Evaluierung der linguistischen Leistungsfähigkeit von Translation Memory-Systemen – Ein Erfahrungsbericht –

Uwe Reinke
Universität des Saarlandes

1 Abgrenzung von TM-, MÜ- und IR-Systemen

Eines der wichtigsten Ergebnisse des Treffens des GLDV-Arbeitskreises 'Maschinelle Übersetzung' im Februar 1999 (vgl. den URL http:// www.heeg.de/~uta/AK-Protokoll-TM-1.html) war m.E. die Feststellung, daß sich Kriterien zur Evaluierung von Systemen zur maschinellen Übersetzung kaum für die Untersuchung von Translation Memory-Systemen eignen. Diese Erkenntnis ist zwar durchaus nicht neu [Rei94], wurde aber m.W. von Vertretern der MÜ bisher nicht in dieser Deutlichkeit formuliert.

	TM-Systeme	MÜ-Systeme	IR-Systeme
Input	– AS-Segmente[1]	– AS-Segmente	– Suchanfragen (natürlichsprachig oder 'Abfragesprache')
Verarbeitungsprozeß	– Retrieval-Prozeß (AS-Segmente als Suchanfrage) – Suchanfragen durch zu übersetzende AS-Segmente vorgegeben ((teil-)automatisch) – Zugriff auf AS/ZS-Segmentpaare einer TM-Datenbank bzw. eines alignierten AS/ZS-Textpaares – Retrieval-Mechanismus ermittelt im TM die relevanten Datensätze auf der Basis der Suchanfrage	a) Direkter Ansatz: Ersetzen von AS-Wörtern durch ZS-Wörter unter Zuhilfenahme morphologischer, lexikalischer und ggf. einfacher syntaktischer Informationen b) Transfer-Ansatz: Verwendung von linguistischen Regelwerken (Grammatiken) und Systemwörterbüchern – Analyse: Erzeugt abstrakte AS-Struktur – Transfer: Übertragung in abstrakte ZS-Struktur – Synthese: Erzeugt ZS-Oberfläche	– Retrieval-Prozeß – Suchanfrage wird vom Benutzer formuliert (manuell) – Zugriff auf Dokumente (natürlichsprachige Texte) – Retrieval-Mechanismus ermittelt die für die Suchanfrage relevanten Dokumente
Output	– Der Suchanfrage 'möglichst ähnliche' AS-Segmente und deren ZS-Entsprechungen	– $ZS_{MÜ}$ Segmente[2]	– Referenzen, die auf potentiell relevante Dokumente mit möglichst hoher 'Ähnlichkeit' zur Suchanfrage verweisen
Evaluierung	– Bewertung der Retrieval-Leistung (intralingual); zentrale Begriffe: Recall, Precision	– Bewertung der Übersetzungsqualität (interlingual); zentrale Begriffe: Fehleranalyse, Verständlichkeit	– Bewertung der Retrieval-Leistung; zentrale Begriffe: Recall, Precision

Tab. 1: Abgrenzung von TM-, MÜ- und IR-Systemen.[1,2]

Die Notwendigkeit eigener Evaluierungskriterien wird insbesondere dann offensichtlich, wenn es um die linguistische Performanz der Systeme geht. Im Gegensatz zu MÜ-Systemen ist eine Untersuchung der linguistischen Leistungsfähigkeit von TM-Systemen zunächst einzelsprachspezifisch, da diese keine eigenen Übersetzungen erstellen und in erster Linie als Retrieval-Programme zu verstehen sind. Wie Tab. 1 verdeutlicht, sind die Verarbeitungsprozesse von TM-Systemen den Verarbeitungsprozessen von Information Retrieval-Systemen (IR) weitaus ähnlicher als denen von MÜ-Systemen. Entsprechend ergeben sich Parallelen und Unterschiede bei den Anforderungen an eine Evaluierung der linguistischen Performanz der verschiedenen Systemtypen.

2 Existierende Vorschläge zur Evaluierung der Retrieval-Leistung von TM-Systemen

In der Literatur finden sich bisher mit Ausnahme der Vorschläge, die die EAGLES-Arbeitsgruppe in ihrem Abschlußbericht zur Evaluierung von Systemen zur Verarbeitung natürlicher Sprache vorgelegt hat [EAGLES96], m.W. keine detaillierteren Überlegungen zu einem systematischen Kriterienkatalog für die Bewertung von TM-Systemen.

Bei der Untersuchung der Retrieval-Leistung von TM-Systemen kann zunächst unterschieden werden zwischen dem Retrieval von ‘exakten Entsprechungen’, bei denen Suchanfrage und AS-Seite des Suchergebnisses übereinstimmen (*exact matches*), und ‘unscharfen Entsprechungen’, bei denen sich Suchanfrage und AS-Seite des Suchergebnisses voneinander in mehr oder weniger starkem Maße unterscheiden (*fuzzy matches*).[3] Entsprechend werden im Bericht der EAGLES-Gruppe zwei verschiedene Benchmark-Tests vorgeschlagen.

Der Ablauf des Benchmark-Tests für *exact matches* stellt sich wie folgt dar [EAGLES96:154]:

(1) Zusammenstellen eines Korpus *T* mit Texten des gleichen Texttyps und des gleichen Sachgebiets

(2) Anlegen eines neuen TM mit einem Teil der Texte aus *T*

(3) Anwenden des TM auf andere Texte aus *T*

(4) Ermitteln des Anteils übersetzter sowie korrekt übersetzter Segmente und Berechnung von Recall- und Precision-Werten.

