machen.
Gelegentlich das Netzkabel auf Schadstellen prüfen und ggf.durch ein neues ersetzen, wenn es im Rasierer zu locker sitzt.
trimmer = Langhaarschneider ist zugeschaltet (zum kontrollierten Trimmen von Schnauzbart und Haaransatz) (c).
combi shave = Kombirasur (gleichzeitige Rasur mit Kurzhaarschneider und Scherfolie).
Der Kurzhaarschneider schneidet den 3-Tage-Bart oder längere «Problemhaare», während die Scherfolie für eine glatte Rasur sorgt (b).
foil shave = Rasur nur mit Scherfolie.
Wir empfehlen, vor dem Waschen zu rasieren, da die Haut nach dem Waschen leicht aufgequollen ist.
Für eine gründlichere Rasur mit weniger Hautreizung sollten Sie beide Teile gleichzeitig auswechseln.
Dieser Rasierer ist mit einem Schwingkopf und Dreifach-Schersystem mit Integral-Schneider ausgestattet.
Das Dreifach-Schersystem mit der Doppel-Scherfolie und dem Integral-Schneider sorgt automatisch für eine permanente Combi-Rasur.
Der Integral-Schneider kürzt längere Haare, die Doppel-Scherfolie mit dem darunterliegenden Duo-Klingenblock rasiert sie glatt – alles in einem Zug.
Sollte der Ladevorgang während des Ladens unterbrochen werden, blinkt das Stecker-Symbol in der Anzeige und fordert so zum Weiterladen auf.
Der Rasierer wurde mit der Einschaltsperre auf Stellung «lock» ausgeliefert.
Einschaltsperre vor der ersten Rasur auf Stellung «•» stellen.
Um ein unbeabsichtigtes Einschalten des Gerätes zu vermeiden (z.B. auf Reisen), die Einschaltsperre auf Stellung «lock» schieben.
Ob ein- oder ausgeschaltet – das Batteriesymbol ist in der grün umrandeten LCD-Anzeige zu sehen.

Der Rasierer ist ein- oder ausgeschaltet, jedoch nicht am Netz angeschlossen.
In der rot umrandeten LCD-Anzeige blinkt das Batteriesymbol – und zeigt so die Restkapazität an – bis die Akku-Einheit völlig leer ist.
Das direkte Rasieren über das Netzkabel bei Spannungen unterhalb 100 V (Auto, Boot usw.) ist nicht möglich.
Zur Rasur an engen Gesichtspartien (z.B. unter der Nase) kann der Schalter auf Schaltstufe «2» geschoben werden, um den Schwingkopf in Winkelstellung zu arretieren.
Der Langhaarschneider ist hoch ausfahrbar und dient zum kontrollierten Konturenschneiden und zum Trimmen von Haaransätzen.
Nur original Braun Teile kaufen.
Beim Triple-Schermagazin auf den Schriftzug achten.
Nur so besteht die Gewähr für eine optimale Rasur.
Zum Laden aus Sicherheitsgründen nur das Braun Ladekabel Nr.5-001-687 verwenden (auf der Packung vermerkt).
Wir stehen Ihnen hier mit unserer Braun Infoline auch für weitere Fragen zu unseren Produkten gern zur Verfügung.
Zwei 1,5 Volt Mignon-Batterien.
Für optimale Leistung Alkali-Mangan-Batterien verwenden (Typ LR 6, AM 3, MN 1500 oder size AA alkaline, z.B. Duracell).
Batterien polrichtig gemäß Markierung einsetzen.
In dieser Position ist der Rasierer automatisch gegen unbeabsichtigtes Einschalten (z.B. auf Reisen) geschützt.
Der Langhaarschneider erlaubt kontrolliertes Trimmen von Schnurrbart und Koteletten.
Langhaarschneider zuschalten:
Wechselanzeige für Scherteile/Reset
(Scherfolie (1) und Klingenblock (2): Ersatzteile-Nr.5000)
Die Wechselanzeige leuchtet noch während der nächsten 7 Rasuren, um Sie an den Scherteile-wechsel zu erinnern.
Danach erfolgt ein automatisches Reset der Anzeige.
Wenn Sie die Scherteile (Scherfolie und Klingenblock) gewechselt haben, drücken Sie die Re-set-Taste (9) mindestens 3 Sekunden lang, um die Wechselanzeige manuell zurückzustellen.
Dabei blinkt die Wechselanzeige zunächst noch und erlischt, sobald das Reset abgeschlossen ist.
Die Wechselanzeige kann zu jeder Zeit manuell zurückgesetzt werden.
Zum kontrollierten Trimmen von Schnauzbart und Haaransatz schieben Sie den ausfahrbaren Langhaarschneider (4) nach oben.
1.Rasieren Sie sich immer, bevor Sie Ihr Gesicht waschen.
Mit den Tasten «sensitive» 6 und «intensive» 7 können Sie Ihren Rasierer auf Ihre persönlichen Bedürfnisse anpassen, die in den verschiedenen Gesichtspartien unterschiedlich sein können.
Die drei möglichen Einstellungen werden mit der Kontrollleuchte im Ein-/Ausschalter 5 ange-zeigt:
• «Intensive» = dunkelblau (mit viel Power bei starkem Bartwuchs)
• «Normal» = hellblau
• «Sensitive» = weiß (für gründliche Rasur auch in empfindlichen Bereichen des Gesichts und Nackens)
Für die gründlichste und schnellste Rasur empfehlen wir die Einstellung «intensive».

Durch Drücken der « + » oder « – » Taste können Sie Ihre bevorzugte Einstellung wählen.
Beim Wiedereinschalten ist automatisch die zuletzt gewählte Einstellung aktiv.
• Für die Rasur an engen Gesichtspartien (z.B. unter der Nase) schieben Sie die «lock» Taste 3 nach hinten, um den Schwingkopf in Winkelstellung zu fixieren.
Schutzkappe (1) abnehmen.
Wenn Sie sich einige Tage nicht rasiert haben, können Sie auch den Langhaarschneider (4) verwenden, um längere Haare zunächst vorzukürzen und dann mit der Scherfolie gründlich auszurasieren.
Langhaarschneider (4) ausfahren (a).
Er ist nicht nur für das großflächige Trimmen konzipiert, sondern auch ideal für das Formen und Stylen von Koteletten, Oberlippen- und Teilbärten.
Häufiger Einsatz des Langhaarschneiders kann die Akku-Kapazität herabsetzen.
(Scherfolie und Klingenblock: 2000 Series)
(Der Rasierer und der Netzstecker können sich leicht erwärmen.)
• Danach wird die Aufladung des Rasierers nach jeder Reinigung automatisch im Clean&Charge Reinigungsgerät erfolgen (siehe Abschnitt B).
Wir empfehlen die manuelle Reinigung nur durchzuführen, wenn das Clean&Charge Reinigungsgerät nicht einsetzbar ist (z.B. auf Reisen).
Dieses Gerät enthält Nickel-Cadmium Akkus.
Braun Clean&Charge wurde zum Reinigen, Laden und Aufbewahren Ihres Rasierers entwickelt.
Da die Reinigungsflüssigkeit einen geringen Ölanteil aufweist, werden beim Reinigungsvorgang auch die Schmieranforderungen des Schersystems erfüllt.
• Um ein Auslaufen der Reinigungsflüssigkeit zu vermeiden, achten Sie beim Aufstellen des Geräts auf einen sicheren Stand.
Das Gerät darf mit eingesetzter Kartusche nicht gekippt, nicht heftig bewegt und in keiner Weise transportiert werden.
• Das Gerät sollte weder in Spiegelschränken noch über Heizungen aufbewahrt, noch auf empfindlichen (polierten oder lackierten) Flächen abgestellt werden.
Das Gerät enthält leicht entzündliche Flüssigkeit.
Von Zündquellen fernhalten, in der Nähe des Gerätes nicht rauchen.
Die Reinigungskartusche nicht neu füllen.
Verwenden Sie ausschließlich die original Reinigungskartusche von Braun.
• Lift-Taste (1) drücken, um das Gehäuse-Oberteil anzuheben
• Halten Sie die Kartusche auf einer ebenen festen Unterlage.
Ziehen Sie den Verschluss vorsichtig ab und schieben Sie die Kartusche bis zum Anschlag in das Bodenfach.
Nach dem Drücken der Lift-Taste zum Öffnen des Gehäuses einige Sekunden warten, bevor die gebrauchte Kartusche herausgenommen wird.
Schließen Sie vor dem Entsorgen der gebrauchten Kartusche die Öffnungen mit dem Verschluss der neuen Kartusche, denn die gebrauchte Kartusche enthält verschmutzte Reinigungsflüssigkeit.
Um Platz einzusparen, ist der Sockel des Geräts an der Rückseite verstellbar.
Stellen Sie nach jeder Rasur den ausgeschalteten Rasierer kopfüber in das Gerät.
• Drücken Sie die Start-/Entriegelungs-Taste (3), um den Rasierer anzuschließen und den Prozess zu starten.
(ungefähr 5 Minuten Reinigung, 4 Stunden Trocknen).

• Der laufende Reinigungsprozess sollte nicht unterbrochen werden, weil dann der Rasierer nicht trocken und zur Benutzung ungeeignet ist.
Muss dennoch abgebrochen werden, Start-/Entriegelungs-Taste (3) drücken.
Wenn die Füllstands-Anzeige (4) die «min»-Markierung erreicht hat, reicht die Reinigungsflüssigkeit in der Kartusche noch für ca. 5 Reinigungsvorgänge.
Bei täglicher Verwendung sollte die Kartusche ca. alle 4 Wochen getauscht werden.
• Aus hygienischen Gründen enthält die Reinigungsflüssigkeit Alkohol, der sich nach dem Öffnen der Kartusche langsam verflüchtigt.
Daher sollte eine Kartusche, falls sie nicht täglich verwendet wird, nach ca. 8 Wochen ausgetauscht werden.
• Gehäuse gelegentlich mit einem feuchten Tuch abwischen, insbesondere die Mulde, in der der Rasierer sitzt.
Gerätehalter (nicht in allen Versionen enthalten)
Gerätehalter mit Schrauben und Dübeln so auf einer glatten Wand befestigen, dass die Standfläche für das Clean&Charge waagerecht ist.
Wenn der Akku leer ist, sollte immer der kombinierte Reinigungs-/Ladeprozess «clean» + «charge» gewählt werden (nicht Reinigung «clean» allein).
Reinigungskartusche einlegen
Verwendung des Clean&Charge
Der Schwingkopf darf nicht arretiert werden (Schalter «head lock» ganz nach unten schieben).
Schwingkopf muss beweglich bleiben.
• Wählen Sie mit der Wahltaste (4) die gewünschte Funktion:
«clean» Reinigung (ca. 15 min)
«clean» + «charge» Reinigung/Laden (ca. 75 min)
«charge» Laden (ca. 60 min)
Beim Drücken der Wahltaste leuchten die entsprechenden Leuchten auf.
(Bei Netzrasierern kann nur die Funktion «clean» gewählt werden.)
Wurde keine Funktion angewählt, wird automatisch «clean» gestartet.
• Nach Ablauf eines Prozesses erlischt die entsprechende Leuchte.
Danach lässt die Reinigungswirkung erheblich nach, und die Kartusche muss ausgetauscht werden.
• Wenn die «cartridge»-Leuchte konstant leuchtet, ist die Reinigungsfunktion gesperrt, und es muss eine neue Kartusche eingesetzt werden.
Erst dann kann der nächste Reinigungsprozess gestartet werden.
Ihr Braun Syncro wird mit dem Braun Clean&Charge Reinigungsgerät geliefert (siehe separate Bedienungsanleitung).
Die «Smart Logic» Elektronik in Ihrem Rasierer analysiert Ihr persönliches Rasierprofil und passt daran die Pflege der Akku-Einheit an, um eine optimale Leistung zu gewährleisten.
• Von diesem Zeitpunkt an sorgt «Smart Logic» in Ihrem Rasierer für eine optimale Pflege der Akku-Einheit.
Sollten Sie z.B. den Akku nie komplett entladen, wird dies von der «Smart Logic» Elektronik erkannt und Ihr Akku durch einen kompletten Entlade-/Ladezyklus neu formatiert (etwa alle 6 Monate, nur wenn erforderlich).
Bei Vollladung leuchten alle fünf Ladekontroll-Leuchten (pro Leuchte 20%), vorausgesetzt, der Rasierer ist am Netz angeschlossen oder eingeschaltet.
2 = Rasur mit fixiertem Schwingkopf in Winkelstellung (z.B. Rasur an engen Gesichtspartien wie unter der Nase)

1 = Rasur mit beweglichem Schwingkopf
3 = Der Langhaarschneider ist zugeschaltet (zum kontrollierten Trimmen von Schnauzbart und Haaransatz).
Es sorgt automatisch für alle Reinigungs- und Schmierbedürfnisse Ihres Rasierers.
Schalten Sie den Rasierer ca. 5–10 Sekunden lang ein, damit der Rasierstaub herausrieseln kann.

Anhang G

Für das iMem-Forschungsprojekt relevante MPRO-Merkmale mit ihren Werten und jeweiliger Erläuterung (siehe auch Reinke 2004: 402f., Rösener 2005: 209ff., Maas et al. 2009: 4ff.)

Merkmal	**Wert**	**Erläuterung**
ori	Beliebige Zeichenkette	Originalform des Wortes
ls	Beliebige Zeichenkette	lexikalische Struktur und Basis der Ableitung eines Wortes
nb	*sg* (Singular) *plu* (Plural)	Numerus
g	*m* (maskulin) *f* (feminin) *n* (neutrum)	Genus
case	*nom* (Nominativ) *gen* (Genitiv) *dat* (Dativ) *acc* (Akkusativ)	Kasus
tns	*pres* (Präsens) *past* (Imperfekt)	Tempus
vtyp	*fiv* (finites Verb) *inf* (Infinitiv) *imperativ* (Imperativ) *izu* (Infinitiv mit *zu*) *ptc1* (Partizip 1) *ptc2* (Partizip 2)	Verbtyp
deg	*base* (Positiv) *comp* (Komparativ) *sup* (Superlativ) *aller* (verstärkter Superlativ)	Steigerungsform
ptc	*1* (Partizip 1) *2* (Partizip 2)	adjektivische Partizipialform
c	*noun* (Substantiv) *verb* (Verb) *adj* (Adjektiv in attributiver Verwendung) *adv* (Adjektiv in prädikativer oder adverbialer Verwendung) *vpref* (Verbpräfix) *fromto* (Intervall)	Wortart

Merkmal	**Wert**	**Erläuterung**
c	*z* (Zahl) *sgml* (SGML-Tag) *w* (Funktionswort)	Wortart
sc (nur, wenn c=w)	*adv* (nicht von einem Adjektiv abgeleitetes Adverb) *art* (Artikel) *card* (Kardinalzahl) *cit* (Anführungszeichen) *clause* (gängige Abkürzung) *comma* (Komma) *compar* (komparative Konjunktion) *conj* (Konjunktion) *coord* (koordinierende Konjunktion) *dem* (Demonstrativartikel) *enum* (Aufzählungszeichen) *interr1* (substantivisches Fragepronomen) *interr2* (adjektivisches Fragepronomen) *itj* (Interjektion) *leer* (Auslassungspunkte) *noun* (spezielles Substantiv) *p* (Präposition) *part* (Partikel) *pers* (Personalpronomen) *poss* (Possessivpronomen) *post* (Postposition) *pred* (Prädikativum) *pron* (Pronomen) *punct* (Satzzeichen) *quant* (Mengenangabe) *refl* (Reflexivpronomen) *rel* (Relativpronomen) *relposs* (possessivisches Relativpronomen) *siehe* (Ausdruck des Verweises) *slash* (Schrägstrich) *subj* (subordinierende Konjunktion) *um_zu* (Infinitivkonjunktion) *unknow*n (unbekannte/nicht codierbare Unterwortart) *wh_adv* (Frageadverb) *webadresse* (Internet- oder E-Mail-Adresse) *zu_inf* (infinitivisches *zu*)	Unterwortart

