LDV-FORUM

Forum der Gesellschaft für Linguistische Datenverarbeitung (GLDV)

LDV-Forum 16 (1999) 1/2

Zeitschrift für Computerlinguistik und Sprachtechnologie

GLDV-Journal for Computational Linguistics and Language Technology

Offizielles Organ der GLDV

Herausgeber

Prof. Dr. Gerhard Knorz; Gesellschaft für Linguistische Datenverarbeitung

Anschrift: Fachhochschule Darmstadt, Fachbereich Information und Dokumentation, Haardtring 100, D-64289 Darmstadt; Tel: (06151) 16-8499; Fax: (06151) 16-8980; e-mail: knorz@www.iud.fh-darmstadt.de

Redaktion

Gerhard Knorz

Wissenschaftlicher Beirat

Prof. Dr. W. Hoeppner (hoeppner@unidui.uni-duisburg.de);
Prof. Dr. Gerhard Knorz; Prof. Dr. Winfried Lenders (lenders@uni-bonn.de); ➢

Editorial

Wie Gerd Knorz, der Herausgeber des LDV-Forum, bereits in Heft 1998/2 ankündigte, ist das vorliegende Doppelheft aus den Aktivitäten des Arbeitskreises „Maschinelle Übersetzung“ der GLDV hervorgegangen. Der Arbeitskreis hat sich 1999 dem thematischen Schwerpunkt „Maschinengestützte Übersetzung“ gewidmet. Mit dem Ziel einer Evaluierung maschineller Übersetzungswerkzeuge für den professionellen Einsatz fand im Rahmen der GLDV-Jahrestagung 1999 eine Sektion zum Thema „Leistungsfähigkeit und Einsatzmöglichkeiten von Translation-Memory-Systemen“ statt. Die Beiträge der vorliegenden Ausgabe spiegeln die unterschiedlichen Facetten der Veranstaltung wider und schließen mit einem Ausblick auf die geplanten Evaluierungsaktivitäten.

Interessierte Leser, die entweder Anregungen oder Kritik an dem geplanten Verfahren haben, sind herzlich aufgefordert, diese an Rita Nübel (rita@iai.uni-sb.de) und Uta Seewald-Heeg (seewald@heeg.de), die Leiterinnen des Arbeitskreises „Maschinelle Übersetzung“, zu richten. Neben konstruktiven Vorschlägen zum Verfahren der Evaluierung sind alle Interessierten eingeladen, an der Evaluierung selbst aktiv mitzuwirken.

Eine anregende Lektüre der folgenden Beiträge wünschen Ihnen

Uta Seewald-Heeg und Rita Nübel

➢ Prof. Dr. Ulrich Schmitz (e-mail: ulrich.schmitz@uni-essen.de) **Erscheinungsweise** 2 Hefte im Jahr, halbjährlich zum 31. Mai und 31. Oktober. Preprints und redaktionelle Planungen sind laufend und aktuell unter der Adresse http://www.iud.fh-darmstadt.de/iud/wwmeth/publ/ldvforum/menu1.htm einsehbar. **Bezugsbedingungen** Für Mitglieder der GLDV ist der Bezugspreis des LDV-Forum im Jahresbeitrag mit eingeschlossen. Jahresabonnements können zum Preis von DEM40,- (incl. Versand), Einzelexemplare zum Preis von DEM20,- (zuzügl. Versandkosten) bestellt werden: LDV-Forum, c/o Dr. Bernhard Schröder, Poppelsdorfer Allee 47, 53115 Bonn **Fachbeiträge** Unaufgefordert eingesandte Fachbeiträge werden vor Veröffentlichung von mindestens 2 ReferentInnen begutachtet. Manuskripte sollten deshalb möglichst frühzeitig eingereicht werden und bei Annahme zur Veröffentlichung in jedem Fall elektronisch und zusätzlich auf Papier übermittelt werden. Artikel sind bevorzugt einzureichen in den Formaten Microsoft Word für Windows® oder Word Perfect® für Windows. Eine Dokumentvorlage für Word für Windows® kann unter der Adresse ftp://www.iud.fh-darmstadt.de/iud/wwmeth/publ/ldvforum/ldvforum.dot heruntergeladen werden. Sie enthält die wichtigsten Styles. **Rubriken** Die namentlich gezeichneten Beiträge geben ausschließlich die Meinung der AutorInnen wieder. Einreichungen sind – wie bei Fachbeiträgen – an den Herausgeber zu übermitteln. **Druck und Vertrieb** GLDV **Satz** Kurt Thomas, Bonn **Auflage** 400 Exemplare **Anschrift der GLDV** Prof. Dr. R. Hausser Universität Erlangen-Nürnberg Abteilung für Computerlinguistik Bismarkstraße 12 D-91054 Erlangen; e-mail: rrh@linguistik.uni-erlangen.de.

Einleitung

Rita Nübel (IAI Saarbrücken)
Uta Seewald-Heeg (Hochschule Anhalt)

Maschinelle Übersetzungswerkzeuge haben in Übersetzungsunternehmen, vor allem im Bereich der Lokalisierung, inzwischen einen festen Platz erhalten. Auch freiberuflich arbeitende Übersetzer setzen sich zunehmend mit neuen Technologien auseinander, die eine Effizienzsteigerung ihrer Übersetzungsleistung versprechen. In diesem Zusammenhang rücken vor allem satzspeicherbasierte Systeme, sogenannte Translation-Memory-Systeme, in das Zentrum des Interesses; denn wie die Ergebnisse einer 1998 vom Arbeitskreis „Maschinelle Übersetzung" der GLDV durchgeführten Evaluierung der linguistischen Performanz vollautomatischer maschineller Übersetzungssysteme nahelegen [Nübel/Seewald-Heeg98], muss beim Einsatz maschineller Übersetzungssysteme für eine qualitativ hochwertige Übersetzung von Dokumenten mit einer relativ aufwendigen Postedition gerechnet werden, die den Nutzen der maschinellen Übersetzung (MÜ) vielfach relativiert, zum Teil sogar in das Gegenteil verkehrt. Die Frage nach den Leistungsmerkmalen von Translation Memories (TM) ist daher sowohl bei zahlreichen industriellen Nutzern als auch bei potentiellen Anwendern von zentralem Interesse, zumal die Produktpalette bei diesem Systemtyp in den vergangenen Jahren ähnlich wie im Bereich der MÜ-Produkte deutlich angewachsen ist.

Für den Arbeitskreis „Maschinelle Übersetzung" war das Anlass, sich 1999 mit TM-Systemen auseinanderzusetzen und Möglichkeiten zu erörtern, die Leistungsfähigkeit von TMs für den professionellen Einsatz zu bewerten.

Anders als bei vollautomatischen Übersetzungssystemen basiert die TM-Technologie auf der Wiederverwendung bereits übersetzter Textsegmente, die im Verlauf vorausgehender Übersetzungsprozesse als Satzpaare gespeichert werden. Da TMs ohne Daten geliefert werden, sind sie ausschließlich für professionelle Nutzer geeignet, bei deren Übersetzungstätigkeit Texte mit einem gewissen Wiederholungsanteil anfallen. Eine Evaluierung der Leistungsfähigkeit von TMs muss daher die Trefferrate berücksichtigen, mit der Textsegmente im bereits vorhandenen Material (Referenzmaterials) aufgefunden werden. Um dies zu bewerkstelligen, sind geeignete Testdaten und Bewertungsmethoden erforderlich. Linguistisch orientierte Qualitätskriterien, die für die Bewertung der Übersetzungsqualität von MÜ-Systemen verwendet werden, wie z.B. Grammatikalität, Lesbarkeit,

Stil oder Verständlichkeit, sind hier nicht von Bedeutung, da TMs selbst keine Übersetzung erzeugen. Für die Bewertung von TMs ist vielmehr zentral, inwieweit sie aus einer möglichen Vielfalt ähnlicher, bereits archivierter Übersetzungen die besten bzw. „ähnlichsten" Kandidaten auswählen, die als Basis für die jeweils aktuelle Übersetzung eines Satzes dienen können. In diesem Zusammenhang ist zunächst der Begriff der Ähnlichkeit zu klären, den auch Reinke in seinem Artikel zur Evaluierung der linguistischen Leistungsfähigkeit von Translation-Memory-Systemen problematisiert.

In einem ersten Treffen des Arbeitskreises „Maschinelle Übersetzung" tauschten Teilnehmer aus der Übersetzerbranche sowie aus Forschung und Entwicklung ihre in unterschiedlichen Zusammenhängen gesammelten Erfahrungen mit TM-Systemen aus. Aufgrund des hier bekundeten Interesses, die Performanz von TMs vor allem auch hinsichtlich ihrer Retrieval-Leistung im Rahmen eines unabhängigen Gremiums wie des GLDV-Arbeitskreises zu untersuchen, fand auf der GLDV-Jahrestagung 1999 in Frankfurt ein Workshop statt, auf dem TM-Systeme aus Nutzer- und Entwicklerperspektive zunächst vorgestellt und Ideen für eine Evaluierung diskutiert wurden.[1]

Aufgrund der unterschiedlichen Provenienz der Workshop-Teilnehmer ist es gelungen, mit dem vorliegenden Band verschiedene Facetten der Beschäftigung mit TM-Technologie und ihre gegenseitigen Berührungspunkte zu präsentieren:

Im ersten Beitrag werden von H. Elsen die typischen Aufgaben eines im Bereich der Lokalisierung tätigen Übersetzungsunternehmens skizziert und die technologiespezifischen und organisatorischen Anforderungen aus Managementsicht beschrieben. Der Beitrag verdeutlicht, dass ohne den professionellen Einsatz von Sprachtechnologiewerkzeugen wie TMs heute kaum ein Übersetzungsdienstleister in der Lage ist, wirtschaftlich erfolgreich zu arbeiten. Da durch den Einsatz von TMs die Produktivitätssteigerung unterstützt werden soll, haben potentielle Nutzer sowohl ein großes Interesse an der Transparenz von Funktionalitäten solcher Werkzeuge als auch an verlässlichen Aussagen zur Systemperformanz. Diese beiden Fragen bestimmten schwerpunktmäßig auch die Beiträge des Workshops und der anschließenden Diskussion (siehe „Ausblick") und werden in den nachfolgend kurz vorgestellten Beiträgen dieses Bandes aus unterschiedlichen Perspektiven diskutiert.

U. Seewald-Heeg und R. Nübel betrachten in ihrem gemeinsamen Beitrag zwei Systeme, die die Funktionalität der vollautomatischen Übersetzung mit jener der Satzarchivierung (im Sinne einer einfachen TM-Funktionalität) verbinden.[2] Der

Beitrag geht zunächst der Frage der Leistungsfähigkeit der Satzarchive beim Abgleich von Sätzen mit dem archivierten Referenzmaterial nach und mündet in die Frage, inwieweit sich die Satzarchive der untersuchten Systeme als Erweiterung des Wörterbuchs für komplexe lexikalische Strukturen eignen.

Die beiden nachfolgenden Beiträge stellen die zentralen Funktionsweisen zweier auf dem Markt verbreiteter TM-Systeme vor: Der Beitrag von H. Bohn illustriert die wichtigsten Funktionen der Translator's Workbench von Trados, und J. Klein beschreibt die Funktionsweise von Transit, dem TM-Werkzeug der Firma STAR.

Erfahrungen mit verschiedenen TM-Systemen (TranslationManager von IBM und Translator's Workbench von Trados) aus Sicht eines professionellen Anwenders dokumentiert A. Pesch in ihrem Beitrag.

U. Reinke diskutiert in seinem ersten Beitrag Ideen für eine Optimierung von TMs durch eine intelligente Verwendung bereits vorhandener Wissensquellen wie paralleler Korpora und Terminologie. K. Lee et al. stellen in ihrem Beitrag das Forschungssystem MIRAC vor, mit dem multilinguale Dokumente (semi-)automatisch aligniert und auf ihre Konsistenz auf Termebene sowie auf einer semantisch abstrakteren Ebene überprüft werden.

Die Evaluation von TM-Systemen ist Gegenstand des zweiten Beitrags von U. Reinke, der im Zusammenhang mit der Frage nach der Bewertung der linguistischen Leistungsfähigkeit von TMs zum einen die Verwendbarkeit von Methoden, die für die Evaluation von *Information Retrieval* Systemen angewendet werden, erörtert, zum anderen darstellt, wie bestimmte Eingabemodifikationen die Trefferraten der TMs beeinflussen und schließlich beispielhaft illustriert, wie die Berücksichtigung linguistischer Eigenschaften beim Abgleich von Sätzen mit dem Referenzmaterial die Retrievalleistung verbessern kann.

Der „Ausblick" am Ende fasst die Abschlussdiskussion zusammen, in der Ideen und Anforderungen an eine TM-Evaluation erörtert und in ein Evaluierungsszenario eingearbeitet wurden, das die folgenden für die Evaluierung erforderlichen Schritte präzisiert.

Literatur

[Nübel/Seewald-Heeg98] Nübel, R. und Seewald-Heeg, U. (Hrsg.) (1998): *Evaluation of the Linguistic Performance of Machine Translation Systems*. Proceedings of the KONVENS-98 Workshop in Bonn. St. Augustin: Gardez! Verlag.

ANMERKUNGEN

[1] Präsentationen einzelner Beiträge können über die Programmübersicht des Workshops unter folgendem URL eingesehen werden:
http://www.heeg.de/~uta/AK-Programm-Jahrestagung.htm.

[2] Personal Translator 2000 Office plus von Linguatec und IBM sowie Langenscheidts T1, eine Entwicklung aus dem Haus Lernout & Hauspie.

Werkzeuge haben ihren Sinn im Nutzen

Harald Elsen
Delta International CITS GmbH

1 Einleitung

In diesem Artikel soll das Umfeld eines Lokalisierungsunternehmens skizziert werden, also eines Unternehmens, in dem Werkzeuge zur Verarbeitung natürlicher Sprache professionell zur Steigerung der Wertschöpfung eingesetzt werden.

2 Das Dienstleistungsspektrum

Im Gegensatz zur traditionellen Übersetzungsbranche haben Lokalisierungsunternehmen einen sehr dezidierten Fokus:

- Technische Dokumentation
- Online-Dokumentation (z.B. auch Online-Hilfe)
- Software
- Collaterals[1]
- Websites

Hinsichtlich der Konsistenz – und somit der Qualitätssicherung – ist es von Vorteil, wenn alle diese produktbegleitenden Medien von einem Dienstleister in eine Zielsprache lokalisiert[2] werden. Vorteilhaft wirkt sich ferner aus, wenn ein Dienstleister mehrere Sprachen über Niederlassungen in den Zielmarktländern[3] anbieten und somit inhaltliche und prozeßorientierte Konsistenz gewährleisten kann. Das hat zwei Vorteile:

- Der Hersteller braucht nur einen Projektmanager auf seiner Seite, der über einen Kontakt bei seinem Dienstleister mehrere Sprachen abdeckt, und nicht für jede Sprache einen. Das erspart ihm Personalkosten.
- Der Dienstleister kann seine Effizienz steigern, da – bei guter Organisation (s.u.) – ein Fehler nur einmal gemacht wird und nicht erneut bei jeder weiteren Sprache, in die übersetzt wird. Dies eröffnet ihm Spielräume bei der Preisgestaltung.

Somit erklären sich auch die derzeit häufigen Akquisitionen von mittleren Lokalisierungunternehmen durch einige große Unternehmen dieser Branche, die als *Multilingual Vendor (MLV)* auf dem Markt auftreten.[4]

Ziel der Dienstleister ist es, dem Kunden soviel Produktionsschritte abzunehmen wie möglich, am besten die komplette Produktionskette, d.h. schon die Erstellung der Dokumentation (*Authoring*) als ersten Schritt in der Produktionskette, bis hin zur Lieferung der fertig gedruckten Handbücher und/ oder der verpackten CDs – im Dienstleistungsjargon als *Full Service* bezeichnet.

3 Der Markt

Jeder Markt wird an seiner Entwicklung und seinen Entwicklungsmöglichkeiten gemessen. Folgende Aussagen von Repräsentanten der Lokalisierungsbranche skizzieren die gegenwärtige Situation:

(1) „Die 10 größten Unternehmen decken weniger als 10% des Lokalisierungsgeschäfts ab."

(2) „Die Branche verzeichnet 25–30% Wachstum pro Jahr."

(3) „Durch Einsatz von Translation-Memory-Systemen wird eine Kostenersparnis von bis zu 50% erwartet."

(4) „Manche Kunden erwarten einen jährlichen Preisnachlaß von ca. 5%."

(5) „Die Bearbeitungszeiträume werden immer kürzer bis hin zur simultanen Erstellung eines Produkts in allen Zielsprachen."

(6) „Die Qualitätsanforderungen steigen."

Zu Aussage (1): Hier zeigt sich, daß ein großes Potential für Einzelübersetzer und klein- bis mittelständische Unternehmen besteht, Übersetzungsaufträge aus diesem Bereich zu übernehmen .

Zu Aussage (2): Solche Wachstumszahlen sind recht gut für einen Markt.

Zu Aussage (3): Diese Aussage relativiert Aussage (2). Hatte vorher ein Projekt ein Volumen von ungefähr 100.000 Wörtern bei einem Wortpreis von 0,38 DM, dann bedeutete das einen Umsatz von 38.000 DM. Durch die Verwendung von Translation-Memory-Systemen kann es somit – bezogen auf den Umsatz

eines einzelnen Projektes – zu einer Umsatzeinbuße von 30–50% kommen. Dem steht jedoch lediglich ein Wachstumspotential von 25–30% gegenüber.

Zu Aussage (4): Geht man von einem gesunden Dienstleistungsunternehmen aus, dann kann bei einem Preisnachlaß dieser Größe für das Unternehmen noch ein Gewinn (nach Abzug aller Kosten) von maximal 10% erwirtschaftet werden. Die Forderung des Kunden verursacht somit eine nicht unerhebliche Reduktion des Gewinns. Da der Kostendruck in der Lokalisierungsbranche sehr hoch ist, und das nicht zuletzt aufgrund der großen Anzahl an Mitbewerbern, bleiben nur begrenzte finanzielle Mittel für Investitionen (Hard- und Software, Ausbildung etc.) übrig.

Zu Aussage (5): Die Bearbeitungszeiträume werden aufgrund des Marktdrucks auf die Produkthersteller immer kürzer. Wer sein Produkt vor der Konkurrenz auf den Markt bringt, sichert sich die größten Marktanteile oder bewahrt seine gute Marktposition. Früher galt das lediglich für die lokalen Märkte, die zunehmend zu einem globalen Markt verschmelzen. Früher wurde ein Produkt erst lokalisiert, nachdem das Originalprodukt fertiggestellt war. Heute geht der Trend mehr zum *SimShip (Simultaneous Shipment)*.[5] Das bedeutet für den Dienstleister, daß er nicht ein fertiges Produkt bearbeitet, sondern bis zur Fertigstellung jederzeit mit Änderungen aller Art rechnen muß und diese möglichst sofort einzuarbeiten hat – eine Herausforderung an die konzeptionelle Modularisierung der Produktbausteine (bei Dokumentation z.B. Textbausteine) und an Versionskontrollsysteme.

Zu Aussage (6): Hier stellt sich die Frage: Was ist Qualität? Große Firmen, die um die Bedeutung von Lokalisierung, Internationalisierung, Globalisierung etc. wissen, haben eigene Qualitätssicherungsabteilungen, die stichprobenhaft die Arbeit der von ihr beauftragten Lokalisierungsunternehmen überprüfen, bewerten und statistisch erfassen.[6] Gute Qualität liegt nur dann vor, wenn die Übersetzung orthographisch korrekt und der Stil der Textsorte angemessen und an die Zielkultur adaptiert ist, wobei zielkulturelle und stilistische Adaptionen wesentlich von der *Corporate Identity* des Kunden abhängen.[7]

Zu den Aussagen (4)–(6): Preis, Bearbeitungszeitrahmen und Qualität sind keine unabhängigen Größen. Alle drei stehen in Proportion zueinander, was ein Trainer für Projektmanagement einmal folgendermaßen als das „First Law of Project Management“ zusammengefaßt hat:

$$f(\text{Preis, Bearbeitungszeitrahmen, Qualität}) = \text{konstant}$$

4 Die Organisation

Jedes Unternehmen paßt sich von seiner Organisation her an die Marktanforderungen an, damit es überlebensfähig ist. Die drei Hauptkriterien für die Organisation eines erfolgreichen Lokalisierungsunternehmens sind:

- *On-Demand*-Ressourcen
- Funktionalität und Virtualität
- Kommunikations- und Wissensmanagement

Wann und wie oft ein Produkt aktualisiert oder neu erstellt wird, unterliegt zwar meist Zyklen, die aber alle ihre eigene Frequenz haben. Somit ist kein gleichmäßiger Bedarf an Lokalisierung gegeben (*peak and valley business*). Oft erfolgen zudem Ankündigung und Auftragsvergabe von Seiten der Produkthersteller sehr kurzfristig, was die Vorbereitungszeit für das Lokalisierungsprojekt sehr stark verkürzt. So muß das Lokalisierungsunternehmen einen Stamm von freien Mitarbeitern aufbauen, der bei Bedarf aktiviert werden kann, wobei eines der Hauptprobleme in diesem Zusammenhang die Verfügbarkeit gut ausgebildeter freier Ressourcen ist.

Jedes zu lokalisierende Produkt stellt andere Anforderungen an das Wissen und die Fähigkeiten der Teammitglieder, was das Problem der Verfügbarkeit zusätzlich verschärft. Somit ist ein Ressourcenmanagement von vitaler Bedeutung.

Hat man letztendlich ein geeignetes Team aufgebaut, muß die Kommunikation und der Transfer des benötigten Wissens gut geplant, durchgeführt und ständig kontrolliert werden. Dies ist von vitaler Bedeutung, da es sich oft um geographisch verteilte Teams (virtuelle Teams) handelt, die gemeinsam an einem Projekt arbeiten, was dazu führt, daß bereits ein kleiner Fehler, der aufgrund eines Mißverständnisses zustande kommt, den Erfolg des Projekts gefährden kann.[8]

Das Management hat auch dafür Sorge zu tragen, daß nach Beendigung des Projekts das projektspezifische Wissen nicht zusammen mit der Auflösung des Projektteams verloren geht. Dieses Wissen kann bzw. muß bei späteren Wiederholungsprojekten, oder Projekten mit ähnlichen Anforderungen, wiederverwendet werden können.

5 Die Mitarbeiter

Jede Organisation ist nur so gut wie ihre Mitarbeiter. In diesem Zusammenhang muß festgestellt werden, daß es – sieht man einmal von Weiterbildungsveranstaltungen ab – kaum Ausbildungsinstitutionen für diese Branche gibt (bis auf wenige Ausnahmen, die aber alle nicht in Deutschland angesiedelt sind).[9] Die daraus resultierenden Folgen für ein Lokalisierungsunternehmen sind daher unter anderem:

- Lange Einarbeitungszeit von 3–6 Monaten
- Überdurchschnittlich hohe Ausbildungskosten
- Zusatzbelastung der „erfahrenen" Mitarbeiter als Ausbilder
- Ineffizienter Einsatz von Werkzeugen
- Qualitätsmängel in der Produktion durch fehlende Erfahrung

Hat man erst einmal einen effizienten Stamm von freien und festen Mitarbeitern aufgebaut, stellt sich ein neues Problem: Die gut eingearbeiteten Mitarbeiter sind potentielle Kandidaten für Abwerbungen durch Mitbewerber. So liegt die Mitarbeiter-Fluktuationsrate in Irland (mit Dublin als Hochburg der Lokalisierungsbranche[10]) z.B. bei bis zu 30%. Der Verlust erfahrener Mitarbeiter bedeutet stets:

- Verlust von Wissen
- Verlust von operativer Intelligenz[11]

Besonders empfindlich trifft es ein Lokalisierungsunternehmen, wenn ganze Teams abgeworben werden, oder zumindest die sogenannten *Key Players*, wodurch gezielt operative Intelligenz eingekauft wird.

Die Nachfrage nach Profis ist immens, da es nur verhältnismäßig wenige wirkliche Profis gibt.

Eine Grundvoraussetzung für jeden ist der sichere Umgang mit einem PC, gängigen Office-Paketen und dem Internet mit all seinen Formen der Datenübertragung (E-Mail, FTP, HTTP etc.). Die Zeiten der Schreibmaschine und des Faxgerätes sind vorbei.

6 Materialien und Werkzeuge

Was einen Profi zu einem großen Teil ausmacht, ist seine Kompetenz im Umgang mit den ihm zur Verfügung stehenden Werkzeugen. Die hier relevanten Fragen sind also:

- Welches Material läßt sich mit welchem Werkzeug am effizientesten bearbeiten, ohne daß Material verloren geht?
- Wie sollten Werkzeuge beschaffen und das Material aufbereitet sein, damit im nächsten Produktionsschritt ebenfalls Werkzeuge eingesetzt werden können?

Betrachten wir zunächst einmal das Material.

6.1 Das Material

Das zu lokalisierende quellsprachige Material besteht aus Texten mit vollständigen Sätzen, Phrasen und einzelnen Wörtern. Sätze und Phrasen sind meistens die grundlegenden Übersetzungseinheiten.[12] Einzelne Wörter und Phrasen bezeichnen thematisierte Objekte oder Kernbegriffe, die als Terminologie eingestuft werden können und, wenn möglich, vor dem Produkt zu lokalisieren sind.

Somit läßt sich über den Text hinaus weiteres Material spezifizieren:

- Terminologie
- Archive von Übersetzungseinheiten (*Translation Memories*)

Die Terminologie wird in allen Produktionsschritten (*Authoring*[13], Übersetzung, Linguistische Qualitätssicherung) benötigt und unterliegt somit einer laufenden Veränderung und Anpassung, für die der Einsatz von Werkzeugen unumgänglich ist.

Translation Memories werden in den Produktionsschritten Übersetzung und linguistische Qualitätssicherung eingesetzt. Effiziente Algorithmen zur Ermittlung des *best match* sind hier unumgänglich.

6.2 Die Werkzeuge

Die wichtigsten Werkzeuge sind Terminologie- und Übersetzungswerkzeuge. Terminologie-Kandidaten lassen sich durch maschinelle Verfahren extrahieren, wobei die Effizienz von der Strategie und den Basisdaten des Werkzeugs abhängt. Die gewonnene oder zur Verfügung gestellte Terminologie kann in speziell dafür vorgesehenen Datenbanksystemen verwaltet werden. Solche Datenbanksysteme sind Bestandteil der gängigen Translation-Memory-Systeme und einiger Übersetzungssysteme und werden in den Arbeitsprozeß eingebunden, damit die Konsistenz und Qualität der Übersetzung erhöht wird.

Wie oben beschrieben, kann der effiziente Einsatz von Translation-Memory-Systemen auf einen einzelnen Auftrag bezogen zu einem Umsatzverlust von 30-50% führen. –Neue Werkzeuge erfordern oft auch neue Prozesse, deren Einführung und Integration in den Workflow oft höhere und kaum kalkulierbare Kosten verursachen als die eigentlichen Investitionskosten für die Software.

Eines der größten Mankos für die Anwender ist das Fehlen von Standards, was zu Inkompatibilitäten zwischen verschiedenen auf dem Markt angebotenen Systemen führt. Das Translation-Memory-System B, das unter bestimmten Gesichtspunkten als optimal eingeschätzt wird, arbeitet nur mit dem Terminologie-Werkzeug A zusammen, wobei A hinsichtlich der Terminologieverwaltung möglicherweise keine ausreichenden Funktionalitäten bietet.

Einer nachteiligen Eigenschaft, die gegenwärtig alle Systeme charakterisiert, ist sicherlich künftig Abhilfe zu schaffen: Derzeit handelt es sich bei allen Produkten um lokale Anwendungen, die nur im Einzelplatz- oder Arbeitsgruppen-Szenario einsetzbar sind. Client/Server-Anwendung über TCP/IP (Internet) wären für die virtuellen und geographisch verteilten Teams das Ende der endlosen *Update-and-Merging*-Schwierigkeiten[14] bei längeren Projekten.

7 Zusammenfassung

Die Branche hat

- gerade erst ihre Adoleszenz erreicht,
- keine Jobeinsteiger, die eine ausreichende Vorbildung besitzen,
- an zeit- und existenzkritischem Material zu arbeiten, da sie zum überwiegenden Teil von der schnellebigen Softwareindustrie bestimmt ist,

- aufgrund der weltweiten Globalisierungsbestrebungen in den nächsten Jahren einen immens steigenden Bedarf an Dienstleistungen zu befriedigen,
- nur „70er-Jahre"-Werkzeuge für ihre Dienstleistungen zur Verfügung, die sich zudem in einem „High-Tech-Produkte-Krieg" befinden.

Investitionen in Ausbildung und Werkzeuge haben gute Chancen, auf fruchtbaren Boden zu fallen, besonders wenn man einen Blick in die Zukunft wagt, die zunehmend von *Online Gisted Translation (online draft-quality translation)*, *Multi Lingual Information Retrieval* etc. gekennzeichnet sein wird.

Literatur

[Esselink98] Esselink, Bert (1998): *A Practical Guide to Software Localization*. Amsterdam, Philadelphia: John Benjamins B.V.

ANMERKUNGEN

1 Die branchenübliche Bezeichnung für Werbematerial, Verpackung etc.

2 Terminus technicus für das Anpassen eines Produkts an den lokalen Zielmarkt.

3 Einer Regel in der Lokalisierungsbranche zufolge wird jede Übersetzung und Bearbeitung immer im Land der Zielsprache von Muttersprachlern durchgeführt, damit eine zeitgemäße und somit marktgerechte Version entsteht.

4 Interessenten an den neuesten Entwicklungen in der Lokalisierungsbranche können sich mit der Nachricht *SUB NEWS-L <Put your name & organization here>* an *listserv@multilingual.com* in die einschlägige Mailing-List eintragen.

5 Eine branchenübliche Bezeichnung, mit der die simultane Auslieferung eines Produkts in mehreren Sprachen gemeint ist.

6 Diese Auswertungen wirken sich direkt auf die Preisverhandlung und die Vergabe von Aufträgen aus.

7 Qualität ist also kein absoluter Wert, sondern wird idealerweise vor dem Beginn des Projektes definiert und u.U. während der Durchführung ständig justiert. Der Slogan „Qualität ist, was der Kunde als solche bezeichnet" zeigt, welchem Druck auch hier die Dienstleister ausgesetzt sind.