Dieses Szenario enthält m.E. einige Unklarheiten:

- Auswahl der Texte: Um zu einem ausreichenden Maß an *exact matches* zu gelangen, reicht es nicht aus, wenn die Texte des Korpus hinsichtlich Sachgebiet und Texttyp übereinstimmen. Vielmehr sollten Textpaare aus 'Updates' (Überarbeitungen, Aktualisierungen etc.) und 'Originalen' (Ursprungstexten) sowie deren Übersetzungen herangezogen werden.
- Bewertung des Ergebnisses: Wie zuvor dargestellt, kann Übersetzungsqualität nicht Gegenstand der Evaluierung von TM-Systemen sein, da diese Systeme selbst keine Übersetzungen erstellen, sondern lediglich dazu dienen, 'Übersetzungseinheiten' zu speichern und zu suchen.[4] Die 'Korrektheit der Übersetzungen' kann also kein Bewertungskriterium für solche Systeme sein, sondern allenfalls die Relevanz der gefundenen AS/ZS-Segmentpaare.[5]

Legt man die im EAGLES-Bericht gegebene Definition von '*exact match*' zugrunde, so scheint die Untersuchung der Retrieval-Leistung bei 'exakten Entsprechungen' eher trivial.[6] Von größerer Bedeutung und aus linguistischer Sicht interessanter sind demgegenüber vor allem Kriterien für die Evaluierung der Fuzzy Match-Algorithmen. Der EAGLES-Bericht skizziert für einen Benchmark-Test für *fuzzy matches* folgendes Szenario [EAGLES96:155]:[7]

(1) Erstellen eines TM aus einem authentischen Text

(2) Erstellen von Test-Suites durch systematische Modifikation des Textmaterials; Typen von Modifikationen: Satzzeichen, 'Konstanten' (Zahlen, Eigennamen), Segmentlänge, Wortwahl (Ersetzungen, Auslassungen, Hinzufügungen), Syntax (Satzstellung, grammatische Konstruktionen)

(3) Ermitteln des Recall-Wertes nach Durchführung der Modifikationen.

Insgesamt bleibt der EAGLES-Bericht bei den Evaluierungskriterien von Fuzzy Match-Algorithmen eher vage und bietet nur wenig Hilfestellung. Explizit genannt werden einige einfache Modifikationen wie die Veränderung von Satzzeichen oder Zahlen und Eigennamen, die für TM-Systeme i.d.R. ebensowenig eine Schwierigkeit darstellen wie das Ersetzen einzelner Wörter (vgl. [Rei94], [Rös/War97]). Komplexere syntaktische und semantische Veränderungen werden im EAGLES-Bericht jedoch nicht näher differenziert.

3 Parameter zur Evaluierung der Retrieval-Leistung von TM-Systemen

3.1 Anwendung von Kennwerten des IR

Zwei für die Bewertung der Effektivität von Retrieval-Systemen zentrale Kennwerte sind *Recall* und *Precision*. Will man diese beiden Merkmale für die Evaluierung der Retrieval-Leistung von TM-Systemen nutzen, so können in Anlehnung an [Salt/McG87:175] folgende Definitionen zugrunde gelegt werden:

$$\text{Recall:} \qquad R = \frac{\text{Zahl der nachgewiesenen relevanten AS / ZS - Segmentpaare}}{\text{Zahl aller relevanten AS / ZS - Segmentpaare der Datenbasis}}$$

$$\text{Precision:} \qquad P = \frac{\text{Zahl der nachgewiesenen relevanten AS / ZS - Segmentpaare}}{\text{Zahl aller nachgewiesenen AS / ZS - Segmentpaare}}$$

Als ein Vergleichswert, der einfacher zu handhaben ist, als separate Recall- und Precision-Werte wird häufig auch das sog. ‘F-Measure’ [vRij79] bevorzugt. Es handelt sich hierbei um das harmonische Mittel aus Recall und Precision:

$$F = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

Als weitere, zu Recall und Precision komplementäre Kennwerte können außerdem *Silence* (Anteil der nicht nachgewiesenen relevanten Segmentpaare an der Menge aller relevanten Segmentpaare der Datenbasis) und *Noise* (Anteil der nachgewiesenen irrelevanten Segmentpaare an der Menge der nachgewiesenen Segmentpaare) ermittelt werden.

Carroll [Car92] nennt neben Recall und Precision zwei weitere Evaluierungsparameter:

- die Reihenfolge/Gewichtung der Treffer bei Anfragen mit mehr als einem Suchergebnis (*correctness of order*)
- die Konsistenz der Ähnlichkeitswerte (*consistency*).

Bei der Betrachtung der Reihenfolge geht es um die Frage, ob die Match-Werte der Ergebnisse einer Suchanfrage den Grad der Ähnlichkeit zwischen Suchanfrage und Treffer widerspiegeln. Das Konsistenzkriterium untersucht, ob ein System bei vergleichbaren Suchanfragen und vergleichbaren Ergebnissen identische Match-Werte aufweist.

3.2 Der Begriff der Relevanz

Bei der Ermittlung der Kenngrößen für die Bestimmung von Retrieval-Effektivität erweist sich der Begriff der *Relevanz* als zentral. In der Informationswissenschaft wird Relevanz gemeinhin als Grad der formalen Übereinstimmung zwischen Suchanfrage und nachgewiesenem Dokument bzw. Grad der Übereinstimmung eines Dokuments mit den Informationsbedürfnissen des Nutzers definiert [Salt/McG87:173f.]. Analog wäre dann unter der Relevanz einer TM-Einheit der Grad der formalen Übereinstimmung zwischen dem zu übersetzenden AS-Segment und dem im TM nachgewiesenem AS-Segment bzw. der Grad, mit dem ein im TM nachgewiesenes AS/ZS-Segmentpaar mit den 'Informationsbedürfnissen' des Übersetzers übereinstimmt, zu verstehen.

Formal ließen sich die Unterschiede zwischen 'Suchanfragen' und 'Suchergebnissen' einfach in Form von mehr oder weniger umfangreichen Ersetzungen, Hinzufügungen, Auslassungen und Umstellungen (Verschiebungen) von Zeichenketten beschreiben. Ein 'Treffer' wäre demzufolge umso relevanter, je geringer das Ausmaß dieser Veränderungen ist. Dies entspricht jedoch nicht unbedingt dem 'Informationsbedürfnis' des Übersetzers, das in erster Linie darin besteht, aus der Menge der in einem TM vorhandenen AS/ZS-Segmentpaare jene herauszufinden, die im Vergleich zum aktuell zu übersetzenden AS-Segment identische oder zumindest möglichst ähnliche 'Inhalte' aufweisen, so daß die 'ZS-Seite' der gefundenen TM-Einheit wahrscheinlich mit möglichst geringem Aufwand in die aktuelle Übersetzung eingebettet werden kann.