Anhang H

Beispiele für unterschiedliche Analyseergebnisse durch MPRO für die Wortarten *verb*, *adj*, *adv*, *fromto* und *z*

Übereinstimmende Wortart	**Weitere zu vergleichende Merkmale**	**Beispielsegmente (relevantes Wort markiert)**	**MPRO-Analyse des relevanten Wortes**
c=verb	nb, tns, vtyp	Der Mann isst einen Apfel.	{ori=isst,lu=essen,mlu=essen, c=verb,vtyp=fiv,nb=sg,tns= pres,gra=small,ds=essen,ts= isst,t=essen,ls=essen,s=v}
c=verb	nb, vtyp	Iss einen Apfel!	{ori=Iss,lu=essen,mlu=essen, c=verb,vtyp= imperativ,nb=sg, gra=cap,ns=Iss,ds=essen,ts=iß, t=essen,ls=essen,s=v}
c=verb	vtyp, deg	Der Mann wird einen Apfel gegessen haben.	{ori=gegessen,lu=essen,mlu= essen,c=verb,vtyp=ptc2,deg= base,gra=small,ds=essen,ts= gegessen,t=essen,ls=essen,s=v; manner}
c=verb	vtyp	Der Mann wird einen Apfel gegessen haben.	{ori=haben,lu=haben,mlu= haben,c=verb,vtyp=inf,gra= small,ds=haben,ts=haben,t= haben,ls=haben,s=v}
c=adj (Adjektiv in attributiver Verwendung)	nb, g, case, deg, ptc	Der singende Mann	{ori=singende,lu=singend, mlu=singend,c=adj,sc~=p; pron;rel;subj;um_zu,ptc=1, ehead={nb=sg,infl=weak, case=nom,g=m},deg=base, gra=small,ds=singen,ts= singende,t=singend,ls=singen, s=v}
c=adj (Adjektiv in attributiver Verwendung)	nb, g, case, deg	Der schöne Mann	{ori=schöne,lu=schön,mlu= schön,c=adj,sc~=p;pron;rel; subj;um_zu,ehead={nb=sg, infl=weak,case=nom,g=m}, deg=base,gra=small,ds=schön, ts=schöne,t=schön,ls=schön, s=a}
c=adv (Adjektiv in prädikativer/adverbialer Verwendung)	deg, ptc	Der Mann läuft singend.	{ori=singend,lu=singend,mlu= singend,c=adv,ptc=1,deg= base,gra=small,ds=singen,ts= singend,t=singend,ls=singen, s=v}

Übereinstim-mende Wortart	**Weitere zu vergleichende Merkmale**	**Beispielsegmente (relevantes Wort markiert)**	**MPRO-Analyse des relevanten Wortes**
c=adv (Adjektiv in prädikativer/ adverbialer Verwendung)	deg	Der Mann ist schön.	{ori=schön,lu=schön,mlu=schön,c=adv,deg=base,gra=small,ds=schön,ts=schön,t=schön, ls=schön,s=a}
c=fromto	nb, g, case	3-4 Äpfel	{ori=3-4,lu=3-4,mlu=3-4,c=fromto,sc~=p;rel;subj;um_zu,ehead={nb=plu,infl=null,case=acc;gen;nom,g=m},gra=other,ds=3-4, ls=3-4,s=digits}
c=fromto	–	3-4. Äpfel schälen und essen.	{ori=3-4,lu=3-4,mlu=3-4,c=fromto,gra=other,ds=3-4,ls=3-4,s=digits}
c=z	nb, g, case	3 Äpfel	{ori=3,lu=3,mlu=3,c=z,sc~=p;rel;subj;um_zu,ehead={nb=plu,infl=null,case=acc;gen;nom,g=m},gra=digits,ds=3,ls=3,s= month}
c=z	–	3. schälen	{ori=3,lu=3,mlu=3,c=z,gra=digits,ds=3,ls=3,s=month}

Anhang I

Einige linguistische Phänomene, die als relevant oder nicht relevant erachtet wurden:

Relevant:

1. Orthografische Varianten

   AS_{neu}: Einige praktische Tips

   AS_{TM}/AS_{iMem}: Einige praktische Tipps

2. Unterschiedliche Interpunktion, Nummerierung, Aufzählungszeichen etc.

   AS_{neu}: Alternativ können Sie den Rasierer mit der mitgelieferten Reinigungsbürste reinigen:

   AS_{TM}/AS_{iMem}: Alternativ können Sie den Rasierer mit der mitgelieferten Reinigungsbürste reinigen.

3. Änderung des Tempus

   AS_{neu}: Unsere Produkte wurden hergestellt, um höchste Ansprüche an Qualität, Funktionalität und Design zu erfüllen.

   AS_{TM}/AS_{iMem}: Unsere Produkte werden hergestellt, um höchste Ansprüche an Qualität, Funktionalität und Design zu erfüllen.

4. Änderung des Numerus mit Kongruenz des Artikels, Adjektivs etc.

   AS_{neu}: Den Akku entsorgen.

   AS_{TM}/AS_{iMem}: Die Akkus entsorgen.

5. Änderung der Steigerungsform

   AS_{neu}: Günstige Umgebungstemperatur beim Laden: 15 °C bis 35 °C.

   AS_{TM}/AS_{iMem}: Günstigste Umgebungstemperatur beim Laden: 15 °C bis 35 °C.

6. Änderung der Diathese

   AS_{neu}: Mit dem bewährten Silk·épil Epiliersystem wird das Haar an der Wurzel entfernt und die Haut bleibt wochenlang glatt.

   AS_{TM}/AS_{iMem}: Das bewährte Silk·épil Epiliersystem entfernt das Haar an der Wurzel und die Haut bleibt wochenlang glatt.

7. Kompositazerlegung/Kompositabildung

   AS_{neu}: Mit der Bürste die Scherkopf-Innenseite reinigen.

   AS_{TM}/AS_{iMem}: Mit der Bürste die Innenseite des Scherkopfes reinigen.

8. Synonyme Benennungen

   AS_{neu}: Eine gründliche Reinigung des Epilierkopfes vor jeder Anwendung reduziert weitestgehend diese Infektionsgefahr.

   AS_{TM}/AS_{iMem}: Eine gründliche Reinigung des Epilierkopfes vor jeder Anwendung reduziert weitestgehend dieses Infektionsrisiko.

9. Hinzufügungen/Löschungen/Ersetzungen, die den Sinn nicht sonderlich verändern und/ oder trotz derer eine Inhaltsähnlichkeit schnell erkennbar ist; ggf. mit Kongruenz des Artikels, Adjektivs etc.

   AS_{neu}: Mit der Bürste den Klingenblock und die Innenseite des Scherkopfes reinigen.

   AS_{TM}/AS_{iMem}: Mit der Bürste die Innenseite des Scherkopfes reinigen.

   AS_{neu}: Scherkopf und Klingenblock separat unter fließendes Wasser halten.

   AS_{TM}/AS_{iMem}: Scherkopf und Klingenblock separat unter fließendem Wasser reinigen.

10. Vertauschung von Haupt- und Nebensatz und damit verbundener Änderung der Interpunktion und Groß-/Kleinschreibung

    AS_{neu}: Um eine optimale Epilation zu gewährleisten, halten Sie das Gerät senkrecht zur Haut und führen Sie es ohne Druck mit der Schalterseite gegen den Haarwuchs.

    AS_{TM}/AS_{iMem}: Halten Sie das Gerät senkrecht zur Haut und führen Sie es ohne Druck mit der Schalterseite gegen den Haarwuchs, um eine optimale Epilation zu gewährleisten.

11. Vertauschungen von Wörtern/Phrasen

    AS_{neu}: Bitte lesen Sie die Gebrauchsanweisung vor Gebrauch des Gerätes sorgfältig durch und bewahren Sie sie auf.

    AS_{TM}/AS_{iMem}: Bitte lesen Sie vor Gebrauch des Gerätes die Gebrauchsanweisung sorgfältig durch und bewahren Sie sie auf.

12. Antonyme Aussagen, deren ZS_{TM} durch ausschließliche Änderung der Negation bzw. Affirmation (do → don't, off → on etc.) schnell angepasst werden könnte.

    AS_{neu}: Rasierer einschalten.

    (ZS_{neu}: Turn shaver on.)

    AS_{TM}/AS_{iMem}: Rasierer ausschalten.

    (ZS_{TM}: Turn shaver off.)

13. Paraphrasen, bei denen der Inhalt identisch bzw. ähnlich ist, sich die Originalform der Segmente jedoch stark voneinander unterscheidet.

AS_{neu}: Um die Haut zu entspannen, empfehlen wir, eine Feuchtigkeitscreme nach der Epilation zu verwenden.

AS_{TM}/AS_{iMem}: Nach der Epilation empfehlen wir die Verwendung einer Feuchtigkeitscreme.

AS_{neu}: Schieben Sie den Distanzkamm auf das Gerät, bis er einrastet.

AS_{TM}/AS_{iMem}: Setzen Sie ihn auf den normalen Scherkopf, bis er einrastet.

Nicht relevant:

1. Antonyme Aussagen, bei denen die Bearbeitung des ZS_{TM} längere Zeit in Anspruch nehmen würde als die bloße Änderung der Negation bzw. Affirmation.

AS_{neu}: Rasierer einschalten und den Scherkopf unter fließendes und warmes Wasser halten.

AS_{TM}/AS_{iMem}: Rasierer ausschalten.

2. Ein AS_{TM}/AS_{iMem} besitzt zwar viele Wörter des AS_{neu}, jedoch weist es inhaltlich keine Ähnlichkeit mit dem AS_{neu} auf.

AS_{neu}: Rasierer einschalten.

AS_{TM}/AS_{iMem}: Sie können eine längere Schnittstufe beim Rasierer einschalten, ohne den «Memory»-Schieber zu verändern.

3. AS_{neu} und AS_{TM}/AS_{iMem} sind komplett verschieden (sowohl hinsichtlich der Wörter als auch der Bedeutung).

AS_{neu}: Nach jedem Gebrauch Netzstecker ziehen und den Epilierkopf reinigen.

AS_{TM}/AS_{iMem}: Sie können eine längere Schnittstufe beim Rasierer einschalten, ohne den «Memory»-Schieber zu verändern.

Anhang J

Online-Fragebogen: Alle AS_{neu}-AS_{TM}-/AS_{neu}-AS_{iMem}-Paare mit den jeweiligen ermittelten Match-Werten und den Befragungsergebnissen inklusive etwaiger subjektiver Ähnlichkeitswerte und sonstiger Bemerkungen sowie Informationen zur Berufsgruppenzugehörigkeit und Aufgabenbeschreibung

Aufgabenbeschreibung

Liebe(r) Evaluationsteilnehmer(in),

vielen Dank, dass Sie sich dazu bereit erklären, mir bei meiner Forschung zu helfen.

Im Folgenden werden Ihnen 50 Mal zwei Sätze dargeboten, denen zwei Ähnlichkeitswerte zwischen 0 % (überhaupt keine Ähnlichkeit) und 100 % (komplette Übereinstimmung) zugeordnet sind. Bei Satz A handelt es sich um den neu zu übersetzenden Satz. Satz B ist der im Translation Memory gespeicherte Satz.

Ihre Aufgabe besteht darin, die *Bedeutungs*gleichheit bzw. *Bedeutungs*ähnlichkeit der Satzpaare zu beurteilen, d. h., zu bewerten, inwieweit die beiden Sätze das Gleiche/Ähnliches *aussagen*.

Dafür müssen Sie für jedes Satzpaar angeben, inwieweit die dargebotenen Ähnlichkeitswerte mit Ihrem Ähnlichkeitsempfinden übereinstimmen.

Falls Sie Ihre Auswahl begründen oder sogar Ihren subjektiven Ähnlichkeitswert angeben möchten, steht Ihnen das freie Textfeld zur Verfügung.

Ihre Angaben werden nicht personalisiert!

Fragen

Frage 0		
Sind Sie beruflich im Übersetzungsbereich tätig?		
	Anzahl absolut	**Anzahl relativ**
Ja	61	64,9 %
Nein	20	21,3 %
Keine Angabe	13	13,8 %

Frage 1					
Satz A: Wenn die Akkus leer sind, können Sie das Gerät auch direkt über das Spezialkabel vom Netz betreiben. Satz B: Wenn der Akku leer ist, können Sie das Gerät auch direkt über das Spezialkabel vom Netz betreiben.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
91 % (SDL)	3	34	47	8	2
99 % (iMem)	0	0	19	60	15
Subj. Ähnlichkeitswerte:	95 % (2x)				
Weitere Bemerkungen:	• 99 % impliziert für mich, dass nur eine winzige Kleinigkeit beim Format (Punkt, Komma, Fettschrift, Leerzeichen) anders ist, aber keinesfalls der Inhalt. • Einzahl/Mehrzahl, nicht weiter schlimm, aber nicht perfekt. • *Die Akkus* wird umgangssprachlich anders genutzt. • Spielt die Anzahl der Akkus allerdings eine wichtige Rolle, wären beide Werte zu hoch. • Konflikt Sing.-Plural: Pot. Anwender müsste vermuten, dass es mehrere Akkus gibt, daher einmal zu niedrig (91 %) und einmal zu hoch (99 %). • Mengenmäßig geringe Abweichung, aber wichtig, da Singular und Plural geändert sind. • 99 % entsprechen dem subjektiven Empfinden. • Sinn ist anders.				

Frage 2					
Satz A: Günstige Umgebungstemperatur beim Laden: 15 °C bis 35 °C. Satz B: Die ideale Umgebungstemperatur für das Laden liegt zwischen 15 °C und 35 °C.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
45 % (SDL)	36	31	24	3	0
76 % (iMem)	2	22	30	33	7
Subj. Ähnlichkeitswerte:	60 %, 65 %, 85 %				
Weitere Bemerkungen:	• Alles unter 50 % bedeutet für mich, dass ich nur kleine Satzteile weiterverwenden kann. • Einziger wichtiger Unterschied bei *günstig* gegenüber ideal, subjektiv.				

Frage 3					
Satz A: Mit der Bürste die Scherkopf-Innenseite reinigen. Satz B: Mit der Bürste die Innenseite des Scherkopfes reinigen.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
70 % (SDL)	38	44	11	1	0
96 % (iMem)	3	17	53	14	7
Subj. Ähnlichkeitswerte:	–				
Weitere Bemerkungen:	• Würde durch die angeschlossene Terminologiedatenbank aufgedeckt (heute Standard); daher 4 % Abzug ausreichend. • Aussage bleibt identisch. • Semantisch so gut wie gleichrangig. • Formulierung anders, Inhalt identisch.				

Frage 4					
Satz A: Gemäß nationaler oder lokaler Bestimmungen geben Sie die leeren Batterien zur umweltgerechten Entsorgung beim Handel oder entsprechenden Sammelstellen ab. Satz B: Geben Sie die Akku-Einheit zum Schutz der Umwelt gemäß nationaler oder lokaler Bestimmungen beim Handel oder bei entsprechenden Sammelstellen ab.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
51 % (SDL)	25	44	22	3	0
85 % (iMem)	1	10	36	31	16
Subj. Ähnlichkeitswerte:	70 %, 88 %				
Weitere Bemerkungen:	• Zwischen Akkumulator und Batterie besteht ein erheblicher Unterschied. • Ich habe zuerst überlegt, ob 90 % treffender wären, da sich die einzelnen Satzteile abgesehen von der Reihenfolge sehr gleichen. Aber ich finde den Unterschied Batterien/Akku-Einheit zu gravierend. • Aussage (bis evtl. auf Bezeichnung Akkus vs. Batterien) identisch. • Fehlendes *leer*, Batterie ≠ Akku, *Schutz der Umwelt ≠ umweltgerechte Entsorgung*. • Batterien != Akku-Einheit.				