[8] Ein Projekt ohne Gewinn können sich Unternehmen in der Regel nicht öfter als zweimal erlauben, ohne in Liquiditätsprobleme zu geraten.

[9] Für das Selbststudium eignet sich [Esselink98], der eine Einführung in die Praxis der Lokalisierung bietet.

[10] Irland ist aufgrund seiner Sonderstellung in der Lokalisierung mittlerweile zum zweitgrößten Software-Exporteur der Welt (nach den USA) aufgestiegen.

[11] Das Wissen über das Zusammenspiel von einzelnen fachlichen Kompetenzen.

[12] Das trifft nicht auf alle Sprachen zu. Bei asiatischen Sprachen werden oft Paragraphen als Übersetzungseinheiten gewählt.

[13] Terminologie sollte schon zum Zeitpunkt der Produkterstellung definiert und idealerweise auch bereits übersetzt sein, damit das Ausgangsprodukt in sich schon konsistent erstellt wird. Zudem ist sie unerläßlich bei der Verwendung von *Controlled Language* im Kontext der maschinellen Übersetzung, durch deren Einsatz gute Ergebnisse erzielt werden können.

[14] Einzelne Translation Memories werden im Verlauf von Projekten öfters zusammengeführt (*merging*) und als aktualisierte Version (*update*) an die Übersetzer zur weiteren Verwendung zurückgesendet.

Translation-Memory-Module automatischer Übersetzungssysteme

Uta Seewald-Heeg (Hochschule Anhalt)
Rita Nübel (IAI Saarbrücken)

1 Systemtypen

Neben vollautomatisch arbeitenden Übersetzungswerkzeugen, bei deren Einsatz der Humanübersetzer (oder allgemein Benutzer) während des Übersetzungsprozesses nicht in den maschinellen Ablauf eingreifen kann,[1] und satzspeicherbasierten Systemen, d.h. Translation Memories (TM),[2] die selbst keinerlei Übersetzung vornehmen, sondern den Übersetzer bei der Recherche nach Textsegmenten unterstützen, die bereits im Zuge anderer Übersetzungen oder im Rahmen der Übersetzung des gerade in Arbeit befindlichen Dokuments übersetzt wurden, ist inzwischen ein Systemtyp entwickelt worden, der die Vorteile eines satzspeicherbasierten Systems mit denen eines vollautomatischen Übersetzungssystems verbindet. Systeme dieses Typs, die also Systemkomponenten auf unterschiedlichen Verarbeitungsstufen integrieren, bezeichnen wir nachfolgend als Integrierte Systeme.[3] Konfigurationen integrierter Systeme lassen sich durch die Kombination eines Translation-Memory-Systems mit einem vollautomatischen Übersetzungssystem erzeugen, wie dies beispielsweise bei der Kombination von Transit, dem TM der Firma Star, mit dem Übersetzungssystem Logos geschehen ist.[4]

Integrierte Systeme sind aber nicht nur das Ergebnis der Zusammenführung zweier sonst auch selbständig einsetzbarer ‚Großsysteme', sondern existieren mittlerweile in der Gestalt vollautomatischer Übersetzungssysteme, die um eine Translation-Memory-Komponente erweitert wurden.

Systemkonzeptionen dieser Art sind eine unmittelbare Konsequenz der eingeschränkten Leistungsfähigkeit vollautomatisch arbeitender maschineller Übersetzungssysteme. Wie die von [Nübel/Seewald-Heeg98] zusammengetragenen Ergebnisse der Evaluation der linguistischen Performanz vollautomatischer maschineller Übersetzungssysteme gezeigt hat, liegt der Prozentsatz der von den Systemen vollständig korrekt übersetzten Sätze ohne vorherige Systemoptimierung durch Hinzuladen benutzerdefinierter Wörterbucheinträge oder Zusatzwörterbücher bei den leistungsfähigsten Systemen durchschnittlich nur knapp über 30%. Für qualitativ hochwertige Übersetzungen ist beim Einsatz solcher Systeme

daher ein relativ hoher Posteditionsaufwand erforderlich. Die Erweiterung der Systeme durch ein Satzarchiv kann zu einer Effizienzsteigerung führen, da mit einem Satzarchiv neben der automatisch erstellten Übersetzung nun auch die Gesamtheit der bereits posteditierten oder manuell übersetzten Texte zur Verfügung gestellt wird.

2 Übersetzungssysteme mit Übersetzungsspeicher-Modul

Unter den kommerziellen Übersetzungssystemen verfügen gegenwärtig die Systeme T1 von Langenscheidt und Personal Translator plus (PT plus) von Linguatec über ein Übersetzungsspeicher-Modul.

Das Übersetzungsspeicher-Modul von T1 ist wie jenes von PT plus eine Datenbank, die in diesem Zusammenhang meist als Übersetzungsarchiv oder Satzarchiv bezeichnet wird. Das Übersetzungsarchiv enthält Satzpaare, die jeweils aus einem ausgangssprachlichen Satz und dessen Übersetzung in die jeweilige Zielsprache bestehen. Die Übersetzung eines Ausgangssatzes ist entweder das Ergebnis der automatischen Übersetzung oder aber eine vom Übersetzer erstellte bzw. von ihm posteditierte Übersetzung. Der Benutzer kann sich jede im Satzarchiv enthaltene Übersetzung eines Satzes anzeigen lassen, wobei die Suche sowohl vom ausgangssprachlichen Text als auch von dessen Übersetzung aus möglich ist.

Nachfolgend werden zunächst einige Merkmale der Benutzerschnittstelle des Satzarchivs von T1, anschließend jene von PT plus dargestellt und einer Evaluierung unterzogen.

2.1 Das Übersetzungsspeicher-Modul von T1

Das Satzarchiv von T1 lässt sich in verschiedene Module untergliedern, so dass bestimmte Satzarchivmodule beispielsweise in Abhängigkeit von der Fachrichtung und Textsorte eines zu übersetzenden Textes oder abhängig vom Auftraggeber ausgewählt werden können, während andere für einen bestimmten Übersetzungsauftrag unberücksichtigt bleiben, das System als Referenzmaterial bei der Suche nach im Archiv enthaltenen Sätzen also nur in den ausgewählten Modulen nachschlägt.

Das Übersetzungsarchiv von T1 lässt sich sowohl in Kombination mit der automatischen Übersetzung als auch ohne diese einsetzen. Beim kombinierten Einsatz des Satzarchivs mit der maschinellen Übersetzung kann der Benutzer spezifizieren, ob Sätze, die in der im Text vorliegenden Form oder einer ähnlichen Form im Übersetzungsarchiv gespeichert sind, nicht mehr maschinell übersetzt und deren Übersetzung automatisch in den Zieltext übernommen werden sollen. Soll das Übersetzungsarchiv dagegen separat eingesetzt werden, muss vom Benutzer die Option „T1 Übersetzung – Nie“ aktiviert werden (vgl. Abb. 1).

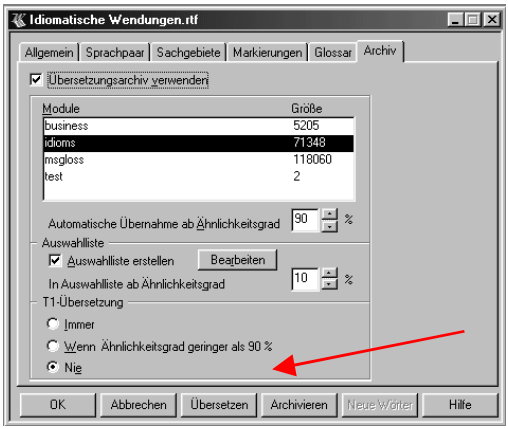


Abb. 1: Benutzerschnittstelle des Satzarchivmoduls von T1.

In einem als „Auswahlliste“ bezeichneten Bildschirmfenster des Übersetzungsarchivs werden bei der Bearbeitung der Suchergebnisse neben dem zu übersetzenden Satz bis zu drei im Satzarchiv als gleich oder ähnlich identifizierte Sätze bzw. deren Übersetzung angezeigt. Der Inhalt der einzelnen Teilfenster ist editierbar, und per Mausklick lässt sich die im Archiv gefundene und gegebenenfalls vom Benutzer bearbeitete Übersetzung in den Zieltext übernehmen.

2.2 Das Übersetzungsspeicher-Modul von PT plus

Ähnlich wie T1 bietet PT plus die Möglichkeit, neben der automatischen Übersetzung auch ein vom Benutzer definiertes Satzarchiv mit dem darin enthaltenen zweisprachigen Referenzmaterial zu verwenden. Der Benutzer kann sein Satzarchiv entweder mit den Ergebnissen der automatischen Übersetzungskomponente von PT plus füllen oder mit einer manuell nacheditierten Version dieser Übersetzungen. Die Größe der jeweils definierten Übersetzungseinheiten ist sowohl bei T1 als auch bei PT plus nicht an Phrasen- oder Satzgrenzen gebunden, sondern kann bei der Bearbeitung des Satzarchivs individuell bestimmt werden.[5] Die möglichen Parameter zur Optimierung der Übersetzung bzw. zur Einschränkung des Suchraums bei der Satzarchivfunktion (Auswahl der Satzarchivmodule, Benutzer usw. [vgl. Abb. 2]) sind mit jenen bei T1 vergleichbar.

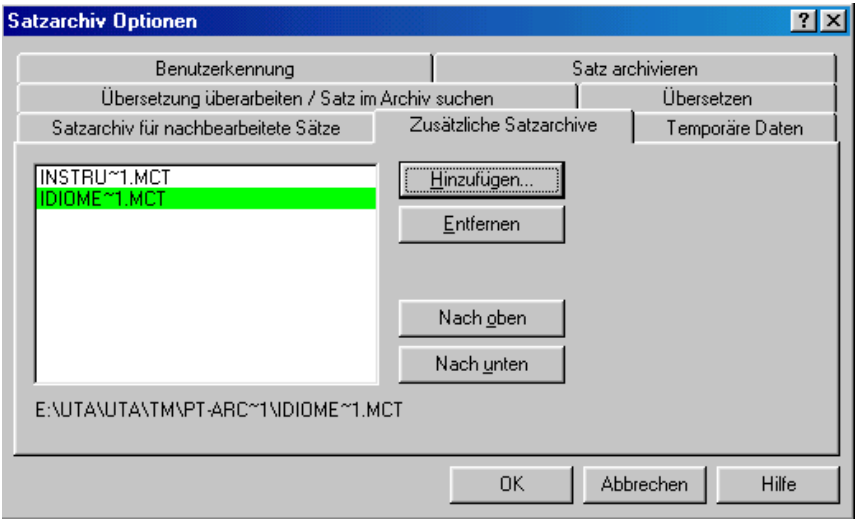


Abb. 2: Benutzerschnittstelle des Satzarchivmoduls von PT plus 2000.

3 Nachschlagen im Satzarchiv

Die wesentliche Leistung eines Satzarchivmoduls beruht auf effizienten Suchalgorithmen, die einen Satz eines Quelltextes mit den entsprechenden quellsprachlichen Einträgen der Datenbank vergleichen. Satzspeichersysteme zeichnen sich dadurch aus, dass die Systeme auf eine Suchanfrage nicht nur vollständige Übereinstimmungen, sogenannte 100%-Matches, liefern, sondern aufgrund der jeweils unterschiedlichen Realisierung von Fuzzy-Match-Algorithmen auch Sätze finden, die Abweichungen gegenüber dem zu übersetzenden Text aufweisen können. Dabei ist der maximale Grad der Abweichung zwischen zu übersetzendem Satz und im Satzspeicher enthaltenen Satz vom Benutzer jeweils in Form eines Prozentwertes anzugeben, so dass bei der Suche des Satzes 1.1 aus Beispiel (1) im Satzarchiv auch ein Satz wie 1.2 gefunden werden kann.

(1) 1.1: *This chapter gives an introduction to X11 operations.*
1.2: *This chapter is an introduction to X11 operations.*

Da sich satzspeicherbasierte Systeme in der Bewertung der Übereinstimmung zwischen zu übersetzendem Satz und im Satzarchiv enthaltenen Referenzsätzen zum Teil deutlich voneinander unterscheiden, sollen die Suchergebnisse der Systeme T1 und PT plus nachfolgend anhand eines kleinen Testkorpus dokumentiert und einander gegenübergestellt werden.

3.1 Das Testmaterial

Um die grundlegende Funktionsweise der beiden Übersetzungssysteme beim Zugriff auf ihr Satzspeichermodul bzw. bei der Suche (Retrieval) in dem jeweiligen Satzarchiv zu überprüfen, wird zunächst ein Satzarchiv angelegt, das nur zwei Referenzsätze enthält, eine Kapitelüberschrift und der erste Satz des dazugehörigen Kapitels (vgl. Bsp. (2)). Durch die Beschränkung des Satzarchivs auf nur zwei Sätze ist es möglich, die Suchergebnisse der Systeme bei Abweichungen von zu übersetzenden Sätzen gegenüber dem Referenzmaterial im Satzarchiv bereits auf der Basis eines kleinen Textkorpus zu systematisieren. Das als Testmaterial verwendete Textkorpus umfasst 34 Testsätze, von denen sieben den ersten Referenzsatz (vgl. Tab. 1) und die übrigen 27 Sätze den zweiten Referenzsatz des Satzarchivs (vgl. Tab. 2) modifizieren.

(2) Chapter 2 Getting Started
This chapter is an introduction to X11 operations.

Das Verhalten der Systeme bei der Suche in einem besonders großen Satzarchiv wurde im vorliegenden Test nicht untersucht, da die von den Systemen vorgenommene Bewertung der Ähnlichkeit zwischen Testsatz und Referenzsatz im Vordergrund stand.

Referenzsatz			
Chapter 2 Getting Started			
Nr.	**Testsatz**	**Match-Wert*) in %**	
		T1	**PT**
1	Chapter 2 Getting started	100	100
2	Chapter 2 Getting started	75	100
3	Chapter 2 Getting Started**.**	75	100
4	Chapter **1** Getting Started	75	95
5	Chapter 2 **Ø**	50	0
6	**Ø** Getting Started	50	0
7	**Ø** Getting **s**tarted	25	0

*Tab. 1: Retrievalergebnisse von T1 und PT plus. Die mit *) markierte Spalte enthält die vom jeweiligen System ermittelte Übereinstimmung zwischen Test- und Referenzsatz.*

Referenzsatz			
This chapter is an introduction to X11 operations.			
Nr.	**Testsatz**	**Match-Wert in %**	
8	This chapter is an introduction to X11 operations.	100	100
9	This chapter **gives** an introduction to X11 operations.	87	91
10	This **manual** is an introduction to X11 operations.	87	86
11	This chapter is an introduction to X11 operations **Ø**	87	100
12	This chapter is an introduction to X11 **functions**.	87	82
13	This chapter is an introduction to X11 **Ø**.	75	87
14	This chapter is an introduction to X11 **Ø**	87	87
15	This chapter is an **XXX** to X11 operations.	87	82
16	This chapter is an **overview over** X11 operations.	75	71
17	This chapter is **a list of** X11 operations.	62	75
18	This chapter is **a brief** introduction to X11 operations.	77	89
19	This chapter is **a brief** introduction to X11 **Ø**.	62	80
20	**The following** chapter is an introduction to X11 operations.	77	78
21	**Chapter 2** is an introduction to X11 operations.	75	84
22	**Chapter 2** is **a brief** introduction to X11 operations.	55	76
23	**Chapter 2 of this manual** is an introduction to X11 operations.	54	86
24	This chapter is an introduction to X11 **operations and X11 functions**.	72	77
25	This chapter is an introduction to **Ø** operations **of X11**.	66	93
26	This **Ø** is an introduction to X11 operations.	87	91
27	**See the** introduction to X11 operations.	50	75
28	X11 operations are introduced in this chapter.	12	53
29	X11 is **introduced in** this chapter.	25	0
30	chapter This an is to introduction operations X11.	75	91

31	is an introduction This chapter to operations. X11	87	74
32	X11 introduction This chapter is an to operations.	**100**	87
33	X11 introduction This **an** chapter is **an** to operations.	**100**	97
34	X11 This **an** chapter is **an** to operations.	**100**	86

*Tab. 2: Retrievalergebnisse von T1 und PT plus. **Ø** markiert die Auslassungen in den Testsätzen relativ zum Referenzsatz.*

4 Ergebnisse des Retrievals im Übersetzungsspeicher-Modul von T1

Wie aus den in Tabelle 1 und in Tabelle 2 dargestellten Retrievalergebnissen hervorgeht, steht die Bewertung der Ähnlichkeit eines zu übersetzenden Satzes relativ zu einem im Archiv enthaltenen Referenzsatz bei T1 im Zusammenhang mit der Anzahl der zwischen Test- und Referenzsatz übereinstimmenden Wörter.

Im Unterschied zur automatischen Übersetzung, wo in der ersten Analysephase, der Phase der Segmentierung eines Satzes in seine Bestandteile, die Interpunktionszeichen von den ihnen vorangehenden Zeichenketten separiert werden, leistet der Retrieval-Algorithmus von T1 diese Segmentierung nicht mehr, so dass Sätze wie Satz (3) aus Tab. 1 oder Satz (11) aus Tab. 2 relativ zum Referenzsatz nach der prozentualen Bewertung des Systems bereits deutlich voneinander abweichen. Obschon die hier vorliegenden Abweichungen in Satz (3) lediglich durch den zusätzlichen und in Satz (11) alleine durch den fehlenden Punkt am Satzende verursacht werden, gibt T1 bei Satz (3) als Match-Wert 75% an, während das System die Ähnlichkeit von Satz (11) zum Referenzsatz mit 87% bemisst. Der Grund hierfür ist die Tatsache, dass der Match-Algorithmus des T1-Satzarchivs jeden Bestandteil eines Satzes zu gleichen Teilen in die Bewertung für die Übereinstimmung zweier Sätze einfließen lässt.

Da der Abgleich zwischen zwei Sätzen allein auf der Basis eines Zeichenkettenvergleichs erfolgt, erhalten Sätze wie Satz (28) und Satz (29) (Tab. 2) einen sehr niedrigen Match-Wert, obschon sie inhaltlich dem Referenzsatz sehr ähnlich sind. Der in Satz (28) und (29) ausgedrückte Sachverhalt wird im Unterschied zum Referenzsatz im Passiv ausgedrückt, was an der Satzoberfläche zu veränderten Wortformen und damit zu einem niedrigen Match-Wert führt.

Beim Abgleich der zu übersetzenden Sätze berücksichtigt T1 nicht die Reihenfolge der Wörter in einem Satz, sondern überprüft lediglich das Auftreten der einzelnen Wörter im Referenzmaterial. Dieses Verfahren bzw. die Tatsache, dass Sätze als ungeordnete Mengen von Wörtern begriffen werden, führt dazu, dass beliebige Permutationen von Wörtern, bis hin zu syntaktisch völlig sinnlosen Kombinationen, als mit dem Referenzmaterial vollständig identisch, d.h. mit einer Ähnlichkeit von 100% bewertet werden (vgl. Tab. 2, Sätze (32) bis (34)). In Anbetracht dieses Ergebnisses muss bezweifelt werden, ob die Systemfunktion des T1, mit der Sätze ab einem bestimmten Match-Wert, also beispielsweise bei einem Ähnlichkeitsgrad von 100%, automatisch in die Übersetzung übernommen werden können, sinnvoll eingesetzt werden kann.

Obschon die Match-Ergebnisse von T1 mit einem Wert von 100% an dieser Stelle besonders auffallen, stellt sich auch in Bezug auf andere Systeme die generelle Frage nach der Bewertung dessen, was von Systemseite als vollständig identisch, also als 100%-Match, identifiziert wird. In diesem Zusammenhang ist daher auch die Frage berechtigt, ob die Ergebnisse des Personal Translator beim Abgleich der Sätze (32) bis (34), der hier jeweils Werte von 87%, 97% und 86% liefert, oder auch jene des Translation-Memory-Systems von Trados, der Translator's Workbench, die hier Werte von 67%, 71% und 75% errechnet, aufschlussreicher sind, handelt es sich bei den fraglichen Testsätzen doch um Wortaneinanderreihungen, in denen zwar jedes Wort des betreffenden Referenzsatzes enthalten ist und deren Länge die des Referenzsatzes nicht unterschreitet, deren Reihenfolge jedoch nichts mit der Abfolge der Wörter im Referenzsatz des Satzarchivs zu tun hat.

Wie die Sätze (35) und (36) aus Tab. 3 belegen, führt beispielsweise die veränderte syntaktische Position der Negationspartikel *not* in (36) zu einer Umkehrung der in Satz (35) ausgedrückten Bedeutung, was seinen Niederschlag in der betreffenden Übersetzung finden muss. Einen vergleichbaren Effekt hat die Verdopplung der Negationspartikel *not* in Satz (38) im Vergleich zu (37). Eine Übernahme der Übersetzung von (37) für Satz (38) hätte somit wie schon im vorhergehenden Fall einen schwerwiegenden Übersetzungsfehler zur Folge. – Selbst wenn man davon ausgeht, dass im realen Übersetzungsfall nur selten Beispielsätze auftreten, die zwar alle Elemente des Referenzsatzes, jedoch in anderer Abfolge enthalten, muss bereits an dieser Stelle unterstrichen werden, dass ein 100%-Match unter linguistischen Gesichtspunkten eine zu hinterfragende Größe ist und – das zeigen die Ergebnisse – insbesondere bei der Suche im Satzarchiv von T1 als sehr

Referenzsatz *I do not drink wine, but I drink beer.*		
Nr.	**Testsatz**	**Match-Wert**
35	I **do not** drink wine, but I drink beer.	**100**
36	I drink wine, but I **do not** drink beer.	**100**
Referenzsatz: *It's true that I do not drink wine.*		
Nr.	**Testsatz**	**Match-Wert**
37	It's true that I do **not** drink wine.	**100**
38	It's **not** true that I do **not** drink wine.	**100**

Tab. 3: 100%-Matches von T1.

problematisch einzustufen ist. Sätze mit dieser Bewertung dürfen unter keinen Umständen ungeprüft in eine Übersetzung übernommen werden.

4.1 Der Suchalgorithmus im T1-Satzarchiv

Der für das Satzarchiv von T1 realisierte Suchalgorithmus lässt sich aus den Beispielen (Tab. 1 und 2) mit *Reverse-Engineering*-Techniken rekonstruieren. Er ist verhältnismäßig einfach. Aus seiner Struktur lässt sich seine Unzulänglichkeit unmittelbar ableiten. Neben der Anzahl der zwischen Testsatz und Referenzsatz übereinstimmenden Wörter hängt der jeweils vom System gelieferte Match-Wert von der Länge der miteinander verglichenen Sätze ab. Wie unter anderem aus Satz (24) ersichtlich ist, spielt aber nicht allein die Länge des Referenzsatzes, der als Vergleichsbasis herangezogen wird, eine Rolle. Vielmehr wird für die Ermittlung des Match-Wertes die Länge des jeweils längeren der beiden miteinander verglichenen Sätze herangezogen. Dieser Wert wird als Quotient angewendet auf die Anzahl der Wörter, die Referenzsatz und zu übersetzender Satz gemeinsam haben (siehe Abb. 1).[6]

4.2 Das Satzarchiv als „Wörterbuch-Erweiterung“ für Mehrwortausdrücke

Wenn nun – wie in Abschnitt 4.1 dargestellt – der Match-Algorithmus des Satzarchivs in T1 als zu trivial erscheint, als dass das Satzarchiv-Modul im professionellen Einsatz Funktionen eines integrierten Translation Memory übernehmen könnte, so stellt sich die Frage, ob es gleichsam als Erweiterung des Wörterbuchs fungieren kann, und zwar für Mehrwortausdrücke, deren Aufnahme im Wörterbuch nicht möglich ist. Zwar ist in T1 vorgesehen, dass Substantive, die aus mehreren Wörtern zusammengesetzt sind (Mehrwortausdrücke), in das Lexikon aufgenommen werden können, doch ist die Aufnahme von Mehrwortausdrücken auf diese Wortart beschränkt. Die Eingabe eines verbalen Idioms wie z.B. eng. *rain cats and dogs* (dt. *in Strömen gießen*) bleibt dem Benutzer mit dem Hinweis „Keine gültige Eingabe für diese Kategorie. Bitte ändern Sie die Eingabe oder die Kategorie.“ verwehrt.

Der Aufnahme von Mehrwortausdrücken in das Satzarchiv des T1 stehen dagegen keinerlei Einschränkungen entgegen. Sie ist in der gleichen Form möglich, wie dies auch mit beliebigen Sätzen oder anderen Satzfragmenten geschehen kann. Im Lieferumfang von T1 ist bereits ein Satzarchivmodul „idioms“ enthalten, in dem als idiomatische Wendung unter anderem *it's raining cats and dogs* aufgelistet ist. Verwendet man nun bei der Übersetzung eines Textes, in dem beispielsweise die Sätze (3), (4) oder (5) auftreten, das Satzarchivmodul „idioms“, so wird für Satz (3) eine Ähnlichkeit von 60%, für die Sätze (4) und (5) eine Ähnlichkeit von jeweils 40% zu dem Referenzidiom *it's raining cats and dogs* berechnet. Der Prozentsatz von 60% Ähnlichkeit, der als Match-Wert für Satz (3) ermittelt wird, resultiert aus der Tatsache, dass in Satz (3) im Unterschied zur Eintragung im Satzarchiv der Satzanfang mit einem Großbuchstaben beginnt und der Satz darüber hinaus mit einem Satzzeichen (Punkt) beschlossen wird. Aufgrund der Konzeption des Match-Algorithmus (siehe den vorhergehenden Abschnitt) werden in den Sätzen (4) und (5) lediglich die Wörter *cats* und *and* als mit dem Referenzmaterial übereinstimmend identifiziert, so dass bei einer Satzlänge von 5 Wörtern ein Prozentsatz von lediglich 40% ermittelt wird.

(3) It's raining cats and dogs.

(4) It rains cats and dogs.

(5) It rained cats and dogs.

translate: aString

„Ermittelt den Match-Wert eines zu übersetzenden Satzes (testWords) relativ zu einem Referenzsatz (refWords)."

```
| testWords result refWords matchValue string index wordsFound |
index := aString indexOf: $. ifAbsent: [aString size].
string := aString copyFrom: 1 to: index.
testWords := T1Scanner new scanTokens: string.
result := Dictionary new.
Reference
  keysDo:
        [:ref |
        refWords := T1Scanner new scanTokens: ref.
        wordsFound := 0.
        testWords do: [:word | (refWords includes: word)
                        ifTrue: [wordsFound :=
                          wordsFound + 1]].
        matchValue := (wordsFound / (testWords size
                    max: refWords size)) asFloat * 100.0.
        matchValue >= self threshold ifTrue: [result at:
                    ref put: matchValue]].
^result
```

Abb. 1: Smalltalk-Implementierung der Funktionsweise des in T1 realisierten Match-Algorithmus: Gezählt werden die Wörter des zu übersetzenden Satzes (testWords), die in einem Referenzsatz (refWords) enthalten sind, und zwar unabhängig von ihrer Position und der Häufigkeit ihres Auftretens. Dieser Algorithmus, der syntaktische Strukturen völlig außer Acht lässt, liefert die in Tab. 1, 2 und 3 dargestellten Ergebnisse.

Bei einer idiomatischen Wendung, die ein flektierendes Satzglied enthält, jedoch nur in einer Form im Satzarchiv abgelegt ist, liegt die Ähnlichkeit immer unterhalb von 100%. Ist die betreffende Wendung schließlich noch in einen Satz eingebettet, der im Normalfall über die idiomatische Wendung hinaus zusätzliche Wörter enthält, so sinkt der Match-Wert in Abhängigkeit von der Zahl der im Satz zusätzlich auftretenden Wörter. Entsprechend wird die Wendung *the most important thing*, für die im Archiv die Übersetzung „das A und O" angegeben ist, in einem

Satz wie (6) nur mit einem Match-Wert von 60% identifiziert, obschon sie hier in derselben Form wie im Archiv auftritt.

*(6) He recognized **the most important thing** of the story.*

Berücksichtigt man, dass erfahrene Benutzer von TMs in der Regel Match-Werte empfehlen, die oberhalb von 70% liegen, um die benötigte Zeit für die Sichtung der vom System gefundenen Referenzsätze auf einen ökonomisch vertretbaren Rahmen zu beschränken, so sprechen die von T1 berechneten Werte für sich selbst.

Die genannten Beispiele belegen somit, dass das Satzarchiv mit dem in T1 implementierten Match-Algorithmus als Erweiterung des Wörterbuchs für Mehrwortausdrücke nicht geeignet ist.

4.1 Bewertung des Suchalgorithmus des T1-Satzarchivs

Der für das Satzarchiv von T1 realisierte Suchalgorithmus ist einfach und leicht nachvollziehbar. Er weist jedoch schwerwiegende Mängel auf, weil er beim Vergleich von Test- und Referenzsatz in zahlreichen Fällen zu niedrige, in anderen Fällen zu hohe Match-Werte liefert. Für den praktischen Einsatz zu niedrige Match-Werte werden an Stellen ermittelt, an denen im Testsatz ein Interpunktionszeichen auftritt, wo im Referenzsatz ein solches nicht vorhanden ist (vgl. Satz (3), Tab.1 oder Satz (10), Tab. 2). Da das Interpunktionszeichen als Bestandteil des ihm vorangehenden Wortes interpretiert wird, wirkt sich die unzureichende Segmentierung in besonderem Maße beim Abgleich kurzer Sätze aus (siehe Satz (3), Tab.1). Da die Abfolge der Elemente im Satz beim Abgleich mit dem Referenzmaterial nicht berücksichtigt wird, zudem auch keine Kontrolle über im Testsatz mehrfach auftretende Wörter erfolgt, werden für entsprechende Sätze (vgl. Satz (35) und (37) aus Tab. 3) zu hohe Match-Werte (100%) ermittelt, weshalb Übersetzungen von 100%-Matches unter keinen Umständen ungeprüft übernommen werden dürfen.