3.3 Der Begriff der Ähnlichkeit

Mit dem Begriff der Ähnlichkeit ist ein weiterer komplexer Begriff angesprochen, der für die Abgrenzung von 'relevanten' und 'irrelevanten' Suchergebnissen von zentraler Bedeutung ist. Die Schwierigkeit bei der Beurteilung von Fuzzy Match-Algorithmen besteht letztlich vor allem darin, einen für diesen Zweck angemessenen Ähnlichkeitsbegriff zu finden und geeignete Ähnlichkeitskriterien

zu definieren, die es ermöglichen, 'relevante' und 'irrelevante' Untersuchungsmerkmale (z.B. für den Aufbau von Test Suites) voneinander zu abzugrenzen. Dabei sollte man beim Erstellen eines entsprechenden Kriterienkatalogs versuchen, jene Maßstäbe anzulegen, anhand derer ein Übersetzer bestimmt, ob sein 'Informationsbedürfnis' befriedigt wurde. Wie die späteren Beispiele zeigen werden, sind diese Kriterien wesentlich komplexer als der an der Oberfläche operierende Zeichenkettenvergleich der meisten TM-Systeme.

Zum Zweck einer ersten Annäherung an den Ähnlichkeitsbegriff könnten in Anlehnung an Begriffe der traditionellen Linguistik *formale*, *semantische* und *pragmatische Ähnlichkeit* unterschieden werden. Eine Unterscheidung von *formaler* und *semantischer Ähnlichkeit* findet sich z.B. auch in der Kognitionspsychologie im Zusammenhang mit Experimenten zum Lernen und Erinnern sprachlicher Einheiten ([Hall71], [USG96]). *Formale Ähnlichkeit* bezeichnet „similarity in terms of common environmental properties" [Hall71:131]. Ein solches 'Umgebungsmerkmal' ist im Zusammenhang mit Experimenten zum Lernen und Erinnern sprachlicher Einheiten z.B. die Anzahl der Buchstaben, die zwei zu lernende Einheiten gemeinsam haben. Formale Ähnlichkeit beschränkt sich also auf Merkmale, die unmittelbar an der Oberfläche der zu vergleichenden Objekte abzulesen sind. Demgegenüber bezieht sich semantische Ähnlichkeit auf den Inhalt der sprachlichen Zeichen. Die kognitive Psychologie unterscheidet hier u.a. zwischen *Bedeutungsähnlichkeit* (Substituierbarkeit der verglichenen sprachlichen Ausdrücke) und *konzeptueller Ähnlichkeit* (Zugehörigkeit der Inhalte zu gleichen Klassen oder Kategorien) ([Hall71], [USG96]).[8] Im wesentlichen operieren die Retrieval-Mechanismen heutiger TM-Systeme auf der Basis von formaler – genauer orthographischer – Ähnlichkeit. Bei den Ähnlichkeitsurteilen von Humanübersetzern stehen demgegenüber semantische und pragmatische Aspekte im Vordergrund.

3.3.1 Bedeutungsähnlichkeit

Neben *exact matches* sind für den Übersetzer vor allem solche TM-Einheiten von vorrangigem Interesse, die Paraphrasen des zu übersetzenden AS-Segments darstellen. In solchen Fällen besteht ebenfalls die Möglichkeit, daß die im TM abgelegte Übersetzung ohne oder mit geringen Veränderungen in den Zieltext übernommen werden kann. Wie das Beispiel in Tab. 2 verdeutlicht, können die von den TM-Systemen ermittelten Ähnlichkeitswerte bei komplexen Paraphrasen extrem niedrig sein, so daß dem Übersetzer solche TM-Einheiten bei entsprechend vorgegebenem Schwellwert gar nicht angeboten werden.[9]

Zu übersetzender AS-Satz ('Update')	TM-Satz ('Original')	Match-Wert (%) Trados TWB
Zurückweisung führt zum sofortigen Löschen der anklopfenden Verbindung.	Das Zurückweisen hat zur Folge, daß die anklopfende Verbindung sofort gelöscht wird.	30
Wird die anklopfende Verbindung zurückgewiesen, so wird diese sofort gelöscht.		39
Weist der Mobilfunkteilnehmer die anklopfende Verbindung zurück, so wird diese sofort gelöscht.		42

Tab. 2: Beispiel für niedrige Match-Werte bei verschiedenen Paraphrasen.

Identische Inhalte liegen nicht nur bei Paraphrasen sondern z.B. auch bei Explikationen bzw. Implikationen vor.[10] Versteht man den Inhalt einer Aussage als Summe expliziter und impliziter Informationen, so bleibt dieser bei einer Veränderung im Explizitheitsgrad des Ausdrucks gleich [vPol88:24f]. Im Gegensatz zu Paraphrasen dürften sich die Unterschiede aber wesentlich häufiger auch auf die Übersetzung auswirken, obwohl man auch hier natürlich nicht ohne weiteres davon ausgehen kann, daß sich die Explizitheitsunterschiede zwischen 'Original' und 'Update' in gleichem Umfang in den entsprechenden Übersetzungen widerspiegeln. Das konstruierte Beispiel in Abb. 1 soll verdeutlichen, daß man sich „[d]as Verhältnis zwischen explizitem und komprimiertem bzw. implikativem Ausdruck […] als eine breite Skala relativer Möglichkeiten vorstellen [kann]" [vPol88:28]. Eine TM-Einheit ist für den Übersetzer wahrscheinlich jedoch nur dann von Interesse, wenn die Explizitheitsunterschiede relativ gering sind (in Abb. 1 also z.B. nicht: zu übersetzende Einheit ist c) und das TM enthält a)).