Frage 5					
Satz A: Um die Haut zu entspannen, empfehlen wir, nach der Epilation eine Feuchtigkeitscreme zu verwenden. Satz B: Um die Haut zu entspannen, empfehlen wir die Verwendung einer Feuchtigkeitscreme nach der Epilation.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
66 % (SDL)	48	37	8	1	0
95 % (iMem)	0	14	57	17	6
Subj. Ähnlichkeitswerte:	85 %				
Weitere Bemerkungen:	• Generell kommt es bei diesen TM-Ähnlichkeitswerten ja nicht nur auf den Inhalt an (der hier meist gut getroffen ist), sondern vor allem auch auf Syntax/Grammatik etc. • Aussage bleibt identisch. • Nur syntaktische Variante • Die Bedeutung ist gleich, aber der Übersetzer muss darüber entscheiden, ob bei der Textsorte entsprechend eine verbale oder substantivische Konstruktion sinnvoller ist. Von daher wird der Match-Wert 95 % dem eigentlichen Arbeitsaufwand nicht gerecht.				

Frage 6					
Satz A: Reinigen unter Wasser Satz B: Reinigen mit Wasser					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
77 % (SDL)	9	26	31	25	3
93 % (iMem)	0	1	25	36	32
Subj. Ähnlichkeitswerte:	0 % (2x)				
Weitere Bemerkungen:	• Ein Bootsrumpf wird unter Wasser gereinigt. Das Deck wird mit Wasser gereinigt. Es handelt sich um zwei völlig verschiedene Aussagen. • Deutlich nicht das Gleiche: unter Wasser impliziert einen laufenden Wasserhahn, mit Wasser eher einen feuchten Schwamm o. ä. • Unterschied von unter/mit kann, abhängig vom Kontext, bedeutend sein. • Reinigen unter Wasser impliziert *unter fließendem Wasser*, was bei *Reinigen mit Wasser* nicht der Fall ist. Deshalb wären 93 % nach meinem subjektiven Empfinden zu hoch. • *Reinigen unter Wasser* ist mehrdeutig. • Geringfügiger Unterschied, ob unter Wasser (eingetaucht) oder mit Wasser (Objekt wird überall mit Wasser benetzt). • Syntaktisch ist der Ähnlichkeitswert von 99 % m. E. gerechtfertigt, inhaltlich jedoch nicht. • Semantisch klarer Unterschied sollte auch im Ähnlichkeitswert widergespiegelt werden. • Eine unscheinbare Präposition, die einen großen Unterschied macht.				

Frage 7					
Satz A: Durch regelmäßiges Reinigen verbessern Sie die Rasierleistung. Satz B: Regelmäßiges Reinigen verbessert die Rasierleistung.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
69 % (SDL)	33	48	12	1	0
92 % (iMem)	0	19	44	26	5
Subj. Ähnlichkeitswerte:	95 %				
Weitere Bemerkungen:	• Aussage identisch • Nur syntaktische Variante				

Frage 8					
Satz A: • Der bewegliche Schwingkopf und die flexiblen Scherfolien sorgen automatisch für eine optimale Anpassung an die Gesichtsform. Satz B: Der bewegliche Scherfolienrahmen passt sich automatisch der Gesichtsform an und sorgt für eine gründliche und sanfte Rasur.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
30 % (SDL)	14	36	35	9	0
76 % (iMem)	0	3	15	41	35
Subj. Ähnlichkeitswerte:	30 %, 50 %, 55 %, 60 %, 65 %–70 %				
Weitere Bemerkungen:	• Die Aussagen ähneln sich inhaltlich fast gar nicht. Auf der Wortebene gibt es ein paar Überschneidungen. Satz A: Zwei Parameter bewirken die Anpassung ans Gesicht. Satz B: Ein Parameter bewirkt die Anpassung ans Gesicht und die Folge davon ist die tolle Rasur. • Der Schwingkopf kommt gar nicht vor, dennoch wird etwas Ähnliches ausgesagt. • Viel kann man aus dem oberen Satz nicht verwerten, dann braucht man ihn eigentlich auch nicht als Vorschlag aus dem TM. • Kleine Zusatzinfo in Satz B, nicht in Satz A enthalten (gründliche und sanfte Rasur) und umgekehrt (beweglicher Schwingkopf und flexible Folie vs. bewegliche Scherfolienrahmen → kann als ein Teil aufgefasst werden, sind aber 2).				

Frage 9					
Satz A: Gelegentlich Flüssigseife (ohne Scheuermittel) verwenden. Satz B: Gelegentlich Flüssigseife verwenden.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
63 % (SDL)	14	38	28	10	4
84 % (iMem)	0	11	32	33	18
Subj. Ähnlichkeitswerte:	90 %, 95 %				
Weitere Bemerkungen:	• Einschränkungen sind extrem wichtig und können nicht einfach weggelassen werden. • Wichtiger Zusatz – Änderung muss klar erkennbar sein.				

Frage 10					
Satz A: Klingenblock gründlich mit dem Bürstchen reinigen. Satz B: Klingenblock gründlich mit der Bürste reinigen.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
78 % (SDL)	18	60	15	1	0
99 % (iMem)	0	1	40	48	5
Subj. Ähnlichkeitswerte:	90 %–94 %, 95 %				
Weitere Bemerkungen:	• Der Vorgang ist der gleiche, aber die Mittel zum Reinigen sind verschieden. Aus morphologischer Sicht sind beide Sätze aber fast identisch, da Bürstchen auch auf Bürste zurückgeht. • Diminutiv nur wichtig, wenn mehrere Bürsten versch. Größe enthalten sind und pro Bürste ein Zweck zugeordnet ist – dies ist hier aber nicht erkennbar.				

Frage 11					
Satz A: Führen Sie den Haarschneider langsam gegen die Haarwuchsrichtung. Satz B: Halten Sie die Haut gestrafft (B) und führen Sie das Gerät langsam gegen die Haarwuchsrichtung.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
43 % (SDL)	11	31	37	14	1
83 % (iMem)	0	2	11	34	47
Subj. Ähnlichkeitswerte:	30 %, 50 %, 60 %–70 %, 66 %				
Weitere Bemerkungen:	• Es wird genau die Hälfte der Aussage wiedergegeben, daher subj. Ähnlichkeitswert: 50 % • Wesentliche Zusatzinfo und Verweis auf Abbildung (in Satz B) machen etwas weniger als die Hälfte der Aussage aus.				

Frage 12					
Satz A: Netzkabel nicht um das Gerät wickeln. Satz B: Wickeln Sie das Netzkabel nicht um das Gerät.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
69 % (SDL)	38	42	12	2	0
96 % (iMem)	1	9	55	24	5
Subj. Ähnlichkeitswerte:	80 %, 90 %				
Weitere Bemerkungen:	• Wenn Kunde für die Übersetzung imperativischen Infinitiv wünscht, dann wird dies mit 96 % deutlich genug als Abweichung angegeben. • Je nachdem welcher Imperativ gewünscht ist, muss der Übersetzer trotz Bedeutungsgleichheit den Satz komplett neu schreiben. • Auch wenn es quasi gleich ist – Bildunterschriften und Freitext sollten nicht gleichbehandelt werden.				

Frage 13					
Satz A: Einige praktische Tips Satz B: Einige praktische Tipps					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
90 % (SDL)	20	59	11	3	1
99 % (iMem)	1	8	68	11	6
Subj. Ähnlichkeitswerte:	100 %				
Weitere Bemerkungen:	• Alte vs. neue Rechtschreibung ändert doch nix am Inhalt. • Alte vs. neue Rechtschreibung – nahezu kein Unterschied • Alte Rechtschreibung kann auch mit obsoleten Zielformulierungen einhergehen.				

Frage 14					
Satz A: Vermeiden Sie jedoch unmittelbar nach der Haarentfernung die Verwendung von Substanzen, die Hautreizungen hervorrufen können, wie z.B. alkoholhaltige Deodorants. Satz B: Da die Haut in diesem Bereich nach der Epilation besonders empfindlich ist, sollten Sie keine hautreizenden Substanzen, wie z.B. alkoholhaltige Deodorants, verwenden.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
45 % (SDL)	33	28	26	7	0
71 % (iMem)	2	17	34	28	13
Subj. Ähnlichkeitswerte:	25 %, 85 %–90 %				
Weitere Bemerkungen:	• Ca. 85 bis 90 % als angemessen empfunden, da nur kleine, relativ unwesentliche zusätzliche Erklärung enthalten.				

Frage 15					
Satz A: Bitte lesen Sie die Gebrauchsanweisung vor Gebrauch des Gerätes sorgfältig durch und bewahren Sie sie auf. Satz B: Bitte lesen Sie vor Gebrauch des Gerätes die Gebrauchsanweisung sorgfältig durch und bewahren Sie sie auf.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
75 % (SDL)	31	55	8	0	0
99 % (iMem)	1	6	65	15	7
Subj. Ähnlichkeitswerte:	85 %–90 %, 90 %, 100 % (2x)				
Weitere Bemerkungen:	• Lediglich Reihenfolge/Priorisierung geändert. • 100 %-Match – exactly the same words but just differently ordered!				

Frage 16					
Satz A: Prüfen Sie, ob die Spannungsangabe mit Ihrer Netzspannung übereinstimmt. Satz B: Prüfen Sie, ob die auf dem Transformator angegebene Spannung mit Ihrer Netzspannung übereinstimmt.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
71 % (SDL)	5	44	33	12	0
96 % (iMem)	0	1	14	55	24
Subj. Ähnlichkeitswerte:	90 %				
Weitere Bemerkungen:	• Aber nur ein wenig zu hoch. Ich denke *auf dem Transformator* ist nicht ganz unwichtig. • Nur geringfügige Zusatzinfo (Transformator als Ort der Angabe)				

Frage 17					
Satz A: Wenn die Akkus leer sind, können Sie das Gerät auch direkt über das Spezialkabel vom Netz betreiben. Satz B: Dieses Akku-/Netzgerät lässt sich jedoch auch direkt über das Spezialkabel vom Netz betreiben, falls die Akkus leer sind.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
30 % (SDL)	50	35	8	1	0
85 % (iMem)	0	14	46	20	14
Subj. Ähnlichkeitswerte:	40 %, 50 %–60 %				
Weitere Bemerkungen:	• Satz B ist ziemlich unklar im Vergleich.				

Frage 18					
Satz A: Mit der Bürste den Klingenblock und die Innenseite des Scherkopfes reinigen. Satz B: Mit der Bürste den Klingenblock und die Scherkopf-Innenseite reinigen.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
79 % (SDL)	31	54	9	0	0
97 % (iMem)	1	14	62	14	3
Subj. Ähnlichkeitswerte:	90 %, 99 %				
Weitere Bemerkungen:	–				

Frage 19					
Satz A: Die verbleibende Ladung reicht noch für 2–3 Rasuren. Satz B: Dann reicht die Ladung noch für ca. 2 bis 3 Rasuren.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
58 % (SDL)	43	42	6	3	0
82 % (iMem)	1	22	54	13	4
Subj. Ähnlichkeitswerte:	85 %–94 %, 90 %				
Weitere Bemerkungen:	–				

Frage 20					
Satz A: Wenn sich der Distanzkamm mit Haaren zusetzt, sollten Sie ihn abnehmen und säubern. Satz B: Sie sollten ihn abnehmen und säubern, wenn sich der Distanzkamm mit Haaren zusetzt.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
30 % (SDL)	74	14	6	0	0
99 % (iMem)	1	3	32	37	21
Subj. Ähnlichkeitswerte:	87 %–93 %, 95 %				
Weitere Bemerkungen:	• Satz B: ziemlich unklar. • In Satz B könnte sich *ihn* auf etwas anderes als den Distanzkamm beziehen. • Satzordnung sollte genau überdacht werden, aber semantisch ist kein Unterschied.				

Frage 21					
Satz A: Dieses Gerät darf am Ende seiner Lebensdauer nicht mit dem Hausmüll entsorgt werden. Satz B: Dennoch sollten Sie im Interesse der Rohstoff-Rückgewinnung das Gerät am Ende seiner Lebensdauer nicht mit dem Hausmüll entsorgen.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
48 % (SDL)	21	22	35	15	1
90 % (iMem)	0	1	17	30	46
Subj. Ähnlichkeitswerte:	35 %, 75 %				
Weitere Bemerkungen:	• Darf vs. sollte im Interesse. Daher so 35 %. • Neue Information im 2. Satz				

Frage 22					
Satz A: (Sollte der Bartschneider nach dem Einschalten nicht sofort laufen, ca. 1 Minute bei Schalterstellung «off» laden.) Satz B: Sollte es jedoch wegen der leeren Akku-Einheit beim Einschalten nicht laufen, reicht 1 Minute Ladezeit (Schalterstellung «0») für den Gebrauch direkt am Stromnetz.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
35 % (SDL)	19	28	36	10	1
74 % (iMem)	0	6	19	34	35
Subj. Ähnlichkeitswerte:	20 %, 65 %				
Weitere Bemerkungen:	–				

Frage 23					
Satz A: Die Entsorgung kann über den Braun Kundendienst erfolgen. Satz B: Die Entsorgung kann über den Braun Kundendienst oder lokal verfügbare Rückgabe- und Sammelstellen erfolgen.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
61 % (SDL)	8	21	43	22	0
98 % (iMem)	0	0	5	33	56
Subj. Ähnlichkeitswerte:	85 %				
Weitere Bemerkungen:	–				

Frage 24					
Satz A: Beste Ergebnisse erzielen Sie, wenn Sie die Haut mit der anderen Hand straff ziehen. Satz B: Um ein optimales Ergebnis zu erzielen, können Sie die Haut mit einer Hand glatt ziehen.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
53 % (SDL)	39	39	15	1	0
77 % (iMem)	3	30	40	16	5
Subj. Ähnlichkeitswerte:	85 % (3x)				
Weitere Bemerkungen:	–				

Frage 25					
Satz A: Das handgehaltene Teil ist von der Anschlussleitung zu trennen, bevor es im Wasser gereinigt wird. Satz B: Gerät ist von der Anschlussleitung zu trennen, bevor Sie es unter Wasser reinigen.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
61 % (SDL)	8	34	34	17	1
87 % (iMem)	0	5	29	40	20
Subj. Ähnlichkeitswerte:	85 % (3x)				
Weitere Bemerkungen:	• Satz A: Handgehaltenes Teil = Handset? Klingt gruselig. Und *unter Wasser reinigen* klingt, als bräuchte man einen Taucheranzug und ein tiefes Becken dafür. • Zwei völlig unterschiedliche Aussagen. Reinigen unter Wasser geht nur mit Taucheranzug, reinigen mit Wasser geht immer (über oder unter Wasser).				

Frage 26					
Satz A: Klingenblock gründlich mit dem Bürstchen reinigen. Satz B: Klingenblock (2) gründlich mit dem Bürstchen (7) reinigen (i).					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
69 % (SDL)	37	43	13	1	0
91 % (iMem)	0	23	44	20	7
Subj. Ähnlichkeitswerte:	95 %, 98 % (2x)				
Weitere Bemerkungen:	–				

Frage 27					
Satz A: Wenn Sie den bodycruZer regelmäßig mit Wasser reinigen, sollten Sie wöchentlich einen Tropfen Leichtmaschinenöl auf dem Langhaarschneider verteilen. Satz B: Klingenblock (2) gründlich mit dem Bürstchen (7) reinigen (i).					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
30 % (SDL)	13	28	40	12	1
75 % (iMem)	0	0	13	41	40
Subj. Ähnlichkeitswerte:	60 %				
Weitere Bemerkungen:	–				

Frage 28					
Satz A: Nach dem Epilieren empfehlen wir die Verwendung einer Feuchtigkeitscreme. Satz B: Nach der Epilation empfehlen wir die Verwendung einer Feuchtigkeitscreme.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
85 % (SDL)	12	69	13	0	0
99 % (iMem)	0	6	65	20	3
Subj. Ähnlichkeitswerte:	90 %, 95 %				
Weitere Bemerkungen:	• *Epilieren* and *Epilation* mean exactly the same thing.				