Bewertet man das Satzarchiv von T1 unter dem Gesichtspunkt von *Recall* und *Precision*, Parameter, die zur Bewertung von Retrievalsystemen herangezogen werden (siehe auch [Reinke99]), lässt sich die Leistungsfähigkeit des Suchalgorithmus in T1 wie folgt zusammenfassen:

Der Recall, d.h. die Zahl der vom Suchalgorithmus als Basis für die Übersetzung eines Testsatzes ermittelten Kandidaten im Verhältnis zu den tatsächlich im Referenzmaterial enthaltenen relevanten Kandidaten, ist zu gering, weil (a) Satzzeichen als Wortbestandteile interpretiert werden, (b) Teilsätze nicht identifiziert werden, da die Vergleichseinheit ein ganzer Satz ist, und schließlich (c) wegen des Ausklammerns einer linguistischen Analyse der Satzarchivdaten, mit der flektierende Bestandteile erkannt und mit anderen Flexionsformen desselben Paradigmas abgeglichen werden könnten. Gleichzeitig ist aber auch die Precision zu gering, und zwar in solchen Fällen, in denen sich die Vergleichsdaten relativ zum Referenzmaterial wie in Satz (33) aus Tab. 2 in der Abfolge oder der Häufigkeit der im Satz enthaltenen Wörter unterscheiden.

5 Der Suchalgorithmus von PT plus

Im Gegensatz zu T1 ist die Retrievalstrategie von PT plus weniger transparent und nachvollziehbar und um einiges komplexer, was sich auch in den Retrievalergebnissen (siehe Tabellen 1 und 2 im vorherigen Abschnitt) widerspiegelt. Beispielsweise werden Modifikationen der Interpunktion oder auch Groß- und Kleinschreibung „robuster“ gehandhabt als bei T1 und bei der prozentualen Bewertung ignoriert (vgl. Sätze (1) – (3) in Tab. 1). Hierbei ist allerdings zu erwähnen, dass sich dieses Verhalten bei unterschiedlichen Satzendezeichen ebenfalls verändert. Während das Hinzufügen oder Weglassen des ‘Punkts’ keinen Einfluss auf das tatsächliche Retrieval und die Angabe der Trefferquote hat, verändert sich sowohl das Retrievalergebnis als auch die vom System kalkulierte Trefferquote beispielsweise beim Hinzufügen eines Fragezeichens (7):

(7) Referenzsatz: *Chapter 2 Getting Started*
Testsatz: *Chapter 2 Getting Started?*
Match-Wert des PT: 98%

Auch bei der Veränderung von als Ziffern dargestellten Zahlen im Testmaterial vergibt PT plus großzügigere Trefferraten als T1 (vgl. (8)), so dass zu vermuten ist, dass auch bei der Modifikation von reinen Zifferndarstellungen bei PT plus andere Gewichtungen angesetzt werden, was allerdings durch spezifische Tests mit entsprechendem Datenmaterial noch zu belegen ist (siehe hierzu auch Kapitel 6):

(8) Referenzsatz: *Chapter 2 Getting Started*
Testsatz: *Chapter 1 Getting Started*
Match-Wert des T1: 75%
Match-Wert des PT: 95%

Ähnlich hohe Bewertungen wie bei T1 erhalten auch bei PT plus Modifikationen, die syntaktisch unsinnig sind (siehe Satz (9)), so dass die Aussagefähigkeit der 100%-Bewertungen durch das System – wie dies auch schon bei der Beschreibung der Ergebnisse des T1 im vorherigen Kapitel festgestellt worden ist – in Frage gestellt werden muss:

(9) Referenzsatz: *This chapter is an introduction to X11 operations.*
Testsatz: *Chapter This an is to introduction operations X11.*
Match-Wert des PT: 91%

5.1 Mehrwortausdrücke im PT plus Satzarchiv

Im Gegensatz zu T1 erlaubt es die Wörterbuchschnittstelle von PT plus, Mehrwortausdrücke als benutzerdefinierte Lexikoneinträge für alle Wortarten zu definieren. Ein Eintrag für das obige Beispiel *rain cats and dogs* ist beispielsweise zulässig. Schwierigkeiten mit Mehrwortausdrücken treten bei PT plus in solchen Fällen auf, in denen ein Mehrwortausdruck neben dem verbalen ein nichtverbales Element enthält, das in Abhängigkeit vom Satzsubjekt variiert, wie z.B. *put one's oar in* (dt. *seinen Senf dazu geben*), das bei einem Subjekt in der 3. Person Singular Maskulinum *als he puts **his** oar in* und bei der 2. Person Singular als *you put **your** oar in* erscheint.

Für die Untersuchung der Satzarchivfunktionalität von PT plus mit idiomatischen Mehrwortausdrücken wurde daher ein mit den T1-Testdaten identisches Experiment durchgeführt. Hierzu mussten zunächst diese Testdaten in einer Archivdatei aufbereitet werden, da das idiomatische Wörterbuch des PT plus nur offline konsultiert werden kann und nicht online wie bei T1 als Referenzquelle verwendbar ist.

Bei der Betrachtung der Ergebnisse fällt auf, dass in einigen Fällen trotz teilweiser Übereinstimmung von modifizierter Eingabe und Satzarchivdaten der Match-Wert mit 0% angegeben und somit kein Übersetzungsvorschlag des be-

treffenden Fragments aus dem Archiv angezeigt wird (vgl. die Beispiele (10) bis (12)):

(10) Referenzsatz: *read from cover to cover*
Testsatz: *She reads the book from cover to cover.*
Match-Wert des PT: 0%

(11) Referenzsatz: *he told us everything*
Testsatz: *He will tell us everything*
Match-Wert des PT 0%

(12) Referenzsatz: *bask in the sun*
Testsatz: *Yesterday I basked in the sun*
Match-Wert des PT 0%

Die Tatsache, dass die Modifikationen teilweise *innerhalb* der idiomatischen Wendungen aufgrund von Flexionsänderungen und Auslassungen oder Komplementierung auftreten, ist alleine keine ausreichende Erklärung für diese Ergebnisse. Gegenproben mit Testsatz (10), der abermals modifiziert wurde ((13) bis (17)), zeigen, dass noch andere Kriterien, wie z.B. die Zeichenlänge der Modifikation, eine Rolle spielen:

(13) Referenzsatz: *read from cover to cover*
Testsatz: *read XXX from cover to cover*
Match-Wert des PT: 87%

(14) Referenzsatz: *read from cover to cover*
Testsatz: *read XXXX from cover to cover*
Match-Wert des PT: 83%

(15) Referenzsatz: *read from cover to cover*
Testsatz: *I wanted to read XXXX from cover to cover*
Match-Wert des PT: 67%

(16) Referenzsatz: *read from cover to cover*
Testsatz: *She wanted to read XXXX from cover to cover*
Match-Wert des PT: 0%

(17) Referenzsatz: *read from cover to cover*
Testsatz: *She wants to read XXXX from cover to cover*
Match-Wert des PT: 65%

Anhand der Match-Werte, die beim Abgleich von Testsatz (10) bis (17) mit im Satzarchiv enthaltenen idiomatischen Wendungen ermittelt wurden, zeigt sich, dass auch das Satzarchiv von PT plus nur in eingeschränktem Maße als Erweiterung des Wörterbuchs brauchbar ist. Hier kann es lohnenswert sein, entsprechende Wendungen, die mit hoher Frequenz auftreten, in ein spezifisches Benutzerwörterbuch einzutragen. Die Frage nach dem Kosten-Nutzen-Effekt einer solchen Aktivität im Vergleich zur Verwendung des Satzarchivs (Vor- und Nachbereitungsaufwand) erfordert weitergehende Untersuchungen und Testprozeduren, die nicht Gegenstand der hier beschriebenen Arbeiten waren.

5.1 Bewertung des Suchalgorithmus des PT plus Satzarchivs

Wie die im vorhergehenden Kapitel illustrierten Beispiele zeigen, ist der Suchalgorithmus des PT plus nicht ohne weiteres nachvollziehbar. Vor allem Testbeispiel (10) und seine Modifikationen (13) bis (17) machen deutlich, dass offensichtlich unterschiedliche Faktoren für die Berechnung der Match-Werte eine Rolle spielen, die sich wiederum je nach Gewichtung gegenseitig zu beeinflussen scheinen. Während die prozentualen Angaben zur Retrievalleistung in (13) und (17) eine scheinbar logische Abhängigkeit von der Länge der Modifikation sowohl innerhalb als auch außerhalb der archivierten Wendung reflektieren, weichen (10) und (16) von dieser Kalkulation ab. Schließlich zeigt auch die unterschiedliche Bewertung von (16), für das der Suchalgorithmus einen Match-Wert von 0% ermittelt, im Vergleich zu (15), das einen Match-Wert von 67% erhält, obschon sich die beiden Sätze lediglich durch das Personalpronomen in Subjektposition – *she* versus *I* – unterscheiden, dass die Berechnung der Ähnlichkeit zwischen Test- und Referenzmaterial häufig nicht plausibel ist.[7]

Somit bleibt festzustellen, dass die Match-Werte des Satzarchivs von PT plus unabhängig von der Struktur der archivierten Daten häufig sehr willkürlich wirken. Die Brauchbarkeit der Satzarchivfunktion hinsichtlich einer Wörterbucherweiterung für idiomatische Mehrwortausdrücke ist – ähnlich wie bei T1 – beschränkt.

6 Leistungsfähigkeit der Übersetzungsspeicher-Module

Vergleicht man abschließend die Match-Werte der beiden hier untersuchten Systeme beim Abgleich der Testdaten, so ergeben sich bei der Bewertung der Ähnlichkeit Unterschiede bis zu 50% zwischen T1 und PT plus (vgl. die Beispiele (18) bis (20)).

(18)	Referenzsatz:	*Chapter 2 Getting Started*
	Testsatz:	*Chapter 2*
	Match-Wert (T1):	50%
	Match-Wert (PT):	0%
(19)	Referenzsatz:	*Chapter 2 Getting Started*
	Testsatz:	*Chapter 2 Getting Started.*
	Match-Wert des T1:	75%
	Match-Wert des PT:	100%
(20)	Referenzsatz:	*This chapter is an introduction to X11 operations.*
	Testsatz:	*This chapter is **a brief** introduction to X11 operations.*
	Match-Wert des T1:	77%
	Match-Wert des PT:	89%

Die Unterschiede in der Bewertung der Ähnlichkeit zwischen Testsatz und im Übersetzungsarchiv enthaltenem Referenzsatz machen deutlich, dass weitergehende systematische Erhebungen erforderlich sind, um fundierte Einschätzungen über die Aussagekraft von Match-Werten einzelner Systeme machen zu können. Hierzu sind neben einem größer angelegten Testkorpus, das systematisch

Modifikationen zwischen Testsatz und Referenzsatz an unterschiedlichen Positionen erfasst, auch authentische Testmaterialien erforderlich, die idealerweise in verschiedenen Textversionen vorliegen sollten. Solange solche Daten nicht vorliegen, muss eine Einschätzung der Leistungsfähigkeit der einzelnen Übersetzungsspeicher und somit auch die hier vorliegende Bewertung vorläufigen Charakter haben.

Literatur

[Nübel/Seewald-Heeg98] Nübel, R. und Seewald-Heeg, U. (eds.) (1998): *Evaluation of the Linguistic Performance of Machine Translation Systems*. Proceedings of the KONVENS-98 Workshop in Bonn. St. Augustin: Gardez! Verlag.

[Andrés-Lange98] Andrés-Lange, C. (1998): 'Tying the Knot. How Baan wed machine translation to translation memory – and survived the honeymoon', in: *Language International* 10/5, 34-36.

[Reinke94] Reinke, U. (1994): 'Zur Leistungsfähigkeit integrierter Übersetzungssysteme', in: *Lebende Sprachen* 3, 97-104.

[Reinke99] Reinke, U. (1999): 'Evaluierung der linguistischen Leistungsfähigkeit von Translation Memory-Systemen - Ein Erfahrungsbericht', in: *LDV Forum*, vorliegendes Heft.

[Twents97] Twents, A. (1997): *Terminologieerkennung in integrierten Übersetzungssystemen am Beispiel des Französischen*. Saarbrücker Studien zu Sprachdatenverarbeitung und Übersetzen. FR 8.6, Universität des Saarlandes, Saarbrücken.

ANMERKUNGEN

[1] Unter den kommerziell verfügbaren vollautomatischen Systemen zählen dazu Logos, Personal Translator plus von Linguatec, Power Translator Pro von Globalink, Systran PROfessional von Systran, T1 Professional von Langenscheidt oder Transcend von HEI-Soft, die im Rahmen einer von Nübel und Seewald-Heeg [Nübel/Seewald-Heeg98] dokumentierten Evaluation auf ihre linguistische Leistungsfähigkeit hin untersucht wurden.

[2] Die Zahl der auf dem Markt angebotenen TM-Systeme ist in den vergangenen Jahren bedeutend gestiegen. Produkte wie SDLX von SDL, Déjà Vu von Atril, Eurolang Optimizer, Joust von Alpnet, Transit von STAR, Translation Manager von IBM, Translator‘s Workbench von Trados oder ZERES*TRANS* von Zeres gehören in diese Systemklasse.

[3] Eine hiervon abweichende weitergefasste Definition des Terminus „integrierte Systeme“ liefert [Reinke94], auf die sich auch [Twents97] bezieht.

[4] Eine entsprechende Systemkonfiguration ist z.B. bei der Firma Baan im Einsatz. Vgl. [Andres-Lange98].

[5] Allerdings wird für den Abgleich von Eingabe- und Referenzmaterial die Suche jeweils bis zur Satzgrenze durchgeführt, unabhängig von der Größe der archivierten Übersetzung.

[6] Abb. 1 zeigt den Kern des Retrieval-Programmes von T1, wie er anhand der Match-Ergebnisse beim Abgleich des Testkorpus mit dem Referenzmaterial (siehe Tab. 1) mittels der objektorientierten Programmiersprache Smalltalk (VisualWorks 5i) nachgestellt werden konnte.

[7] Inwieweit diese Ergebnisse eine Konsequenz der in Retrieval-Software vielfach berücksichtigten Übereinstimmung von Trigrammen (hier: Buchstabenfolgen der Länge ‘3’) oder anderen n-Grammen sind, die in unterschiedlicher Gewichtung in die Berechnung von Match-Werten eingehen, kann im Rahmen der vorliegenden Arbeit nicht beantwortet werden.

Translator‘s Workbench: Funktionalität – Entwicklungen – Kundenanforderungen

Hartmut Bohn
TRADOS GmbH, Stuttgart

1 Über TRADOS

TRADOS ist spezialisiert auf Software und Dienstleistungen rund um den professionellen Übersetzerarbeitsplatz. Als Anbieter von Lösungen für den optimierten Übersetzungs-Workflow kann die TRADOS GmbH, Stuttgart auf inzwischen 15 Jahre Erfahrung in der Übersetzungsbranche zurückblicken. Mit mehr als 25.000 Installationen weltweit hat sich TRADOS als Marktführer im Bereich Software für maschinengestützte Übersetzung etabliert.

TRADOS-Kunden kommen aus allen Bereichen der Wirtschaft (Maschinenbau, Automobilindustrie, Chemie, Software, Banken, Versicherungen, ...) und des öffentlichen Sektors (EU, Ministerien, nationale und internationale Organisationen). Zudem setzen sehr viele große und kleinere Übersetzungs- und Software-Lokalisierungsdienstleister sowie zahlreiche freiberuflich tätige Übersetzer TRADOS-Produkte ein.

Viele TRADOS-Kunden sind weltweit tätig und legen daher Wert auf einen Ansprechpartner vor Ort. Daher bietet TRADOS von inzwischen zehn Niederlassungen aus Vertrieb, Beratung und technische Unterstützung in Europa, USA und Asien an.

2 TRADOS-Produkte

TRADOS-Produkte optimieren den Übersetzungsprozess auf unterschiedlichen Ebenen:

- Gut recherchierte und gepflegte Terminologie bildet die Grundlage einer jeden qualitativ hochwertigen Übersetzung. Mit *MultiTerm’95 Plus!* bietet TRADOS eine flexible Lösung für die Erfassung, Nutzung und Pflege von Terminologie. Mit *MultiTerm Dictionary* und *Multi-*

Term Web Interface besteht zudem die Möglichkeit, eine mit *MultiTerm* erstellte Terminologie auf CD-ROM oder im Internet zu publizieren.

- Mit *TRADOS Translator's Workbench* bietet TRADOS ein datenbankgestütztes Translation-Memory-System, das schon einmal erstellte Übersetzungen reproduziert und damit Doppelarbeiten überflüssig macht. Durch die Integration von *MultiTerm'95 Plus!* in *Translator's Workbench* ist zudem auch für Neuübersetzungen terminologische Konsistenz gewährleistet.
- Konvertierungslösungen und Filter für die Bearbeitung unterschiedlicher Dateiformate sowie das Alignment-Werkzeug *WinAlign*, das die Erstellung einer Übersetzungsdatenbank aus vorhandenen Übersetzungen erlaubt, ergänzen die Produktpalette.

3 Kundenanforderungen

Die gegenwärtige Funktionalität und die zukünftige Entwicklung der TRADOS-Produkte ergeben sich aus den Anforderungen der TRADOS-Kunden. Dabei lassen sich grundsätzlich zwei Kundengruppen mit zum Teil gegensätzlichen Interessen unterscheiden: Einerseits Übersetzer, die ihre Übersetzungen mit Unterstützung der TRADOS-Produkte anfertigen. Andererseits die Ersteller von zu übersetzenden Dokumenten, die somit als Auftraggeber auf dem Übersetzungsmarkt fungieren.

3.1 Anforderungen von Übersetzern

Übersetzer wollen sich in der Regel auf ihre eigentliche Arbeit – das Übersetzen – konzentrieren. Ihre Forderungen an ein Übersetzungswerkzeug lauten daher:

- Integration in die vertraute Arbeitsumgebung
- Übersetzung unabhängig vom Ausgangs- und Zielformat des Dokumentes
- Entlastung bei der Terminologiesuche
- Entlastung bei Routinearbeiten

3.2 Anforderungen der Übersetzungs-Auftraggeber

Die Auftraggeber von Übersetzungsaufträgen verfolgen beim Einsatz von Übersetzungswerkzeugen vor allem folgende Ziele:

- Übersetzung beliebiger Formate (Word, FrameMaker, Interleaf, HTML, SGML, ...)
- Minimierung des Nachbearbeitungsaufwandes, besonders bei aufwendigen DTP-Formaten
- Beschleunigung des Übersetzungsprozesses
- Erhöhung der Konsistenz und Qualität der Übersetzungen
- Kosteneinsparung durch Wiederverwenden schon vorhandener Übersetzungen

3.3 Interessenkonflikt

Einige der Anforderungen der beiden Seiten scheinen zunächst schwer miteinander vereinbar zu sein. So etwa die Anforderung an den Übersetzer, beliebige Formate zu übersetzen – also unterschiedlichste Programme und Editoren zu beherrschen, und dabei gleichzeitig auch noch den DTP-Nachbearbeitungsaufwand für den Auftraggeber zu minimieren.

Erschwert wird die Situation durch die Tatsache, daß auf dem Übersetzungsmarkt eine Entwicklung hin zu schwierig zu übersetzenden Formaten (*HTML* und – in besonderem Maße – *SGML*) zu beobachten ist. Hier versucht TRADOS durch intelligente Lösungen zu vermitteln.

4 TRADOS-Lösungen

Durch die direkte Integration von *Translator's Workbench* in MS Office (Word, PowerPoint) ist TRADOS in der Lage, etwa 80% des weltweiten Übersetzungsvolumens direkt im ursprünglichen Dateiformat abzudecken. Der Übersetzer arbeitet in der vertrauten Office-Umgebung und kann in vollem Umfang auf die Funktionalität von Word bzw. PowerPoint zurückgreifen.

Die datenbankgestützte Translation-Memory-Technologie läßt Doppelarbeiten hinfällig werden: Ein einmal übersetzter Satz wird in der Datenbank abgelegt und muß nicht ein zweites Mal übersetzt werden. Die aktive Terminologieerken-

nung durch *MultiTerm* erspart den Griff zur Terminologieliste und garantiert terminologisch konsistente Übersetzungen.

Für die gängigsten DTP-Formate, FrameMaker und Interleaf, lautet die von TRADOS verfolgte Strategie: Trennung von Layout und Übersetzung. Eine durchdachte Konversionslösung – der *S-Tagger* für FrameMaker und Interleaf – zieht den zu übersetzenden Text aus FrameMaker- und Interleaf-Dokumenten heraus, ordnet ihn in logischer Reihenfolge an und bereitet ihn zur Übersetzung mit *Translator's Workbench* auf. Nach der Übersetzung findet eine komplette Überprüfung der Format-Tags statt, bevor aus dem übersetzten Text das Zieldokument im FrameMaker- bzw. Interleaf-Format generiert wird. Sämtliche Formate bleiben erhalten, die DTP-Nachbearbeitung beschränkt sich in der Regel auf das Anpassen von Seitenumbrüchen.

Für viele andere Formate (z.B. PageMaker, QuarkXPress, RC, ...) bestehen Filter, die eine problemlose Bearbeitung mit *Translator's Workbench* erlauben.

Mit *TagEditor* stellt TRADOS einen speziellen Editor für getaggte Formate, insbesondere HTML und SGML, zur Verfügung und nimmt damit diesen schwer bearbeitbaren Formaten ihren Schrecken. HTML- und SGML-Dateien können direkt in *TagEditor* eingelesen werden. Tags sind während der Bearbeitung vor dem Überschreiben geschützt. Zudem findet auf Wunsch eine strikte Überprüfung der Tag-Konsistenz schon während der Übersetzung statt. Mit der Preview-Funktion läßt sich schon während der Übersetzung ein Blick auf das Resultat werfen: Ein HTML- bzw. SGML-Browser wird innerhalb des *TagEditor* aktiviert und stellt ausgangs- und zielsprachliche Version dar.

Die Bedienoberfläche ist programmübergreifend einheitlich gestaltet. Egal ob in Word, in *TagEditor* oder in PowerPoint: Symbolleisten, Menü und Tastaturkürzel sind identisch.

Translator's Workbench verfügt zudem über eine API-Schnittstelle, die es Anwendern erlaubt, die Datenbankfunktionen von externen Anwendungen aus anzusprechen. Der Integration in bestehende Systeme und Formate steht somit nichts im Wege.

Weitere wichtige Eigenschaften sind das Datenbankpflegemodul, die Projektmanagement-Werkzeuge (Analyse, Vorübersetzung, Import externer Übersetzungen). Durch Unterstützung des offenen TMX-Standards (*Translation Memory eXchange*, Level 1) ist der Austausch von Daten mit anderen Systemen gewährleistet.

5 Ausblick

Um auch in Zukunft den Kundenanforderungen gerecht zu werden, ist eine genaue Beobachtung der Entwicklungen auf dem Dokumentations- und Übersetzungsmarkt erforderlich. Die Zielrichtung zukünftiger Entwicklungen für TRADOS ist deshalb:

- Unterstützung offener Standards wie TMX oder MARTIF;
- Beobachtung der Entwicklungen im Bereich SGML/XML;
- Integration des Dokumentations- und Übersetzungsprozesses in Dokumentenmanagementsysteme;
- Integration von *Translator's Workbench* in weitere Anwendungen.

Professionelles Übersetzen mit STAR Transit

Judith Klein
STAR Deutschland GmbH

1 Translate it !

Transit (**Trans**late **it**) ist ein professionelles Translation-Memory (TM) System entwickelt auf der Basis der langjährigen praktischen Erfahrung von STAR, eines der größten und erfolgreichsten Übersetzungsunternehmens der Welt. Die Kunden von STAR kommen aus der Automobilindustrie, der Informationstechnologie und dem Bereich Maschinenbau. Transit, das seit 1994 kommerziell als Produkt vertrieben wird, kommt Ende 1999 als 32-Bit-Version mit Unicode- und XML-Unterstützung als Version 3.0 auf den Markt.[1]

Vorrangiges Ziel von Transit ist die Steigerung der Übersetzungsproduktivität durch automatisches Vorübersetzen auf der Grundlage geprüfter manuell angefertigter Übersetzungen. In Kombination mit integrierter Terminologieprüfung, ausgereifter übersetzungsspezifischer Funktionalität beim computergestützten manuellen Übersetzen und vollkompatiblen Schnittstellen zu den gängigen Textverarbeitungs- und Desktop-Publishing (DTP) Systemen bietet Transit daher eine gute Lösung für die im Übersetzungsgeschäft aktuellen Aufgaben:

- Die Übersetzung von Dokumenten, die mit unterschiedlichen Textverarbeitungs- und DTP-Systemen erstellt wurden, kann mit einer einzigen Software bearbeitet werden, wobei aufwendige Layoutaufgaben, die mit dem Übersetzungsvorgang primär nichts zu tun haben, nicht erforderlich sind.

- In der technischen Dokumentation muß die Übersetzung sehr umfangreicher Texte bei immer kürzern Lieferzeiten zu kostengünstigen Konditionen angefertigt werden. Da häufig aktualisierte Versionen bereits existierender Dokumente zu übersetzen sind und die technischen Texte selbst überdurchschnittlich viel sprachliche Wiederholungen aufweisen, können viele Passagen früherer Übersetzungen wiederverwendet werden. Während das Wiederverwenden bereits angefertigter Übersetzungen ohne TM-Technologie (über „Copy & Paste“) zwar möglich, aber fehleranfällig und sehr zeitaufwendig ist,

greift Transit automatisch auf bereits vorhandenes Übersetzungsmaterial zu.

- Die Qualität der Übersetzung, die auf sprachlicher Klarheit sowie inhaltlicher und terminologischer Korrektheit basiert, ist ein Kernfaktor in der technischen Dokumentation, da Fehler in Bedienungsanleitungen und Service-Handbüchern zu teuren Produkt- und Folgeschäden führen können. Die automatische Wiederverwendung manuell angefertigter und geprüfter Übersetzungen sowie die automatische Terminologieprüfung mittels der integrierten Terminologiekomponente TermStar dienen der Qualitätssicherung und unterstützen die terminologische Konsistenz der Übersetzung.

2 Arbeitsweise von Transit

Die komplette Version des Transit-Softwarepakets umfasst den übersetzungsspezifischen Transit-Editor mit integrierten Import- und Exportmechanismen zur Bearbeitung von Dokumenten unterschiedlicher Dateiformate[2], das vollintegrierte Terminologieverwaltungssystem TermStar, das Translation-Memory als Kernstück des Systems, sowie ein interaktives Alignment-Tool zum Konvertieren existierender Übersetzungen in Transit-Format. Transit operiert im Gegensatz zu anderen TM-Systemen nicht auf einem Datenbanksystem, sondern arbeitet mit einem effizienten Dateiverwaltungssystem. Die ausgangssprachliche Datei und die zielsprachliche Datei bilden ein *Sprachpaar*, wobei das Dateikürzel die jeweilige Sprache anzeigt. Liegt ein fertig übersetztes Sprachpaar in Transit-Format vor, dient es dem Translation-Memory als so genanntes *Referenzmaterial* für künftige Übersetzungsprojekte.

In Abbildung 1 ist der Übersetzungsprozess in Transit am Beispiel der englisch-deutschen Übersetzung eines Worddokuments dargestellt: Beim Import wird die Formatierungsinformation des englischen Worddokuments (Text.doc) herausgefiltert und in einer separaten Datei (Text.cod) abgelegt. Nach dem Filtern entsteht ein Transit-Sprachpaar (Text.eng, Text.ger), in dem die Formatierungsinformation in XML-Format kodiert ist[3] und im Text durch Platzhalter (*Tags*) ersetzt ist. Im zweiten Schritt wird der Text in so genannte Segmente (Überschriften, Sätze, usw.) unterteilt, für die anschließend eine Vorübersetzung mit Zugriff auf das projektspezifische Referenzmaterial durchgeführt wird: die Übersetzung identischer Texteinheiten wird aus dem Translation-Memory automatisch in die ziel-

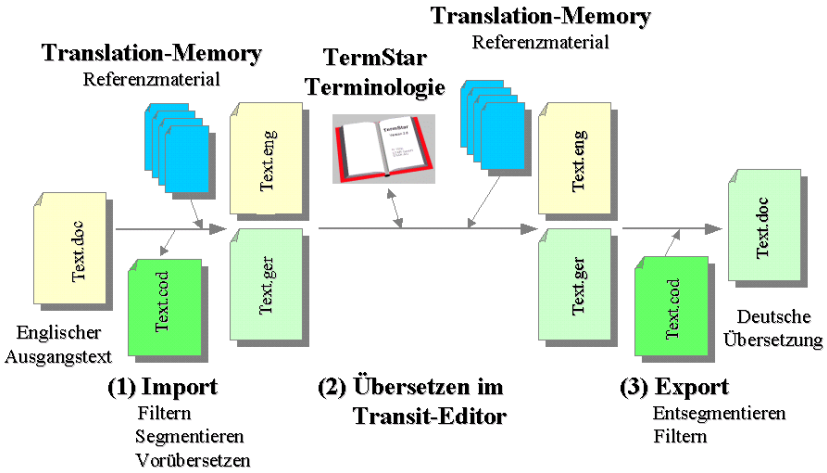


Abb. 1: (1) Import eines englischen Worddokuments, (2) Übersetzen im Transit-Editor, (3) Export der deutschen Transit-Datei ins Word-Format.

sprachliche Datei (Text.ger) eingefügt, so dass sie nach dem Import bereits vollständig übersetzte Segmente enthält. Im Transit-Editor wird der deutsche Text mit Zugriff auf die projektspezifischen Terminologiewörterbücher von TermStar und die Übersetzungsvorschläge aus dem Translation-Memory fertig bearbeitet. Anschließend wird die zielsprachliche Datei aus Transit exportiert, wobei die Segmentierung aufgehoben und die Formatierungsinformation wieder hinzugefügt wird. Als Ergebnis liegt die deutsche Übersetzung im Wordformat vor.