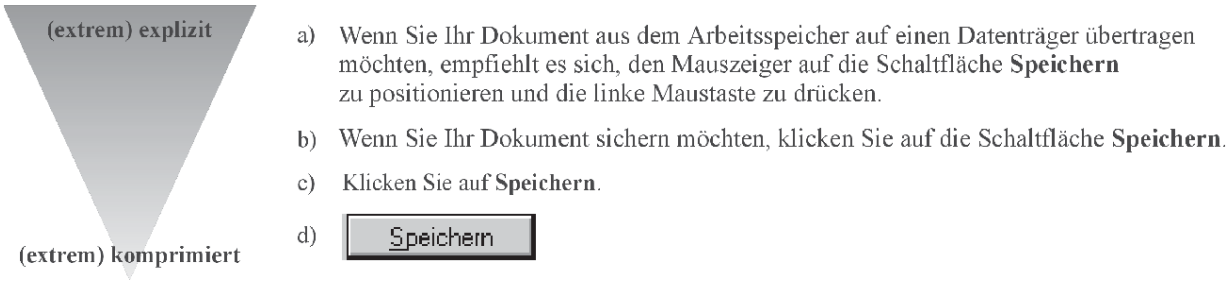

Abb. 1: Oberflächenveränderung durch Implikation/Explikation - der Inhalt (die Nachricht) bleibt unverändert.

3.3.2 Pragmatische Ähnlichkeit

In den bisher angeführten Beispielen für ‘identische Inhalte’ wurden pragmatische Merkmale wie Sender-Empfänger-Beziehung und Kommunikationsebenen (fachintern, fachextern, interfachlich) bewußt invariant gehalten. Variiert man nun diese Parameter bei gleichbleibendem Inhalt, so ergeben sich wie im folgenden Beispiel zweifelsohne ebenfalls ‘ähnliche’ Aussagen:

(1a) Aus der Produktdokumentation eines Mobilfunktelefons:

Entweder Sie beenden das erste Gespräch und nehmen das zweite an. Oder Sie unterbrechen Ihr erstes Gespräch, ohne es zu beenden, um mit dem zweiten Anrufer zu telefonieren.

(1b) Aus einer an Anbieter von Mobilfunkdiensten gerichteten Leistungsbeschreibung einer Mobilfunkanlage:

Wenn der A-Teilnehmer die neu ankommende Verbindung annimmt, kann er die aktive Verbindung entweder freigeben, oder auf Halten setzen, bevor er auf die anklopfende Verbindung antwortet.

Sofern nicht bestimmte grundlegende Unterschiede zwischen den Kulturen der AS und ZS bestehen, dürften sich solche pragmatischen Ungleichheiten i.d.R. in vollem Umfang auf den zu erstellenden Zieltext auswirken. TM-Segmente, die gegenüber einem zu übersetzenden AS-Segment in erster Linie pragmatische Unterschiede aufweisen, sind für den Übersetzer bestenfalls dann von Interesse, wenn diese Unterschiede an der Textoberfläche nur geringe lexikalische und syntaktische Differenzen bewirken. Dies könnte z.B. dann zutreffen, wenn sich zwei Sätze identischen Inhalts auf der Ausdrucksseite ausschließlich durch die Sender- bzw. Produktspezifik der verwendeten Terminologie unterscheiden (verschiedene Sender bei gleicher Kommunikationsebene):[11]

(2a) MICROSOFT WORD:

Mit Druckformaten können Sie das Formatieren Ihrer Texte in beträchtlichem Ausmaß automatisieren.

(2b) WORDPERFECT:

Mit der Style-Funktion können Sie das Formatieren Ihrer Texte in be trächtlichem Ausmaß automatisieren.

Solche Konstellationen sind aber insgesamt eher unrealistisch, da sich - wie das Original des WORDPERFECT-Beispiels zeigt - die Beschreibungen identischer Sachverhalte bei verschiedenen Sendern i.d.R. kaum nur im Hinblick auf senderspezifische Terminologie unterscheiden dürften.

3.3.3 Konzeptuelle Ähnlichkeit

Ähnlichkeiten bestehen schließlich auch zwischen Sätzen, deren *Inhalte* variieren:

(3a) Klicken Sie auf die Stelle, an der die Tabelle eingefügt werden soll.

(3b) Klicken Sie auf die Stelle, an der die Grafik eingefügt werden soll.

Unter dem Aspekt der *formalen Ähnlichkeit* unterscheidet sich Beispiel 3 kaum vom folgenden Satzpaar, bei dem lediglich die Ausdrucksseite modifiziert wird, der Inhalt jedoch gleich bleibt:[12]

(4a) Sondernummern sind Nummern, die einem Diensteanbieter zugewiesen werden und im gesamten Mobilvermittlungsstellenbereich gültig sind.

(4b) Diensteanbieternummern sind Nummern, die einem Diensteanbieter zugewiesen werden und im gesamten Mobilvermittlungsstellenbereich gültig sind.

Im Unterschied zu Sätzen, die sich ausschließlich auf der Ausdrucksseite unterscheiden, sind bei ‘verwandten’ Inhalten jedoch in jedem Fall Anpassungen des im TM gefundenen ZS-Segments erforderlich.

‘Inhaltsverwandtschaften’ beruhen natürlich nicht nur, wie in dem sehr einfachen Beispiel 3, auf Kohyponomie. Vielmehr sind hier auch andere semantische Relationen wie Hyperonymie/Hyponymie, Kontradiktion oder Antonymie einzubeziehen.

4 Erfahrungsbericht über erste eigene empirische Untersuchungen

Im folgenden werden einige Ergebnisse eigener, an einem kleinen Korpus authentischer Texte vorgenommener Untersuchungen geschildert, bei denen es allerdings weniger um einen Vergleich der Leistung verschiedener TM-Systeme ging, als um die Frage, inwieweit sich quantitative Kenngrößen des IR für die Evaluie-

rung solcher Systeme eignen. Im Mittelpunkt stand ferner das Interesse, die in realen Texten vorkommenden Modifikationen zu typisieren, und herauszufinden, bei welchen Typen TM-Systeme Retrieval-Schwierigkeiten aufweisen, um später Vorschläge zur Optimierung der Systeme durch Integration linguistischer Komponenten entwickeln zu können.