Frage 29					
Satz A: Für die Haarentfernung an der Wurzel sind das normale Reaktionen, die auch rasch wieder abklingen. Satz B: Das sind normale Reaktionen, die auch rasch wieder abklingen.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
59 % (SDL)	11	40	39	4	0
88 % (iMem)	0	3	30	50	11
Subj. Ähnlichkeitswerte:	–				
Weitere Bemerkungen:	–				

Frage 30					
Satz A: Epilation von Achselbereich und Bikinizone Satz B: Epilation im Achselbereich und in der Bikinizone					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
64 % (SDL)	36	51	6	1	0
90 % (iMem)	2	27	49	12	4
Subj. Ähnlichkeitswerte:	92 %, 95 %, 95 %				
Weitere Bemerkungen:	–				

Frage 31					
Satz A: Das Gerät ist aus hygienischen Gründen nicht zum gemeinsamen Gebrauch mit anderen Personen gedacht. Satz B: Aus hygienischen Gründen sollten Sie das Gerät nicht gemeinsam mit anderen Personen benutzen.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
45 % (SDL)	51	36	6	1	0
81 % (iMem)	3	40	33	13	5
Subj. Ähnlichkeitswerte:	90 % (5x)				
Weitere Bemerkungen:	–				

Frage 32					
Satz A: Im Garantiefall senden Sie das Gerät mit Kaufbeleg bitte an einen autorisierten Braun Kundendienstpartner. Satz B: Im Garantiefall senden Sie bitte das vollständige Gerät mit der ausgefüllten Garantiekarte einem unserer autorisierten Servicehändler oder an eine Braun Kundendienststelle.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
52 % (SDL)	10	27	31	26	0
80 % (iMem)	0	2	19	37	36
Subj. Ähnlichkeitswerte:	70 %–78 %				
Weitere Bemerkungen:	–				

Frage 33					
Satz A: Straffen Sie die Haut (B) und führen Sie das Gerät langsam gegen die Haarwuchsrichtung. Satz B: Halten Sie die Haut gestrafft (B) und führen Sie das Gerät langsam gegen die Haarwuchsrichtung.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
91 % (SDL)	2	39	44	9	0
98 % (iMem)	0	4	33	48	9
Subj. Ähnlichkeitswerte:	70 %				
Weitere Bemerkungen:	• Auch wenn der erste Satzteil unterschiedlich ausgedrückt ist, sagt er genau dasselbe aus. • Die Prozentangaben sind so nahe beieinander, dass man beide als richtig bezeichnen kann.				

Frage 34					
Satz A: Um die Haut zu entspannen, empfehlen wir, nach der Epilation eine Feuchtigkeitscreme zu verwenden. Satz B: Nach der Epilation empfehlen wir die Verwendung einer Feuchtigkeitscreme.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
36 % (SDL)	56	30	6	2	0
86 % (iMem)	0	11	47	27	9
Subj. Ähnlichkeitswerte:	70 %, 86 %–90 %, 90 %				
Weitere Bemerkungen:	–				

Frage 35					
Satz A: Es ist im Handel oder beim Braun Kundendienst erhältlich. Satz B: Zubehörteile (Rasierfolie, Klingenblock) sind beim Händler oder Braun Kundendienst erhältlich.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
49 % (SDL)	13	31	31	19	0
76 % (iMem)	0	4	30	27	33
Subj. Ähnlichkeitswerte:	25 %, 80 %, 85 %				
Weitere Bemerkungen:	• Da *Es* kein Plural ist und nicht klar ist, wofür *Es* steht, sind die Sätze nicht besonders bedeutungsähnlich. • Ein entscheidender Teil fehlt im oberen Satz.				

Frage 36					
Satz A: Durch regelmäßiges Reinigen erhalten Sie eine optimale Rasierleistung. Satz B: Durch regelmäßiges Reinigen verbessern Sie die Rasierleistung Ihres Rasierers.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
60 % (SDL)	11	49	22	12	0
81 % (iMem)	0	10	41	28	15
Subj. Ähnlichkeitswerte:	45 %, 70 %, 90 %				
Weitere Bemerkungen:	• Eine Verbesserung muss keine Optimierung erreichen. • *Verbessern* und *erhalten* sind zu unterschiedlich für einen höheren Wert. • Andere Aussage				

Frage 37					
Satz A: Bitte lesen Sie die Gebrauchsanweisung vor Gebrauch des Gerätes sorgfältig durch und bewahren Sie sie auf. Satz B: Lesen Sie bitte vor der ersten Anwendung die Gebrauchsanweisung vollständig und sorgfältig durch.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
38 % (SDL)	37	37	14	6	0
79 % (iMem)	0	4	44	28	18
Subj. Ähnlichkeitswerte:	25 %, 50 % (2x)				
Weitere Bemerkungen:	–				

Frage 38					
Satz A: Um eine optimale Epilation zu gewährleisten, halten Sie das Gerät senkrecht zur Haut und führen Sie es ohne Druck mit der Schalterseite gegen den Haarwuchs. Satz B: Um eine optimale Epilation zu gewährleisten, halten Sie das Gerät senkrecht zur Haut.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
55 % (SDL)	4	15	44	30	1
85 % (iMem)	0	1	7	39	47
Subj. Ähnlichkeitswerte:	40 %, 45 %, 90 %				
Weitere Bemerkungen:	• Mehr als die Hälfte der Aussage ist komplett neu.				

Frage 39					
Satz A: Danach 2 Stunden aufladen. Satz B: Danach wieder voll aufladen.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
66 % (SDL)	6	14	46	26	2
87 % (iMem)	0	2	15	39	38
Subj. Ähnlichkeitswerte:	45 %, 50 %, 55 %				
Weitere Bemerkungen:	• 2 Stunden besagen nicht, dass das Gerät dann voll aufgeladen ist. • Starke Diskrepanz syntaktische/inhaltliche Ähnlichkeit				

Frage 40					
Satz A: Das Netzkabel darf nicht um das Gerät gewickelt werden. Satz B: Wickeln Sie das Netzkabel nicht um das Gerät.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
48 % (SDL)	48	37	8	1	0
94 % (iMem)	0	7	47	29	11
Subj. Ähnlichkeitswerte:	75 %, 90 %, 99 %				
Weitere Bemerkungen:	–				

Frage 41					
Satz A: Drücken Sie die Entriegelungstaste (3), um das Schneidsystem (2) zu öffnen. Satz B: Schiebeschalter ausfahren und die Entriegelungstaste (4) drücken, um das benutzte Rasiersystem auszuwerfen.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
50 % (SDL)	1	19	26	45	3
75 % (iMem)	0	0	14	29	51
Subj. Ähnlichkeitswerte:	30 %, 35 % (2x)				
Weitere Bemerkungen:	• Zwei verschiedene Aussagen • Wesentliche Information unterschlagen.				

Frage 42					
Satz A: Führen Sie den Haarschneider langsam gegen die Haarwuchsrichtung. Satz B: Halten Sie den Distanzkamm flach auf dem Haar, parallel zur Kopfhaut und führen Sie den Haarschneider langsam gegen die Haarwuchsrichtung.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
37 % (SDL)	13	26	45	9	1
99 % (iMem)	0	0	0	13	81
Subj. Ähnlichkeitswerte:	50 % (2x)				
Weitere Bemerkungen:	• Satz A viel unausführlicher • Ich würde eher ca. 50 % sagen, da der erste Satz doppelt so viele Informationen enthält wie der zweite. • About 50 % I would say, first sentence omits first half of second.				

Frage 43					
Satz A: Um die maximale Kapazität der Akku-Einheit zu erhalten, sollte das Gerät ca. alle 6 Monate durch regulären Gebrauch entladen werden. Satz B: Um die optimale Leistung und Lebensdauer der Akku-Einheit zu erhalten, sollte dieser Lade-/Entladevorgang alle sechs Monate wiederholt werden.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
54 % (SDL)	12	35	29	17	1
75 % (iMem)	0	11	35	31	17
Subj. Ähnlichkeitswerte:	40 %				
Weitere Bemerkungen:	–				

Frage 44					
Satz A: Trennen Sie das Gerät von der Anschlussleitung, bevor Sie es unter Wasser reinigen. Satz B: Gerät ist von der Anschlussleitung zu trennen, bevor Sie es unter Wasser reinigen.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
64 % (SDL)	27	54	11	2	0
95 % (iMem)	0	10	57	15	12
Subj. Ähnlichkeitswerte:	50 %				
Weitere Bemerkungen:	–				

Frage 45					
Satz A: • Der bewegliche Schwingkopf und die flexiblen Scherfolien sorgen automatisch für eine optimale Anpassung an die Gesichtsform. Satz B: Das bewegliche Schersystem sorgt dabei automatisch für eine optimale Anpassung der Doppel-Scherfolie und des Integral-Schneiders an die Gesichtsform.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
40 % (SDL)	27	39	18	10	0
84 % (iMem)	0	3	26	40	25
Subj. Ähnlichkeitswerte:	30 %, 60 %, 70 %, 73 %, 75 %, 75 %–80 %				
Weitere Bemerkungen:	• Semantisch gleich, nur anders formuliert.				

Frage 46					
Satz A: Grundsätzlich raten wir aber, das Gerät von Kindern fern zu halten. Satz B: Halten Sie das Gerät von Kindern fern.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
53 % (SDL)	19	43	26	6	0
81 % (iMem)	0	9	41	32	12
Subj. Ähnlichkeitswerte:	70 %–75 %				
Weitere Bemerkungen:	–				

Frage 47					
Satz A: Es ist im Handel oder beim Braun Kundendienst erhältlich. Satz B: Im Handel oder bei Braun Kundendienststellen ist das Braun Clean&Charge erhältlich.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
38 % (SDL)	36	38	18	2	0
90 % (iMem)	0	1	21	43	29
Subj. Ähnlichkeitswerte:	70 %, 85 %				
Weitere Bemerkungen:	• Kommt auf den Satz davor an! Kann 80 % stimmen, kann 95 % stimmen.				

Frage 48					
Satz A: Straffen Sie die Haut (B) und führen Sie das Gerät langsam gegen die Haarwuchsrichtung. Satz B: Setzen Sie das Rasiersystem (3) auf die gestraffte Haut und rasieren Sie sanft gegen die Haarwuchsrichtung.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
46 % (SDL)	23	36	25	10	0
81 % (iMem)	0	10	32	32	20
Subj. Ähnlichkeitswerte:	65 %–70 %, 85 %				
Weitere Bemerkungen:	• Second sentence is much more descriptive.				

Frage 49					
Satz A: Wenn sich der Distanzkamm mit Haaren zusetzt, sollten Sie ihn abnehmen und säubern. Satz B: Um zu vermeiden, dass sich der Distanzkamm bei längerer Anwendung mit Haaren zusetzt, sollte er zwischendurch ausgeschüttelt und gereinigt werden.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
47 % (SDL)	13	28	30	20	3
76 % (iMem)	0	10	22	34	28
Subj. Ähnlichkeitswerte:	35 % (2x), 60 %				
Weitere Bemerkungen:	• Es besteht ein wesentlicher Unterschied zwischen dem Verhindern eines Zustands und der Beseitigung desselben. • Aussage bleibt im Grunde identisch.				

Frage 50					
Satz A: Bei allen Formen der Epilation, bei denen die Haare an den Wurzeln entfernt werden, kann es zu kleinen Hautverletzungen und zum Einwachsen von Haaren kommen. Satz B: Bei allen Formen der Epilation an der Wurzel kann es zu kleinen Hautverletzungen und zum Einwachsen von Haaren kommen.					
	viel zu niedrig	**zu niedrig**	**genau richtig**	**zu hoch**	**viel zu hoch**
74 % (SDL)	12	43	34	5	0
91 % (iMem)	0	11	37	33	13
Subj. Ähnlichkeitswerte:	60 %, 95 %				
Weitere Bemerkungen:	• In Satz A fehlt eine Information. • Für dieses Satzpaar und grundsätzlich: Es ist nicht egal, welcher der beiden Sätze im TM enthalten ist und welcher zu übersetzen ist. Denn ein fairer Match-Wert misst die anfallende Übersetzungsarbeit. Hat man einen längeren Satz im TM und einen kürzeren im zu übersetzenden Text, so kann der Match-Wert 99 % betragen, denn überflüssige Wörter sind schnell gelöscht. Anders rum wäre der Match-Wert 50 % oder noch weniger fair, wenn im zu übersetzenden Satz viele zusätzliche Wörter zu übersetzen sind! Der Match-Wert soll also m. E. die Edit Distance messen, und zwar unidirektional, und nicht wie üblich die pure richtungsunabhängige Token-/Trigrammdifferenz zwischen zwei Sätzen.				

Anhang K

Häufigkeiten der gewählten Bewertungsstufen für die Bewertung des Nachbearbeitungsaufwandes der besten Matches beider Systeme für ausgewählte ZS_{TM} bzw. ZS_{iMem} zur Erstellung der Übersetzung eines AS_{neu}

Frage 1					
AS_{neu}:	Alternativ können Sie den Rasierer mit der mitgelieferten Reinigungsbürste reinigen:				
AS_{TM}:	Alternativ können Sie den Rasierer mit der mitgelieferten Reinigungsbürste reinigen.				
ZS_{TM}:	Alternatively, you may clean the shaver using the cleaning brush provided.				
AS_{iMem}:	Alternativ können Sie den Rasierer mit dem mitgelieferten Reinigungsbürstchen reinigen:				
ZS_{iMem}:	Alternatively, you may clean the shaver using the small cleaning brush provided:				
	1	**2**	**3**	**4**	**5**
ZS_{TM}	5	1	0	0	0
ZS_{iMem}	0	6	0	0	0

Frage 2					
AS_{neu}:	Alternativ können Sie den Rasierer mit der gelieferten Bürste reinigen:				
AS_{TM}:	Alternativ können Sie den Rasierer mit der mitgelieferten Bürste reinigen.				
ZS_{TM}:	Alternatively, you may clean the shaver using the brush provided.				
AS_{iMem}:	Alternativ können Sie den Rasierer mit der mitgelieferten Bürste reinigen:				
ZS_{iMem}:	Alternatively, you may clean the shaver using the brush provided:				
	1	**2**	**3**	**4**	**5**
ZS_{TM}	4	2	0	0	0
ZS_{iMem}	5	1	0	0	0

Frage 3					
AS_{neu}:	Schäden durch unsachgemäßen Gebrauch, normaler Verschleiß sowie Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen.				
AS_{TM}:	Schäden durch unsachgemäßen Gebrauch, normaler Verschleiß, Verbrauch sowie Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen.				
ZS_{TM}:	Damage due to improper use, normal wear, use as well as defects that have a negligible effect on the value or operation of the appliance.				
AS_{iMem}:	Schäden durch unsachgemäßen Gebrauch, normaler Verschleiß und Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen.				
ZS_{iMem}:	Damage due to improper use, normal wear and defects that have a negligible effect on the value or operation of the appliance.				
	1	**2**	**3**	**4**	**5**
ZS_{TM}	1	5	0	0	0
ZS_{iMem}	5	1	0	0	0

Frage 4					
AS_{neu}:	Wir empfehlen die Verwendung einer Feuchtigkeitscreme nach der Epilation.				
AS_{TM}:	Um die Haut zu entspannen, empfehlen wir die Verwendung einer Feuchtigkeitscreme nach der Epilation.				
ZS_{TM}:	To relax the skin we recommend applying a moisture cream after epilation.				
AS_{iMem}:	Nach der Epilation empfehlen wir die Verwendung einer Feuchtigkeitscreme.				
ZS_{iMem}:	After epilation, we recommend applying a moisture cream.				
	1	**2**	**3**	**4**	**5**
ZS_{TM}	0	6	0	0	0
ZS_{iMem}	5	1	0	0	0