3 Übersetzen mit Transit

Der übersetzungsspezifische Transit-Editor (vgl. Abb.2) arbeitet mit fünf Fenstern, deren Anordnung benutzerspezifisch definierbar ist: (1) Ausgangstext-Fenster, (2) Zieltext-Fenster, (3) Fuzzy-Index Fenster mit Übersetzungsvorschlägen, (3) Wörterbuchfenster und (4) Notizfenster mit Informationen zum Status der Segmente (z.B. übersetzt, nicht übersetzt etc.). Der Transit-Editor basiert auf einer Emulation des weitverbreiteten Textverarbeitungssystems WinWord[4] und bietet eine Vielzahl von Tastenkombinationen zum effizienten manuellen Übersetzen an. Globale Editieroperationen, wie z.B. gezieltes Suchen/Ersetzen, können für alle

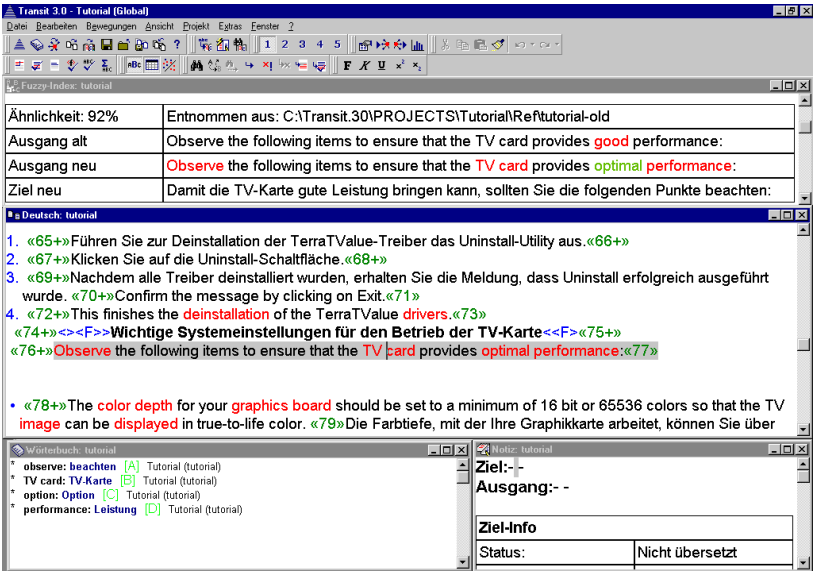


Abb. 2: Transit-Editor 3.0 mit Zieltext-Fenster, Fuzzy-Index Fenster, Wörterbuchfenster und Notizfenster. Das Ausgangstext-Fenster liegt im Hintergrund.

Dateien eines Übersetzungsprojekts gleichzeitig durchgeführt werden und vereinfachen so die beim Übersetzen anfallenden Routinearbeiten. Ab Version 3.0 wird der Text als Quasi-WYSIWYG angezeigt, so dass die für die Übersetzung notwendigen Formatierungsinformationen (z.B. die Spaltenbreite einer Tabelle, die für zielsprachliche Formulierungen beachtet werden sollte) direkt sichtbar sind.

Eine detaillierte Übersicht über die Übersetzung (wie viele Segmente automatisch vorübersetzt wurden, ob ein Segment auf der Basis eines Übersetzungsvorschlages übersetzt wurde, etc.) kann jederzeit abgerufen werden. Zur Feinabstimmung von Übersetzungseinheiten (Konkordanzen) können mit Hilfe eines ausgereiften Filtermechanismus genau diejenigen Segmente angezeigt (bzw. ausgeblendet) werden, die z.B. ein bestimmtes Wort oder eine bestimmte Phrase enthalten. Integrierte Prüffunktionen erlauben es, den Text auf korrekte Orthographie[5] und projektspezifische Fachterminologie zu überprüfen.

3.1 Interaktion zwischen Transit und Translation-Memory

Das Translation-Memory von Transit besteht aus Referenzdateien, den Transit-Sprachpaaren, die projektspezifisch zusammengestellt werden können. Neben Referenztexten, d.h. Übersetzungen, die zuvor mit Transit erstellt wurden oder über das Alignment-Tool ins Transit-Format umgewandelt wurden, werden auch Segmente aus dem aktuellen Übersetzungsprojekt als Referenzbasis verwendet: Sobald eine Texteinheit in Transit übersetzt wurde, dient diese Übersetzung sofort im Transit-Editor als Übersetzungsvorschlag für nachfolgende Segmente. Aus den Referenztexten kann ein projektspezifischer Referenzextrakt generiert werden, aus dem alle redundanten Segmente entfernt wurden, wobei Übersetzungsvarianten jedoch erhalten bleiben. Außerdem kann ein projektspezifisches Fuzzy-Match-Sprachpaar erzeugt werden, das alle diejenigen ausgangssprachlichen Segmente (plus deren Übersetzung) enthält, die zu den Segmenten des aktuellen Ausgangstextes einen zuvor festgesetzten Ähnlichkeitsgrad aufweisen.

Auf die Translation-Memory-Komponente wird zweifach zugegriffen: beim automatischen Vorübersetzen während des Imports und beim Übersetzen im Transit-Editor. Während des Imports werden alle Segmente, für die im Referenzmaterial identische Segmente (*Exact Match*) gefunden werden, automatisch vorübersetzt und brauchen nicht mehr bearbeitet zu werden. Außerdem kann optional eine teilweise Vorübersetzung durchgeführt werden, bei der Transit geringe Abweichungen (symmetrische Zahlen- und Formatierungsdifferenzen oder benutzerdefinierte Ausnahmen) in der Übersetzung automatisch anpaßt, die anschließend nur überprüft werden müssen (vgl. Abb.3).

Ausgang alt	Windows 95 or Windows 98
Ausgang neu	Windows 98 or Windows 2000
Ziel neu	Windows ›98‹ oder Windows ›2000‹

Abb. 3: Die Zahlen 95 und 98 wurden automatisch an die Zahlen des aktuellen Segments angepasst.

Beim manuellen Übersetzen im Transit-Editor liefert das Translation-Memory automatisch Übersetzungsvorschläge und bietet sie im Fuzzy-Index-Fenster an. Die Unterschiede zwischen ausgangssprachlichem Referenzsegment und dem aktuellen Segment werden farblich hervorgehoben und erleichtern somit das Bearbei-

ten des Übersetzungsvorschlages (siehe Abb.4): Die Wörter *drivers and* sind im aktuellen Segment hinzugekommen und werden in „Ausgang neu“ farblich markiert. Die Zahlenunterschiede zwischen „Ausgang alt“ und „Ausgang neu“ sind ebenfalls farblich hervorgehoben und wurden automatisch im Übersetzungsvorschlag angepaßt.

Ähnlichkeit: 84%	Entnommen aus: D:\transit-30\PROJECTS\Tutorial\Ref\tutorial-old
Ausgang alt	CD-ROM drive (to install the software under Windows 95 or Windows 98)
Ausgang neu	CD-ROM drive (to install the drivers and software under Windows 98 or Windows 2000)
Ziel neu	CD-ROM Laufwerk (zur Installation der Software unter Windows ›98‹ oder Windows ›2000‹)

Abb. 4: Übersetzungsvorschlag im Fuzzy-Index-Fenster.

Der beste Übersetzungsvorschlag, d.h. derjenige der den höchsten Ähnlichkeitsgrad aufweist, wird zuerst angezeigt. Weitere Vorschläge können per Tastendruck angefordert werden. Das Referenzsprachpaar kann jederzeit (z.B. zum Korrigieren der Referenzübersetzung oder zum Einsehen des Kontextes des Referenzsegments) jederzeit im Transit-Editor geöffnet werden. Sobald der Übersetzungsvorschlag angepasst ist, wird er per Tastendruck in den Zieltext übernommen.

3.2 Interaktion zwischen Transit und TermStar

Das integrierte Terminologieverwaltungssystem TermStar dient der effizienten Nutzung von Fachterminologie beim Übersetzen mit Transit. Für das aktuelle Sprachpaar führt Transit automatisch in allen projektspezifischen Wörterbüchern eine Hintergrundsuche durch und hebt die ausgangssprachlichen Wörter, für die ein Eintrag gefunden wurde, im Text farblich hervor. Die gefundenen Fachbegriffe werden im Wörterbuchfenster des Transit-Editors zu einem einzigen, großen, virtuellen Wörterbuch zusammengefasst. Zu jedem Segment werden im Wörterbuchfenster die Fachbegriffe und die zugehörige(n) Übersetzung(en) angezeigt, wobei zu jedem Eintrag angegeben ist, aus welchem Wörterbuch er stammt (vgl. Abb.2). Die Übersetzung eines Fachbegriffs kann per Tastendruck direkt in den Zieltext übernommen werden. Durch das interaktive Zusammenspiel zwischen TermStar und Transit kann automatisch überprüft werden, ob ein Fachbegriff tatsächlich mit einer in den Projektwörterbüchern vorhandenen Übersetzung übersetzt wurde. Im Transit-Editor können neue Fachbegriffe über die Schnelleingabe-Maske oder per Mausklick Terminologie direkt erfaßt werden, so dass sie für die weitere Übersetzung sofort zur Verfügung stehen.

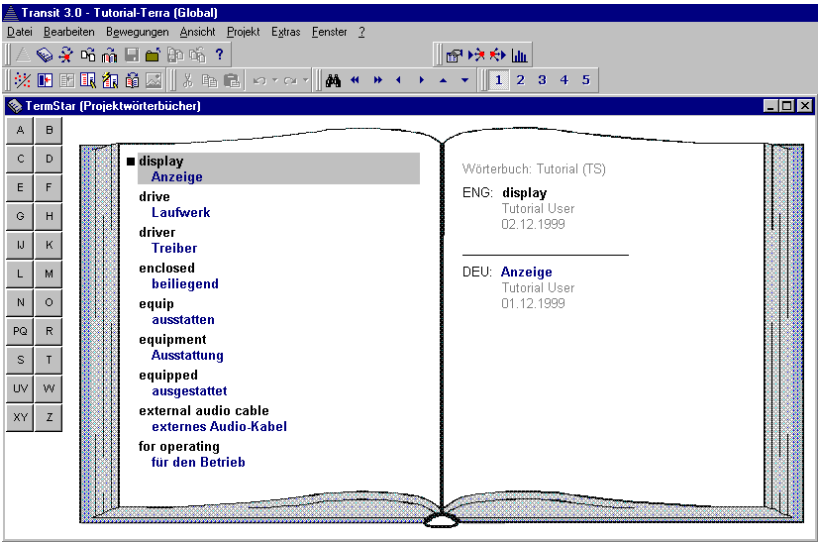


Abb. 5: TermStar Arbeitsplatz. Das Layout der Wörterbücher ist individuell konfigurierbar.

TermStar selbst ist ein professionelles Datenbanksystem für die multilinguale Terminologiearbeit, das in einem benutzerfreundlichen Buchdesign (vgl. Abb. 5) eine Vielzahl leistungsfähiger Funktionen zur Terminologiepflege und zum Terminologieabruf bietet:[6]

- **Komplexes Annotationsschema**: Jeder Fachbegriff kann mit einer Vielzahl unterschiedlicher Informationen (Querverweise, Definition, Wortart, Eintragsdatum, Benutzerkennung, graphische Illustrationen) versehen (annotiert) sein, wobei Auswahllisten und Vorgabewerte die Konsistenz terminologischer Information sowie die sachgebiets- oder kundenorientierte Klassifikation der Einträge ermöglichen.
- **Effiziente Abrufmechanismen**: Neben Navigiermechanismen (z.B. über die Buchstabenleiste) und benutzerfreundlicher Hyperlink-Technologie, über die Querverweise zu verwandten Fachbegriffen direkt angesteuert werden können, bietet TermStar effiziente Suchfunktionen (u.a. über reguläre Ausdrücke) zum gezielten Abrufen spezieller Datensät-

ze. Außerdem ist es möglich, über benutzerdefinierbare Filter nur bestimmte Datensätze aller zu einem Projekt geladenen Wörterbücher anzeigen zu lassen.

- **Import-/Export-Schnittstellen**: Existierende Terminologielisten können in (neue oder bereits bestehende) TermStar-Wörterbücher integriert werden, wobei eine detaillierte Kontrollfunktion zur Vermeidung von Doppeleinträgen bereitgestellt wird. Über benutzerspezifisch konfigurierbare Exportmechanismen können (vollständige) Wörterbücher aus TermStar exportiert werden oder Auszüge eines Wörterbuchs, die zuvor über die oben genannten Filter definiert wurden, in ein neues TermStar-Wörterbuch umgewandelt werden.

4 Kennzeichen von Transit

Transit zeichnet sich durch folgende Kriterien besonders aus:

Unterstützte Sprachen: Neben west- und osteuropäischen Sprachen[7] werden auch asiatische Sprachen, wie Japanisch, Chinesisch, Thailändisch oder Koreanisch, unterstützt. Die Version 3.0 wird zusätzlich Arabisch (und später auch Hebräisch) unterstützen. Transit 3.0 basiert auf Unicode und erleichtert damit die Handhabung unterschiedlicher Zeichensätze.

Unterstützte Dateiformate: Vollintegrierte Import- und Exportmechanismen erlauben die reibungslose Bearbeitung einer Vielzahl gängiger Dateiformate, ohne dass zusätzliche Konvertierungstools benötigt werden. Eingebettete Objekte eines unterstützten Dateiformats (z.B. MS-Powerpoint-Graphik in einem Word-Dokument) können in Transit bearbeitet werden.

Performanz der TM-Komponente: Im Gegensatz zu Datenbank-basierten TM-Systemen treten Performanzprobleme beim Zugriff auf das Datei-basierte Translation-Memory von Transit wesentlich weniger auf. Die Retrieval-Geschwindigkeit, die u.a. von Größe und Verwaltung des TM abhängt, kann in Transit über die benutzerspezifische Zusammenstellung des Referenzmaterials direkt positiv beeinflußt werden.

Transparenz der TM-Komponente: Das Prinzip des Dateiverwaltungssystems erlaubt es, nur genau diejenigen Übersetzungen als Referenzbasis auszuwählen, die für das aktuelle Übersetzungsprojekt relevant sind und den ausgewählten Referenzdateien zudem eine bestimmte Priorität zuzuweisen. Da das Referenzmaterial im Translation-Memory in Dateiformat vorliegt, kann ein Referenz-Sprachpaar im Transit-Editor jederzeit bearbeitet werden, und der Kontext einer Übersetzungseinheit ist direkt einsehbar.

Effektivität der TM-Komponente: Aufgrund der XML-basierten Kodierung der Formatierungsinformation in Transit-Sprachpaaren werden die Unterschiede zwischen Segmenten, die aus unterschiedlichen Dateiformaten in Transit importiert werden, reduziert. Dadurch kann das Referenzmaterial im Translation Memory noch effizienter zwischen Übersetzungsprojekten genutzt werden, deren ausgangssprachliche Texte in verschiedenen Dateiformaten erstellt wurden. Ab Version 3.0 erlaubt Transit den Austausch von Referenzmaterial über TMX (Translation Memory Exchange Format).

Interaktives Alignment-Tool: Existierende Übersetzungen in Dateiformaten, die Transit unterstützt, können über das Alignment-Tool zu Transit-Sprachpaaren aufbereitet und in das Translation-Memory integriert werden. Der Alignment-Prozess läuft weitgehend automatisch ab und verlangt nur dann ein Eingreifen, wenn Strukturunterschiede zwischen Ausgangstext und Übersetzung vorliegen.

5 Schlussbemerkung

Grundlage einer qualitativ hochwertigen TM-Komponente sind die Qualität der darin verfügbaren, manuell erstellten Übersetzungen und die Qualität des Retrieval-Algorithmus zum Auffinden der besten Referenzübersetzungen. Da ein Translation-Memory-System (im Gegensatz zu einem automatischen Übersetzungssystem) nicht selbständig übersetzt, sondern nur auf vorliegende Übersetzungen zurückgreift, ist die Korrektheit der Übersetzungseinheiten, die entweder im Rahmen der Vorübersetzung automatisch eingefügt wurden oder im Transit-Editor angeboten werden, gewährleistet, sofern die zuvor manuell angefertigte Übersetzung fehlerfrei ist.

Transit soll und kann qualifizierte Übersetzerinnen und Übersetzer nicht ersetzen, sondern dient ihnen als professionelle Arbeitshilfe. Die Gesamtheit der Transit-Funktionalitäten bildet eine vollständige, integrierte Übersetzungsumgebung, die bei gleichzeitiger Verbesserung der Übersetzungsqualität umfassende Möglichkeiten zur Produktivitätssteigerung bietet.

Danksagung

Mein Dank gilt meinen Kolleginnen und Kollegen von der Transit Abteilung der STAR Deutschland GmbH, insbesondere Christiane Gläser, Bettina Steudel und Oliver Rau, die mit ihrem umfassenden Fachwissen viele wesentliche Punkte zu diesem Artikel beigetragen haben.

Literatur

[Andrés Lange98] Andrés Lange, C. (1998): 'Tying the Knot. How Baan wed machine translation to translation memory – and survived the honeymoon', in: *Language International*, 34-36.

[Benis99] Bensi, M. (1999): 'Translation Memory. From Q_1 to R_1', in: *Bulletin of the Institute of Translation and Interpretation*, 4/99, 4-19.

[Steudel95] Steudel, B. (1995): *Management von Änderungsübersetzungen. Untersuchung zur Einführung eines Übersetzungsspeichers in einem industriellen Sprachendienst.* Diplomarbeit.

ANMERKUNGEN

[1] Die STAR Firmengruppe (http://www.star-group.net) mit Hauptsitz in der Schweiz wurde vor 15 Jahren gegründet und hat sich vor allem auf den Bereich technische Dokumentation spezialisiert. Das Transit System wird seit Ende der 80er Jahre von der STAR Deutschland GmbH entwickelt und ist seit 1990 bei STAR intern im Einsatz.

[2] Neben den Dateiformaten Winword, Wordperfect, Amipro, Windows Resource-Dateien, Windows Hilfe-Dateien, Interleaf, FrameMaker, PageMaker, SGML, HTML, Ventura Publisher, QuarkXPress, C-Quellcode oder andere Quellcodes, ASCII, und Ansi-basierten Dateiformaten wird Transit 3.0 zudem Powerpoint und Excel unterstützen. Nicht alle Filter sind Bestandteil des Transit-Standardpakets.

[3] Die XML-konforme Kodierung wird ab Version 3.0 von Transit angeboten.

[4] Emulationen sind Nachbildungen der Texteditoren, d.h. bestimmte Tastenkombinationen der Editoren (wie z.B. STRG + S zum Speichern) haben im Transit-Editor die selbe Funktion.

[5] Das Transit-Standardpaket umfasst zwei (von insgesamt 18) Rechtschreibprüfprogrammen. Für Deutsch kann nach der alten und nach der neuen deutschen Rechtschreibung geprüft werden.

[6] TermStar kann als integraler Bestandteil von Transit genutzt werden, ist aber auch als eigenes Softwarepaket erhältlich.

[7] Deutsch, Englisch, Französisch, Italienisch, Spanisch, Portugiesisch, Niederländisch, Schwedisch, Dänisch, Norwegisch, Finnisch, Griechisch, Türkisch, Indonesisch, Russisch, Tschechisch, Polnisch, Ungarisch, Rumänisch, Bulgarisch, Slowenisch, Ukrainisch.

Erfahrungen im Einsatz mit Translation-Memory-Systemen

Antje Pesch
DELTA International CITS GmbH

1 Einleitung

Im Rahmen der Arbeiten des Arbeitskreises Maschinelle Übersetzung zum Thema Leistungsfähigkeit und Einsatzmöglichkeiten von Translation Memory (TM) Systemen soll der folgende Artikel einen Überblick über die praktischen Erfahrungen eines Übersetzungsunternehmens beim Einsatz von TM-Systemen geben.

Unter einem TM sei im folgenden (entsprechend der am meisten verbreiteten TM-Architektur) eine Datenbank zu verstehen, in welcher Übersetzungseinheiten abgespeichert werden. Eine Übersetzungseinheit setzt sich zusammen aus einem ausgangssprachlichen Satz und seiner zielsprachlichen Entsprechung.

TM Systeme

- basieren auf dem Prinzip des Recycling: sie garantieren die Wiederverwertbarkeit der Übersetzungseinheiten schon während des Übersetzungsprozesses sowie ihre Konservierung für Folgeprojekte;
- arbeiten in der Regel satzbasiert[1];
- ermöglichen dem Übersetzer – im Gegensatz zum passiven Charakter der maschinellen Übersetzung – einen interaktiven Arbeitsprozeß: das System bietet Übersetzungseinheiten aus der Datenbank an, die der Übersetzer annehmen, editieren oder ablehnen kann.

TM-Systeme lassen sich bezüglich ihrer Funktionsweise in datenbank- und sprachpaarbezogene Systeme untergliedern. Als Vertreter auf der Basis von Datenbanken organisierter Systeme können beispielsweise die Translator's Workbench von TRADOS, der TranslationManager von IBM und Déjà Vu von ATRIL genannt werden. Das System Transit von STAR hingegen arbeitet sprachpaarbezogen.

Im folgenden sollen einige Eigenschaften der Translator's Workbench von TRADOS – dem derzeitig marktführenden TM-System – und des Transla-tionManager von IBM – einem TM-System, das die Delta International CITS GmbH oft einsetzt und dessen Vertrieb sie unterstützt – gegenübergestellt werden, um bestimmte Erfahrungen aus dem praktischen Umgang mit den TM-Systemen zu veranschaulichen.

2 TM-Systeme in der Praxis

Der Einsatz von TM-Systemen betrifft vor allem Übersetzer, den technischen Support und das Projektmanagement.

2.1 Erfahrungen der Übersetzer

Für den Übersetzer, der mit einem TM-System arbeitet, ist zunächst die Benutzerfreundlichkeit des jeweiligen Systems von entscheidender Bedeutung. Weiterhin ist es für ihn von Vorteil, während des Übersetzungsvorgangs auf das TM und die Terminologieerkennung zugreifen zu können.

2.1.1 Benutzerfreundlichkeit

- **Arbeitsumgebung**: Die Schnittstelle der Translator's Workbench zu den Textverarbeitungsprogrammen Word für Windows bzw. WordPerfect ermöglicht dem Übersetzer die Arbeit in einer vertrauten Umgebung. Ferner arbeitet die Workbench nach dem WYSIWYG-Prinzip.

  Beim TranslationManager muß der Übersetzer im systemeigenen Editor arbeiten. Dessen Aufbau ist für den Übersetzer zunächst fremd und verlangt eine gewisse Einarbeitung. Zudem arbeitet der TranslationManager nicht nach dem WYSIWYG-Prinzip.

- **Shortcuts**: Die Workbench akzeptiert einerseits die dem Übersetzer aus den Textverarbeitungsprogrammen bekannten Shortcuts und bietet ihm andererseits die Möglichkeit, die Funktionen der Workbench mit spezifischen Shortcuts zu bedienen.

Anders beim TranslationManager: das System akzeptiert die bekannten Shortcuts nicht, sondern verwendet dem Übersetzer völlig fremde Shortcuts. Teilweise geraten die im TranslationManager definierten Shortcuts sogar in Konflikt mit den Shortcuts der Textverarbeitungsprogramme.

- **Bearbeitungsfunktionen** (z. B. Suchen und Ersetzen): Die Translator's Workbench stellt dem Übersetzer zum einen die aus dem Textverarbeitungsprogramm bekannte Funktion zur Verfügung und bietet ihm zum anderen eine spezifische Funktion (*File/Maintenance/Global Changes*), die es ihm erlaubt, Änderungen auf Datei-Ebene vorzunehmen.

Beim TranslationManager ist das Verfahren komplizierter: Möchte der Übersetzer Änderungen während des Übersetzungsvorgangs vornehmen, so ist er gezwungen, in den *Post-Editing-Style* zu wechseln. Globale Änderungen sind hier sowohl auf Datei- als auch auf Folder-Ebene[2] möglich – die Datei bzw. sämtliche Dateien des Folders müssen jedoch vorher vollständig bearbeitet worden sein.

Die Translator's Workbench bietet dem Übersetzer aufgrund ihrer Arbeitsumgebung zunächst einen intuitiveren Zugang, jedoch ist TRADOS gezwungen, bei jeder Veränderung der Textverabeitungsprogramme, die Workbench entsprechend anpassen zu müssen. Der TranslationManager hat hinsichtlich der Benutzerfreundlichkeit derzeit noch einige Defizite.

2.1.2 Zugriff auf das TM

- **Minimum Match Value**: Die Translator's Workbench bietet dem Übersetzer die Möglichkeit, den *Minimum Match Value*[3] sowohl für das TM (Abb. 1, gegenüberliegenden Seite) als auch für die Terminologieerkennung selber einzustellen (Abb. 2).

  Er muß sich dabei natürlich bewußt sein, daß ein niedrigerer Minimum Match Value in der Regel eine geringere Qualität der Übersetzung impliziert. Jedoch kann es für den Übersetzer dennoch interessant sein, von diesen „schlechteren" Übersetzungs- bzw. Terminologievorschlägen profitieren zu können.

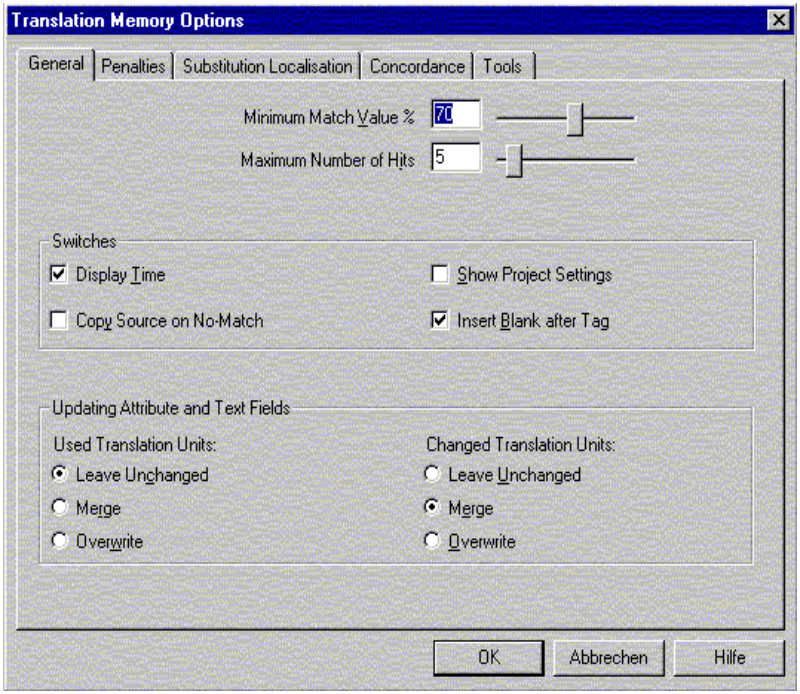


Abb. 1: Angabe des „Minimum Match Value“ für das TM.

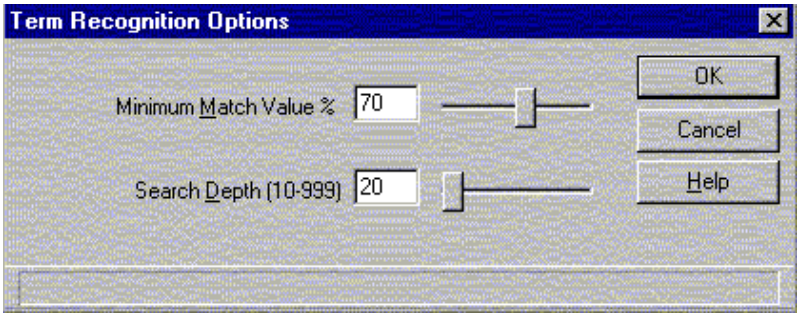


Abb. 2: Angabe des „Minimum Match Value“ für die Terminologieerkennung.

Der TranslationManager bietet diese Möglichkeit nicht. Der Minimum Match Value liegt bei 70 %, kann aber nicht verändert werden.

- **Penalties**: Die Translator's Workbench erlaubt es dem Übersetzer, sogenannte Penalties[4] individuell einzustellen.

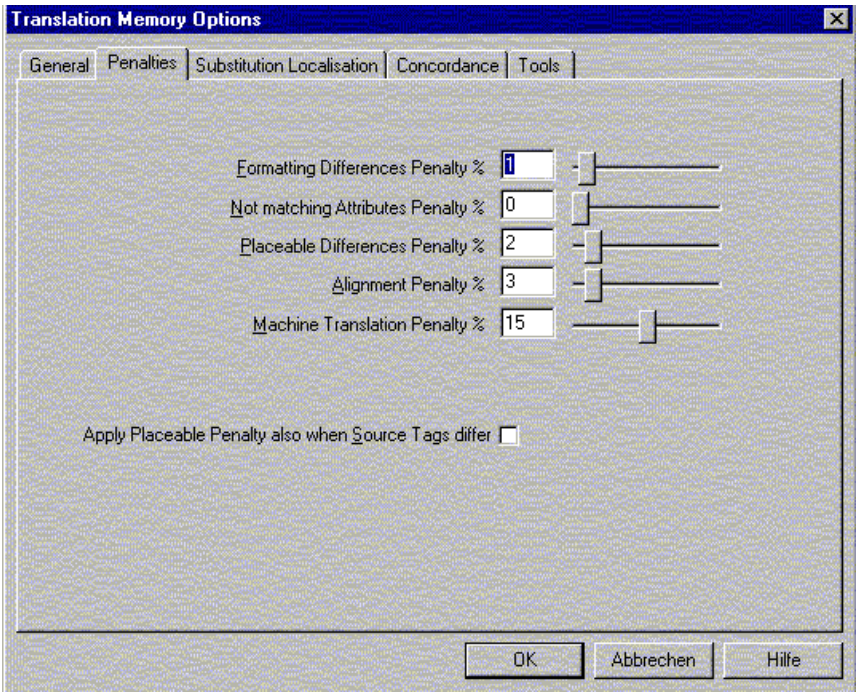


Abb. 3: Individuell spezifizierbare Abzüge für Match-Werte.

Beim TranslationManager ist dies nicht möglich: Weder gibt es eine Differenzierung von Abzügen für Kandidaten aus unterschiedlichen Referenzmaterialien, noch kann man Penalties individuell einstellen.