Das verwendete Korpus besteht aus fünf Textpaaren. Die Texte sind Teile der deutschsprachigen Leistungsbeschreibung einer Mobilfunkanlage.[13] Sie wenden sich an die Anbieter von Mobilfunkdienstleistungen, d.h. an Fachleute, und zählen zur Textsorte '(fachinterne) Produktinformation'. Jedes der fünf Textpaare besteht aus einem 'Originaltext' und einem 'Update', wobei die 'Updates' jeweils eine aktuellere Version des Produkts beschreiben. Jedes Textpaar stellt ein bestimmtes Leistungsmerkmal der Mobilfunkanlage dar. Die 'Originaltexte' umfassen insgesamt 876 Segmente (ca. 9.900 Wörter), die 'Updates' 898 Segmente (ca. 11.100 Wörter). Die Texte wurden im ASCII-Format ohne jegliche Formatierung zur Verfügung gestellt.[14]

Um jene Stellen eines Textpaares zu ermitteln, die sich inhaltlich einander zuordnen lassen, wurde zunächst jedes 'Update' mit seinem 'Original' verglichen. Zur Unterstützung dieser Aufgabe wurde die Alignment-Komponente eines kommerziellen TM-Systems verwendet. Solche Werkzeuge werden gewöhnlich dazu benutzt, AS- und ZS-Texte zu synchronisieren, d.h. AS- und ZS-Entsprechungen einander zuzuordnen. Da die Beziehungen zwischen 'Originaltext' und 'Update' wesentlich komplexer sein können, als zwischen AS- und ZS-Text, ist es selbstverständlich, daß die mit dem Alignment-Werkzeug erzielten Ergebnisse zahlreiche manuelle Korrekturen erforderten. So mußten all jene Stellen entfernt werden, die keine inhaltliche Entsprechung besaßen, d.h. im 'Update' neu hinzugekommen waren oder im Vergleich zum 'Originaltext' ausgelassen wurden. Ferner wurden absolut invariante (d.h. semantisch und syntaktisch identische) Textstellen entfernt, so daß letztendlich jene Segmentpaare übrigblieben, die semantische und/oder syntaktische Modifikationen aufwiesen. Auf diese Weise wurden 126 Segmentpaare (AS_{Org},AS_{Upd}) mit einem Gesamtumfang von 3.835 Wörtern aus dem Korpus extrahiert. Aus den 126 'Original-Segmenten' AS_{Org} wurden anschließend Translation Memories für die in den Untersuchungen verwendeten Systeme erzeugt[15] und mit den 126 'Update-Segmenten' (AS_{Upd}) Suchen in den TMs durchgeführt. Die folgende Tabelle faßt die Retrieval-Ergebnisse zusammen:

	IBM TranslationManger	Star Transit	Trados TWB
Anzahl der relevanten Segmente im TM	126	126	126
Anzahl der gefundenen Segmente	52	125	124
Anzahl der gefundenen relevanten Segmente	52	84	97
Recall	52/126=0,41	84/126=0,67	97/126=0,77
Precision	52/52=1	84/125=0,67	97/124=0,78
Silence	1-0,41=0,59	1-0,67=0,33	1-0,77=0,23
Noise	1-1=0	1-0,67=0,33	1-0,78=0,22
F-Measure	0,58	0,67	0,77

Tab. 3: Testergebnisse für die im Teilkorpus 'Updates' gegenüber dem Teilkorpus 'Originaltexte' modifizierten Segmente.

Solche Zahlen könnten bei einem Systemvergleich unter Verwendung authentischer Texte, wie er etwa im EAGLES-Bericht vorgeschlagen wird, sehr schnell zu vorschnellen und unberechtigten Urteilen führen. Betrachtet man jedoch die 126 Segmentpaare des Tests genauer, so wird deutlich, daß mehr als 50 % sehr komplexe Unterschiede aufweisen. Tab. 4 zeigt zwei typische Beispiele:

'Original'	'Update'
Es ist auch möglich, eine neue TMSI zuzuweisen (TMSI-Realloction), z.B. beim Gesprächsaufbau oder bei jeder Aktualisierung der Aufenthaltsregistrierung.	Nach einer bestimmten Anzahl von Zugriffen oder wenn ein bestimmtes Ereignis, wie die Aktualisierung der Aufenthaltregistrierung stattfindet, kann dem einzelnen Mobilteilnehmer eine neue TMSI zugewiesen werden.
Die MWD enthalten die Adressen der SMS-Einheiten, die Kurzinformationen für die spätere Zustellung speichern.	MWD ist eine Liste mit bis zu 7 SMS-SC-Adressen, in denen Short Messages gespeichert sind, um zu einem späteren Zeitpunkt dem Mobilteilnehmer B übertragen zu werden.

Tab. 4: Beispiele für komplexe Modifikationen.

Um eine bessere Vergleichs- und Beschreibungsgrundlage zu erhalten, wurden daher aus den 66 Segmentpaaren mit mehrfachen Veränderungen insgesamt 189 Muster mit jeweils nur einer Veränderung abgeleitet. Hierbei wurde z.B. die Hinzufügung eines Begriffs in einer Aufzählung ebenso als *eine* Modifikation gewertet, wie die Hinzufügung einer neuen Aussage in Form einer Satzverknüpfung.

	IBM TranslationManger	Star Transit	Trados TWB
Anzahl der relevanten Segmente im TM	189	189	189
Anzahl der gefundenen Segmente	145	189	187
Anzahl der gefundenen relevanten Segmente	145	168	179
Recall	145/189=0,77	168/189=0,89	179/189=0,95
Precision	145/145=1	168/189=0,89	179/187=0,96
Silence	1-0,77=0,23	1-0,89=0,11	1-0,95=0,05
Noise	1-1=0	1-0,89=0,11	1-0,96=0,04
F-Measure	0,87	0,89	0,95

Tab. 5: Ergebnisse des Tests mit Segmentpaaren, die nur eine Modifikation enthalten.