Frage 5					
AS_{neu}:	Jede durch Haarentfernung entstandene Kleinstverletzung birgt die Gefahr der Entzündung durch das Eindringen von Bakterien, unter anderem durch das Gleiten des Gerätes über die Haut.				
AS_{TM}:	Es kann vorkommen, dass sich die Haut durch das Eindringen von Bakterien entzündet (z.B. wenn das Gerät über die Haut gleitet).				
ZS_{TM}:	In some cases inflammation of the skin could occur when bacteria penetrate the skin (e.g. when sliding the appliance over the skin).				
AS_{iMem}:					
ZS_{iMem}:					
	1	**2**	**3**	**4**	**5**
ZS_{TM}	0	1	4	1	0
ZS_{iMem}	0	0	0	0	6

Frage 6					
AS_{neu}:	Führen Sie den Haarschneider langsam gegen die Haarwuchsrichtung.				
AS_{TM}:	Führen Sie den Haarschneider vorsichtig gegen die Haarwuchsrichtung.				
ZS_{TM}:	Move the clipper carefully against the direction of hair growth.				
AS_{iMem}:	Halten Sie den Distanzkamm flach auf dem Haar, parallel zur Kopfhaut und führen Sie den Haarschneider langsam gegen die Haarwuchsrichtung.				
ZS_{iMem}:	Keep the distance comb flat on the hair, parallel to the head, and slowly move the clipper against the direction of hair growth.				
	1	**2**	**3**	**4**	**5**
ZS_{TM}	0	6	0	0	0
ZS_{iMem}	0	5	1	0	0

Frage 7					
AS_{neu}:	Rasierer einschalten und den Scherkopf unter fließendes, warmes Wasser halten.				
AS_{TM}:	Rasierer einschalten und den Scherkopf unter fließendes und warmes Wasser halten.				
ZS_{TM}:	Turn on the shaver and hold the shaver head under warm and running water.				
AS_{iMem}:	Rasierer einschalten und den Scherkopf unter warmes, fließendes Wasser halten.				
ZS_{iMem}:	Turn on the shaver and hold the shaver head under warm running water.				
	1	**2**	**3**	**4**	**5**
ZS_{TM}	2	4	0	0	0
ZS_{iMem}	3	3	0	0	0

Frage 8					
AS_{neu}:	Schäden durch unsachgemäßen Gebrauch, normaler Verschleiß und Verbrauch sowie Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen.				
AS_{TM}:	Schäden durch unsachgemäßen Gebrauch, normaler Verschleiß, Verbrauch sowie Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen.				
ZS_{TM}:	Damage due to improper use, normal wear, use as well as defects that have a negligible effect on the value or operation of the appliance.				
AS_{iMem}:	Schäden durch unsachgemäßen Gebrauch, normaler Verschleiß sowie Verbrauch und Mängel, die den Wert oder die Gebrauchstauglichkeit des Gerätes nur unerheblich beeinflussen.				
ZS_{iMem}:	Damage due to improper use, normal wear and use as well as defects that have a negligible effect on the value or operation of the appliance.				
	1	**2**	**3**	**4**	**5**
ZS_{TM}	3	3	0	0	0
ZS_{iMem}	4	2	0	0	0

Frage 9					
AS_{neu}:	• Der bewegliche Schwingkopf und die flexiblen Scherfolien sorgen automatisch für eine optimale Anpassung an die Gesichtsform.				
AS_{TM}:	• Der bewegliche Schwingkopf und die flexiblen Scherfolien sorgen automatisch für eine optimale Anpassung an die Gesichtsform				
ZS_{TM}:	• The pivoting shaver head and floating foils automatically adjust to every contour of your face				
AS_{iMem}:	Der bewegliche Schwingkopf und die flexiblen Scherfolien sorgen automatisch für eine optimale Anpassung an die Gesichtsform.				
ZS_{iMem}:	The pivoting shaver head and floating foils automatically adjust to every contour of your face.				
	1	**2**	**3**	**4**	**5**
ZS_{TM}	4	2	0	0	0
ZS_{iMem}	3	3	0	0	0

Frage 10

AS_{neu}: Grundsätzlich raten wir aber, das Gerät von Kindern fern zu halten.

AS_{TM}: Halten Sie das Gerät von Kindern fern.
ZS_{TM}: Keep the appliance out of the reach of children.

AS_{iMem}: Das Gerät von Kindern fern zu halten raten wir aber grundsätzlich.
ZS_{iMem}: Keeping the appliance out of the reach of children is what we generally recommend.

	1	**2**	**3**	**4**	**5**
ZS_{TM}	1	3	2	0	0
ZS_{iMem}	3	2	1	0	0

Frage 11

AS_{neu}: Wenn sich der Distanzkamm mit Haaren zusetzt, sollten Sie ihn abnehmen und säubern.

AS_{TM}: Um zu vermeiden, dass sich der Distanzkamm mit Haaren zusetzt, sollten Sie ihn abnehmen und säubern.
ZS_{TM}: To avoid clogging the distance comb with hair, remove and clean it.

AS_{iMem}: Sie sollten ihn abnehmen und säubern, wenn sich der Distanzkamm mit Haaren zusetzt.
ZS_{iMem}: Remove and clean it if too much hair is caught in the distance comb.

	1	**2**	**3**	**4**	**5**
ZS_{TM}	2	3	1	0	0
ZS_{iMem}	1	4	1	0	0

Frage 12

AS_{neu}: Zwingen Sie den Haarschneider nicht schneller durch das Haar, als das Gerät schneiden kann.

AS_{TM}: Zwingen Sie den Haar- und Bartschneider nicht schneller durch das Haar, als das Gerät schneiden kann.
ZS_{TM}: Do not force the appliance through the hair faster than it can be cut.

AS_{iMem}: Führen Sie langsame und kontrollierte Bewegungen aus, zwingen Sie den Haarschneider nicht schneller durch das Haar, als das Gerät schneiden kann.
ZS_{iMem}: Use a slow and controlled movement, do not force the clipper through the hair faster than the clipper can cut it.

	1	**2**	**3**	**4**	**5**
ZS_{TM}	4	2	0	0	0
ZS_{iMem}	0	4	2	0	0

Frage 13					
AS_{neu}:	Gelegentlich Flüssigseife (ohne Scheuermittel) verwenden.				
AS_{TM}:	Gelegentlich Flüssigseife verwenden.				
ZS_{TM}:	You may also use liquid soap.				
AS_{iMem}:	Gelegentlich Seife verwenden (Flüssigseife auf natürlicher Basis ohne Scheuermittel)				
ZS_{iMem}:	A natural based soap may also be used provided it contains no particles or abrasive substances.				
	1	**2**	**3**	**4**	**5**
ZS_{TM}	0	5	1	0	0
ZS_{iMem}	1	1	3	0	1

Frage 14					
AS_{neu}:	Wenn Sie den bodycruZer regelmäßig mit Wasser reinigen, sollten Sie wöchentlich einen Tropfen Leichtmaschinenöl auf dem Langhaarschneider verteilen.				
AS_{TM}:	Wenn Sie den Scherkopf unter Wasser reinigen, sollten die Scherteile nach jeder Reinigung geschmiert werden.				
ZS_{TM}:	If you clean the shaver head under running water, lubricate it after each cleaning.				
AS_{iMem}:	Verteilen Sie etwas Leichtmaschinenöl oder Vaseline auf der Scherfolie und dem Langhaarschneider.				
ZS_{iMem}:	Apply some light machine oil or vaseline to the shaver foil and the metal parts of the long hair trimmer.				
	1	**2**	**3**	**4**	**5**
ZS_{TM}	0	1	2	3	0
ZS_{iMem}	0	1	1	4	0

Frage 15					
AS_{neu}:	Diese Batterien gewährleisten eine Rasierleistung von ca. 60 Minuten.				
AS_{TM}:	Eine Volladung reicht ca. 60 Minuten.				
ZS_{TM}:	Fully charged, the appliance gives about 60 minutes of shaving.				
AS_{iMem}:	Die «Smart Logic» Elektronik in Ihrem Rasierer analysiert Ihr persönliches Rasierprofil und passt daran die Pflege der Akku-Einheit an, um eine optimale Leistung zu gewährleisten.				
ZS_{iMem}:	The « Smart Logic » electronics inside your shaver analyzes your personal shaving patterns, and then adapts the battery care to ensure optimal performance				
	1	**2**	**3**	**4**	**5**
ZS_{TM}	0	4	2	0	0
ZS_{iMem}	0	0	1	3	2

Literaturverzeichnis

Abir, Eli; Klein, Steve; Miller, David; Steinbaum, Michael (2002): „Fluent Machines' EliMT System". In: Richardson, Stephen D. (Hg.): *AMTA 2002, Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California, USA.* Lecture Notes in Computer Science, Bd. 2499. Heidelberg, Berlin: Springer Verlag, 216–219.

Abouelhoda, Mohamed Ibrahim; Kurtz, Stefan; Ohlebusch, Enno (2004): „Replacing suffix trees with enhanced suffix arrays". In: *Journal of Discrete Algorithms*, Bd. 2. Amsterdam: Elsevier B.V., 53–86.

Abouelhoda, Mohamed Ibrahim; Ohlebusch, Enno; Kurtz, Stefan (2002): „Optimal Exact String Matching Based on Suffix Arrays". In: Laender, Alberto H. F.; Oliveira, Arlindo L. (Hg.): *SPIRE 2002, Proceedings of the 9th International Symposium on String Processing and Information Retrieval, 11–13 September, Lisbon, Portugal*. Lecture Notes in Computer Science, Bd. 2476. Heidelberg, Berlin: Springer Verlag, 31–43.

Across Systems GmbH (2014): *Anwender-Handbuch – ‚Across im Überblick' v6*. Stand: 18.06.2014. Dokumentation zum Translation-Memory-System Across v6.
URL: http://www.across.net/fileadmin/sector/support/documentation/user_manual_reference_guide_v60_de.pdf (13.08.2014)

Alegria, Iñaki; Casillas, Arantza; Díaz de Ilarraza, Arantza; Igartua, Jon; Labaka, Gorka; Lersundi, Mikel; Mayor, Aingeru; Sarasola, Kepa; Saralegi, Xabier; Laskurain, Bittor (2008): „Mixing approaches to MT for Basque: selecting the best output from RBMT, EBMT and SMT". In: Alegria, Iñaki; Màrquez, Lluís; Sarasola, Kepa (Hg.): *Proceedings of the MATMT2008 workshop: Mixing Approaches to Machine Translation, Donostia-San Sebastian, Spain*. 27–34.

ALPAC (1966): *Language and machines: computers in translation and linguistics*. A Report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences.

Aluru, Srinivas (2004): „Suffix Trees and Suffix Arrays". In: Mehta, Dinesh P.; Sahni, Sartaj (Hg.): *Handbook of Data Structures and Applications*. Boca Raton: Chapman & Hall/CRC, 29-1–29-22.

Arnold, Douglas J. (2003): „Why translation is difficult for computers". In: Somers, Harold L. (Hg.): *Computers and Translation: A Translator's Guide*. Amsterdam, Philadelphia: John Benjamins Publishing Company, 119–142.

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

Arthern, Peter J. (1979): „Machine translation and computerized terminology systems: a translator's viewpoint". In: Snell, Barbara M. (Hg.): *Translating and the Computer: Proceedings of a Seminar, 14 November 1978, London, England*. Amsterdam: North-Holland Publishing Company, 77–108.

ATRIL Language Engineering (1993–2003): *Déjà Vu X Professional User's Guide*. Dokumentation zum Translation-Memory-System Déjà Vu X Professional.

Aziz, Wilker; Rios, Miguel; Specia, Lucia (2011): „Shallow Semantic Trees for SMT". In: *Proceedings of the 6th Workshop on Statistical Machine Translation, 30–31 July 2011, Edinburgh, Scotland, UK*. 316–322.

Azzano, Dino (2009): „CAT und MÜ – Getrennte Welten?". In: Seewald-Heeg, Uta; Stein, Daniel (Hg.): *Journal for Language Technology and Computational Linguistics (JLCL).*

Maschinelle Übersetzung von der Theorie zur Anwendung – Machine Translation – Theory and Applications, Bd. 24, Nr. 3. Köthen, München: Hochschule Anhalt, Ludwig-Maximilians-Universität München, 19–36.

Azzano, Dino; Reinke, Uwe; Sauer, Melanie (2011): „Ansätze zur Verbesserung der Retrieval-Leistung kommerzieller Translation-Memory-Systeme". In: Hedeland, Hanna; Schmidt, Thomas; Wörner, Kai (Hg.): *Multilingual Resources and Multilingual Applications, Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011, 28–30 September 2011, Universität Hamburg, Hamburg, Germany*. Arbeiten zur Mehrsprachigkeit, Folge B, Nr. 96–2011. Hamburg: Sonderforschungsbereich 538, Hamburger Zentrum für Sprachkorpora, Universität Hamburg, 123–128.

Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier (1999): *Modern Information Retrieval*. Essex: Pearson Education Limited.

Benjamin, Bryce; Gerber, Laurie; Knight, Kevin; Marcu, Daniel (2003): „Language Weaver: The Next Generation of Machine Translation". In: *MT Summit IX, Proceedings of the 9th Machine Translation Summit, 23–27 September 2003, New Orleans, USA*. 445–446.

Biçici, Ergun; Dymetman, Marc (2008): „Dynamic Translation Memory: Using Statistical Machine Translation to Improve Translation Memory Fuzzy Matches". In: Gelbukh, Alexander F. (Hg.): *CICLing 2008, Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics, 17–23 February 2008, Haifa, Israel.* Lecture Notes in Computer Science, Bd. 4919. Heidelberg, Berlin: Springer Verlag, 454–465.

Blatt, Achim (1998): „EURAMIS Alignment and Translation Memory Technology". In: *T&T – Terminologie et Traduction*, Bd. 1.1998. Luxemburg: Office des publications officielles des Communautés européennes, 74–101.

Bloodgood, Michael; Strauss, Benjamin (2014): „Translation Memory retrieval methods". In: *EACL 2014, Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 26–30 April 2014, Gothenburg, Sweden.* 202–210.

Bowker, Lynne (2002): *Computer-Aided Translation Technology: A Practical Introduction*. Didactics of Translation Series. Ottawa: University of Ottawa Press.

Bowker, Lynne; Barlow, Michael (2008): „A comparative evaluation of bilingual concordancers and translation memory systems". In: Yuste Rodrigo, Elia (Hg.): *Topics of Language Resources for Translation and Localisation*. Amsterdam, Philadelphia: John Benjamins Publishing Company, 1–22.

Brosius, Felix (1998): *SPSS 8 – Professionelle Statistik unter Windows*. Bonn: mitp Verlags GmbH & Co. KG.

Brown, Peter E.; Cocke, John; Della Pietra, Stephen A.; Della Pietra, Vincent J.; Jelinek, Fredrick; Lafferty, John D.; Mercer, Robert L.; Roossin, Paul S. (1990): „A Statistical Approach to Machine Translation". In: *Computational Linguistics*, Bd. 16, Nr. 2. Cambridge, USA: MIT Press, 79–85.

Brown, Peter E.; Cocke, John; Della Pietra, Stephen A.; Della Pietra, Vincent J.; Jelinek, Fredrick; Mercer, Robert L.; Roossin, Paul S. (1988): „A Statistical Approach to French/ English Translation". In: *Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, 12–14 June 1988, Carnegie Mellon University, Center for Machine Translation, Pittsburgh, Pennsylvania, USA.*

Bruckner, Christine; Plitt, Mirko (2001): „Evaluating the operational benefit of using machine translation output as translation memory input". In: *MT Summit VIII, Proceedings of the 8th Machine Translation Summit, Workshop on MT Evaluation, 18–22 September 2001, Santiago de Compostela, Spain.* 61–65.

Büttcher, Stefan; Clarke, Charles L. A.; Cormack, Gordon V. (2010): *Information Retrieval: Implementing and Evaluating Search Engines.* Cambridge, MA, USA; London: The MIT Press.

Callison-Burch, Chris; Bannard, Colin; Schroeder, Josh (2005): „A Compact Data Structure for Searchable Translation Memorys". In: *EAMT 2005, Proceedings of the 10th European Association for Machine Translation Conference 'Practical Applications of Machine Translation', 30–31 May 2005, Budapest, Hungary.* 59–65.