- **Konkordanzsuche**: Beim Arbeiten mit der Workbench kann der Übersetzer von der Konkordanzsuche, der Suche nach Begriffen oder Phrasen im Kontext, profitieren. Er kann dabei den Minimum Match Value wiederum individuell einstellen.

  Der TranslationManager bietet diese Funktion nicht.

2.1.3 Zugriff auf Terminologie

Beide Systeme verfügen über eine aktive Terminologieerkennung.

Bei der Translator's Workbench ist es das externe Terminologieverwaltungstool *Multiterm*, welches über eine Schnittstelle in den Übersetzungsprozeß eingebunden werden kann. Die Tatsache, daß es sich bei *Multiterm* noch um eine 16-Bit-Version handelt, was sich durch hohe Suchzeiten bei großen Glossaren bemerkbar macht, bedeutet ein gewisses Defizit, da die Übersetzer dazu tendieren, die Terminologieerkennung zu deaktivieren, was wiederum Mängel in der Konsistenz der Übersetzung zur Folge haben kann.

Der TranslationManager besitzt eine interne Dictionary-Komponente, die für jeden Folder die Auswahl und hierarchische Organisation mehrerer Dictionaries erlaubt.

2.2 Erfahrungen des technischen Support

Der technische Support interessiert sich zunächst dafür, welche Dateiformate mit dem jeweiligen TM-System bearbeitet bzw. welche Konvertierungen eventuell notwendig werden können. Weiterhin arbeitet er mit den Analyse- und Vorübersetzungsfunktionen der Systeme und ist zuständig für die Aktualisierung des TM.

2.2.1 Dateiformate und Konvertierungen

Mit der Workbench können ausschließlich rtf-Dateien bearbeitet werden. Bei anderen Dateiformaten sind Konvertierungen notwendig (z. B. mit dem S-Tagger bei Interleaf- bzw. FrameMaker-Dateien), die immer ein gewisses Risiko beinhalten.[5]

Der TranslationManager erlaubt ein formatunabhängiges Arbeiten, da er für jedes Dateiformat über ein spezifisches Markup verfügt.

2.2.2 Analyse- und Vorübersetzungsfunktion

Die Analysefunktion der Translator's Workbench liefert zum einen die Wort- und Segmentzahl und zum anderen die Anzahl bzw. den prozentualen Anteil an Wiederholungen sowie *No*, *Fuzzy* und *Full Matches*.

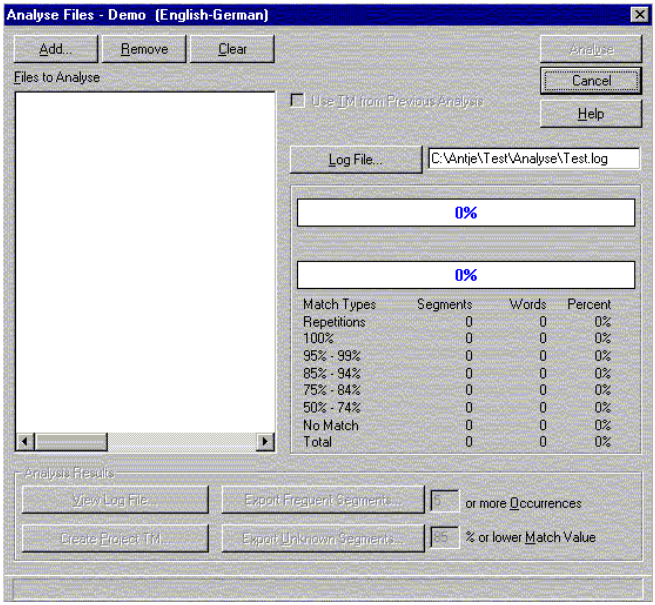


Abb. 4: Dateianalyse mit der Translator's Workbench.

Bei der Vorübersetzung kann der Minimum Match Value individuell eingestellt werden. In der Regel werden allerdings nur Full Matches ersetzt. Es besteht die Möglichkeit, zum einen häufig vorkommende, andererseits aber auch unbekannte Segmente zu exportieren und auf maschinellem Wege übersetzen zu lassen.

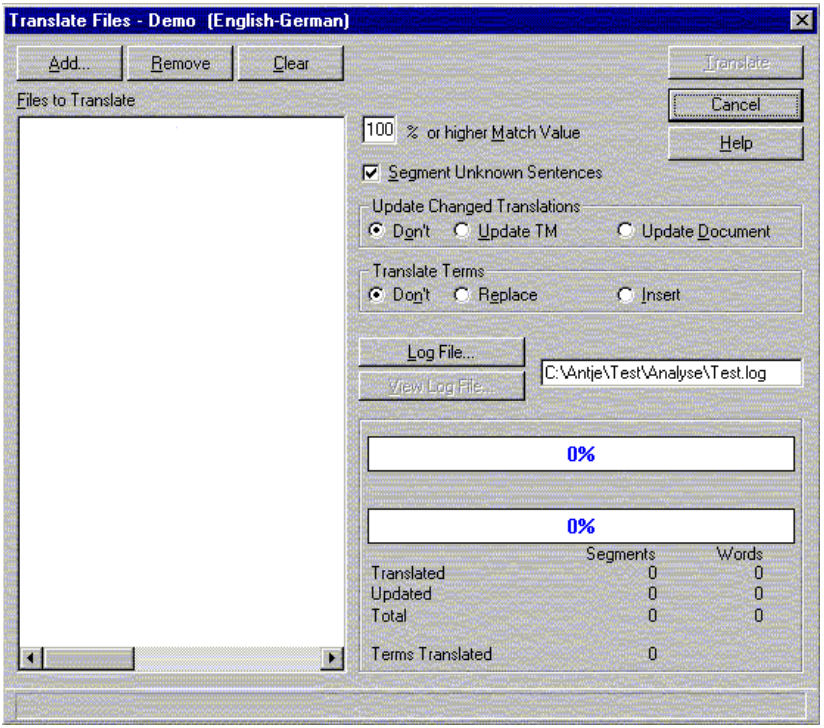


Abb. 5: Einstellungen für die Vorübersetzung.

Beim TranslationManager liefert die WordCount-Funktion die Wortzahl sowie die Anzahl an *No*, *Fuzzy* und *Full Matches*. Die Segmentierung wird erst durch die Analysefunktion durchgeführt, wobei gleichzeitig eine Vorübersetzung durchgeführt werden kann. Es werden ausschließlich Full Matches[6] ersetzt.

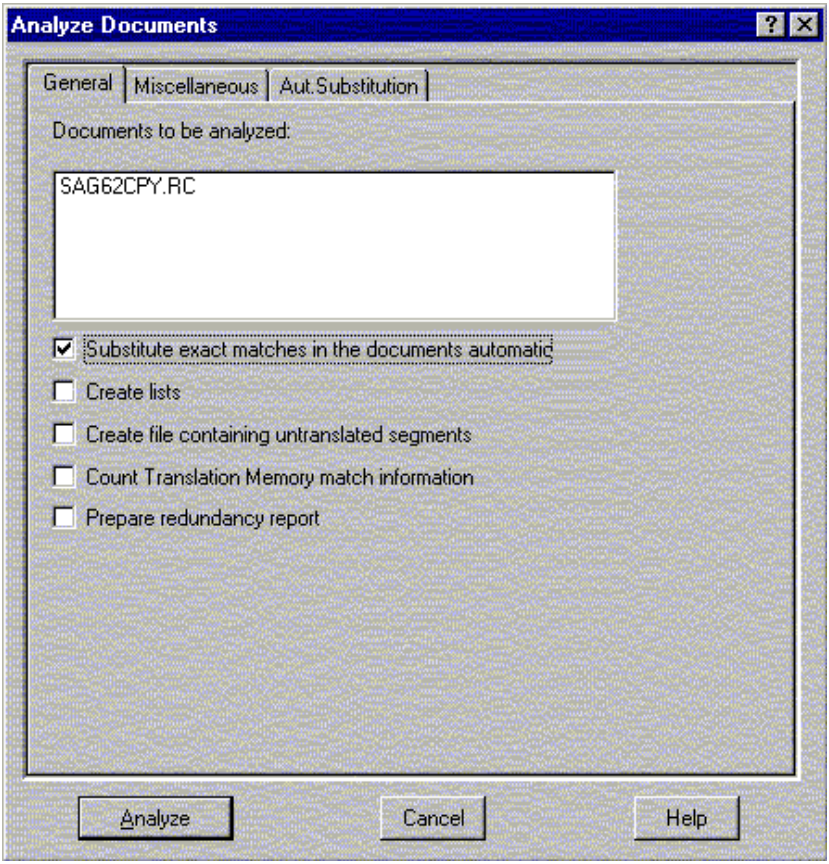


Abb. 6: Dateianalyse mit dem TranslationManager.

2.2.3 Aktualisierung des TM

Der Prozeß der Aktualisierung des TM wird bei der Workbench *CleanUp* genannt. Beim CleanUp werden die ausgangssprachlichen Segmente aus dem vollständig übersetzten Text entfernt und zusammen mit ihren zielsprachlichen Entsprechungen als Übersetzungseinheit im TM abgespeichert. Dabei werden – dif-

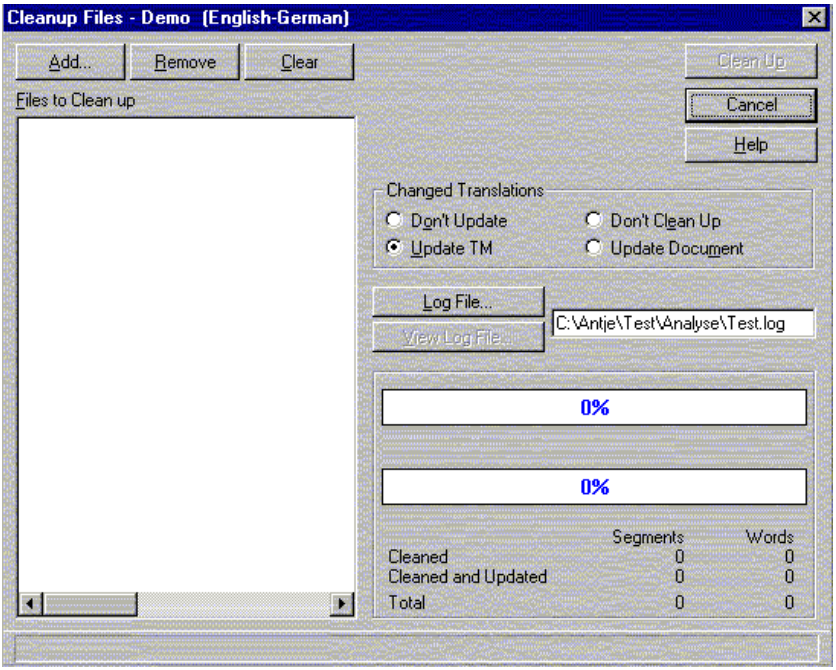


Abb. 7: Aktualisierung des TM bei der Translator's Workbench.

ferenziert nach den den Übersetzungseinheiten zugeordneten projektspezifischen Attributen – schon vorhandene Übersetzungseinheiten überschrieben.

Beim TranslationManager ist der Prozeß um einiges aufwendiger. Die einzelnen Translation Memories müssen zunächst mit dem Befehl *Build archive TM* auf die Aktualisierung vorbereitet werden. Sie werden dann in einem neuen TM zusammengeführt. Sämtliche Einträge dieses TM müssen schließlich – zusätzlich differenziert nach den unterschiedlichen Markups – durchgesehen und unifiziert werden.

2.3 Erfahrungen des Projektmanagements

Die praktischen Erfahrungen des Projektmanagements beim Einsatz von TM-Systemen sind eher indirekter Art. Im Projektmanagement treffen die unterschiedlichen Motivationen der Kunden einerseits und der Übersetzer andererseits aufeinander. Es ist die Aufgabe des Projektmanagers, zwischen dem Interesse der Kunden, die Zeit- und Geldaufwand der Übersetzung möglichst gering halten wollen und dem Interesse der Übersetzer, die eine angemessene Bezahlung und die Anerkennung des kreativen Aspektes ihrer Tätigkeit verlangen, zu vermitteln.

Der Konflikt äußert sich vor allem in der Frage der Bezahlung von 100%-Matches. Die Kunden sind oft nicht bereit, für diese zu bezahlen, da sie meinen, der Übersetzer brauche sie nicht zu bearbeiten. Die Praxis zeigt allerdings, daß es keine 100%-Matches gibt. Es gibt in der Übersetzung keine mathematischen Äquivalenzen, sondern es handelt sich bei jeder Übersetzungseinheit um eine semantische Einheit, die im Kontext überprüft werden muß. Logische Konsequenz für den Übersetzer ist es jedoch, diese notwendige Überprüfung nicht vorzunehmen, da sie ihn Zeit kostet, die ihm nicht bezahlt wird. Dies geht zu Lasten der Qualität der Übersetzung. Außerdem kann bei den Übersetzern Demotivation hervorgerufen werden, denn ihre vorher kreative Übersetzeraufgabe wird reduziert auf ein simples Bestätigen der 100%-Matches.[7]

3 Fazit

Viele Faktoren beeinflussen die Entscheidung für oder gegen den Einsatz von TM-Systemen bzw. die Auswahl eines bestimmten TM-Systems.

Von Bedeutung ist zunächst die Akzeptanz des TM-Systems bei den Übersetzern, die vor allem von der Benutzerfreundlichkeit des Systems beeinflußt wird. Der technische Support interessiert sich für die Vor- und Nachteile im Vergleich der technischen Leistungen, primär bezüglich der unterstützten Dateiformate und hinsichtlich der Möglichkeiten der Pflege und Aktualisierung des TM.

In der Praxis steht der Benutzerfreundlichkeit und dem technischen Leistungsvermögen eines TM-Systems jedoch immer die wirtschaftliche Betrachtung gegenüber.

ANMERKUNGEN

[1] Die Segmentierung des zu übersetzenden Textes nach Sätzen und damit die Bildung der Übersetzungseinheiten aus in sich schlüssigen Satzeinheiten hat sich in der Praxis bewährt. Dennoch besteht die Möglichkeit, bei asiatischen Sprachen sogar die Notwendigkeit, den zu übersetzenden Text nach Absätzen zu segmentieren.

[2] Bei TranslationManager werden die Dateien in Foldern (Ordnern) organisiert.

[3] Der *Minimum Match Value* ist ein prozentualer Schwellenwert, ab welchem Übersetzungseinheiten bzw. Terminologie dem Übersetzer als Vorschlag angezeigt werden. Im allgemeinen wird ein Wert von 70 % empfohlen.

[4] Penalties sind prozentuale Abzüge vom Match-Wert auf Grund von Unterschieden in der Formatierung, den Attributen oder den *Placeables* (nicht zu übersetzende Texteinheiten wie Zahlen oder Datumsangaben) bzw. auf Grund von Alignment oder maschineller Übersetzung.

[5] Für die Bearbeitung von HTML-Dateien hat TRADOS inzwischen den Tag-Editor entwickelt.

[6] Der TranslationManager unterscheidet bei den Full Matches nach dem Kontext in *exact context match* und *last exact context match* bzw. *Full Matches* aus sogenannten *joined segments*.

[7] In verschärfter Form existiert diese Problematik im Falle von maschineller Übersetzung, wo den Übersetzern lediglich die Aufgabe des Post-Editing bleibt und es für sie außerdem meist weit schwieriger ist, die falsch übersetzten Sätze zu korrigieren als diese selber zu übersetzen.

Überlegungen zu einer engeren Verzahnung von Terminologiedatenbanken, Translation Memories und Textkorpora

Uwe Reinke
Universität des Saarlandes

1 Interaktion der Komponenten von Translation Memory-Systemen

TM-Systeme besitzen im wesentlichen zwei 'Wissensquellen': Terminologiedatenbank und Referenzmaterial, d.h. eine TM-Datenbank oder eine maschinenlesbare Sammlung von Texten und ihren Übersetzungen. Bislang ist das Maß an Interaktion zwischen diesen beiden 'Wissensquellen' allerdings eher gering. So werden Terminologiedatenbanken lediglich benutzt, um in den zu übersetzenden Texten Termini zu identifizieren und die entsprechenden zielsprachlichen Benennungen zur Verfügung zu stellen. Keines der verfügbaren kommerziellen Systeme setzt vorhandene Terminologie ein, um die Retrieval-Leistung seiner TM-Komponente zu verbessern.

Ähnlich stellt sich die Situation bei der Verwendung vorhandener maschinenlesbarer Texte und ihrer Übersetzungen dar. Diese werden in erster Linie als 'Ansammlung von Übersetzungseinheiten' behandelt. Die in ihnen 'eingebettete' Terminologie bleibt ungenutzt. Zwar ist die rechnergestützte Terminologiegewinnung noch ein sehr junges Forschungsfeld, dennoch finden sich in der Literatur einige interessante Ansätze, die bislang allerdings nicht in kommerzielle TM-Systeme Eingang gefunden haben und lediglich in einigen meist nicht-kommerziellen Konkordanzwerkzeugen und Programmen zur Terminologieextraktion implementiert sind.

Im folgenden möchte ich einige Möglichkeiten aufzeigen, wie die Retrieval-Leistung von TM-Systemen durch Nutzung der vorhandenen terminologischen Informationen und Parallelkorpora[1] verbessert werden könnte. Dabei werde ich zwischen 'expliziten' und 'impliziten' terminologischen Informationen unterscheiden. 'Explizite terminologische Informationen' sind die für andere Komponenten des TM-Sy-

stems direkt zugänglichen Termini der Terminologiedatenbank. 'Implizite terminologische Informationen' sind die für die Komponenten des TM-Systems nicht zugänglichen, in die Texte des Parallelkorpus eingebetteten Termini.

2 Optimierung von TM-Systemen durch bessere Nutzung expliziter terminologischer Informationen

Derzeit finden sich in der Literatur die beiden folgenden Vorschläge:

- 'Generalisierung' der im TM gespeicherten Übersetzungseinheiten mit Hilfe der in der Terminologiedatenbank vorhandenen Benennungen
- Verwendung von Termini zur Alignierung von Einheiten unterhalb der Satzebene.

2.1 Explizite terminologische Informationen zur Generalisierung der Übersetzungseinheiten im Translation Memory

In einem Aufsatz, der in einer Sonderausgabe der Zeitschrift *Machine Translation* zur rechnergestützten Humanübersetzung erschienen ist, schlagen Langé et al. [LGD97] vor, die Übersetzungseinheiten eines TM mit Hilfe der in der Terminologiedatenbank des TM-Systems verfügbaren Termini zu 'abstrahieren' und bekannte Termini durch Variablen zu ersetzen. Das Ziel dieses Ansatzes besteht darin, den Recall von TM-Komponenten – d.h. die 'Fähigkeit', relevante Übersetzungseinheiten zu finden - zu verbessern. Das folgende Beispiel aus dem erwähnten Aufsatz verdeutlicht den Grundgedanken:

(1) Proceed with *installation checking*.

(2) Proceed with *customization.*

(3) Proceed with *X*.

Da die kursiv markierten Termini mit ihren zielsprachlichen (ZS) Entsprechungen in der Terminologiedatenbank vorhanden sind, reicht es nach Langé et al. aus, wenn das TM das abstrahierte Segment (3) enthält. Langé et al. gehen davon aus, daß:

> "a sentence that has been skeletonized to include variable parts is more general, and should therefore be found more frequently in the translation memory than fully instantiated sentences" [LGD97:46].

Allerdings haben Tests von TM-Systemen gezeigt, daß einfache paradigmatische Modifikationen – d.h. Änderungen, die (nahezu) keinen Einfluß auf Syntax und Länge einer Übersetzungseinheit besitzen[2] – i.d.R. nicht zu Retrieval-Problemen führen (vgl. z.B. [Rei94], [Rös/War97]). Der einzige Vorteil einer 'Generalisierung' besteht möglicherweise darin, daß sich auf diese Weise der Umfang von TM-Datenbanken reduzieren läßt. Dies setzt jedoch voraus, daß die zu bearbeitenden Texte einen hohen Anteil an syntaktisch identischen ausgangssprachlichen (AS) Sätzen enthalten. Andererseits nennen Langé et al. selbst bereits eine Reihe von Schwierigkeiten, die bei einer 'Generalisierung' von TM-Einheiten gelöst werden müßten:

- Benennungsüberschneidung: Auf eine Wortfolge der Übersetzungseinheit 'passen' mehrere Termini (z.B. '*Install the receiving antenna support.* → *receiving antenna* vs. *antenna support*')
- Terminologische Varianten: Identifikation verschiedener Instanzen einer Benennung (z.B. morphosyntaktische Varianten)
- Auswahl von ZS-Termini bei Synonymie
- Kongruenzprobleme: Angleichung von Kasus und Numerus beim Ersetzen der Platzhalter.

Zu Recht weisen Langé et al. darauf hin, daß diese Schwierigkeiten für jede Art der Termextraktion und -erkennung typisch sind und daher eigentlich sowieso in den entsprechenden Komponenten rechnergestützter Übersetzungshilfen gelöst werden müßten [LGD97:49]. M.E. liegt das entscheidende Problem der von Langé et al. vorgeschlagenen 'Generalisierung' jedoch vielmehr in den vielfältigen Unterschieden der Oberflächenrepräsentationen in AS und ZS. Zur Verdeutlichung mögen die folgenden Sätze dienen, die mögliche deutsche Übersetzungen von (1) und (2) darstellen:

(1a) *Überprüfung der Installation* fortsetzen.
⇒ *X* fortsetzen.

(1b) Setzen Sie die *Überprüfung der Installation* fort.
⇒ Setzen Sie *X* fort.

(1c) *Überprüfen* Sie als nächstes die *Installation*.
⇒ ??

(1d) *Überprüfung des Leitungssystems fortsetzen.*
⇒ *X* fortsetzen.

(2a) Restliche *benutzerdefinierte Einstellungen festlegen*.
⇒ Restliche *X* (??)

(2b) *Legen* Sie die restlichen *benutzerdefinierten Einstellungen fest*.
⇒ ??

(2c) *Legen* Sie als nächstes die *benutzerdefinierten Einstellungen fest*.
⇒ ??

Erstens können AS/ZS-Einheiten natürlich nur dann in der von Langé et al. vorgeschlagenen Form abstrahiert werden, wenn sich Termini auf 'AS- und ZS-Seite' gleichermaßen durch Variablen ersetzen lassen. Andererseits können „[s]prachliche Repräsentationen eines Begriffes […] von Sprache zu Sprache zwischen Fachwendung, Mehrwortbenennung und einfacher Benennung variieren" [Schm96:200]. Wird also z.B. ein in der AS durch ein einfaches Substantiv oder durch eine Mehrwortbenennung repräsentierter Begriff in der ZS durch eine Verbalphrase wiedergegeben (siehe (1c), (2a-c)), scheidet eine Generalisierung der Übersetzungseinheit spätestens dann aus, wenn diskontinuierliche Strukturen verwendet werden.

Zweitens müssen Strategien für den Umgang mit Polysemien und Homonymien gefunden werden. Dies gilt sowohl für Termini (vgl. die Wiedergabe von *installation checking* in (1a-c) vs. (1d)) als auch für 'Nicht-Termini' (vgl. die Wiedergabe von *proceed* (1a,b,d) vs. (1c) und (2a,b) vs. (2c)).

2.2 Explizite terminologische Informationen zur Erkennung von Satzfragmenten

Der Aufsatz von Langé et al. enthält einen weiteren Vorschlag zur Nutzung expliziter terminologischer Informationen, der sich einer für die (computerlinguistische) Weiterentwicklung von TM-Systemen m.E. weitaus wichtigeren Frage zuwendet, nämlich der Aufgabe des Retrievals von Übersetzungseinheiten unterhalb der Satzebene. Hierzu müssen zumindest zwei Probleme gelöst werden:

- die Identifikation von ‘Satzfragmenten’ in AS und ZS mit Hilfe eines geeigneten Segmentierungsverfahrens
- die Alignierung der identifizierten Fragmente, wobei es sich bei diesen Fragmenten um Einheiten verschiedenster Strukturebenen (Teilsätze, Phrasen unterschiedlicher Komplexität) handeln kann.

Für die Alignierung auf Satzebene werden häufig sog. ‘Anker’ verwendet, d.h. Zeichenketten, die auf der ‘AS-Seite’ und der ‘ZS-Seite’ eines Satzpaares identisch oder sehr ähnlich sind. Hierzu zählen u.a. Datumsangaben, Zahlen, Eigennamen sowie die sogenannten ‘cognates’, d.h. AS- und ZS-Wörter, „[that] share ‘obvious’ phonological or orthographic and semantic properties, with the result that they are likely to be used as mutual translations“ [SFI92:71]. In TM-Systemen scheint es naheliegend, zusätzlich auch die in den Terminologiedatenbanken der Systeme verfügbaren Termini als ‘Anker’ einzusetzen. Eben dies schlagen Langé et al. für die Alignierung von Satzfragmenten vor, wobei sie für die vorausgehende Segmentierung der Sätze jedoch ein sehr einfaches Verfahren vorsehen, das sich auf einige wenige Heuristiken beschränkt:

> “we envisage that it *[d.h. ein aus Phrasen/Satzfragmenten bestehendes TM; U.R.]* will be triggered only in simple cases, for example when the splitting of a sentence into two bricks is made easier by the presence of an unambiguous marker such as a conjunction, or punctuation marks” [LGD97:43].

Daß diese auf einfachen Oberflächenmarkierungen beruhende Segmentierung nicht ausreicht, um aus Parallelkorpora Einheiten unterhalb der Satzebene zu extrahieren, mag das folgende Beispiel verdeutlichen, das der deutschsprachigen Leistungsbeschreibung eines Mobilfunksystems und seiner englischen Übersetzung entnommen wurde:

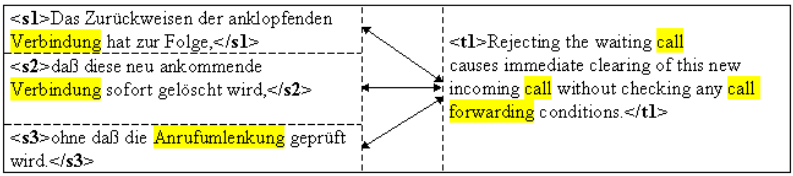
AS: Das Zurückweisen der anklopfenden Verbindung hat zur Folge, daß diese neu ankommende Verbindung sofort gelöscht wird, ohne daß die Anrufumlenkung geprüft wird.

ZS: Rejecting the waiting call causes immediate clearing of this new incoming call without checking any call forwarding conditions.

In der Terminologiedatenbank seien die folgenden Termini enthalten, die für den Alignierungsprozeß als ‘Anker’ zur Verfügung stehen:

Anklopfen	⇔	call waiting
Anrufumlenkung	⇔	call forwarding
Verbindung	⇔	call; connection

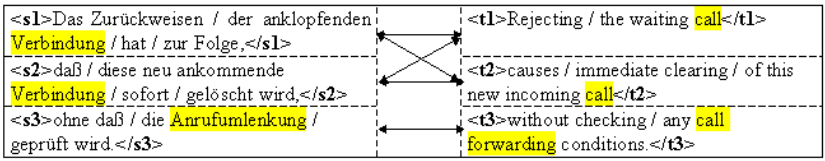
Bei Sprachen mit einem vergleichsweise geringen Anteil an eindeutigen Oberflächenmarkierungen führt das von Langé et al. vorgeschlagene Verfahren wohl eher selten zu einer für den Übersetzer brauchbaren Alignierung von Satzfragmenten. In dem angeführten Beispiel erhält man folgendes Resultat:



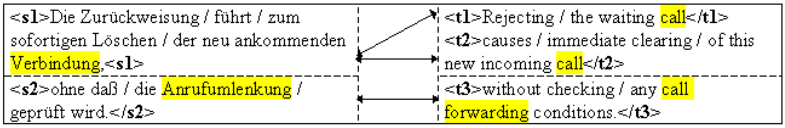
Für die Satzsegmentierung in TM-Systemen bedarf es also eines Verfahrens, das in der Lage ist, auch solche Teilsatzgrenzen zu erkennen, die nicht explizit an der Satzoberfläche markiert sind. Ein Beispiel für einen entsprechenden Formalismus ist das bereits in den 80er Jahren als Segmentierungsmodul für das MÜ-System SUSY entwickelte zweistufige regelbasierte Parser-Konzept PHRASEG [Schm86]. Aufbauend auf den Ergebnissen einer vorangegangenen morphosyntaktischen Analyse sowie einer Wortartendisambiguierung werden auf der ersten Stufe jene Wortklassen zu Einheiten verbunden, zwischen denen keine Teilsatzgrenzen auftreten können. Auf der zweiten Stufe werden diese sogenannten 'Phrasings' dann zu teilsatzwertigen Einheiten zusammengefaßt. Das Verfahren wurde für die Sprachen Deutsch, Englisch und Französisch implementiert und kann ohne weiteres auf andere Sprachen ausgeweitet werden. Eine detaillierte Beschreibung enthält [Schm86]. Dort wird auch darauf hingewiesen, daß „das Verfahren zusammen mit einer morphologischen Analyse und der Wortklassenvereindeutigung als eigenständige syntaktische Analyse innerhalb eines Systems zur computergestützten Übersetzung verwendet werden [kann]" [Schm86:156].[3]

Die folgende Tabelle zeigt, wie das Satzpaar des Beispiels unter Anwendung der PHRASEG-Regeln segmentiert würde. Die Segmentierung des deutschen Satzes in drei Teilsätze bleibt dabei aufgrund der vorhandenen Oberflächenmarkierungen unverändert. Der englische Satz wird jedoch ebenfalls in drei Teilsätze zerlegt. Das Beispiel macht aber auch deutlich, daß die verfügbaren Benen-

nungen aus der Terminologiedatenbank für eine korrekte Zuordnung der Satzfragmente nicht unbedingt ausreichend sind.