Die Ergebnisse des zweiten Testlaufs scheinen bei den drei Systemen für die aus dem Textkorpus extrahierten Testsätze eine insgesamt hohe Retrieval-Effektivität nachzuweisen. Betrachtet man jedoch die Match-Werte der einzelnen Beispiele, so wird deutlich, daß die entsprechenden Werte oftmals sehr niedrig und inkonsistent sind (Tab. 6).

'Update'	'Original'	'Match-Wert' (%)[16]	
		Star Transit	Trados TWB
Die HLR initiiert das Löschen der alten Mobilitätsdaten.	Die HLR löscht die alten Mobilitätsdaten.	33	67
Die Notrufnummer ist eine international festgelegte Nummer.	Notrufnummern sind international festgelegte Nummern.	39	45
Wenn der Teilnehmer A eine neue Verbindung zum Teilnehmer C aufbauen will, sendet die MS des Mobilteilnehmers die Anforderung auf Halten an die MSC.	Die MS des Mobilteilnehmers sendet die Anforderung auf Halten an die MSC.	-	37

Tab. 6: Beispiele für niedrige bzw. inkonsistente Match-Werte bei 'einfachen' Modifikationen.[16]

5 Fazit und Ausblick

Zusammenfassend läßt sich feststellen, daß eine bloße 'quantitative Evaluierung' unter Verwendung gängiger informationswissenschaftlicher Kenngrößen wie Recall und Precision allein nicht sehr aussagekräftig ist. Eine 'qualitative Evaluierung' erfordert andererseits eine Typologie von 'Ähnlichkeitsmerkmalen', um z.B. geeignete Test Suites aufbauen zu können. Orientiert man sich bei der Erstellung einer solchen Typologie an authentischen Texten, so muß berücksich-

tigt werden, daß nicht alle Veränderungen, die zwischen ‘Original’ und ‘Update’ vorgenommen wurden, unbedingt für die Untersuchung von TM-Systemen relevant sein müssen. So könnten TM-Systeme sicherlich bereits heute effizienter eingesetzt werden, wenn die Modifikationen in AS-Texten auf ein inhaltlich und stilistisch nötiges Minimum beschränkt würden. Entsprechende kontrastive Untersuchungen verschiedener Versionen authentischer Texte wären daher vermutlich auch im Hinblick auf die Entwicklung und Anwendung kontrollierter Sprachen von Interesse.

Eine genauere Betrachtung der aus dem Textkorpus extrahierten Segmentpaare (AS_{Org},AS_{Upd}) zeigt, daß Retrieval-Probleme bei TM-Systemen vor allem bei einer Häufung verschiedener Modifikationen (z.B. auch bei mehrfachen morphosyntaktischen Modifikationen) sowie bei stark variierenden Segmentlängen auftreten. ‘Oberflächenunterschiede’, die auf Aspekte der Formen- und Wortbildung zurückzuführen sind, könnten vermutlich bereits durch vergleichsweise einfache Mittel (Lemmatisierung, Einbeziehung morphologischer Strukturen (Derivationsmuster)) ausgeglichen werden. Dies gilt ebenso für einfachere syntaktische Phänomene. Die Beispiele in Tab. 7 stellen den Oberflächenformen von ‘Original’ und ‘Update’ jeweils die durch eine morphologische Analyse ‘normalisierten’ Zeichenketten gegenüber.[17]

	‘Original’	‘Update’	‘Match-Wert’ (%) Trados TWB
	Formenbildung		
1 a)	Sondernummern sind im gesamten Mobilvermitthungsstellenbereich gueltig.	Eine Sondernummer ist im gesamten Mobilvermitthungsstellenbereich gueltig.	72
1 b)	besonder nummer gesamt mobil ver mitteln stelle bereich gueltig	besonder nummer gesamt mobil ver mitteln stelle bereich gueltig	100
	Wortbildung		
2 a)	In diesem Fall werden die Bedingungen fuer die Umlenkung des Anrufs nicht geprueft.	In diesem Fall werden die Anrufumlenkungsbedingungen nicht geprueft.	64
2 b)	fallen werden be dingen um lenken an rufen pruefen	fallen werden an rufen um lenken be dingen pruefen	95
	Syntax		
3 a)	Die internationale Mobilfunkgeraetekennung wird fuer die Identifikation des Mobilteilnehmers an die MSC gesendet.	Die internationale Mobilfunkgeraete-kennung wird an die MSC gesendet, um den Mobilteilnehmer zu identifizieren.	68
3 b)	international mobil funk geraet kennen identifzieren mobil teil nehmen MSC senden	international mobil funk geraet kennen MSC senden mobil teil nehmen identifzieren	91

Tab. 7: Ausgleich von ‘Oberflächenunterschieden’ zwischen ‘Original’ und ‘Update’ durch

Nutzung morphologischer Analysen.

Bei stark variierenden Satzlängen von ‘Original’ und ‘Update’ sind jedoch in jedem Fall aufwendigere Verfahren erforderlich, die eine Extraktion von Teilsegmenten (Satzfragmenten) ermöglichen (s. hierzu [Rei99]).

Literatur

[Car92] Carroll, J. (1992): *Repetitions Processing using a Metric Space and the Angle of Similarity*. Technical Report No. 90/3. Manchester: Centre for Computational Linguistics, UMIST.

[EAGLES96] Expert Advisory Group on Language Engineering Standards (1996): *Evaluation of Natural Language Processing Systems*. Final Report (First phase). URL: ftp://issco-ftp.unige.ch/pub/ewg96.ps (25.07.99).

[Hall71] Hall, J. (1971): *Verbal learning and retention*. Philadelphia et al.: J.B. Lippincott.