Carbonell, Jaime; Klein, Steve; Miller, David; Steinbaum, Michael; Grassiany, Tomer; Frey, Jochen (2006): „Context-based Machine Translation". In: *AMTA 2006, Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, August 2006, Cambrige.*19–28.

Carl, Michael; Schmidt-Wigger, Antje (1998): „Shallow Post Morphological Processing with KURD". In: Powers, David M. W. (Hg.): *NeMLaP3/CoNLL98, Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 257–265.

Carl, Michael; Way, Andy (2003): *Recent advances in example-based machine translation*. Dordrecht: Kluwer Academic Publishers.

Carroll, Jeremy J. (1992): *Repetitions Processing using aMetrie Space and the Angle of Similarity*. CCL/UMIST Report No. 90/3. Manchester: Centre for Computational Linguistics, UMIST.

Carstensen, Kai-Uwe (2012): *Sprachtechnologie – Ein Überblick.* Version 2.1. Webveröffentlichung.
URL: http://www.kai-uwe-carstensen.de/Publikationen/Sprachtechnologie.pdf (31.03.2014)

Carstensen, Kai-Uwe; Ebert, Christian; Ebert, Cornelia; Jekat, Susanne; Klabunde, Ralf; Langer, Hagen (2010): *Computerlinguistik und Sprachtechnologie – Eine Einführung*. 3. Auflage. Heidelberg: Spektrum Akademischer Verlag.

CASMACAT (2014): *Benutzeranleitung zur CASMACAT Home Edition*. Webseite zum Open-Source-Projekt CASMACAT.
URL: http://www.casmacat.eu/index.php?n=UserGuide.HomeEdition (22.02.2015)

Castilho Monteiro de Sousa, Sheila; Aziz, Wilker; Specia, Lucia (2011): „Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles". In: *RANLP 2011, Proceedings of the International Conference on Recent Advances in Natural Language Processing, 12–14 September 2011, Hissar, Bulgaria.* 97–103.

Chatterji, Sanjay; Roy, Devshri; Sarkar, Sudeshna; Basu, Anupam (2009): „A Hybrid Approach for Bengali to Hindi Machine Translation". In: *ICON 2009, Proceedings of the 7th International Conference on Natural Language Processing, 14–17 December 2009, Hyderabad, India*. 83–91.

Chatzitheodorou, Konstantinos (2015): „Improving translation memory fuzzy matching by paraphrasing". In: *RANLP 2015, Proceedings of the Workshop on Natural Language Processing for Translation Memories (NLP4TM), 11 September 2015, Hissar, Bulgaria.* 24–30.

Cordts, Sönke (2012): *Datenqualität in Datenbanken*. Heide: mana-Buch.

Cormen, Thomas H.; Leiserson, Charles E.; Rivest, Ronald L.; Stein, Clifford (2001): *Introduction to Algorithms*. Second Edition. Cambridge, MA, USA; London: The MIT Press.

Dandapat, Sandipan; Morrissey, Sara; Kumar Naskar, Sudip; Somers, Harold L. (2010): „*Statistically Motivated Example-based Machine Translation using Translation Memory*". In: *ICON 2010, Proceedings of the 8th International Conference on Natural Language Processing, 8–11 December 2010, Kharagpur, India.*

Dara, Aswarth; Dandapat, Sandipan; Groves, Declan; van Genabith, Josef (2013): „TMTprime: A Recommender System for MT and TM Integration“. In: *HLT-NAACL 2013, Proceedings of the 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Demonstration Session, 10–12 June 2013, Atlanta, Georgia, USA.* 10–13.

DIN 2331 (1980): *Begriffssysteme und ihre Darstellung*. Berlin, Köln: Beuth Verlag GmbH.

Dudenredaktion (2006): *Duden – Deutsches Universalwörterbuch*. 6. überarbeitete und erweiterte Auflage. Mannheim, Leipzig, Wien, Zürich: Dudenverlag.

EAGLES (1996): *Evaluation of Natural Language Processing Systems.* Final *Report. EAGLES DOCUMENT EAG-EWG-PR.2. Version of October 1996.*
URL: http://www.issco.unige.ch/en/research/projects/ewg96/ewg96.html (03.04.2014)
http://www.issco.unige.ch/en/research/projects/ewg96/node157.html#SECTION001043000000000000000 (03.04.2014)

E DIN 2342 (2004–2009): *Begriffe der Terminologielehre – Entwurf*. Belin, Köln: Beuth Verlag GmbH.

Elita, Natalia; Gavrila, Monica (2006): „Enhancing Translation Memorys with Semantic Knowledge“. In: *CESCL, Proceedings of the 1st Central European Student Conference in Linguistics, 29–31 May 2006, Budapest, Hungary.*

Erk, Katrin; Priese, Lutz (2008): *Theoretische Informatik*. 3. Auflage. Heidelberg, Berlin: Springer Verlag.

Esplà-Gomis, Miquel; Sánchez-Martínez, Felipe; Forcada, Mikel L. (2011): „Using machine translation in computer-aided translation to suggest the target-side words to change“. In: *MT Summit XIII, Proceedings of the 13th Machine Translation Summit, 1–23 September 2011, Xiamen, China.* 172–179.

Esselink, Bert (2000): *A Practical Guide to Localization*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Europäische Kommission (2009): *Geschichte des Übersetzungsdienstes der Europäischen Kommission*. Luxemburg: Amt für Veröffentlichungen der Europäischen Union.

Europäisches Parlament (o.J.): *Zusatzthema zu Modul 3 Sekundärrecht der EU – Alles was Recht ist: der Acquis communautaire*.
URL: http://www.europarl.europa.eu/brussels/website/media/modul_03/Zusatzthemen/Pdf/Acquis.pdf (31.03.2014)

European Commission (2014): *Language Technology Resources – DGT-Translation Memory*.
URL: https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory?search (31.03.2014)

Farach, Martin (1997): „Optimal suffix tree construction with large alphabets“. In: *FOCS '97, Proceedings of the 38th IEEE Symposium on Foundations of Computer Science, 19–22 October, Miami Beach, Florida, USA.* 137–143.

Federico, Marcello; Bertoldi, Nicola; Cettolo, Mauro; Negri, Matteo; Turchi, Marco; Tromebetti, Marco; Cattelan, Alessandro; Farina, Antonio; Lupinetti, Domenico; Martines, Andrea; Massidda, Alberto; Schwenk, Holger; Barrault, Loïc; Blain, Frederic; Koehn, Philipp; Buck, Christian; Germann, Ulrich (2014): „The Matecat Tool“. In: Tounsi, Lamia; Rak, Rafal (Hg.): *COLING 2014, Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations, 23–29 August 2014, Dublin, Ireland.* 129–132.

Fitschen, Arne (2004): *Ein Computerlinguistisches Lexikon als komplexes System*. Dissertation. Stuttgart: Institut für maschinelle Sprachverarbeitung, Philosophisch-Historische Fakultät der Universität Stuttgart.

URL: http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/IMSLex/fitschendiss.pdf (04.05.2012)

Flanagan, Kevin (2014): *Filling in the gaps: what we need from TM subsegment recall*. URL: http://www.kftrans.co.uk/lift/FillingInTheGaps.pdf (26.02.2015)

Forster, Richard (2006): *TransTech-Lerneinheit: Satz- und Phrasenähnlichkeit*. Zürich: Universität Zürich.
URL: http://kitt.cl.uzh.ch/clab/satzaehnlichkeit/tutorial/tt_sensim.pdf (31.07.2014)

Frederking, Robert; Nirenburg, Sergei (1994): „Three heads are better than one". In: *ANLP 1994, Proceedings of the 4th Conference on Applied Natural Language Processing*. 95–100.

Frederking, Robert; Nirenburg, Sergei; Farwell, David; Helmreich, Stephen; Hovy, Eduard; Knight, Kevin; Beale, Stephen; Domashnev, Constantine; Attardo, Donalee; Grannes, Dean; Brown, Ralf (1994): „Integrating Translations from Multiple Sources within the Pangloss Mark III Machine Translation System". In: *Technology Partnerships for Crossing the Language Barrier: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas, Columbia, Maryland*. 73–80.

Freigang, Karl-Heinz (2000): „Maschinelle Übersetzung". In: Schmitz, Klaus-Dirk; Wahle, Kirsten (Hg.): *Softwarelokalisierung*. Tübingen: Stauffenburg Verlag, 167–180.

Gentner, Dedre; Markman, Arthur B. (1994): „Structural Alignment in Comparison: No Difference Without Similarity". In: *Psycological Science*, Bd. 5, Nr. 3. Washington, D.C.: American Psychological Society, 152–158.

Gentner, Dedre; Markman, Arthur B. (1995): „Similarity is like analogy: Structural alignment in comparison". In: Cacciari, Cristina (Hg.): *Similarity in language, thought and perception*. Brüssel: BREPOLS, 111–147.

Gervais, Daniel (2002): „The Full-Text Multilingual Corpus: Breaking the Translation Memory Bottleneck". In: *Translating and the Computer 24, Proceedings of the 24th International Conference on Translating and the Computer, 21–22 November 2002, London, England*. London: Aslib.

Good, Robert L. (1988): „Automated Lookup: AutoTerm of ALP Systems". In: Vasconcellos, Mureil (Hg.): *Technology as Translation Strategy*. American Translators Association, Scholarly Monograph Series II. Amsterdam, Philadelphia: John Benjamins Publishing Company, 87–91.

Gotti, Fabrizio; Langlais, Philippe; Macklovitch, Elliott; Bourigault, Didier; Robichaud, Benoit; Coulombe, Claude (2005): „3GTM: A Third-Generation Translation Memory". In: *CLiNE 2005, Proceedings of the 3rd Computational Linguistics in the North-East Workshop, 26 August 2005, Gatineau, Québec, Canada*.

Grönroos, Mickel; Becks, Ari (2005): „Bringing Intelligence to Translation Memory Technology". In: *Translating and the Computer 27, Proceedings of the 27th International Conference on Translating and the Computer, 24–25 November 2005, London, England*. London: Aslib.

Gupta, Rohit; Orăsan, Constantin (2014): „Incorporating Paraphrasing in Translation Memory Matching and Retrieval". In: Tadić, Marko; Koehn, Philipp; Roturier, Johann; Way, Andy (Hg.): *EAMT 2014, Proceedings of the 17th European Association for Machine Translation Conference, 16–18 June 2014, Dubrovnik, Croatia.* 3–10.

Gusfield, Dan (1997): *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge, New York, Melbourne: Cambridge University Press.

Haapalainen, Mariikka; Majorin, Ari (1994): *GERTWOL: Ein System zur automatischen Wortformenerkennung deutscher Wörter*. Bericht Lingsoft Inc.

URL: https://files.ifi.uzh.ch/cl/volk/LexMorphVorl/Lexikon04.Gertwol.html#Einl (09.10.2014)

Haller, Johann (1996): *MULTILINT*. Saarbrücken: IAI.
URL: http://www.iai-sb.de/docs/multilint.pdf (05.10.2013)

Haller, Johann (2000): „Sprachtechnologie für die Automobilindustrie“. In: Wilss, Wolfram (Hg.): *Weltgesellschaft – Weltverkehrssprache – Weltkultur*. Tübingen: Stauffenburg Verlag, 250–263.

Haller, Johann (2007): „Elektronischer Tutor – Intelligente Werkzeuge für computerunterstütztes Fremdsprachen-Lernen“. In: Roche, Jörg (Hg.): *Fremdsprachen lernen medial: Entwicklungen, Forschungen, Perspektiven*. Kommunikation und Kulturen, Bd. 5. Berlin: LIT Verlag, 72–88.

Haller, Johann; Ripplinger, Bärbel; Maas, Heinz D.; Gastmeyer, Manuela (2001): *Automatische Indexierung von wirtschaftswissenschaftlichen Texten – ein Experiment*. IAI, Hamburgisches Welt-Wirtschafts-Archiv.
URL: http://www.iai-sb.de/docs/0012-gastmeyer.pdf (12.04.2010)

He, Yifan; Ma, Yanjun; van Genabith, Josef; Way, Andy (2010a): „Bridging SMT and TM with Translation Recommendation“. In: Hajic, Jan; Carberry, Sandra; Clark, Stephen (Hg.): *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 11–16 July 2010, Uppsala, Sweden*. 622–630.

He, Yifan; Ma, Yanjun; Way, Andy; van Genabith, Josef (2010b): „Integrating N-best SMT Outputs into a TM System“. In: Huang, Chu-Ren; Jurafsky, Dan (Hg.): *COLING 2010, 23rd International Conference on Computational Linguistics, Poster Volume, 23–27 August 2010, Beijing, China.* Chinese Information Processing Society of China 2010, 374–382.

He, Yifan; Ma, Yanjun; Way, Andy; van Genabith, Josef (2011): „Rich Linguistic Features for Translation Memory-Inspired Consistent Translation“. In: *MT Summit XIII, Proceedings of the 13th Machine Translation Summit, 1–23 September 2011, Xiamen, China.* 456–462.

Hearne, Mary; Way, Andy (2011): „Statistical Machine Translation: A Guide for Linguists and Translators“. In: *Language and Linguistics Compass*, Bd. 5, Nr. 5. Oxford: Blackwell Publishing Ltd., 205–226.

Henrich, Andreas (2008): *Information Retrieval 1 – Grundlagen, Modelle, Anwendungen*. Version 1.2 (Rev: 5727, Stand: 7. Januar 2008). Bamberg: Otto-Friedrich-Universität, Lehrstuhl für Medieninformatik, 2001–2008.
URL: http://www.uni-bamberg.de/fileadmin/uni/fakultaeten/wiai_lehrstuehle/medieninformatik/Dateien/Publikationen/2008/henrich-ir1-1.2.pdf (09.04.2014)

Heyn, Matthias (1998): „Translation Memories: Insights and Prospects“. In: Bowker, Lynne; Cronin, Michael; Kenny, Dorothy; Pearson, Jennifer (Hg.): *Unity in Diversity? Current Trends in Translation Studies.* Manchester: St. Jerome Publishing, 123–136.

Hoffmann, Erika (1982): „Stages in the life cycle of LEXIS“. In: Snell, Barbara M. (Hg.): *Term banks for tomorrow's world: Translating and the Computer 4, Proceedings of the 4th International Conference on Translating and the Computer, 11–12 November 1982, London, England*. London: Aslib, 186–191.

Hutchins, W. John (1988): „Recent developments in Machine Translation: a review of the last five years“. In: Maxwell, Dan; Schubert, Klaus; Witkam, Toon (Hg.): *New directions in machine translation.* Dordrecht: Foris Publications Holland, 7–64.

Hutchins, W. John (1995): **„**A new era in machine translation**“**. In: *Translating and the Computer 16: Proceedings of the 16th International Conference on Translating and the Computer*, Bd. 47, Nr. 10. London: Aslib, 211–219.

Hutchins, W. John (1998): „The origins of the translator's workstation". In: Somers, Harold L. (Hg.): *Machine Translation*, Bd. 13, Nr. 4. Dordrecht: Kluwer Academic Publishers, 287–307.

Hutchins, W. John; Somers, Harold L. (1992): *An introduction to Machine Translation*. London, San Diego, New York, Boston, Sydney, Tokio, Toronto: Academic Press.

IAI (2011): *AUTINDEX (Automatische Indexierung und Klassifizierung)*. Projektbeschreibung. Webseite des IAI.
URL: http://www.iai-sb.de/iai/index.php/AUTINDEX-Automatische-Indexierung-und-Klassifizierung.html (05.10.2013)

IAI (o.J.): Dokumentation zu den MPRO-Merkmalen ori, lu, mlu, state, ls, t, ts, c, case, deg, ds, ehead, g, gra, gs, infl, nb, ptc und sc. [unveröffentlicht].

Isabelle, Pierre (1993): „Machine-Aided Human Translation and the Paradigm Shift". In: *MT Summit IV, Proceedings of the 4th Machine Translation Summit, 20–22 July 1993, Kobe, Japan*. 177–179.