Das Beispiel verdeutlicht ferner, daß sich AS- und ZS-Fragmente nicht immer im Verhältnis 1:1 entsprechen. So können z.B. einander zuzuordnende AS- und ZS-Fragmente der Stufe 1 (‘Phrasings’; durch ‘/ ’ getrennt) Teil unterschiedlicher Fragmente der Stufe 2 (‘Teilsätze’) sein (z.B. die Verbalphrasen *hat zur Folge* und *causes* in <s1> bzw. <t2>). Die einander zugeordneten Teilsätze $<s_i>$ und $<t_j>$ dürfen also nicht ohne weiteres als Übersetzungseinheiten verstanden werden. Darüber hinaus dürften häufig auch stufenübergreifende Zuordnungen zwischen ‘Phrasing’-Ebene und Teilsatzebene vorkommen. So entspricht in der folgenden Abwandlung des bisherigen Beispiels das AS-Fragment *die Zurückweisung* dem ZS-Teilsatz <t1>. Hier scheint eine 1:2-Alignierung von <s1> mit <t1> und <t2> sinnvoll.



3 Optimierung von TM-Systemen durch bessere Nutzung impliziter terminologischer Informationen

TM-Systeme könnten die in den vorhandenen maschinenlesbaren Texten und ihren Übersetzungen verfügbaren impliziten terminologischen Informationen auf mindestens zweierlei Weise nutzen. Beispielsweise ließen sich die Texte als Basis zur automatisierten Gewinnung zweisprachiger Terminologie verwenden, um den Bestand der Terminologiedatenbank zu erweitern, oder sie könnten der Extraktion zweisprachiger ‘Anker’ für die Alignierung von TM-Einheiten unterhalb der Satzebene dienen und so die Retrieval-Leistung der TM-Komponente verbessern.

3.1 Implizite terminologische Informationen zur Erweiterung der Terminologiedatenbank

Die übersetzungsvorbereitende Terminologiearbeit, d.h. das 'Füllen' einer Terminologiedatenbank mit der für eine qualitativ hochwertige Übersetzung notwendigen Terminologie, stellt eine der zeitaufwendigsten Arbeitsschritte des Übersetzungsprozesses dar. Angesichts knapper Zeitvorgaben und fehlender linguistischer Datenverarbeitungsverfahren bzw. unzureichendem Wissen über den sinnvollen Einsatz bereits verfügbarer Verfahren erhalten Übersetzer oftmals bestenfalls Wortlisten mit fragmentarischen Informationen, die keinesfalls dazu dienen können, die terminologische Konsistenz des ZS-Textes zu sichern.

Andererseits ist gerade in den letzten Jahren ein wachsendes Maß an angewandten Forschungsarbeiten zu verzeichnen, die sich mit der rechnergestützten Extraktion von Terminologie oder der automatisierten Erstellung von Wörterbüchern aus maschinenlesbaren Textkorpora beschäftigen und verschiedene Werkzeuge und Methoden zur Vereinfachung dieser arbeitsintensiven Prozesse hervorgebracht haben.

Verfahren zur Terminologieextraktion lassen sich grob nach statistischen, linguistischen und hybriden Ansätzen unterteilen [Drou97]. Nach der Anzahl der beteiligten Sprachen können monolinguale und bilinguale Ansätze unterschieden werden.[4]

Bilinguale Extraktionsverfahren basieren auf Parallelkorpora und verwenden entweder ein probabilistisches 'Übersetzungsmodell' (rein statistische Verfahren; siehe hierzu auch [BPPM93]) oder berechnen Assoziationen zwischen potentiellen AS-Termini und ihren Entsprechungen 'auf der ZS-Seite' des Korpus, wobei die AS-Termkandidaten zunächst anhand linguistischer Muster identifiziert werden (hybride Verfahren). Dabei werden zuvor i.d.R. die beiden Seiten des Parallelkorpus auf Satzebene aligniert, um dann anhand von Kookkurrenzwerten in den alignierten Einheiten die wahrscheinlichsten Übersetzungskandidaten zu ermitteln (vgl. z.B. [Dai94]).[5]

Betrachtet man die Leistungsfähigkeit der Extraktionsverfahren, so scheinen die Ergebnisse einiger statistischer Verfahren im Vergleich zu linguistischen Ansätzen einen sehr viel höheren Anteil an Noise zu enthalten (vgl. z.B. [HJKH96:148]).[6] Ferner können Werkzeuge, die ausschließlich statistische Methoden verwenden, i.d.R. keine Mehrwortbenennungen extrahieren (vgl. z.B. [Rapp96] und [Bro97]). Andererseits sind rein linguistische Verfahren nicht nur

per Definition sprachabhängig, sondern auch nicht ohne weiteres in der Lage, einfache, ausschließlich aus Stammwörtern bestehende Benennungen zu identifizieren [LBBL96]. Darüber hinaus enthalten auch die Resultate linguistischer Extraktionsverfahren z.T. einen relativ hohen Noise-Anteil [Pea98], was vermutlich auf eine zu starke Verallgemeinerung der Wortbildungsmuster fachsprachlicher Benennungen zurückgeführt werden kann. So verifiziert Pearsons Arbeit z.B. die Hypothese, daß fachsprachliche Wortbildungsmuster innerhalb einer Sprache von Fachgebiet zu Fachgebiet sowie zwischen verschiedenen Kommunikationsebenen variieren können.

Als zentrale Anwendungsfelder für Software zur Termextraktion nennen [LBBL96] die Bereiche Übersetzen, Terminologiearbeit und Dokumentenmanagement und weisen darauf hin, daß diese drei Bereiche sehr unterschiedliche Benutzeranforderungen aufweisen. Starke Unterschiede in den Anforderungen an eine Termextraktionssoftware bestehen jedoch bereits innerhalb des für diesen Aufsatz relevanten Bereichs des Übersetzens. Die Anwendungsmöglichkeiten reichen hier von der automatischen Generierung zweisprachiger Wörterbücher für beispielbasierte MÜ-Systeme bis zur rechnergestützten Erstellung von Terminologiesammlungen und projektspezifischen Glossaren für die computergestützte Humanübersetzung. Während die für die Humanübersetzung nötige Genauigkeit der Glossare i.d.R. eine umfangreiche manuelle Aufbereitung der maschinell gewonnenen Ergebnisse erfordert, sind für die Generierung von Systemwörterbüchern zur Unterstützung des Alignierungsprozesses in beispielbasierten MÜ-Systemen sehr viel niedrigere Precision-Werte akzeptabel (vgl. [Bro97]).

Insgesamt ist offensichtlich, daß die eindeutige Identifikation von Termini in maschinenlesbaren Korpora und das Erstellen begriffsorientierter Terminologiedatenbanken Aufgaben sind, ”that must, in all cases, be carried out by humans during the last stages” [LBBL96:294]. Hierfür nimmt man sich in der Übersetzungsbranche aber leider nur selten Zeit. Immerhin könnte die in Parallelkorpora eingebettete Terminologie jedoch zumindest zur Unterstützung des Retrievals in TM-Systemen genutzt werden. Im folgenden soll daher untersucht werden, ob sich ein einfaches statistisches Verfahren zur Extraktion von Übersetzungskandidaten [Rapp96] zur Gewinnung von ‘Ankern’ und somit zur Unterstützung der Alignierung von Satzfragmenten eignet.[7]

3.2 Implizite terminologische Informationen zur Erkennung von Satzfragmenten

Rapps Verfahren zur Extraktion von Übersetzungskandidaten setzt ein auf Satzebene aligniertes und lemmatisiertes Parallelkorpus voraus und geht von den beiden folgenden Annahmen aus:

(4) Extrahiert man aus dem Korpus alle Satzpaare, in denen eine AS-Wortform *s* vorkommt, und bestimmt die Häufigkeit aller ZS-Wortformen in den extrahierten Sätzen, so weisen gebräuchliche Übersetzungen (nach häufigen Funktionswörtern) die größte Häufigkeit auf.

(5) Die Häufigkeiten von AS-Wörtern im AS-Teilkorpus und ihren ZS-Entsprechungen im ZS-Teilkorpus sind idealerweise annähernd identisch. M.a.W.: Der Quotient aus beiden Häufigkeiten beträgt idealerweise eins.

Diese Annahmen lassen sich in der folgenden Formel ausdrücken [Rapp96:106]:

$$a_t = \begin{cases} f_{st} \cdot f_s / f_t & \text{für } f_s \leq f_t \\ f_{st} \cdot f_t / f_s & \text{für } f_s > f_t \end{cases}$$

Die 'Wahrscheinlichkeit' a_t, mit der ein Wort *t* eine Übersetzung des Wortes *s* ist – oder in Rapps Worten die 'Aktivität' von *t* – hängt ab von der Häufigkeit f_{st}, mit der *t* in den alignierten Satzpaaren gemeinsam mit *s* vorkommt, sowie von der Relation zwischen den Korpushäufigkeiten f_s und f_t der Wörter *s* bzw. *t*. Dabei führt die Berücksichtigung von Hypothese b) dazu, daß die ersten Positionen der Rangliste der Übersetzungskandidaten nicht von häufigen Funktionswörtern besetzt werden.

In einem Experiment habe ich das von Rapp beschriebene 'Aktivitätsmaß' auf ein kleines deutsch-englisches Parallelkorpus angewandt, das aus Texten der technischen Leistungsbeschreibung eines Mobilfunksystems besteht. Jede 'Seite' des Korpus umfaßt ca. 10.000 Wörter. Dieser Wert mag unter korpuslinguistischen Gesichtspunkten äußerst niedrig erscheinen, andererseits ist dieser Umfang für kleinere Übersetzungsprojekte durchaus realistisch. Das Korpus wurde in einem vorbereitenden teilautomatischen Arbeitsschritt mit einem kommerziellen Alignment-Werkzeug aligniert. Für die Lemmatisierung wurde MPRO (s.o.)

eingesetzt. Die Häufigkeitswerte wurden mit Hilfe von WORD BASIC-Makros ermittelt.

Die beiden Beispiele in Tabelle 1 und 2 verdeutlichen, daß dieses einfache statistische Verfahren keine Mehrwortbenennungen extrahieren kann. Die Identifikation von ZS-Mehrwortbenennungen ist insbesondere dann schwierig, wenn die einzelnen Komponenten vergleichsweise häufig auch separat oder in anderen Mehrwortbenennungen auftreten (Tabelle 1).

Deutsche Lemmata (s) mit Korpushäufigkeit (f_s)	Rang	Englischer Übersetzungs-kandidat (t)	Häufigkeit von t in den alignierten Sätzen (f_{st})	Korpushäufigkeit von t (f_t)	'Aktivität' von t (a_t)
Kennungs-anforderung 6	1	failure	3	5	2,50
	2	cause	3	8	2,25
	3	correlation	1	6	1,00
	4	previous	2	12	1,00
	5	identity	5	34	0,88
	6	recovery	1	7	0,86
	7	request	5	36	0,83
	8	include	1	8	0,75
	9	interworking	1	4	0,67
	10	identify	1	12	0,50

Tab. 1: Kennungsanforderung ⇔ identity request

Das Beispiel in Tabelle 2 zeigt jedoch auch, daß sich bei ZS-Mehrwortbenennungen zumindest dann brauchbare 'Anker' für die Alignierung von Satzfragmenten ergeben können, wenn die Bestandteile der Mehrwortbenennung nur selten separat oder in anderen Mehrwortbenennungen vorkommen. In dem in Tabelle 2 dargestellten Beispiel könnte die deutsche Benennung *Korrelationstabelle* zusammen mit dem an erster Stelle rangierenden Bestandteil der englischen Benennung *correlation table* als 'Anker' verwendet werden.

Deutsche Lemmata (s) mit Korpushäufigkeit (f_s)	Rang	Englischer Übersetzungs-kandidat (t)	Häufigkeit von t in den alignierten Sätzen (f_{st})	Korpushäufigkeit von t (f_t)	'Aktivität' von t (a_t)
Korrelationstabelle 5	1	correlation	3	6	2,50
	2	table	3	8	1,88
	3	access	2	10	1,00
	4	contact	1	5	1,00
	5	TMSI	8	48	0,83
	6	acknowledgement	1	4	0,80

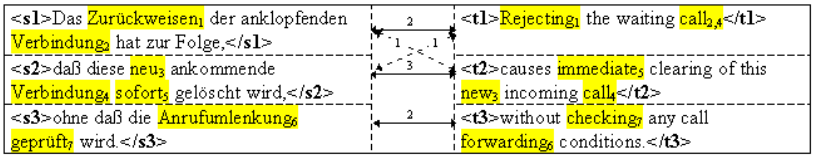
Tab. 2: Korrelationstabelle ⇔ correlation table.

Daß einfache statistische Verfahren zur Extraktion von Übersetzungskandidaten durchaus geeignet sein könnten, um 'Anker' für die Alignierung von Satzfragmenten zu gewinnen, mag ein weiterer Blick auf das bereits mehrfach angeführte deutsch-englische Mobilfunkbeispiel belegen. Tabelle 3 enthält für alle Lemmata des deutschen Satzes die mit Hilfe des zuvor beschriebenen Verfahrens extrahierten Übersetzungskandidaten.

Deutsche Lemmata (s) mit Korpushäufigkeit (f_s)		Englischer Übersetzungskandidat (t)	Häufigkeit von t in den alignierten Sätzen (f_{st})	Korpushäufigkeit von t (f_t)	'Aktivität' von t (a_t)
zurückweisen	9	reject	8	15	4,80
anklopfend	20	accept	9	18	8,10
Verbindung	186	call	225	375	111,60
Folge	4	clearing	2	4	2,00
neu	45	new	40	47	38,30
ankommend	16	incoming	15	21	11,43
sofort	5	immediate	3	8	1,88
löschen	17	cancel	9	13	6,88
Anrufumlenkung	17	forward	15	25	10,20
prüfen	9	check	8	12	6,00

Tab. 3: 'Aktivitäten' für alle 'Rang-1-Übersetzungskandidaten' des Beispiels

Die Tabelle zeigt, daß sich für unser Beispiel in der Mehrzahl der Fälle brauchbare 'Anker' ergeben (grau hinterlegt). Für die übrigen Lemmata konnten keine korrekten Übersetzungen ermittelt werden. Die zusätzlichen 'Anker' führen in unserem Beispiel zu einer eindeutigen Zuordnung der Teilsätze.



4 Einbindung in die bestehenden Arbeitsabläufe

Die vorherigen Abschnitte sollten verdeutlichen, daß das in TM-Systemen verfügbare explizite und implizite terminologische Wissen genutzt werden kann, um Einheiten unterhalb der Satzebene zu alignieren und somit die Performanz von TMs zu erhöhen. Die Nutzung expliziten terminologischen Wissens scheint zwar naheliegend, jedoch dürften die in der Terminologiedatenbank verfügbaren Be-

nennungen nicht immer ausreichen, um brauchbare (eindeutige) Alignierungen zu erhalten. Ein weiterer Schritt könnte daher darin bestehen, Verfahren zur Extraktion von Übersetzungskandidaten in TM-Systeme und Alignierungssoftware zu integrieren, um zusätzliche 'Anker' für den Alignierungsprozeß zu gewinnen. Hier sind jedoch detaillierte Untersuchungen notwendig, um einen Ansatz zu finden, der akzeptable Ergebnisse liefert, ohne allzu zeitaufwendig zu sein.

Darüber hinaus stellt sich die Frage, wie die angesprochenen Verfahren zur Alignierung von Teilsätzen in die Arbeitsabläufe von TM-Systemen eingebettet werden können. Hier sind mindestens drei verschiedene Teilprozesse zu unterscheiden:

(a) vor der eigentlichen Übersetzungsphase:

– die Aufbereitung von bereits übersetztem Textmaterial

(b) während des Übersetzens:

– Aufbereitung des zu übersetzenden Textmaterials

– Aufbereitung von neuen Übersetzungseinheiten, die dem TM hinzugefügt werden sollen.

Muß vor Beginn der eigentlichen Übersetzung zunächst ein TM aus vorhandenen Texten und deren Übersetzungen aufgebaut werden, so müssen die Algorithmen zur Zuordnung von AS- und ZS-Satzfragmenten natürlich in die entsprechenden Alignment-Werkzeuge integriert sein, wobei jedoch Sätze und Satzfragmente in separaten Arbeitsschritten aligniert werden sollten. Auf diese Weise kann das Ergebnis der automatischen Satzzuordnung bei Bedarf zunächst manuell korrigiert werden, bevor die Verarbeitung der Satzfragmente beginnt. Je nach Textumfang und Verarbeitungsmethode kann die Alignierung von Einheiten unterhalb der Satzebene einen hohen Aufwand an Rechnerzeit erfordern. Sofern jedoch das Ergebnis ohne weitere manuelle Korrekturen in das TM übernommen wird, sind längere Rechnerzeiten akzeptabel, da die Alignierung der Fragmente vor der eigentlichen Übersetzungsphase stattfindet und Übersetzer in ihrer eigentlichen Tätigkeit somit nicht behindert werden.

Führt während des Übersetzens die Suche in einem TM nicht zu brauchbaren Ergebnissen (d.h. keine Treffer oder zu geringe Ähnlichkeit zwischen Treffern und Suchanfrage), so könnte der Anwender eine weitere Suche in einem aus Satzfragmenten bestehenden TM durchführen. Eine solche Suche setzt natürlich die morphologische und syntaktische Analyse des zu übersetzenden AS-Materi-

als voraus. Dies könnte entweder vor Beginn der eigentlichen Übersetzung erfolgen, so daß der gesamte AS-Text in einem Arbeitsschritt analysiert wird, oder die linguistischen Analysen werden in die Retrieval-Phase eingebunden, so daß nur jene Sätze des AS-Textes analysiert werden, für die der Anwender eine Suche nach Satzfragmenten auslöst.

Werden in der eigentlichen Übersetzungsphase ZS-Vorschläge aus dem TM modifiziert oder AS-Sätze vollständig neu übersetzt, so werden dem TM neue Übersetzungseinheiten hinzugefügt. Für diese neuen Einheiten muß unmittelbar vor dem Abspeichern natürlich zunächst die Satzsegmentierung sowie die Alignierung der erkannten Fragmente durchgeführt werden. Die hierfür benötigte Verarbeitungszeit dürfte in einem akzeptablen Bereich liegen, da jeweils nur ein AS/ZS-Segmentpaar verarbeitet wird.

Literatur

[Ahm94] Ahmad, K. (1994): Language Engineering and the Processing of Specialist Terminology. In: *Proceedings of Language Engineering Convention*, 6–7 July 1994, CNIT-La Défense. Paris. Edinburgh: European Network in Language and Speech (ELSNET).

[Ahm/Rog92] Ahmad, K./Rogers, M. (1992): Terminology Management: A Corpus-Based Approach. In: *Translating and the Computer 14. Quality Standards and the Implementation of Technology in Translation*. London: Aslib, 33–44.

[BPPM93] Brown, P./Della Pietra, St./Della Pietra, V./Mercer, R. (1993): The Mathematics of Statistical Machine Translation: Parameter Estimation. In: *Computational Linguistics*, 19(2), 263–311.

[Bro97] Brown, R. (1997): Automated Dictionary Extraction for 'Knowledge-Free' Example-Based Translation. In: *MT Yesterday, Today, and Tomorrow. Proceedings of TMI-97*, Santa Fe, July 23-25, 1997, 169–174.

[Car/Schm98] Carl, M./Schmidt-Wigger, A. (1998): Shallow Post Morphological Processing with KURD. In: *Proceedings of NeMLaP-98*. Sydney, Januar 1998.

[Dag/Chu94] Dagan, I./Church, K. (1994): Termight: Identifying and Translating Technical Terminology. In: *Proceedings of COLING 1994*. The 15th International Conference on Computational Linguistics, August 1994. Kyoto, Japan. 34–40.

[Dag/Chu97] Dagan, I./Church, K. (1997): Termight: Coordinating Humans and Machines in Bilingual Terminology Acquisition. In: *Machine Translation*, 12, 89–107.

[Dai94] Daille, B. (1994): *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. Paris: Université de Paris VII, [Dissertation].

[Drou97] Drouin, P. (1997): Une méthodologie d'identification automatique des syntagmes terminologiques: l'apport de la description du non-terme. In: *META*, 42(1), 45–54.

[Gro98] Groß, B. (1998): *Vergleichende Untersuchung von Alignment-Tools*. Saarbrücken: Fachrichtung 8.6, Universität des Saarlandes (Saarbrücker Studien zu Sprachdatenverarbeitung und Übersetzen, herausgegeben von Karl-Heinz Freigang und Uwe Reinke, Band 15).

[HJKH96] Heid, U./Jauss, S./Krüger, K./Hohmann, A. (1996): Term extraction with standard tools for corpus exploration: Experience from German. In: Galinski, Ch./Schmitz, K.-D. (Hrsg.): *TKE '96: Terminology and Knowledge Engineering. Proceedings of the 4th International Congress on Terminology and Knowledge Engineering, 26–28 August 1996, Vienna*. Frankfurt/M.: INDEKS, 139–150.

[LGD97] Langé, M./Gaussier, É./Daille, B. (1997): Bricks and Skeletons: Some Ideas for the Near Future of MAHT. In: *Machine Translation*, 12, 39–51.

[LBBL96] L'Homme, M.-C./Benali, L./Bertrand, C./Lauduique, P. (1996): "Definition of an Evaluation Grid for term extraction software". In: *Terminology*, 3(2), 291-312.

[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.

[Mack/Han96] Macklovitch, E./Hannan, M.-L. (1996): Line 'em up: Advances in Alignment Technology and Their Impact on Translation Support Tools. In: *Expanding MT Horizons. Proceedings of the Second Conference for Machine Translation in the Americas. 2-5 October, 1996*. Montreal, Canada. Washington DC: Association for Machine Translation in the Americas (AMTA), 145–156.

[Pea98] Pearson, J. (1998): *Terms in Context*. Amsterdam, Philadelphia: Benjamins.

[Rapp96] Rapp, R. (1996): *Die Berechnung von Assoziationen: Ein korpuslinguistischer Ansatz*. Hildesheim: Olms.

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

[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, Bad 8)

[Schm86] Schmitz, K.-D. (1986): *Automatische Segmentierung natürlichsprachiger Sätze*. Hildesheim: Olms.

[Schm96] Schmitz, K.-D. (1996): Verwaltung sprachlicher Einheiten in Terminologieverwaltungssystemen. In: Lauer, A./Gerzymisch-Arbogast, H./Haller, J./Steiner, E.: *Übersetzungswissenschaft im Umbruch. Festschrift für Wolfram Wilss zum 70. Geburtstag*. Tübingen: Narr, 197–207.

[SFI92] Simard, M./Foster, G./Isabelle, P. (1992): Using Cognates to Align Sentences in Bilingual Corpora. In: *Proceedings of TMI-92*, Centre d' innovation en technologies de l' information, Montreal, Kanada, 67–81.

ANMERKUNGEN

[1] Unter einem Parallelkorpus soll im folgenden ein *n*-sprachiges Korpus verstanden werden, das entsprechend der Anzahl der beteiligten Sprachen aus *n* Teilkorpora besteht. Dabei stellen die Texte eines der Teilkorpora Ausgangstexte für die Übersetzung in die übrigen *n-1* Sprachen dar (Teilkorpora 2, 3, …, *n*). Zur Problematik des Begriffs 'Parallelkorpus' siehe auch [Pea98:47f.].

[2] Vergleichsweise geringe Längenveränderungen treten bei lexikalischen Ersetzungen nur dann auf, wenn das ersetzende Lexem im Vergleich zum ersetzten Lexem aus einer höheren oder geringeren Anzahl von Wörtern besteht.

[3] Leider ist PHRASEG jedoch nicht mehr implementiert. Andererseits war PHRASEG eines der wenigen sprachdatenverarbeitenden Programme seiner Zeit, das auf einer möglichst strikten Trennung von Algorithmus und linguistischen Regeln beruhte, so daß es mir sinnvoll scheint, das in [Schm86] dokumentierte Regelwerk wiederzuver-

wenden. Dabei nutze ich für die vorausgehenden morphologischen Analysen das am Institut für Angewandte Informationswissenschaft (IAI) in Saarbrücken entwickelte Werkzeug MPRO [Maas96]. MPRO läßt sich weitgehend auf das morphologische Modul des MÜ-Systems SUSY zurückführen und umfaßt Komponenten für die Sprachen Deutsch, Englisch und Französisch. Für die Wortartendisambiguierung sowie für die eigentliche Implementierung der PHRASEG-Regeln wird der ebenfalls am IAI entwickelte Formalismus KURD verwendet, der morphosyntaktisch analysierte Sätze mit Hilfe 'flacher Operationen' (*shallow processing*) weiterverarbeitet [Car/Schm98]. Mögliche Einsatzbereiche von KURD sind z.B. Wortformendisambiguierung, Syntaxprüfung, Stilprüfung oder partielles Parsing. Deutsche und englische Disambiguierungsregeln wurden mir z.T. vom IAI zur Verfügung gestellt, so daß sich meine eigenen Bemühungen im wesentlichen auf die Umsetzung der PHRASEG-Regeln konzentrieren.

[4] Einen knappen Überblick über monolinguale und bilinguale Verfahren zur Extraktion von Termkandidaten enthält [Dag/Chu97]. Detaillierte Darstellungen verschiedener statistischer Maße, die in Verfahren zur Terminologieextraktion eingesetzt werden, finden sich in [Dai94].

[5] Es wird allgemein davon ausgegangen, daß

> "[t]he problems of sentence-alignment, if not entirely resolved, are fairly well understood" [Mack/Han96:147].

Tests von Alignment-Werkzeugen zeigen jedoch, daß die derzeit eingesetzten Verfahren zur Alignierung auf Satzebene bei der Erkennung komplexerer Zuordnungen häufig nicht fehlerfrei arbeiten [Gro98]. Solche komplexeren Zuordnungen sind z.B. Kontraktionen (n:1-Entsprechungen), Expansionen (1:n-Entsprechungen), Auslassungen (1:0-Entsprechungen) oder Hinzufügungen (0:1-Entsprechungen). Eine Extraktionsmethode, die auf den Ergebnissen eines wortbasierten Alignierungsverfahrens beruht, wird in [Dag/Chu94] und [Dag/Chu97] beschrieben.

[6] Heid hat die Ergebnisse seines linguistischen Ansatzes den Resultaten des in [Ahm/Rog1992] und [Ahm94] beschriebenen statistischen Verfahrens gegenübergestellt, das im wesentlichen auf der Hypothese beruht, daß fachsprachliche Benennungen in Fachtextkorpora häufiger auftreten als in einem 'repräsentativen' gemeinsprachlichen Korpus.

[7] Das in [Brow97] beschriebene Verfahren zur Wörterbuchgenerierung dient im übrigen ähnlichen Zwecken. Das aus einem Parallelkorpus gewonnene Wörterbuch unterstützt die Alignierung von Satzfragmenten in einem beispielbasierten MÜ-System und kann daher nicht mit den Systemwörterbüchern 'traditioneller' (regelbasierter) MÜ-Systeme verglichen werden.

Development of a Multilingual Information Retrieval and Check System Based on Database Semantics *

Kiyong Lee (KOREA U), Suk-Jin Chang (SEOUL N. U),
Yun-Pyo Hong (DANKUK U), Key-Sun Choi (KAIST),
Minhaeng Lee (YONSEI U), Jae Sung Lee (ETRI, KOREA),
Jungha Hong (KOREA U), Juho Lee (KAIST),
Junsik Hong (YONSEI U)

1 Introduction

Technical documents for multilateral agreements or international business transactions are normally produced in a bilingual or multilingual form. Being mostly of legal nature, these documents require especially accurate and speedy translations by expert translators. In order to aid these experts, automatic ways of checking translation results (such as a spelling checker) would be highly desirable.

This paper describes the MIRAC system for Multilingual Information Retrieval And Checking. It is designed to find translation errors in multilingual documents, and to evaluate the overall results of translation. Unlike a machine translation or a translation memory system [Volk98, Webb98], the primary function of the MIRAC system is to evaluate previously translated and aligned documents in source and target languages, while dynamically building a database that consists of aligned multilingual texts carrying semantically equivalent content.

MIRAC consists of two components: one is a lexical evaluation module and the other is a semantic evaluation module, based on Hausser's Database Semantics [Hausser99].[1] Instead of aiming at the metric evaluation of machine translation systems, MIRAC directly evaluates translated documents by checking first the consistency of use of lexical terms and then semantic equivalences between source and target documents.

The paper is organized as follows: a brief introduction of Termight, a workbench for technical translators, in section 2; an introductory overview of the MIRAC system, with a description of each of its parts, in section 3; a report on its implementation and experiment in section 4; and concluding remarks in the final section 5.

2 Related Works: Termight

Dagan and Church's Termight [Dagan/Church94] is a workbench for technical translators. It mainly checks the correctness of translated technical terminology. The process is semi-automatic, for it requires a manual listing of technical terms in an original text before their corresponding translations are automatically searched from the translated text.

For listing technical terms, Termight analyzes a document for part-of-speech tagging and finds compound nouns. Out of these compound nouns, technical terms are identified and edited appropriately under a suitable environment provided by the system. Termight's alignment program then automatically locates their corresponding translations, while correct translations are selected manually to build a translation glossary. Here, the workbench helps to find correct translation pairs.

3 Overall Structure of the MIRAC System

The overall architecture of the MIRAC system is shown below.

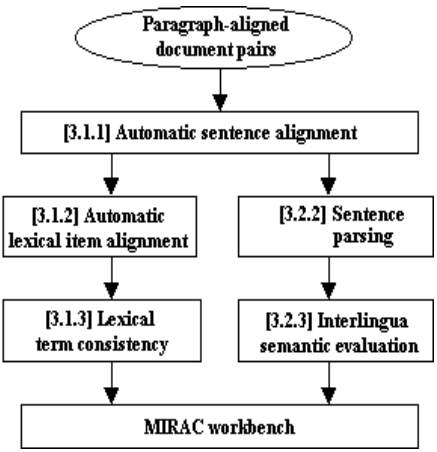


Figure 1: Structure of the MIRAC system.

The MIRAC system deals with multilingual documents, specifically comparing a pair of documents in a source and a target language. As its input, MIRAC takes in paragraph-aligned document pairs in these two languages [Klee/Park97]. Then these pairs of documents are aligned at both the sentential and lexical levels by an automatic alignment program. The use consistency of technical or key terms is also automatically checked by another module. These processes are carried out by a statistical method as pre-processing steps for the evaluation of correct translation [Collier/Ono/ Hira98].