[Kri95] Krings, H. (1995): *Texte reparieren. Empirische Untersuchungen zum Prozeß der Nachredaktion von Maschinenübersetzungen*. Hildesheim: Universität Hildesheim, Institut für Angewandte Sprachwissenschaft [Habilitationsschrift]

[Maas96] Maas, H.-D. (1996): MPRO – Ein System zur Analyse und Synthese deutscher Wörter. In: Hausser R. (Hrsg.): *Linguistische Verifikation, Sprache und Information. Dokumentation zur Ersten Morpholympics 1994*. Tübingen: Niemeyer, 141–166.

[Rei94] Reinke, U. (1994): Zur Leistungsfähigkeit integrierter Übersetzungssysteme. In: *Lebende Sprachen*, 3/94, 97–104.

[Rei99] Reinke, U. (in diesem Heft): Überlegungen zu einer engeren Verzahnung von Terminologiedatenbanken, Translation Memories und Textkorpora.

[Rös/War97] Rösener, Ch./Wargenau, J. (1997): *Terminologie- und Satzerkennung für Englisch und Russisch am Beispiel der Translator’s Workbench von Trados*. Saarbrücken: Fachrichtung 8.6, Universität des Saarlandes (Saarbrücker Studien zu Sprachdatenverarbeitung und Übersetzen, herausgegeben von Karl-Heinz Freigang und Uwe Reinke, Band 8).

[Sag94] Sager, J. (1994): *Language Engineering and Translation: Consequences of Automation*. Amsterdam: Benjamins.

[Salt/McG87] Salton, G./McGill, M. (1987): *Information Retrieval - Grundlegendes für Informationswissenschaftler*. Hamburg, New York: McGraw Hill.

[See/Nüb99] Seewald-Heeg, U./Nübel, R. (in diesem Heft): Translation-Memory-Module in MÜ-Systemen.

[USG96] Ulrich, R./Stapf, K.-H./Giray, M. (1996): Faktoren und Prozesse des Einprägens und Erinnerns. In: Albert, D./Stapf, K.-H. (Hrsg.): *Gedächtnis. Enzyklopädie der Psychologie* (Themenbereich C, Theorie und Forschung: Ser. 2, Kognition; Bd. 4). Göttingen et al.: Hogrefe.

[vRij79] van Rijsbergen, C. J. (1979): *Information Retrieval*. London: Butterworths.

[Vin/Darb95] Vinay, J.-P./Darbelnet, J. (1995): *Comparative Stylistics of French and English. A methodology for translation*. Aus dem Französischen übers. u. bearb. v. J. C. Sager u. M.-J. Hamel. Amsterdam: Benjamins.

[vPol88] von Polenz, P. (1988): Deutsche Satzsemantik: *Grundbegriffe des Zwischen-den-Zeilen-Lesens*. Berlin, New York: de Gruyter.

ANMERKUNGEN

1 AS = Ausgangssprache/ausgangssprachlich, ZS = Zielsprache/zielsprachlich; Segmente sind neben ganzen Sätzen auch Überschriften und Aufzählungspunkte. Sie werden i.d.R. durch Satzzeichen sowie durch Absatzmarken (¶) begrenzt.

2 In seinem Vergleich der Ergebnisse von Humanübersetzung und MÜ bezeichnet Sager die Sprache maschinell generierter Übersetzungen als künstliche, auf der Basis natürlicher Sprachen modellierte Systeme. Er weist darauf hin, daß im Grunde genommen jedes MÜ-System eine eigene Sprache produziert, so daß man eigentlich von 'LOGOS Englisch', 'T1 Englisch' oder 'PT Englisch' sprechen sollte [Sag94:257].

3 Der EAGLES-Bericht [EAGLES96:144] definiert *exact match* als "a perfect character by character match between current source segment and stored source segment". Alle übrigen Retrieval-Ergebnisse – d.h. auch solche Treffer, die Unterschiede hinsichtlich Satzzeichen oder Groß-/Kleinschreibung aufweisen – gelten als *fuzzy matches*.

4 So auch die Definition von TMs im EAGLES-Bericht:

> "a translation memory is a multilingual text archive containing (segmented, aligned, parsed and classified) multilingual texts, allowing *storage and retrieval of aligned multilingual text segments* against various search conditions" [EAGLES96:140; meine Hervorhebung].

[5] Es sei darauf hingewiesen, daß der Begriff der 'Relevanz', auf den ich später noch einmal zurückkommen werde, selbst für *exact matches* nicht unbedingt problemlos ist, da auch bei Identität von zu übersetzendem AS-Segment und TM-Suchergebnis die Notwendigkeit bestehen kann, das vom TM als Übersetzung angebotene ZS-Segment anzupassen. Als Stichworte seien hier lediglich Homographien und insbesondere Unterschiede bezüglich Textkohäsion und -kohärenz sowie Thema-Rhema-Struktur genannt. Für den Übersetzer wird die 'Qualität' eines Treffers sicherlich in entscheidendem Maße dadurch bestimmt, wie gering der Umfang der Modifikationen ist, die vorgenommen werden müssen, um ein ZS-Segment des TM in den aktuellen Text einzupassen. Bewertet werden kann m.E. aber lediglich die eigentliche Retrieval-Leistung des Systems, d.h., ob ein entsprechendes AS-Segment im TM gefunden wurde oder nicht.

[6] Daß ein *exact match* bei einzelnen TM-Systemen allerdings alles andere als ein 'perfect character by character match' sein kann, zeigen die Beispiele in [See/Nüb99], wonach die TM-Komponente des MÜ-Systems 'LANGENSCHEIDTS T1' auch beliebige Umstellungen der Zeichenketten noch als 100%ige Entsprechung akzeptiert.

[7] Daneben werden im EAGLES-Bericht u.a. auch empirische Untersuchungen vorgeschlagen, bei denen der Nutzen von *fuzzy matches* dadurch bestimmt werden soll, daß beobachtet wird, wie Übersetzer solche Retrieval-Ergebnisse verwenden. Ein solches Monitoring-Verfahren wie es beispielsweise von Krings [Kri95] für die Untersuchung des Prozesses der Nachredaktion maschineller Übersetzungen konzipiert und angewandt wurde, müßte zum einen für TM-Systeme zunächst noch entwickelt werden, zum anderen würde durch das Beobachten von Übersetzern natürlich nicht die Retrieval-Leistung des Systems bewertet, sondern vielmehr der Einfluß von TM-Systemen auf das Übersetzungsverhalten untersucht.