Jekat, Susanne; Volk, Martin (2010): „Maschinelle und computergestützte Übersetzung". In: Carstensen, Kai-Uwe; Ebert, Christian; Ebert, Cornelia; Jekat, Susanne; Klabunde, Ralf; Langer, Hagen (Hg.): *Computerlinguistik und Sprachtechnologie – Eine Einführung*. 3. Auflage. Heidelberg: Spektrum Akademischer Verlag, 642–658.

Karp, Richard M.; Rabin, Michael O. (1987): „Efficient randomized pattern-matching algorithms". In: *IBM Journal of Research and Development*, Bd. 31, Nr. 2, 249–260.

Kasai, Toru; Lee, Gunho; Arimura, Hiroki; Arikawa, Setsuo; Park, Kunsoo (2001): „Linear-Time Longest-Common-Prefix Computation in Suffix Arrays and Its Applications". In: Amir, Amihood; Landau, Gad M. (Hg.): *Proceedings of the 12th Annual Symposium on Combinatorial Pattern Matching*. Lecture Notes in Computer Science, Bd. 2089. Heidelberg, Berlin: Springer Verlag, 181–192.

Kay, Martin (1980): „The Proper Place of Men and Machines in Language Translation". In: Somers, Harold L. (Hg.): *Machine Translation*, Bd. 12, Nr. 1–2. Dordrecht: Kluwer Academic Publishers, 3–23.

Koehn, Philipp (2010): *Statistical Machine Translation*. Cambridge, New York: Cambridge University Press.

Koehn, Philipp; Alabau, Vincent; Carl, Michael; Casacuberta, Francisco; García-Martínez, Mercedes; González-Rubio, Jesús; Keller, Frank; Ortiz-Martínez, Daniel; Sanchis-Trilles, Germán; Germann, Ulrich (2013): *CASMACAT – Final Public Report.*
URL: http://www.casmacat.eu/uploads/Deliverables/final-public-report.pdf (22.02.2015)

Koehn, Philipp; Hoang, Hieu; Birch, Alexandra; Callison-Burch, Chris; Federico, Marcello; Bertoldi, Nicola; Cowan, Brooke; Shen, Wade; Moran, Christine; Zens, Richard; Dyer, Chris; Bojar, Ondrej; Constantin, Alexandra; Herbst, Evan (2007): „Moses: Open Source Toolkit for Statistical Machine Translation". In: *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, June 2007, Prague, Czech Republic.* 177–180.

Koehn, Philipp; Senellart, Jean (2010a): „Fast Approximate String Matching with Suffix Arrays and A* Parsing". In: *AMTA 2010, Proceedings of the 9th Conference of the Association for Machine Translation in the Americas, 31 October–4 November 2010, Denver, Colorado, USA.*

Koehn, Philipp; Senellart, Jean (2010b): „Convergence of Translation Memory and Statistical Machine Translation". In: Zhechev, Ventsislav (Hg.): *JEC 2010, Proceedings of the 2nd Joint EM+/CGNL Workshop 'Bringing MT to the User: Research on Integrating MT in the Translation Industry', 4 November 2010, Denver, Colorado, USA.* 21–31.

Kofler, Michael; Eller, Frank; Beyer, Alexander; Schwichtenberg, Holger (2011): *Visual Basic 2010 – Grundlagen, ADO.NET, Windows Presentation Foundation*. München: Addison-Wesley Verlag.

Krollmann, Friedrich (1971): „Linguistic Data Banks and the Technical Translator". In: *Meta : journal des traducteurs/Meta: Translators' Journal*, Bd. 16, Nr. 1–2. Montréal: Les Presses de l'Université de Montréal, 117–124.

Krypczyk, Veikko (2011): „Daten, Daten, Daten – Teil 1: Grundlagen des Datenbankentwurfs, ER-Modell, Normalisierung". In: *Entwickler Magazin*, Nr. 2.2011, 68–74.

Kuhns, Robert J. (2007): *Advanced Leveraging, The New Generation of TMs*. TAUS Report, October 2007. De Rijp: TAUS B.V.

Kunze, Claudia; Lemnitzer, Lothar (2007): *Computerlexikographie – eine Einführung*. Tübingen: Gunter Narr Verlag.

Lagoudaki, Elina (2006): *Translation Memories Survey 2006, Translation Memory systems: Enlightening users' perspective*. November 2006. London: Imperial College London.
URL: http://www3.imperial.ac.uk/pls/portallive/docs/1/7307707.PDF (19.07.2011)

Lagoudaki, Elina (2008): „The Value of Machine Translation for the Professional Translator". In: *AMTA 2008, MT at work: 8th Conference of the Association for Machine Translation in the Americas, Proceedings, 21–25 October 2008, Waikiki, Hawaii, USA.* 262–269.

Lang, Hans W. (2006): *Algorithmen in Java.* 2. Auflage. München: Oldenbourg Wissenschaftsverlag GmbH.

Lehrberger, John; Bourbeau, Laurent (1988): *Machine Translation – Linguistic characteristics of MT systems and general methodology of evaluation*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Lemnitzer, Lothar; Zinsmeister, Heike (2006): *Korpuslinguistik: Eine Einführung*. Tübingen: Gunter Narr Verlag.

Lenders, Winfried (2012): „Wie Computer Sprache lernen". In: Podtergera, Irina (Hg.): *Schnittpunkt Slavistik: Ost und West im wissenschaftlichen Dialog. Festgabe für Helmut Keipert zum 70. Geburtstag. Teil 1: Slavistik im Dialog – einst und jetzt*. Göttingen: Bonn University Press bei V&R unipress, 455–472.

Levenshtein, Vladimir I. (1966): „Binary codes capable of correcting deletions, insertions, and reversals". In: *Soviet Physics – Doklady*, Bd. 10, Nr. 8. College Park, MD, USA: American Institute of Physics, 707–710.

Lingua et Machina (2006): *Similis Manager, Similis Translation Tool, Similis Xeditor – Version 2 – Guide de l'utilisateur*.
URL: http://similis.org/linguaetmachina.www/index.php?afficher=10&sel=43&info=Handbücher (14.03.2012)

Linguistic Data Consortium (1992–2010): *Hansard French/English*. The LDC Corpus Catalog. Philadelphia: University of Pennsylvania.
URL: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20 (23.01.2013)

Lippmann, Erhard O. (1971): „An Approach to Computer-Aided Translation". In: *IEEE Transactions on Engineering Writing and Speech*, Bd. 14, Nr. 1. New York: Institute of Electrical and Electronics Engineers (IEEE), 10–33.

Lobin, Henning (2009): *Computerlinguistik und Texttechnologie*. Paderborn: Wilhelm Fink GmbH & Co. Verlags-KG.

Lohde, Michael (2006): *Wortbildung des modernen Deutschen: ein Lehr- und Übungsbuch*. Tübingen: Narr Francke Attempto Verlag GmbH + Co. KG.

Lonsdale, Deryle (2007): *From ALPS to Alpnet (and beyond)*.
URL: http://www.mt-archive.info/Lonsdale-2007.pdf (28.08.2012)

Ma, Yanjun; He, Yifan; Way, Andy; van Genabith, Josef (2011): „Consistent Translation using Discriminative Learning: A Translation Memory-inspired Approach“. In: *ACL 2011, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 19–24 June, Portland, Oregon, USA.* 1239–1248.

Maas, Heinz D. (1998): *Multilingualität in MPRO*. Saarbrücken: IAI.
URL: http://iai.iai-sb.de/docs/mmpro.pdf (07.02.2012)

Maas, Heinz D.; Rösener, Christoph; Theofilidis, Axel (2009): „Morphosyntactic and semantic analysis of text: The MPRO tagging procedure“. In: Mahlow, Cerstin; Piotrowski, Michael (Hg.): *State of the Art in Computational Morphology: Workshop on Systems and Frameworks for Computational Morphology (SFCM 2009), Zurich, Switzerland, 4 September 2009, Proceedings*. Communications in Computer and Information Science, Bd. 41. Heidelberg, Berlin: Springer Verlag, 76–87.

Macherey, Wolfgang; Och, Franz J. (2007): „An Empirical Study on Computing Consensus Translations from Multiple Machine Translation Systems“. In: *EMNLP and CONLL 2007, Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 2007, Prague.* 986–995.

Macken, Lieve (2009): „In search of the recurrent units of translation“. In: Daelemans, Walter; Hoste, Véronique (Hg.): *Evaluation of Translation Technology*. Brüssel: Academic and Scientific Publishers, 195–212.

Macklovitch, Elliott (2000): „Two Types of Translation Memory“. In: *Translating and the Computer 22, Proceedings of the 22nd International Conference on Translating and the Computer, 16–17 November 2000, London, England*. London: Aslib.

Macklovitch, Elliot; Russell, Graham (2000): „What's been Forgotten in Translation Memory“. In: *AMTA 2000, Envisioning Machine Translation in the Information Future: Proceedings of the 4th Conference of the Association for Machine Translation in the Americas, 10–14 October 2000, Cuernavaca, Mexico.* Lecture Notes in Computer Science, Bd. 1934. Heidelberg, Berlin: Springer Verlag, 137–146.

Manber, Udi; Myers, Gene (1993): „Suffix arrays: A new method for on-line string searches“. In: *SIAM Journal of Computing*, Bd. 22, Nr. 5. Philadelphia: Society for Industrial and Applied Mathematics, 935–948.

Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich (2008): *Introduction to Information Retrieval*. Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi: Cambridge University Press.

Massion, François (2005): *Translation Memory Systeme im Vergleich*. Reutlingen: doculine Verlags-GmbH.

Matusov, Evgeny; Ueffing, Nicola; Ney, Hermann (2006): „Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment“. In: *EACL 2006, Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 3–7 April 2006, Trento, Italy.* 33–40.

McCreight, Edward M. (1976): „A space-economical suffix tree construction algorithm“. In: Coffman Jr., Edward G. (Hg.): *Journal of the ACM*, Bd. 23, Nr. 2. New York: ACM, 262–272.

McNaught, John (1980): „Terminological Data Banks: a model for a British Linguistic Data Bank (LDB)“. In: *Machine aids for translators: Aslib Technical Translation Group conference and exhibition, Proceedings, 20 November 1980, London, England*. 297–308.

McTait, Kevin (2001): „Linguistic Knowledge and Complexity in an EBMT System Based on Translation Patterns“. In: *MT Summit VIII, Proceedings of the 8th Machine Translation Summit, Workshop on MT Evaluation, 18–22 September 2001, Santiago de Compostela, Spain*. 23–34.

Melby, Alan K. (1982): „Multi-level translation aids in a distributed system“. In: Horecký, Ján (Hg.): *COLING 1982, Proceedings of the 9th International Conference on Computational Linguistics, 5–10 July 1982, Prague, Czech Republic*, Bd. 9. Amsterdam: North-Holland Publishing Company, 215–220.

Melby, Alan K. (1992): „The translator workstation“. In: Newton, John (Hg.): *Computers in Translation: A Practical Appraisal*. London: Routledge, 147–165.

Melby, Alan K.; Warner, C. Terry (1995): *The possibility of language: a discussion of the nature of language, with implications for human and machine translation*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Mitkov, Ruslan; Corpas, Gloria (2008): „Improving Third Generation Translation Memory Systems Through Identification of Rhetorical Predicates“. In: *LangTech 2008, Proceedings of the 3rd Language and Speech Technology Conference, 28–29 February, Rome, Italy*.

Moorkens, Joss; O'Brien, Sharon; da Silva, Igor A. L.; de Lima Fonseca, Norma B.; Alves, Fabio (2015): „Correlations of perceived post-editing effort with measurements of actual effort“. In: Lamont, Andrew (Hg.): *Machine Translation*, Bd. 29, Nr. 3–4. Dordrecht: Kluwer Academic Publishers, 267–284.

MultiCorpora R&D Inc. (2011): *Translation Memory*. Webseite der MultiCorpora R&D Inc. URL: http://www.multicorpora.com/en/multitrans-prism/translation-memory/ (14.12.2012)

Nagao, Makoto (1984): „A framework of a mechanical translation between Japanese and English by analogy principle“. In: Elithorn, Alick; Banerji, Ranan (Hg.): *Artificial and Human Intelligence*. Edited Review Papers presented at the International NATO Symposion on Artificial and Human Intelligence, Lyon, 1981. Amsterdam, New York, Oxford: North Holland, 173–180.

Nielsen, Jakob (1993): *Usability Engineering*. San Francisco: Morgan Kaufmann Publishers Inc.

Nirenburg, Sergei; Domashnev, Constantine; Grannes, Dean J. (1993): „Two Approaches to Matching in Example-Based Machine Translation“. In: *TMI 1993, Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation, 14–16 July 1993, Kyoto, Japan*. 47–57.

Nyberg, Eric; Mitamura, Teruko; Huijsen, Willem-Olaf (2003): „Controlled language for authoring and translation“. In: Somers, Harold L. (Hg.): *Computers and Translation: A Translator's Guide*. Amsterdam, Philadelphia: John Benjamins Publishing Company, 245–281.

O'Brien, Sharon (1998): „Practical Experience of Computer-Aided Translation Tools in the Localization Industry“. In: Bowker, Lynne; Cronin, Michael; Kenny, Dorothy; Pearson, Jennifer (Hg.): *Unity in Diversity? Current Trends in Translation Studies.* Manchester: St. Jerome Publishing, 115–122.

OpenTM2 (2010): *Working with language-support files*. Wiki zum Open-Source-TM-System OpenTM2. Stand der Webseite: 27.09.2010.
URL: http://www.beo-doc.de/opentm2wiki/index.php/Working_with_language-support_files (24.02.2015)

Oracle (1993, 2016): *nanoTime*. Dokumentation zur Java-Klasse ‚System'. Java Platform Standard Ed. 7, Webseite.
URL: http://docs.oracle.com/javase/7/docs/api/java/lang/System.html#nanoTime() (15.06.2015)

Planas, Emmanuel (1998): *TELA: Structures et algorithmes pour la Traduction Fondée sur la Mémoire*. Dissertation. Grenoble: Universität Joseph Fourier.

Planas, Emmanuel (2005): „SIMILIS Second-generation translation memory software". In*: Translating and the Computer 27, Proceedings of the 27th International Conference on Translating and the Computer, 24–25 November 2005, London, England*. London: Aslib.

Planas, Emmanuel; Furuse, Osamu (1999): „Formalizing Translation Memories". In: *MT Summit VII, Proceedings of the 7th Machine Translation Summit 'MT in the Great Translation Era', 13–17 September 1999, Kent Ridge Digital Labs, Singapore.* 331–339.

Planas, Emmanuel; Furuse, Osamu (2000): „Multi-level Similar Segment Matching Algorithm for Translation Memories and Example-Based Machine Translation". In: *COLING 2000, Proceedings of the 18th International Conference on Computational Linguistics, 31 July–4 August 2000, Universität des Saarlandes, Saarbrücken, Germany,* Bd. 2. San Francisco: Morgan Kaufmann Publishers Inc., 621–627.

Procter & Gamble Manufacturing GmbH (2011): Webseite der Marke Braun. Bedienungsanleitungen zu Bart- und Haarschneidern, Epiliergeräten, Lady Shavern, Präzisionshaarschneidern und Rasierern.
URL: http://www.service.braun.com/line/SH/S5281/S5281_2_D.pdf (10.09.2013)
http://www.service.braun.com/line/SH/S5281/S5281_1_GB.pdf (10.09.2013)
http://www.service.braun.com/line/SH/S5601/S5601_1_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5602/S5602_4_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5606/S5606_1_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5303/S5303_3_MN.pdf (10.09.2013)
http://www.service.braun.com/line/SH/S5316/S5316_8_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5375/S5375_6_MN_AMEE.pdf (06.09.2013)
http://www.service.braun.com/line/sh/s5395/S5395_18_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5327/S5327_1_MN_AMEE.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5328/S5328_1_MN_AMEE.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5575/S5575_1_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5363/S5363_1_MN_AMEE.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5363/S5363_5_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5780/S5780_1_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5785/S5785_1_MN.pdf (10.09.2013)
http://www.service.braun.com/line/SH/S5302/S5302_1_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5447/S5447_2_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5492/S5492_9_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5503/S5503_2_D.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5604/S5604_3_AP.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5635/S5635_5_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5684/S5684_4_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5693/S5693_2_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5694/S5694_6_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5722/S5722_1_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5732/S5732_1_MN.pdf (06.09.2013)
http://www.service.braun.com/line/SH/S5738/S5738_1_MN.pdf (06.09.2013)

Quasthoff, Uwe (1998): „Deutscher Wortschatz im Internet". In: *LDV-Forum*, Bd.15, Nr. 2, 4–23.