Translations are evaluated in two steps. The first step evaluates the lexical correspondence between pairs of the aligned documents, displaying the results of evaluation in the alignment workbench. The system checks the correctness and consistency of the use of translated terms in the target language. The second step checks the semantic correspondence by a statistical method between the corresponding pairs of terms, phrases, and sentences, again displaying the results of evaluation on the alignment workbench.

3.1 Lexical Evaluation

Lexical items in each pair of aligned sentences are all aligned automatically by a statistical method [Jslee/Kang/Jhlee/Le/Choi97]. The module for lexical evaluation then analyzes them and displays the two lists of original and translated terms on its workbench. The correctness of translation should, however, be checked manually.

3.1.1 Automatic Sentence Alignment

For lexical item alignment, sentences must be aligned first. For this, the Gale/ Church method [Gale/Church91] is used to measure the length of each sentence for statistical calculation. This method, however, needs to be improved by providing ways of using information from dictionaries and also from the feedback of evaluation processes.

For our experiment, the introductory chapter of Negroponte's (1995) *Being Digital*, both in its original English and in its Korean translation, was analyzed. Each sentence and paragraph in the chapter was marked for the experiment. Both the English and the Korean versions were found to contain 20 paragraphs each. The English version contains 81 sentences, but the Korean version contains 86.

The accuracy of alignment is 97.47%. In this short experiment, the statistical method was found to be very fast and efficient but produced a far less satisfactory result on alignment accuracy. For its improvement, the additional use of dictionary information should be helpful [Collier/Ono/Hira98].

3.1.2 Automatic Lexical Item Alignment

Lexical items are statistically aligned on the basis of co-occurrence information [Hull98, Jhlee99]. The overall process is shown in Figure 2.

The intermediate steps are as follows:

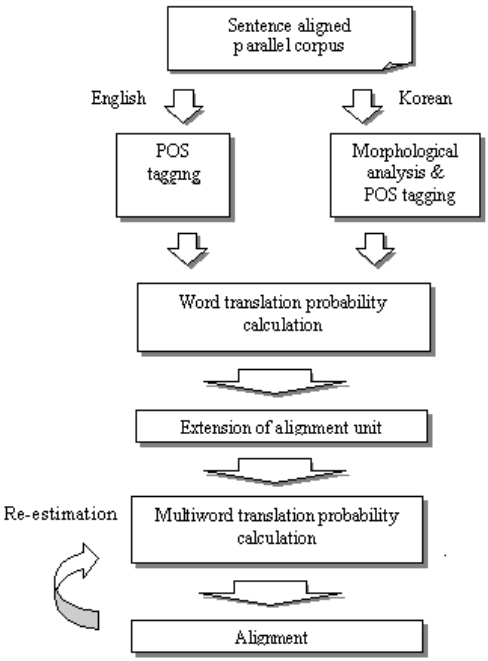


Figure 2: Overall flow of automatic alignment.

(i) For easy alignment, content words are extracted from each language document. In English, content words are nouns, verbs, or adjectives. In Korean, however, only nouns and noun-derived verbs or adjectives are treated as content words because pure verbs and adjectives are rarely used as technical terms.

(ii) The probability of word translation is calculated on the basis of information on bilingual co-occurrence. The basic assumption is that the word translation probability is higher if a word and its translation occur more frequently in aligned sentence pairs. The calculation of translation probability or similarity between two words is usually based on their respective meaning and information as well as their Dice coefficients that provide co-occurrence information.

In this experiment, Dice coefficients are used to calculate word translation probability. The translation probability $C_p(E_i, K_j)$ for an English word E_i and its corresponding Korean word K_j, for instance, is defined as follows:

$$C_p(E_i, K_j) = \frac{2C(E_i, K_j)}{C(E_i) + C(K_j)}$$

$C(E_i)$: the number of segments which contain E_i

$C(E_i, K_j)$: the number of pair segments which contain E_i and K_j in each segments

(iii) This method of calculating word translation probabilities can be extended to the calculation of multiword translation probabilities by including neighboring content words among their alignment units. It is assumed here that sequences of neighboring words can be formed into multi-content words. For an easy implementation, the following four cases are considered here: 1:1, 1:2, 2:1, and 2:2 types of word correspondence.

Each of the 1:2 and 2:1 cases can be extended only if all of the following conditions are satisfied:

$$C_p(E_i, K_j K_{j+1}) \geq \frac{C_p(E_i, K_j) + C_p(E_i, K_{j+1})}{2}$$

$$C_p(E_i E_{i+1}, K_j) \geq \frac{C_p(E_i, K_j) + C_p(E_{i+1}, K_j)}{2}$$

or

$$C_p(E_i, K_j K_{j+1}) \geq \max\{C_p(E_i, K_j), C_p(E_i, K_{j+1})\}$$

$$C_p(E_i E_{i+1}, K_j) \geq \max\{C_p(E_i, K_j), C_p(E_{i+1}, K_j)\}$$

The 2:2 case is a little more complicated. It can be extended only if the following condition is satisfied:

$$
\begin{aligned}
C_p(E_iE_{i+1}, K_jK_{j+1}) &\geq \frac{C_p(E_i, K_j) + C_p(E_{i+1}, K_{j+1})}{2} \text{ and} \\
&\geq \frac{C_p(E_i, K_{j+1}) + C_p(E_{i+1}, K_j)}{2} \text{ and} \\
&\geq \frac{C_p(E_iE_{i+1}, K_j) + C_p(E_iE_{i+1}, K_{j+1})}{2} \text{ and} \\
&\geq \frac{C_p(E_i, K_jK_{j+1}) + C_p(E_{i+1}, K_jK_{j+1})}{2}
\end{aligned}
$$

or

$$
\begin{aligned}
C_p(E_iE_{i+1}, K_jK_{j+1}) \geq \max\{&C_p(E_i, K_j), C_p(E_i, K_{j+1}), C_p(E_{i+1}, K_j), C_p(E_{i+1}, K_{j+1}), \\
&C_p(E_iE_{i+1}, K_j), C_p(E_iE_{i+1}, K_{j+1}), C_p(E_i, K_jK_{j+1}), C_p(E_{i+1}, K_jK_{j+1})\}
\end{aligned}
$$

A distance limit should be imposed to exclude meaningless multi-words that form parts of a content word but are separated by too great a distance.

3.1.3 Lexical Term Consistency Checking

For the accuracy and quality of translation, the consistent use of terms should be checked, especially in technical documents. This is especially so in the case of technical terms and proper nouns; otherwise only confusion will arise.

For example, the term “computer” is usually translated to “khem.phyu.the”, but it can be translated to “khom.phyu.the”, “cen.ca.kyey.san.ki”, “cen.san.ki”, and so on.[2] Someone familiar with the concept of a computer may think they are all the same, but others may not, for “cen.ca.kyey.san.ki” normally refers to a calculator. Another example is the name of a university in Chonju, Korea. It has a unique Korean name that has been translated or, more accurately speaking, romanized into Jeonbug, Chonpuk, or Chonbuk National University, causing great confusion.

In order to evaluate lexical consistency, we need to list lexical items and their corresponding translations. The MIRAC workbench first extracts them from a pair of documents, producing an aligned lexical list. It then analyzes the list to examine the consistency of their use. The results of these analyses can be used to build a translation dictionary for further use in checking the accuracy of translation.

3.2 Semantic Evaluation

The semantic evaluation module of MIRAC requires the parsing of each pair of aligned sentences from both source and target languages. Being implemented within the Malaga system, it analyzes each sentence left-associatively, yielding its result in an attribute-value matrix (AVM) form. These AVMs contain semantic information that constitutes an interlingua (IL), thus allowing the bi-directional translation of one language to another. The semantic evaluation of each pair of source and target sentences is then carried on by comparing their semantic information in the interlingua.

3.2.1 A Theoretical Basis

For semantic evaluation, MIRAC adopts Hausser's Database Semantics [Hausser99].[3] It allows the representation of propositional content and other related semantic information in an abstract interlingua, thus making it possible to evaluate both the consistency of the TL-formulations and their adequacy vis à vis the propositional content.

The technical basis of this evaluation is the transition counters characteristic of database semantics. In current systems, transition counters indicate which navigation through the propositional content is the most recent and which navigations are the most frequent. The purpose of the counters is to ensure that the autonomous navigation underlying conceptualization in language production proceeds without splits and loops.

It is conceivable, however, to employ counters in other applications as well. For example, in order to model the learning of new fashionable formulations, additional counters would be implemented at the level of natural language.

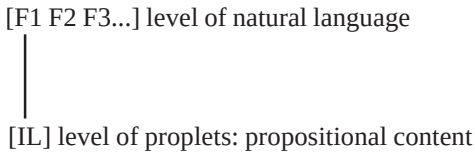


Figure 3: Matching of formulations in NL and proplet levels.

Here alternative formulations, F1, F2, F3, etc., for the same propositional content have values for their frequency in interpretation and production. Based on these frequency values, the speaker could choose a common or a special formulation depending on the utterance situation.

In IL-based MT, the above schema is extended as follows.

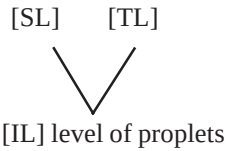


Figure 4: Convergence of SL and TL formulations at the level of proplets.

This schema suggests another possible use of counters: they mark not only alternative formulations of the SL and the TL for frequency relative to corresponding IL propositions, but also their correlation to each other.

The characteristic technical environment of database semantics is especially suited for an efficient implementation of counters. Furthermore, database semantics is special because it treats (i) the IL formally as an unordered set of AVMs and (ii) the interpretation and production procedures alike on the common basis of a time-linear navigation. These structural properties are ideally suited for storing the information specific to translation memory or database.

3.2.2 Sentence Parsing

Although both source and target languages have their own distinct systems, these systems have the same structure with the same theoretical basis, namely Hausser's Left-Associative Grammar [Hausser99], and are all implemented in the same programming language, Malaga. As for Korean, for instance, Lee [Klee99a] implemented its morphological analyzer Komor and Hong/Lee [Hong/ Klee99] its syntactic parser.

3.2.3 Interlingua Semantic Evaluation

In order to allow a bidirectional translation from one language to another and also to evaluate its correspondence, sentences are mapped into the interlingua format of sets of proplets. These are created in the process of morphological and syntactic parsing.[4]

Each pair of parsed sentences in the source and target languages with their semantic content is now checked with the evaluation module for semantic equivalence or identity. The following example shows how the semantic content of sentence (1) is represented in Interlingua.

(1) The European Council shall provide the Union with the necessary impetus for its development and shall define the general political guidelines.

```
[ IL: [ Ref_Ind: ⟨ <"Euro_Council", 1 >, < "Union", 2 >, < "impetus", 3 >,
                   < "development", 4 >, < "guideline", 5 > ⟩
        List:    ⟨ [ Rel: "provide", Arg1: < 1 >, Arg2: < 2 >, Arg3: < 3 >],
          [Rel:    "necessary_for", Arg1: < 3 >, Arg2: < 4 >],
          [Rel:    "general", Arg1: < 5 >],
          [Rel:    "political", Arg1: < 5 >],
          [Rel:    "define", Arg1: < 1 >, Arg2: < 5 >],
           [Rel:   "precedes", Ut: t1, Et: t2] ⟩ ] ]
```

The attribute IL takes as value two complex features, *Ref_Ind* and *List*. The first feature consists of an attribute *Ref-Ind* and its value that lists all of the key terms occurring in sentence (1). The second feature, on the other hand, simply consists of a list of proplets, each representing basic propositional content conveyed by the sentence. In each proplet, a relation *Rel* takes more than one argument *Arg*, while each *Arg* is related through an index number to a key term listed in *Ref_Ind*. The last proplet states that the utterance time *t1* precedes the event time *t2*, thus referring to an event occurring in the future.

Assuming that we have obtained a similar, if not the same, matrix representation for a Korean translation of sentence (1), the evaluation module checks and gives an evaluation point for each of the following items:[5]

(2) Evaluation Items and Scores

items	*scores*
·Proposition-Relation	40(35)
·Reference-Indices	30(25)
·Modification	20(10)
·Tense	10(10)
	100 (80) → **very good**

4 Implementation and Experiment

4.1 Experiment of Lexical Evaluation

The English-Korean parallel corpus, consisting of a 750-page volume on Uruguay Round multilateral agreements, was used for our experiments. The corpus is aligned in segment units in the preprocessing step. Each segment is mostly composed of one single sentence. Some statistical facts about the parallel corpus are given below:

items	English	Korean
segments	4,968	4,968
words (phrases)	139,265	79,290
average length of segments	28.03	15.96
content words	65,844	65,653
unique content words	2,681	3,847

Table 1: Statistics for the parallel corpus.

This table shows that the number of Korean words or word groups occurring in the corpus is smaller than that of English words because compound words are used more frequently in Korean. The average number of words occurring in each

of the English sentences is 28.03, while the average number of words occurring in each of the Korean sentences is 15.96.

The experiment was performed in several steps. Each language document from the parallel corpus is POS-tagged by a language tagger. The tagging process filters the content words. For tagging English, [Brill94]'s method was used. For Korean, an English HMM (Hidden Markov Model) tagger was modified to process Korean sentences [Shin/Han/ Park/Choi95].

The translation probability of content words is calculated on the basis of bilingual co-occurrence information. By extending it to their neighboring words, the translation probability of multi-words is then calculated. This probability is used to align the multi-words and then is recalculated by counting the aligned multi-words.

This process is normally repeated seven times. In order to find meaningful multi-words, positional information is also used. The following graphs (Figure 5) show the change of the number of extracted unique translation pairs and the percentage of each type of correspondence at the last alignment. We can see that they both show a similar trend: the number of translation pairs decreases rapidly at the first re-estimation but decreases very slowly from the second re-estimation and finally remains constant after the fourth re-estimation.

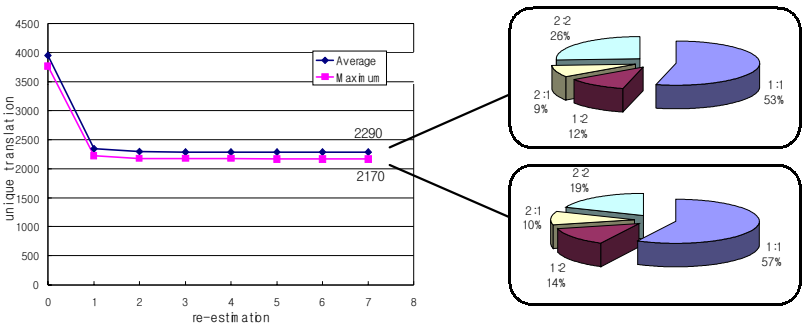


Figure 5: Change in the number of extracted unique translation pairs.

After the fourth re-estimation, there is also no change in translation pairs, while there is a small change in the translation probability. We assume that both of the results converge at the seventh iteration. After the seventh re-estimation, the ratio of the types of corresponding pairs in each case is similar. The number of correspondences of type 1:1 is the largest, type 2:2 the second largest, type 1:2 the next, and type 2:1 the last.

For the measurement of alignment accuracy, two methods were used: the calculation average method for distance limits and the maximum value method. In this experiment, the accuracy was calculated for 100 randomly selected translation pairs by using both methods. They both produced similar test results. The accuracy of alignment at each re-estimation is shown in Figure 6.

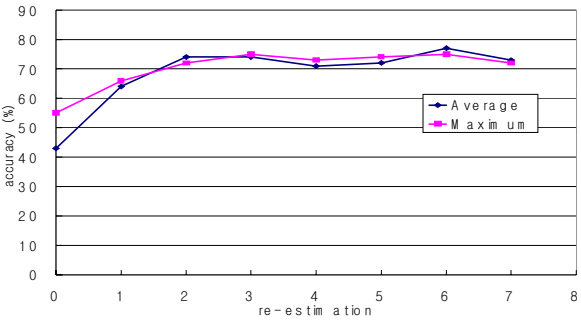


Figure 6: Accuracy of alignment.

The accuracy increases at the beginning of iteration, but stops changing after a certain point. The two curves in Figure 4 show almost the same results. The calculation average method, for instance, produced 2,290 unique translation pairs at the convergence points (from the fifth iteration to the seventh) with an average accuracy of 74%.

Some examples of the extracted translation pairs are shown in Table 2.

	English	Korean
(3)	convention	hyep.yak
(4)	countermeasure	tay.ung co.chi
(5)	working party	cak.ep.pan
(6)	result (of) negotiation(s)	hyep.sang keyl.kwa
(7)	result (of) negotiation(s)	hyep.sang(.uy) keyl.kwa

Table 2: Examples of extracted translation pairs.

The experiment was performed on the Uruguay Round (UR) documents on the world economy and diplomatic affairs. Since these documents deal with problems in a very specialized domain, no translation dictionary of general use provides appropriate translation words. A case in point is the pair "convention" : "hyep.yak" shown in (3). The English word "convention" generally means "cip.hoy" (meeting), "kwan.lyey" (traditional case), "sa.hoy.cek. kwan.swup" (social custom) in Korean. In these documents, "convention" is aligned to the diplomatic term "hyep.yak".

Here, various types of alignment were found. The alignment "convention" : "heyp.yak" in (3) is of type 1:1, while the alignment "countermeasure" : "tay.ung co.chi" in (4) is of type 1:2. The alignment "working party" : "cak.ep.pan" in (5), on the other hand, is of type 2:1. One word or phrase may have two different alignments, too: for example, the phrase "result (of) negotiation(s)" is aligned to "hyep.sang kyel.kwa" in (6) or to "hyep.sang.uy kyel.kwa" in (7). But these two are the same if only content parts are taken into account, for the particle "uy" in Korean is a function word meaning "of". This shows that our method of extending multi-words is effective for finding content multi-words in various lexical forms.[6]

Figure 7 (following page) shows a screen shot of the MIRAC workbench. If a user selects a word in a source language, its translation candidates and their translation probabilities are shown in the workbench. When the user selects one of the translation candidates, some translation examples appear in the main screen. The screen shows the word "committee" being translated to "wi.wen.hoy", displaying its translation candidates and a list of translation examples.

Figure 7: Screenshot of the MIRAC system.

4.2 Experiment of Semantic Evaluation

Here, we randomly selected five English sentences from the UR documents and translated them into two different languages, Korean and German. Then these translations were compared to check how their semantic content was preserved in each of the translations. Table 3 shows the results of the comparison.

	TL : Korean		TL : German	
	Scores	Violations	Scores	Violations
Sentence 1	75	Prop-Rel, Ref-Ind	100	
Sentence 2	85	Prop-Rel	80	Ref-Ind, Mod
Sentence 3	100		100	
Sentence 4	75	Prop_Rel, Ref-Ind	70	Ref-Ind, Mod
Sentence 5	90	Ref-Ind	100	
Average	85		90	

Table 3: Results of semantic evaluation (Korean and German).

The average scores in Table 3 show that the German translation scored higher than the Korean translation. While the Korean translation often failed to capture Prop(osition)-Rel(ation) and Ref(erence)-Ind(ices), the German translation failed to capture Ref-Ind and Mod(ification) in sentences 2 and 4. These results indicated that the translation between typologically similar SL and TL, like English and German, was easier than the translation between typologically dissimilar languages like English and Korean. The latter case even failed to capture such basic relations like Prop-relations.

5 Concluding Remarks

Machine translation is a formidable task. The task of evaluating the results of translation, however, is more tractable. The MIRAC system shows such a possibility by demonstrating how multilingual texts can be systematically aligned for checking the consistency of the use of lexical terms as well as the semantic equivalences between source and target languages.

The MIRAC system evaluates the quality of previously translated documents aligned in source and target languages, while continuously updating a database consisting of such aligned multilingual texts. It thus closely resembles a translation memory system. Nevertheless, its main function is to systematically evaluate the accuracy of translations at both the lexical and the propositional level. An evaluation tool like the MIRAC system is not only useful, but also necessary for building an adequate translation memory or storage system as well as an efficiently running machine translation system. When combined into one coherent system, these three systems of evaluation, memory, and translation can become a constantly or dynamically upgrading integrated machine translation system.

Especially when source and target languages differ from each other structurally, the evaluation of semantic equivalence plays an important role. This, for instance, should be the case, when a non-western language like Korean or Chinese is translated into English or vice versa. Being based on Hausser's Database Semantics [Hausser99], the MIRAC system can adequately represent the semantic content of sentences in both source and target languages in terms of abstract proplets and check their semantic equivalence. Contents, stored in the MIRAC system, can be recycled to evaluate both the consistency of a TL-formulation and its adequacy relative to the propositional content.

Since it is based on Database Semantics, the MIRAC system can also be implemented to be part of a machine translation system. For it can reproduce acceptable sentences in a target language by navigating through a word bank or an arrayed field of proplets that has been built of a source language and then by selecting appropriate sequences of words or proplets. In further research, Database Semantics may thus be extended into an approach to machine translation where translation memory serves not only as an aid to human translation, but as an important component of automatic translation.

References

[Beutel97] Beutel, Bjoern (1997): 'Malaga 4.3', Abteilung Computerlinguistik, Universität Erlangen-Nürnberg, http://www.linguistik.uni-erlangen.de/Malaga.de.html.

[Brill94] Brill, Eric (1994): 'Some Advances in Transformation-Based Parts of Speech Tagging', *Proceedings of the 12th National Conference of Artiificial Intelligence*, 722–727.

[Chang98] Chang, Suk-Jin (1998): 'Translation: Mapping and Evaluation', [written in Korean], *Language and Information 2.1*, 1–41.

[Collier/Ono/Hira98] Collier, Nigel, Kenji Ono, and Hideki Hirakawa (1998): 'An Experiment in Hybrid Dictionary and Statistical Sentence Alignment', *Proceedings of the 17th International Conference on Computational Linguistics*, 268–274.

[Cop/Flick/Mal/Riehe/Sag96] Copestake, Ann, Dan Flickinger, Robert Malouf, Susanne Riehemann, and Ivan A. Sag (1996): 'Translation using Minimal Recursion Semantics', *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven.

[Dagan/Church94] Dagan, Ido and Ken Church (1994): 'Termight: Identifying and Translation Technical Terminology', *Proceedings of the 4th Conference on Applied Natural Language Processing*, 34–40.

[Gale/Church91] Gale, William A. and Kenneth W. Church (1991): 'A Program for Aligning Sentences in Bilingual Corpora', *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 177–184.

[Hausser99] Hausser, Roland (1999): *Foundations of Computational Linguistics: Man-Machine Communication in Natural Language*, Berlin: Springer-Verlag.

[Hong/Klee99] Hong, Jungha and Kiyong Lee (1999): 'Processing Korean Relative Adnominal Clauses', [written in Korean], *Proceedings of the Eleventh Conference on Korean Characters and Korean Information Processing*, 265–271.

[Hull98] Hull, David A. (1998): 'A Practical Approach to Terminology Alignment', *Proceedings of the First Workshop on Computational Terminology*, 1–7.

[Jhlee99] Lee, Juho (1999): *Extraction of English-Korean Compound Noun Translation through Automatic Method*, [written in Korean], Master's thesis, KAIST.

[Jslee/Kang/Jhlee/Le/Choi97] Lee, Jae Sung, Jung-Gu Kang, Juho Lee, Hung Le, Key-Sun Choi (1997): 'Design and Implementation of Alignment Workbench', [written in Korean], *Proceedings of the Ninth Conference on Korean Characters and Korean Information Processing*, 430–435.

[Klee/Park97] Lee, Kiyong and Chongwon Park (1997): 'On Designing a Term Match Checking System for Bilingual Documents', [written in Korean], *Proceedings of the 1997 Spring Conference on Cognitive Science*, 220–229.

[Klee99a] Lee, Kiyong (1999a): *Computational Morphology*, [written in Korean], Seoul: Korea University Press.

[Klee99b] Lee, Kiyong (1999b): 'A Basis of Database Semantics: from Feature Structures to Tables', [written in Korean], *Proceedings of the Eleventh Conference on Korean Characters and Korean Information Processing*, 297–303.

[Lee/Jee/Chung98] Lee, Minhaeng, Kwangsin Jee, So W. Chung (1998): 'A Research on Test Suites for Translation Systems', [written in Korean], *Language and Information* 2.2, 185–220.

[Shin/Han/Park/Choi95] Shin, Jung H., Young S. Han, Young C. Park, and Key S. Choi (1995): 'A HMM Part-of-Speech Tagger for Korean with Word-Phrasal Relations', *Proceedings of Recent Advances in Natural Language Processing*, 439–449.

[Volk98] Volk, Martin (1998): 'The Automatic Translation of Idioms, Machine Translation vs. Translation Memory Systems', in: Nico Weber(ed.), *Machine Translation: Theory, Applications, and Evaluation. An Assessment of the State of Art*, St. Augustin: Gardez Verlag.

[Webb98] Webb, Lynn E. (1998): 'Advantages and Disadvantages of Translation Memory: A Cost/Benefit Analysis', MA Thesis, Monterey Institute of International Studies.

ENDNOTES

* This paper was supported by a NON-DIRECTED RESEARCH FUND, Korea Research Foundation, 1996. This is the final report on a three-year (1996-1999) joint collaborative research project (PI: Prof. Kiyong Lee, Korea University) with overseas consultant Prof. Roland R. Hausser and his research staff at the University of Erlangen-Nürnberg, Germany. We are grateful to these organizations for their financial and technical support. We are also grateful to Chongwon Park, Dr. Kunsik Lee, Dr. Kung-Un Choi, Dr. Si-Jong Ryu, and Koaunghi Un, who participated in the project as research associates in the past two or three years. We also owe our gratitude to the anonymous referees for their constructive comments, Prof. Uta Seewald-Heeg and

Rita Nübel for their critical review and editorial assistance, and finally Mr. Gary Rector for his professional styling and proofreading.

[1] This is an extension of Hausser's Left-Associative Grammar [Hausser99], implemented in Beutel's C-like programming language called Malaga [Beutel97], which accommodates LAG with attributes.

[2] In this paper, Hangul is romanized using the Yale system.

[3] Lee [Klee99c] proposed a slightly different version of Database Semantics by adopting an object-oriented relational model, for it can easily convert AVMs for natural language into table forms and allows the use of SQL for developing a natural language query system.

[4] Here we try to adopt Copestake et al.'s representation schema [Cop/Flick/Mal/ Riehe/ Sag96], which is introduced in their Minimal Recursion Semantics.

[5] A more detailed scheme of evaluation for the MIRAC system is presented in [Chang98] and [Lee/Jee/Chung98].

[6] Just as inflectional endings or prepositions rarely carry any content in English, nominal particles like "uy" carry practically no content in an agglutinative language like Korean or Japanese.

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:

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

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

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)).

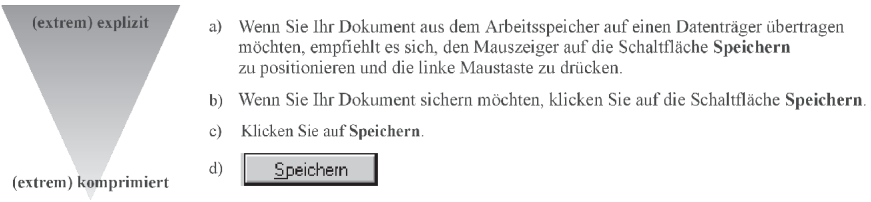


Abb. 1: Oberflächenveränderung durch Implikation/Explikation - der Inhalt (die Nachricht) bleibt unverändert.

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 Mobilvermittlungsstellenbereich gueltig.	Eine Sondernummer ist im gesamten Mobilvermittlungsstellenbereich 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 identifizieren mobil teil nehmen MSC senden	international mobil funk geraet kennen MSC senden mobil teil nehmen identifizieren	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.

Ausblick

Uta Seewald-Heeg (Hochschule Anhalt)
Rita Nübel (IAI Saarbrücken)

1 Evaluierung von Translation-Memory-Systemen

Eine Bewertung der Leistungsfähigkeit von Translation Memories bzw. Satzarchiv-Modulen vollautomatischer Übersetzungssysteme stellt unterschiedliche Anforderungen an das Evaluierungsszenario, das Datenmaterial und die Durchführung der Evaluierung. Voraussetzungen hierzu wurden in der abschließenden Diskussion des vom Arbeitskreis „Maschinelle Übersetzung“ veranstalteten Workshops auf der Basis der Beiträge formuliert, die in diesem Band zusammengestellt wurden, um einzelne Schritte des Evaluierungsvorhabens festzulegen.

2 Evaluationsdesign

Ausgangspunkt der Diskussion um Qualitätsanforderungen und mögliche Testprozeduren bildete das von der EAGLES-Arbeitsgruppe „Assessment and evaluation“ [EAGLES96] entworfene Standardverfahren zur Evaluierung von NLP-Software, das sich an der ISO-Norm 9126 orientiert und im wesentlichen als dreistufiges Verfahren konzipiert ist.[1] Die dort formulierten Qualitätskriterien für natürlichsprachige Systeme umfassen Funktionalität (*functionality*), Zuverlässigkeit (*reliability*), Brauchbarkeit (*usability*), Leistungsfähigkeit (*efficiency*), Wartbarkeit (*maintainability*) sowie Portierbarkeit (*portability*) des jeweils untersuchten Softwareproduktes.

In einem eigens für Translation-Memory-Systeme entworfenen Evaluationsdesign wird darüber hinaus zwischen sogenannten *on-line*-Funktionen und *off-line*-Funktionen von TMs unterschieden. Zu den *off-line*-Funktionen werden die Art der Textanalyse, die Möglichkeiten des Imports und Exports von Text in die TM-Datenbank, die Segmentierung der Texteinheiten bei der Übersetzung und beim Alignment, d.h. dem Import bereits übersetzer Texte und deren quellsprachigen Dokumenten in synchronisierter Form, sowie die für das Alignment selbst zur Verfügung stehenden Funktionen gezählt. Als *on-line*-Funktionen werden demgegenüber Parameter wie die maximale Größe des TM, die Retrieval-Geschwindigkeit bei der Suche im Übersetzungsarchiv sowie die Trefferquote beim Abgleich eines zu übersetzenden Satzes mit den Daten des TM bezeichnet. Dabei geht in

die Bewertung der einzelnen Funktionen stets die Zahl der jeweils benötigen Arbeitsschritte ein.