[8] Als dritter Typ der semantischen Ähnlichkeit wird die *assoziative Ähnlichkeit* (Vergleich der Häufigkeit, mit der sprachliche Ausdrücke mit einem anderen, als Stimulus vorgegebenen Ausdruck assoziiert werden) angeführt. Diese ist jedoch für unsere Zwecke nicht von Belang.

[9] Die Firma TRADOS empfiehlt im Handbuch ihrer TRANSLATOR'S WORKBENCH z.B. eine Untergrenze von 60-75%.

[10] Die Begriffe 'Implikation' und 'Explikation' entstammen ursprünglich der komparativen Stilistik. Die Vertreter der *stylistique comparée* haben mit *implicitation* und *explicitation* Unterschiede in der Explizitheit von AS-Einheit und ZS-Einheit bezeichnet [Vin/Darb95]. Implikation und Explikation sind jedoch Phänomene, die auch beim intralingualen Vergleich von 'Originaltexten' und 'Updates' festzustellen sind.

[11] Das MS WORD Beispiel wurde dem Handbuch *Arbeiten mit Microsoft Word: Textverarbeitungsprogramm, Version 5* (Microsoft Corporation 1989, S. 30.3) entnommen. Das WORDPERFECT-Beispiel entstammt dem *WordPerfect Arbeitsbuch, Version 5.1* (Word-

Perfect Corporation 1989, S. 412) und wurde zu Demonstrationszwecken modifiziert. Es lautet im Original:

> *Mit Hilfe der Style-Funktion von WordPerfect kann das Formatieren von Texten weitestgehend automatisiert werden.*

[12] Entsprechend ergeben sich z.B. mit TRADOS TRANSLATOR'S WORKBENCH Match-Werte von 94% (Beispiel 3) bzw. 95% (Beispiel 4).

[13] Daneben liegen auch die entsprechenden englischen Zieltexte vor.

[14] Die Originaldokumente werden mit ADOBE FRAMEMAKER auf UNIX-Rechnern erstellt. Für die Zwecke dieser Arbeit wurde der Text im ASCII-Format aus den Dokumenten extrahiert.

[15] Es handelt sich hierbei um die TM-Systeme TRADOS TRANSLATOR'S WORKBENCH (Version 1.05), STAR TRANSIT (Version 2.1) und IBM TRANSLATIONMANAGER (Version 2.0) Die Untersuchungen wurden bereits vor längerer Zeit durchgeführt, so daß inzwischen neuere Versionen der verschiedenen TM-Systeme zur Verfügung stehen. Da es hier jedoch nicht um einen Vergleich bzw. Test der Systeme geht, sondern vielmehr um Überlegungen zur Entwicklung von Evaluierungskriterien, spielt die Aktualität der TM-Systeme nur eine geringe Rolle.

[16] IBM TRANSLATIONMANAGER konnte in diesen Vergleich nicht einbezogen werden, da dieses System keine Match-Werte angibt. Dem Anwender wird lediglich mitgeteilt, ob ein Suchergebnis ein *exact match* oder ein *fuzzy match* ist.

[17] Zu diesem Zweck wurde das am INSTITUT FÜR ANGEWANDTE INFORMATIONSWISSENSCHAFT (IAI) in Saarbrücken entwickelte morphologische Analysewerkzeug MPRO verwendet [Maas96]. Das folgende Beispiel zeigt das Analyseergebnis für Satz 1 a) aus Tab. 7:

```
{ori=Sondernummern,c=noun,lu=sondernummer,s=abstract,
t=besonder#nummer,cs=a#n,ts=sonder#nummer,ds=besonder#nummer,
ls=besonder#nummer,ss=a#abstract,w=2,ehead={nb=plu,g=f}}

{ori=sind, lu=sein,c=w,sc=verb,vtys=sein,tns=pres,mode=ind,
per=3;1,nb=plu}

{ori=im, lu=in,c=w,sc=p,ehead={case=dat,nb=sg,g=m;n},pcom}

{ori=gesamten, c=adj,lu=gesamt,endung=en,deg=base,t=gesamt,
cs=a,ts=gesamt, ds=gesamt,ls=gesamt,ss=a,w=1}

{ori=Mobilvermittlungsstellenbereich,c=noun,lu=mobilvermittlungsstellenbereich,
s=domain,t=mobil#vermittlung#stelle#bereich,cs=a#n#n#n,
ts=mobil#vermittlungs#stellen#bereich,
ds=mobil#ver$mitteln~ung#stelle#bereich,
```

```
ls=mobil#ver$mitteln#stelle#bereich,ss=a#ation#loc#domain,w=4,
ehead={case=nom;dat;acc,nb=sg,g=m}}

{ori=gueltig,c=adv,lu=gueltig,deg=base,t=gueltig,cs=a,
ts=gueltig,ds=gueltig,ls=gueltig,ss=a,w=1,lng=germ}

{ori=., lu=stop,c=w,sc=punct}
```

Von den zahlreichen verschiedenen Merkmalen wurde lediglich die morphologische Struktur (ls) verwendet. Gegenüber dem Lemma (lu) bietet diese die Möglichkeit, Derivationsmuster zu berücksichtigen und Oberflächenunterschiede auszugleichen, die z.B. durch Bildung bzw. Auflösung von Komposita entstehen (vgl. z.B. ‘Anrufumlenkungsbedingungen’ vs. ‘Bedingungen für Umlenkung des Anrufs’ in Bsp. 2). Die in Form verschiedener Trennzeichen (‘#’ für Wortstämme, ‘$’ für nicht abtrennbare wortbildende Präfixe und ‘_$’ für abtrennbare wortbildende Präfixe) verfügbaren Derivationsangaben bleiben unberücksichtigt. Die Trennzeichen wurden durch Leerzeichen ersetzt.