Rapp, Reinhard (2002): „A Part-of-Speech-Based Search Algorithm for Translation Memories“. In: *LREC 2002, Proceedings of the 3rd International Conference on Language Resources and Evaluation, 27 May–2 June 2002, Las Palmas de Gran Canaria, Spain.* 466–472.

Reinke, Uwe (1999): „Evaluierung der linguistischen Leistungsfähigkeit von Translation Memory-Systemen: Ein Erfahrungsbericht“. In: *LDV-Forum*, Bd.16, Nr. 1/2, 100–117.

Reinke, Uwe (2004): *Translation Memories: Systeme – Konzepte – Linguistische Optimierung.* Dissertation. Frankfurt am Main: Peter Lang.

Reinke, Uwe (2013): „State of the Art in Translation Memory Technology“. In: Rehm, Georg; Sasaki, Felix; Stein, Daniel; Witt, Andreas (Hg.): *Translation: Computation, Corpora, Cognition. Special Issue on Language Technologies for a Multilingual Europe*, Bd. 3, Nr. 1, 27–48.

Ripplinger, Bärbel (1998): „EMIS – A Multilingual Information System“. In: Farwell, David; Gerber, Laurie; Hovy, Eduard H. (Hg.): *AMTA 1998, Machine Translation and the Information Soup: 3rd Conference of the Association for Machine Translation in the Americas, Proceedings, 28–31 October 1998, Langhorne, PA, USA*. Lecture Notes in Computer Science, Bd. 1529. Heidelberg, Berlin: Springer Verlag, 506–509.

Ripplinger, Bärbel (2000): „MproIR – A Cross-language Information Retrieval Component Enhanced by Linguistic Knowledge“. In: Mariani, Joseph-Jean; Harman, Donna (Hg.): *RIAO 2000, Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval, 12–14 April 2000, College de France, Paris, France.* 1107–1123.

Rösener, Christoph (2005): *Die Stecknadel im Heuhaufen: Natürlichsprachiger Zugang zu Volltextdatenbanken.* Dissertation. Frankfurt am Main: Peter Lang.

Rösener, Christoph (2010): „Computational Linguistics in the Translator's Workflow – Combining Authoring Tools and Translation Memory Systems“. In: *NAACL HLT 2010, Proceedings of the Workshop on Computational Linguistics and Writing, June 2010, Los Angeles, California, USA*. Association for Computational Linguistics 2010, 1–6.

Roth, Tobias (2014): *Wortverbindungen und Verbindungen von Wörtern – Lexikografische und distributionelle Aspekte kombinatorischer Begriffsbildung zwischen Syntax und Morphologie*. Tübingen: Narr Francke Attempto Verlag GmbH + Co. KG.

Runte, Matthias (1999): *Proximitätsmaße – Strategieoptionen unter Missing Values*. Working Paper. Kiel.
URL: http://www.runte.de/matthias/publications/proximitaetsmasse.pdf (04.08.2014)

Sack-Kastl, Elisabeth (2007): „TM-Präsentationen aus der Praxis für die Praxis – Wordfast und Transit“. In: *TransRelations, Mitgliederzeitschrift des BDÜ-Landesverbands Bremen-Niedersachsen e.V.*, Heft 1/2007, April 2007. Göttingen: BDÜ Landesverband Bremen und Niedersachsen e. V., 12–13.

Sánchez-Martinez, Felipe; Forcada, Mikel L.; Way, Andy (2009): „Hybrid Rule-Based – Example-Based MT: Feeding Apertium with Sub-sentential Translation Units“. In: Forcada, Mikel L.; Way, Andy (Hg.): *Proceedings of the 3rd Workshop on Example-Based Machine Translation, 12–13 November 2009, Dublin City University, Dublin, Ireland*. 11–18.

Sauer, Hermann (2002): *Relationale Datenbanken – Theorie und Praxis*. 5. Auflage. München. Addison-Wesley Verlag.

Schäler, Reinhard (2001): „Beyond Translation Memories“. In: *MT Summit VIII, Proceedings of the 8th Machine Translation Summit, Workshop on MT Evaluation, 18–22 September 2001, Santiago de Compostela, Spain.* 49–55.

Schmidt, Paul (1998): „Automatisches Übersetzen“. In: Snell-Hornby, Mary; Hönig, Hans G.; Kussmaul, Paul; Schmitt, Peter A. (Hg.): *Handbuch Translation*. Tübingen: Stauffenburg Verlag, 133–137.

Schmitt, Ingo (2006): *Ähnlichkeitssuche in Multimedia-Datenbanken – Retrieval, Suchalgorithmen, Abfragebehandlung*. München: Oldenbourg Wissenschaftsverlag GmbH.

SDL plc. (2009): *SDL Trados Studio 2009 SP3 – Online Hilfe*.
URL: http://producthelp.sdl.com/SDL%20Trados%20Studio/client_en/SDL_Trados_Studio_Help.htm (04.04.2012)

SDL plc. (2009–2011): *SDL Trados Studio 2011 – Kurzanleitung ‚Dokumente übersetzen und überprüfen'*. Dokumentation zum Translation-Memory-System SDL Trados Studio 2011.

SDL plc. (2009–2013): *Translation Automation API 1.0* (automation of the translation memory related task, translation providers etc.).
URL: http://producthelp.sdl.com/sdk/TranslationMemoryApi/1.0/index.aspx?s=107556%7C20131006104911%7C0x9f187d35dfc7732e014e36d41f179d40a5591733 (06.10.2013)

Seewald-Heeg, Uta (2005): „Der Einsatz von Translation-Memory-Systemen am Übersetzerarbeitsplatz – Aufbau, Funktionsweise und allgemeine Kaufkriterien". In: *MDÜ (Mitteilungen für Dolmetscher und Übersetzer)*, Heft 4–5/2005, 8–38.

Shilon, Reshef; Fadida, Hanna; Wintner, Shuly (2012): „Incorporating Linguistic Knowledge in Statistical Machine Translation: Translating Prepositions". In: *EACL 2012, Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data (Hybrid2012), 23 April 2012, Avignon, France*. 106–114.

Sikes, Richard (2007): „Fuzzy matching in theory and practice". In: *MultiLingual*, Bd. 18, Nr. 6. Sandpoint, ID: MultiLingual Computing Inc., 39–43.

Simard, Michel; Langlais, Philippe (2001): „Sub-sentential Exploitation of Translation Memorys". In: *MT Summit VIII, Proceedings of the 8th Machine Translation Summit, Workshop on MT Evaluation, 18–22 September 2001, Santiago de Compostela, Spain*. 335–339.

Singh, Thoudam D.; Bandyopadhyay, Sivaji (2010): „Manipuri-English Example Based Machine Translation System". In: Alexander Gelbukh (Hg.): *International Journal of Computational Linguistics and Applications (IJCLA*), Bd. 1, Nr. 1–2. Neu Delhi: Bahri Publications, 201–216.

Sjöberg, Lennart (1975): „Models of Similarity and Intensity". In: *Psychological Bulletin*, Bd. 82, Nr. 2. Washington, D.C.: American Psychological Association, 191–206.

Smith, James; Clark, Stephen (2009): „EBMT for SMT: A New EBMT-SMT Hybrid". In: Forcada, Mikel L.; Way, Andy (Hg.): *Proceedings of the 3rd Workshop on Example-Based Machine Translation, 12–13 November 2009, Dublin City University, Dublin, Ireland*. 3–11.

Smith, Ross (2009): „Copyright Issues in Translation Memory Ownership". In: *Translating and the Computer 31, Proceedings of the 31st International Conference on Translating and the Computer, 19–20 November 2000, London, England*. London: Aslib.

Smolej, Vito (o.J.): *OmegaT 3.0 – User's Guide*. Dokumentation zum Open-Source-TM-System OmegaT.
URL: http://ob.nubati.net/ditundat/omegat/docs/en30/appendix.TokenizerPlugin.inOmegaT.html#d0e10802 (26.04.2016)

Somers, Harold L. (2003a): „Translation Memory Systems". In: *Computers and Translation: A Translator's Guide*. Amsterdam, Philadelphia: John Benjamins Publishing Company, 31–47.

Somers, Harold L. (2003b): „An Overview of EBMT". In: Carl, Michael; Way, Andy (Hg.): *Recent advances in example-based machine translation*. Dordrecht: Kluwer Academic Publishers, 3–57.

Somers, Harold L. (2003c): „The translator's workstation". In: *Computers and Translation: A Translator's Guide*. Amsterdam, Philadelphia: John Benjamins Publishing Company, 13–30.

Somers, Harold L. (2003d): „Machine translation: Latest developments". In: Mitkov, Ruslan (Hg.): *The Oxford Handbook of Computational Linguistics*. New York: Oxford University Press, 512–528.

Somers, Harold L.; Fernandez Diaz, Gabriela (2004): „Translation Memory vs. Example-based MT: What's the difference?". In: *International Journal of Translation*, Bd. 16, Nr. 2, 5–33. URL: http://personalpages.manchester.ac.uk/staff/harold.somers/ebmtvstm_ijt.pdf (07.11.2014)

STAR AG (2012): *Transit NXT – Benutzerhandbuch 2012-09*. Ramsen, Schweiz: STAR AG.

Stehouwer, Herman; van Zaanen, Menno (2010): „Finding Patterns in Strings using Suffixarrays". In: *IMCSIT 2010, Proceedings of the International Multiconference on Computer Science and Information Technology, 18-20 October 2010, Wisla, Poland.* 505–511.

Stein, Daniel (2009): „Maschinelle Übersetzung – ein Überblick". In: Seewald-Heeg, Uta; Stein, Daniel (Hg.): *Journal for Language Technology and Computational Linguistics (JLCL). Machinelle Übersetzung von der Theorie zur Anwendung – Machine Translation – Theory and Applications*, Bd. 24, Nr. 3. Köthen, München: Hochschule Anhalt, Ludwig-Maximilians-Universität München, 5–18.

Stock, Wolfgang G. (2007): *Information Retrieval – Informationen suchen und finden*. München: Oldenbourg Wissenschaftsverlag GmbH.

Sung, Wing-Kin (2010): *Algorithms in Bioinformatics: A Practical Introduction*. Boca Raton: Chapman & Hall/CRC Mathematical and Computational Biology Series.

Toral, Antonio; Way, Andy (2014): „Is Machine Translation Ready for Literature?" In: *Translating and the Computer 36, Proceedings of the 36th International Conference on Translating and the Computer, 27–28 November 2014, London, England*. London: AsLing, 174–176.

Tversky, Amos (1977): „Features of Similarity". In: *Psychological Review*, Bd. 84, Nr. 4. Washington, D.C.: American Psychological Association Inc., 327–352.

Tversky, Amos; Gati, Itamar (1978): „Studies of Similarity". In: Rosch, Eleanor; Llyod, Barbara B. (Hg.): *Cognition and categorization*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc., 79–98.

Ukkonen, Esko (1995): „On-line construction of suffix trees". In: *Algorithmica*, Bd. 14, Nr. 3. Heidelberg, Berlin: Springer Verlag, 249–260.

van Slype, Georges (1979): *Critical study of methods for evaluating the quality of machine translation*. Final report. Brüssel: Bureau Marcel van Dijk.

van Zaanen, Menno; Somers, Harold L. (2005): „DEMOCRAT: Deciding between Multiple Outputs Created by Automatic Translation". In: *MT Summit X, Proceedings of the 10th Machine Translation Summit, 12–16 September 2005, Phuket, Thailand*. 173–180.

Vanallemeersch, Tom; Vandeghinste, Vincent (2014): „Improving fuzzy matching through syntactic knowledge". In: *Translating and the Computer 36, Proceedings of the 36th International Conference on Translating and the Computer, 27–28 November 2014, London, England*. London: AsLing, 90–99.

Vossen, Piek (1997): „EuroWordNet: a multilingual database for information retrieval". In: *Proceedings of the 3rd DELOS workshop on Cross-language Information Retrieval, 5–7 March 1997, Zurich, Switzerland.*

Weiner, Peter (1973): „Linear pattern matching algorithms". In: *SWAT '08, Proceedings of the 14th IEEE Symposium on Switching and Automata Theory, 15–17 October 1973* . 1–11.

Weitz, Melanie (2017): „Improving retrieval performance of translation memories using morpho-syntactic analyses and generalized suffix arrays“. In: *Machine Translation*. Dordrecht: Springer Science+Business Media B.V., DOI: 10.1007/s10590-017-9193-3.

Whyman, Edward K.; Somers, Harold L. (1999): „Evaluation Metrics for a Translation Memory System“. In: *Software – Practice and Experience*, Bd. 29, Nr. 14. Hoboken, New Jersey, USA: John Wiley & Sons Ltd., 1265–1284.

Wise, Michael J. (1993): *String Similarity via Greedy String Tiling and Running Karp-Rabin Matching.*
URL: http://luggage.bcs.uwa.edu.au/~michaelw/ftp/doc/RKR_GST.ps (11.08.2014)

Zerfaß, Angelika (2002): „Evaluating Translation Memory Systems”. In: *LREC 2002, Proceedings of the 3rd International Conference on Language Resources and Evaluation, Workshop on Language resources for translation work and research, 27 May 2002, Las Palmas de Gran Canaria, Spain*. 49–52.

Zhechev, Ventsislav; van Genabith, Josef (2010a): „Maximising TM Performance through Sub-Tree Alignment and SMT“. In: *AMTA 2010, Proceedings of the 9th Conference of the Association for Machine Translation in the Americas, 31 October–4 November 2010, Denver, Colorado, USA.*

Zhechev, Ventsislav; van Genabith, Josef (2010b): „Seeding Statistical Machine Translation with Translation Memory Output through Tree-Based Structural Alignment“. In: *SSST-4, Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation, Beijing, China.*

Seit Anfang der 90er Jahre zählen Translation Memorys zu den bedeutsamsten Werkzeugen im Bereich der computergestützten Übersetzung. Dennoch wird in den meisten auf dem Markt verfügbaren Translation-Memory-Systemen nur ein Vergleich der Zeichenketten zwischen einem neu zu übersetzenden Segment mit einem im Translation Memory gespeicherten ausgangssprachlichen Segment durchgeführt, um ähnliche Übersetzungseinheiten aufzufinden. Linguistische Übereinstimmungen bleiben häufig unberücksichtigt, sodass bedeutungsgleiche bzw. -ähnliche Segmente mit unterschiedlicher (morpho-)syntaktischer Struktur nicht oder nur mit einem geringeren Ähnlichkeitswert ausgegeben werden.
In diesem Band wird die Optimierung eines kommerziellen Translation-Memory-Systems durch Hinzuschaltung eines morphosyntaktischen Analyseprogramms beschrieben. Der Fokus liegt dabei auf der Erläuterung des entwickelten Prototyps, bestehend aus der Beschreibung des erweiterten Algorithmus zur Identifikation der längsten gemeinsamen Teilzeichenketten zwischen zwei zu vergleichenden ausgangssprachlichen Segmenten sowie des selbst konzipierten, auf linguistischen Analysen beruhenden Proximitätsmaßes zur Ermittlung von Ähnlichkeitswerten, die das menschliche Ähnlichkeitsempfinden abbilden sollen. Die umfangreiche Evaluierung des entwickelten Systems demonstriert das Ausmaß der erreichten linguistischen Optimierung kommerzieller zeichenkettenbasierter Translation Memorys.