Neben Qualitätsanforderungen an Translation-Memory-Systeme wurden auch Metriken für die Tests festgelegt sowie Anforderungen an adäquates Testmaterial und organisatorische Fragen zur konkreten Durchführung der Evaluierung erörtert.

2.1 Qualitätskriterien für Translation-Memory-Systeme

Als wesentliche Gründe für den professionellen Einsatz von Translation Memories werden

- bessere Übersetzungsqualität durch konsistente Übersetzungen
- Zeiteinsparung durch die Reduktion von Mehrfachübersetzungen
- Wiederverwendbarkeit von Daten

genannt. Hinsichtlich der Wiederverwendbarkeit bereits übersetzter Texte oder Textfragmente spielt vor allem die Retrievalfunktion der Translation Memories eine entscheidende Rolle. Die für die Evaluierung zentralen Qualitätsmerkmale betreffen daher in erster Linie die *on-line*-Funktionen des EAGLES-Evaluationsdesigns.[2]

Unter den *on-line*-Funktionen soll zunächst die Bewertung der Retrievalleistung im Vordergrund der Evaluierungsaktivitäten stehen. Besonderes Augenmerk soll auf die Präzision der Treffer gelegt werden und die Verlässlichkeit der von den Systemen für Retrievals bei modifizierten Eingaben kalkulierten Trefferraten, die mittels *Fuzzy-Match*-Algorithmen ermittelt werden. Wie der Beitrag von Seewald-Heeg und Nübel illustriert, sind beispielsweise die Match-Werte, die von den Satzarchivmodulen der untersuchten maschinellen Übersetzungssysteme berechnet werden, zum einen unplausibel, wenn man die Angaben mit den tatsächlich gelieferten Retrievals vergleicht; zum anderen weichen die Angaben beim Vergleich der beiden Systeme erheblich voneinander ab, was gegen Aussagen über allgemeingültige Schwellenwerte für brauchbare bzw. nicht brauchbare Kandidaten für Übersetzungen spricht. Wenn ein Überseter beispielsweise den häufig als Schwellenwert für brauchbare Retrievals angesehenen Match-Wert von 70% festlegt, ist für ihn entscheidend, ob die beim Retrieval ermittelten Kandidaten oberhalb dieser Grenze tatsächlich ohne unverhältnismäßig hohen zusätzlichen Arbeitssaufwand (z.B. umfangreiche Posteditionsaktionen) für die Er-

stellung einer Übersetzung verwendet werden können, bzw. ob die ermittelten Kandidaten, die unterhalb dieser Grenze liegen, tatsächlich nicht oder nur mit erhöhtem Arbeitsaufwand in die Übersetzung integriert werden können.

Es ist beabsichtigt, neben der Retrievalleistung auch die Alignment-Funktionen der Systeme in die Evaluierung einzubeziehen. Darüber hinaus sollen auch die von den verschiedenen Translation-Memory-Systemen verarbeitbaren Dateiformate berücksichtigt werden, da sie bei der Entscheidung für ein bestimmtes System für zahlreiche Anwender von zentraler Bedeutung sind.

2.2 Evaluierungsmethodologie

Es wird davon ausgegangen, dass bei der Evaluierung keine vorab bereits angelegten TM-Datenbanken verwendet werden. Die Translation Memories werden erst nach und nach aufgebaut, so dass sich der Umfang des Referenzmaterials im Laufe der Evaluierung schrittweise vergrößert.

Für die Evaluierung der *Fuzzy-Match*-Funktion soll zunächst eine Klassifikation möglicher im Text auftretender Modifikationstypen erstellt werden, die in den Testdaten entsprechend reflektiert sein muss. Die Bewertung der Erkennungsleistung der verschiedenen Modifikationen beim Retrieval soll auf einer qualitativen Klassifikation der Retrievalergebnisse erfolgen, die anschließend quantitativ ausgewertet wird. Hierzu muss zunächst ein Bewertungsschema konzipiert werden, das das Maß der Ähnlichkeit zwischen Testsatz und Referenzsatz berücksichtigt.

2.3 Systeme

Die Auswahl der zu evaluierenden Systeme hängt u.a. von ihrer Verfügbarkeit an den einzelnen Evaluierungsstandorten ab, die *nicht* die Standorte der industriellen Anbieter sein werden, um größtmögliche Objektivität und identische Bedingungen für die Durchführung des Evaluierungsvorhabens zu garantieren. Neben den Translation Memories der Firmen Trados und Star, die bereits auf dem Workshop vorgestellt wurden, sollen möglichst viele der auf dem Markt verfügbaren Translation-Memory-Systeme in die Evaluierung einbezogen werden.

3 Durchführung der Evaluierung und Auswertung der Ergebnisse

Im Anschluss an die Zusammenstellung des oben beschriebenen Testkorpus sollen Teile des Korpus mittels der Alignment-Funktionen in die zu evaluierenden Systeme eingelesen werden. In den Testdurchläufen sollen die *Fuzzy-Match*-Funktionen dann jeweils mit den modifizierten Daten bzw. aktualisierten Textversionen getestet und die Retrievalergebnisse entsprechend dem oben beschriebenen Klassifikationsschema bewertet werden.

Es ist geplant, die Ergebnisse der Evaluation abschließend zusammenfassend zu dokumentieren und der Öffentlichkeit zu präsentieren.

Literatur

[**EAGLES96**] EAGLES Evaluation of Natural Language Processing Systems. Final Report. EAGLES Document EAG-EWG-PR.2, Oktober 1996.

ANMERKUNGEN

[1] Vgl. den URL http://www.issco.unige.ch/projects/ewg96/ewg96.html.

[2] Siehe auch das Protokoll der Sitzung des Arbeitskreises „Maschinelle Übersetzung“ vom 19.2.1999, das auf dem URL
http://www.heeg.de/~uta/AK-Protokoll-TM-1.html
publiziert ist.

Roland Hausser

Foundations of Computational Linguistics. Man-Machine Communication in Natural Language

Berlin Heidelberg: Springer, 1999. ISBN 3-540-66015-1

Rezensiert von Winfried Lenders, Bonn

Das vorliegende Werk von Roland Hausser, „Foundations of Computational Linguistics“, ist, auch wenn es sich streckenweise einen solchen Anstrich gibt, weit mehr als ein bloßes Lehrbuch der Computerlinguistik. Das Buch erhebt vielmehr den Anspruch, Sprachtheorie zu bieten, ja darüber hinaus Kommunikationstheorie im Sinne einer Darstellung der theoretischen Grundlagen von Mensch-Maschine-Kommunikation in natürlicher Sprache. Kommunikationstheorie auf der Grundlage von Prinzipien der Computerlinguistik darzustellen, dahinter steht ein besonderes wissenschaftstheoretisches Verständnis von Computerlinguistik: Denn dieser noch relativ jungen Disziplin wird die Aufgabe zugeschrieben, Modelle sprachlicher Kommunikation, Modelle kommunikativen Verhaltens zu entwickeln, und zwar auf dem Wege der Algorithmisierung und Formalisierung. Es geht darum, wie man auch mit Einleitung S. 4 sagen könnten, einen Computer als Kommunikator auszubilden, also ein funktionales Modell der natürlich-sprachlichen Kommunikation aufzustellen. Hausser hat für eine Theorie dieser Art eine besondere Bezeichnung geprägt. Er nennt sie SLIM-Sprachtheorie, eine Abkürzung, hinter der sich folgendes verbirgt:

1. S = *Surface compositional* (Methodologisches Prinzip)

   SLIM verwendet eine oberflächenkompositionale Syntax und Semantik. in der nur konkrete Oberflächen zusammengebaut werden dürfen, unter Verzicht auf Null-Elemente, Identitätsabbildungen oder Transformationen.

2. L = *Linear* (=Empirisches Prinzip)

   SLIM verwendet eine zeitlineare Grammatik, die die empirische Tatsache formalisiert, dass in den natürlichen Sprachen immer ein Wort nach dem anderen geäußert bzw. wahrgenommen wird.

3. I = *Internal* (Ontologisches Prinzip)

SLIM behandelt die sprachliche Interpretation als einen Sprecher-Hörer-internen kognitiven Prozess.

4. M = *Matching* (Funktionales Prinzip)

SLIM behandelt Referenz auf der Grundlage eines Abpassens (*matching*) zwischen wörtlicher Sprachbedeutung und Verwendungskontext.

(Zitiert nach der Rohfassung der noch im Erscheinen begriffenen dt. Ausgabe des Buches)

Durch diese Markierungspunkte ist der Rahmen abgesteckt für ein umfassendes Theoriegebäude, das Hausser in 24 Kapiteln, gleichmäßig verteilt auf 4 große Teile präsentiert. Das Buch als ganzes beruht auf der Idee, Sprache nicht als abstraktes Konstrukt zu behandeln, sondern als Medium, das zusammen mit den Prozessen des Wahrnehmens, des Denken und des Verstehens der Kommunikation zwischen Individuen dient. Zu Beginn eines jeden Teils und eines jeden Kapitels zeigt der Autor selbst dem Leser den roten Faden, der das Buch durchzieht: Der Weg führt von den einfachen zu den komplexen Hilfsmitteln der sprachlichen Kommunikation und von den einfachen zu den komplizierteren Methoden der formalen Beschreibung. Gleichzeitig werden zahllose in der Sprach- und Kommunikationstheorie, der Mathematik und Logik, der Informatik und Physik schon vorliegende Entwürfe zur Sprache und ihrer Beschreibung vorgestellt und besprochen.

Der erste große Teil (Kapitel 1–6) trägt die Überschrift „Theory of Language". Ausgangspunkt sind hier allgemeine Grundlagen der Kommunikation und der Darstellung sprachlicher Zeichen im und für den Computer. Es werden die grundlegenden Aspekte maschineller Sprachanalyse im Rahmen der umfassenderen Aufgabe der Mensch-Maschine-Kommunikation, die Möglichkeiten einer technologischen Umsetzung dieser Forschungen, z.B. in Information Retrieval Systemen und maschinellen Übersetzungssystemen, beschrieben. Ab Kapitel drei wird Sprachtheorie als Kommunikationstheorie dargelegt. In diesen Kapiteln wird durch die „Konstruktion" einer Roboters namens Curious (S. 51: „in terms of constructing a robot named curious") die Komponenten der sprachlichen Kommunikation in einem prototypischen Modell dargestellt, das es – ähnlich wie Winograds SHRDLU-System – mit einer vereinfachten Welt zu tun hat. Im Unterschied zu und in Erweiterung von SHRDLU geht es Curious jedoch auch um das

Wahrnehmen und Erkennen der Welt bzw. um die Art und Weise, wie die „Welt“ als „Umwelt“ oder „Kontext“ des Roboters von diesem wahrgenommen und erkannt werden kann. Der Übergang von Wahrnehmen zum Erkennen wird dabei als eine Art Mustererkennung aufgefasst, durch welche die Parameter eines in der Umgebung wahrgenommene „Objekts“ gleichsam in das Innere des „kognitiven Agenten“ transportiert werden und dort ein ‚Instantiierungs-Konzept‘, I-Konzept genannt, bilden. Dieses wird dadurch „erkannt“, dass es „matcht“ mit einem M-Konzept, das sich als „type“, vielleicht auch „prototype“, innerhalb des kognitiven Agenten befindet. Es wäre hier – wie auch später – zu fragen, wie sich die von Hausser hier eingeführten Begriffe wie I-Konzept, M-Konzept etc. zu verschiedenen älteren kognitiven Modellen verhalten, etwa zu der Theorie der Mentalen Modelle von Johnson-Laird, der ältern Frame-, Script- und Schema-Theorien etc. Abgesehen von dieser Frage gelingt es Hausser jedoch sehr gut, mit Hilfe der beinahe genialen Annahme eines einfachen Roboters die grundlegenden Handlungen beim Wahrnehmen und Erkennen, also bei kognitiven Operationen darzustellen. In den Kapiteln 4 und 5 wird der zunächst sprachlose Curious mit Sprache ausgestattet, die ihn dann – über seine Fähigkeiten als wahrnehmenden, erkennenden und handelnden Agenten hinaus – auch zu sprachlicher Kommunikation befähigt. Hierzu wird – auch unter Hinweis auf Sprachtheorien wie die von Austin, Frege, Grice und Wittgenstein – im Gesamtrahmen der SLIM-Theorie eine Referenztheorie entwickelt. Gemäß dieser Theorie wird der Referenzprozess als eine interne ‚matching procedure‘ zwischen der literalen Bedeutung, d.i. dem M-Konzept, und einem entsprechenden kontextualen Referenten, dem I-Konzept verstanden. Die ‚literale Bedeutung‘ (*literal meaning*) kommt durch Konvention zustande. Hausser unterscheidet also zwei Arten von Bedeutung; er nennt sie meaning_1 und meaning_2: meaning_1 ist die literale Bedeutung, meaning_2 die der Äußerung eines Sprechers zugeordnete Bedeutung, also wohl diejenige, die sich aus dem I-Konzept ergibt. Auch hier stellt sich die Frage, wie diese Referenztheorie, die sich leicht zu einer Verstehenstheorie erweitern ließe, im Vergleich zu anderen Konzeptionen einzuordnen ist (z.B. in Relation zu Winograds Konzept eines operationalen Verstehens und zu den entsprechenden Vorstellungen der Kognitionspsychologie). Für Hausser ist der Unterschied klar: es handelt sich um eine ‚interne‘ matching procedure, während andere Theorien den Vorgang des Referenz (und auch des Verstehens) gleichsam von aussen betrachten, als etwas, das sich zwischen Sprecher und Hörer abspielt, und nicht in ihnen (intern). Genau hier aber wäre wiederum zu prüfen, wie sich Haussers Vorstellungen zu anderen Theorien verhalten, z.B. zu Winograd, der einerseits eine ‚interne‘

Repräsentation seiner (Modell-)welt konstruiert, wenn diese auch gesetzt und nicht wahrgenommen wird, und andererseits eine interne Repräsentation der Äußerungen des Benutzers, die beide aufeinander abgebildet werden müssen, damit es zu Referenz und Verstehen kommen kann.

Die letzten beiden Kapitel dieses Teils sind im Großen und Ganzen dem Zeichencharakter der Sprache, den Zeichentypen und ihrer Interpretation im Rahmen der SLIM-Theorie und den innnertextualen Referenzen und Koreferenzen gewidmet.

Im zweiten Teil „Theory of Grammar" (Kap. 7–12) steht die Sprache als wichtigstes Kommunikationsmittel im Mittelpunkt. Es geht nicht mehr um das einzelne Zeichen, mit dem zu kommunizierende Entitäten bezeichnet werden, oder um das singuläre Ereignis selbst, das kognitiv und sprachlich verarbeitet wird, sondern um Zeichenkomplexe, um die Art und Weise, wie Zeichen zu komplexen Gebilden verknüpft werden und um die wissenschaftliche Beschreibung dieser Objekte und Vorgänge. Es geht – kurz gesagt – um die grundlegenden Methoden der formalen Beschreibung komplexer Sprachgebilde, also um Grammatiktheorie. Was aber heißt Grammatiktheorie? Um dies zu klären, geht Hausser - auf der Basis des immer noch die Diskussion beherrschenden Paradigmas der Generativen Grammatik - auf die Ursprünge des generativen Konkatenierens von Symbolen zurück und unterscheidet unter Gesichtspunkten einer formalen Sprachtheorie drei elementare Formalismen, die Kategoriale Grammatik *(C-Grammar*), die Phrasenstrukturgrammatik (*PS-Grammar*) und die Links-Assosiative Grammatik (*LA-Grammar*). Zunächst werden in den Kapiteln 7 und 8 die beiden ersten Grammatiktypen ausführlich bezüglich ihrer grundsätzlichen Eigenschaften und ihrer Mächtigkeit erörtert. Aufgrund seines sehr systematischen Ansatzes gelingt es Hausser dabei, die wichtigsten Unterschiede und Besonderheiten dieser Grammatiktypen klar zu machen. Ganz nebenbei oder vielmehr folgerichtig werden in diesem Kontext auch die für ein Lehrbuch unverzichtbaren Grundlagen des Arbeitens mit formalen Grammatiken und des Parsing (deklarativ vs. prozedural, kontextfrei vs. kontextsensitiv, top-down vs. bottom-up etc.) erklärt. Die Darstellung mündet in einer Liste von Desiderata besonders der PS-Grammatiken, ja eigentlich kommt Hausser zu einer vernichtenden Beurteilung dieses Grammatikformalismus, und zwar in allen zur Zeit aktuellen Versionen (GPSG, HPSG, LFG), die er allesamt als dem *nativism* verpflichtet sieht, also der These von der Existenz allgemeiner angeborener sprachlicher Grundmechanismen. Bei aller Prominenz, zahlreichen Revisionen, vielen Anhängern etc. sei die Entwicklung der PSG dennoch ein *textbook example* für mangelnde Konvergenz. Ein wichtiges Argument für diese

Sichtweise sei, dass in der Tat die praktischen Systeme der maschinellen Sprachverarbeitung sich entweder zu einer Theorie nur im Sinne eines Lippenbekenntnisses äußern oder auf Theorien gänzlich verzichten. Gründe dafür sind nach Hausser, dass sich auf der Basis des *nativism* keine funktionale Theorie der Kommunikation aufbauen lasse und dass PS-Grammar-Formalismen inkompatibel seien mit „the input-output conditions of the speaker-hearer." (179). Diese Desiderate würden, wie Hausser meint, durch den dritten Grammatik-Formalismus, die LA-Grammatik, aufgehoben, wie in den folgenden Kapiteln 10–12 zunächst allgemein, und dann im dritten Teil des Buches praktisch dargelegt wird.

Der Formalismus der LA-Grammatik (LAG) wurde von Hausser entwickelt und wird hier erstmals in umfassender Form vorgestellt und mit anderen Formalismen (C-Struktur, PS-Struktur) konfrontiert. Er sei „input-output equivalent with the speaker-hearer" (S. 183). Es wird damit ein Grammatiktyp vorgeschlagen, der sprachliche Äußerungen, die wir als Mensschen ‚zeit-linear' produzieren und auch ‚zeit-linear' wahrnehmen, in gleicher Weise ‚zeit-linear' abarbeitet bzw. ableitet. Dabei bleibt die Interpretation eines erkannten Elements, z.B. einer Präpositionalphrase, deren Zuordnung zu einem Verb oder zu einem Nomen, dem Referenzprozess, der in der Kommunikation zwischen Sprecher und Hörer abläuft, überlassen. Auf diese Weise erhält Hausser weit weniger komplexe Strukturen, denn es wird darauf verzichtet, alle möglichen syntaktischen Beziehungen zwischen Konstituenten und semantischen Lesarten darzustellen. Die LAG vermeidet also die in anderen Formalismen immer anzunehmenden vielfältigen Interpretationen von Sätzen; dies sei der natürlichen Kommunikation adäquater. Hauser wird damit zwar der Tatsache gerecht, dass wir uns ‚zeit-linear' sprachlich äußern und sprachliche Äußerungen ‚zeit-linear' wahrnehmen; es sei hier aber offen gelassen, ob damit auch die planerischen oder strategischen Akte, die beim Entwurf und bei der Strukturierung einer Äußerung in ihrer Gesamtheit, noch ehe sie hervorgebracht ist, am Werke sind, gerecht wird.

Der dritte Teil (Kap. 13–18), der mit „Morphology and Syntax" überschrieben ist, konkretisiert dieses methodologische Konzept der LAG in Bezug auf die grundlegenden sprachlichen Phänomene der Formenbildung und Syntax. In schöner Klarheit werden hier sowohl die grundlegenden Begriffe der Morphologie (Kap. 13) und Syntax (Kap. 16) dargelegt, als auch – unter Anwendung der LA-Grammatik – konkret Formalismen zur Erkennung von Wortformen (Kap. 14) und syntaktischen Strukturen (Kap. 17 und 18) vorgestellt. Auf dem Gebiet der Morphologie steht dabei das von Hausser entwickelte System La-MORPH im Zentrum, auf das hier nicht näher eingegangen wird, da es schon früher beschrieben und seine

Funktionstüchtigkeit in praktischen Tests nachgewiesen wurde (vgl. Hausser [Hg.]: Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994. Tübingen: Niemeyer, 1996). In der Syntax strebt Hausser eine reine Oberflächensyntax an, in der es nicht um syntaktische Funktionen oder Rollen geht, sondern nur um die Wohlgeformtheit der Sätze, die sich aus den morphosyntaktischen Merkmalen ergibt, die in den Wortformen angelegt sind. Sätze sind Konstruktionen, zwischen deren Teilen syntagmatische Beziehungen bestehen, und zwar solche der Valenz, der Kongruenz und der Stellung. Nur um diese drei „Grundprinzipien der natürlichsprachlichen Syntax" geht es. Bei Ableitung eines Satzes ist zu garantieren, dass die Valenzbedingungen der Wörter erfüllt sind, Kongruenz zwischen den Wortformen erreicht ist und nur zulässige Wortstellungen erzeugt werden. Jede Wortform ist Valenzträger und hat ‚Valenzstellen', gleichsam Eigenschaften, die auf die nachfolgende Wortform schließen lassen. Diese Eigenschaften müssen in dem Lexikon, auf das die syntaktische Beschreibung zugreift, kodiert sein oder sich aus der vorausgehenden morphologischen Analyse der Wortformen ergeben. Zu diesen ‚Eigenschaften' gehört z.B. die Angabe der zweiten Person bei „du" und „liest", die – wenn diese Wortformen im Text zeitlinear auftreten – zur Überprüfung ihrer Kongruenz dienen. Die Valenzstellen fungieren gleichsam als *slots*, die im Verlauf der zeitlinearen Analyse durch geeignete *filler* ausgefüllt werden müssen. Wie dies im LA-Formalismus umgesetzt wird, zeigt Hausser unter Verwendung algebraisch formulierter Satzmuster. Die Anwendung dieser Formalismen wird schließlich an Fragmenten der deutschen und englischen Syntax veranschaulicht. Dabei beschränkt sich Hausser auf den deklarativen Hauptsatz, und hier vor allem auf die Nominalgruppen als nominale *filler* des verbalen Valenzrahmens, jedoch werden auch diskontinuierliche Sequenzen und Distanzstellungen, z.B. bei Auxiliarkonstruktionen, betrachtet.

Wenn Hausser sich in der Syntax ganz auf Oberflächenkonkatenationen beschränkt, so tut er das aus guten Gründen, die in seiner umfassenderen SLIM-Theorie motiviert sind. Denn die Auswahl der Satzkonstruktion aus einer Vielzahl von Möglichkeiten, die Wahl eines Tempus zum Ausdruck einer bestimmten Befindlichkeit in der Zeit, die Wahl eines Modus, die Wahl gerade ‚dieser' Wortstellung und nicht einer anderen usw., dies alles sind Prozesse, die im Sprachbenutzer ablaufen. Sie finden außerhalb der Syntax in den Teilen des ‚kognitiven Agenten' statt, die Hausser Semantik und Pragmatik nennt.

Diesen Teilen widmet er sich nun im vierten Teil (Kap. 19-24) des Buches, der mit „Semantics and Pragmatics" überschrieben ist. Man könnte meinen, nun folge er der alten semiotischen Trias Syntax, Semantik, Pragmatik, nach der sich so viele

Lehrbücher der Linguistik in ihrem Aufbau richten. Weit gefehlt! Hausser geht vielmehr ganz folgerichtig seinen Weg, wie er in den ersten theoretischen Teilen angelegt ist, weiter, ein Weg, der jetzt von den komplexen sprachlichen Zeichen, den Wortformen und Sätzen, zum Bezeichneten führt, oder, um Haussers Begriffe zu benutzen, der von der Ebene I, der sprachlichen Oberfläche (*language surface*) zur Ebene II, dem semantischen Inhalt (*semantic content*) führt. In konkreten natürlichsprachlichen Systemen wird der Weg von Ebene I nach Ebene II gemeinhin als semantische Interpretation bezeichnet. In diesem Zusammenhang tritt ein Kerngedanke des Buches und seiner leitenden Theorie besonders deutlich hervor: Hausser erläutert in knapper und präziser Form, was den drei Typen semantischer Systeme, dem semantischen System der logischen Sprachen, dem der Programmiersprachen und dem der natürlichen Sprache gemeinsam ist und was sie trennt. Ohne hier zuviel zu verraten: Es wird in einfacher, doch überzeugender Form dargestellt, was semantische Interpretation in künstlichen Sprachen und in der natürlichen Sprache heißt. Dabei werden die Grundkonzepte der modelltheoretischen Semantik von Tarski verständlich erörtert und in ihrer Anwendbarkeit auf natürliche Sprachen geprüft. Indem er weiterhin erläutert, was vom Standpunkt der modelltheoretischen Semantik aus Wahrheit ist, was man in dieser Hinsicht unter Bedeutung eines Ausdrucks zu verstehen hat und in welchem Sinn in diesem Zusammenhang von Ontologie gesprochen werden kann, schlägt Hausser die Brücke zwischen logischer und linguistischer Semantik. Es versteht sich, dass dabei auch die sprachtheoretischen Konzeptionen von Frege, Carnap und anderer Autoren der analytischen Sprachforschung zur Sprache kommen. Es geht nicht um bloße Deskription der Gegebenheiten, also nicht z.B. um die vielfältigen Möglichkeiten der Interpretation eines Satzes, oder um die Möglichkeiten, ein Wort von einem anderen hinsichtlich seiner ‚Bedeutung' zu unterscheiden, sondern um den Prozess der Benutzung von Sprache in kommunikativen Zusammenhängen. Indem Hausser dies auseinanderlegt, wird auch deutlich, auf welche Weise die Computerlinguistik Begriffe wie Bedeutung, Interpretation und Verstehen klären kann.

Eine Konsequenz des modelltheoretischen Ansatzes liegt darin, dass das Wissen von Sprecher und Hörer in der Kommunikation intern repräsentiert sein muss. Hausser schlägt hierfür eine *database* vor, die gleichsam das interne Weltmodell des Sprechers und des Hörers enthält. In dieser sind propositionale Ausdrücke enthalten, von Hausser *proplets* genannte, sowie *woplets*, wortbezogene Ausdrücke, und Begriffe, die sog. *coplets*. Zwischen diesen drei ‚Entitäten' der systeminternen Datenbank und den früher schon definierten Konzepttypen M-con-

cept und I-concept bestehen Beziehungen, die ausführlich erläutert werden. Ebenso wird in einer ganzen Reihe von Modellen vorgeführt, wie sich der Autor den Ablauf der syntaktisch-semantischen Interpretation vorstellt. Auch hier wird, wie schon in anderem Zusammenhang bemerkt, leider kein Bezug zu einschlägigen Modellen der kognitiven Psychologie der 80er und 90er Jahre hergestellt, zu denen ähnliche Überlegungen führten.

Der Einbau in den kommunikativen Zusammenhang soll schließlich in Haussers Gesamtgebäude die Pragmatik leisten. Hausser nennt dies ‚pragmatische Interpretation' (S. 467). Allerdings bestehen hier noch begriffliche Unsicherheiten, denn offenbar bedeutet ‚pragmatische Interpretation' Verschiedenes, je nachdem, ob man sich im ‚hearer mode' oder im ‚speaker mode' befindet (S. 484 ff.). Es hat den Anschein, als ob Pragmatik hier ganz anders verstanden wird, als es seit Morris, Austin und Searle in der Linguistik üblich ist. Auch hier fehlen leider klärende Worte.

Was mit Haussers Buch insgesamt für die Computerlinguistik gewonnen ist, kann man vielleicht am besten verdeutlichen, wenn man einige Leitideen, die das Buch durchziehen, nochmals zusammenfassend beurteilt.

Eine erste Leitidee besteht darin, dass das Buch ein Gesamtkonzept der sprachlichen Kommunikation und ihrer Beschreibung mit algorithmischen Mitteln präsentieren will. Gegenüber anderen Ansätzen ist dabei ungewöhnlich, dass maschinelle Sprachanalyse nicht als „Spiel" angesehen wird, in welchem das analysierende System alle strukturellen und interpretatorischen Möglichkeiten einer Äußerung abzuleiten und Mehrdeutigkeiten aufzulösen hat. Vielmehr findet Sprachanalyse zwischen Kommunikationspartnern statt, von denen einer ein Mensch, der andere auch ein Computer sein mag. Hausser konstatiert, dass wir erfolgreich kommunizieren, mit Menschen und mit Computern. Daraus ist die Aufgabe der Computerlinguistik abzuleiten, nämlich zu klären, wie Sprache dabei funktioniert. Die Auflösung von Problemen, auch solchen der Sprachanalyse selbst, findet im übergeordneten – pragmatischen – Problemlösungsprozess durch den Nutzer (also den Interaktionspartner Mensch) statt, für den die Sprachanalyse die notwendigen Strukturen ermittelt. Ein solches Gesamtmodell ist begrüßenswert. Es liegt damit im Grunde die Ausformulierung einer Idee vor, die schon Ende der 70-er Jahre in Gerold Ungeheuers M-C-Modell der Computerlinguistik entwickelt wurde. Es handelte sich dabei um den Vorschlag, sprachverarbeitende künstliche Systeme nicht als autonome Systeme zu konzipieren, sondern als Systemverbünde mit den Partnern Mensch (M) und Computer (C). Natürlichsprachliche Prozesse sollten in Form problemlösender Interaktionen zwischen Computer

und Mensch stattfinden, also in Form pragmatischer Prozessen, die den Benutzer einbeziehen und für die durch die Maschine ‚nur' die strukturell aufbereiteten Daten bereitgestellt werden. Computerlinguistische Lösungen haben in diesem Modell nichts mit der Beschreibung der angeborenen Sprecher-Hörer-Mechanismen und Kompetenzen zu tun, eine These, die mit Haussers vehement vorgetragenen Argumenten gegen den von ihm so genannten Nativismus übereinstimmt.

Als weitere Leitidee des Buches ist auszumachen, dass gleichsam als Manifestation des vorgetragenen Konzepts ein ‚kognitiver Agent' modelliert wird, der mit Sprache begabt ist und sich in einer Umwelt bewegt. Die ‚Sprachkompetenz' dieses kognitiven Agenten folgt aber nicht den Ideen der ‚nativistischen' Theorien, sondern den Prinzipien der Linksassoziativen Grammatik, seine kognitive Kompetenz folgt denen der modelltheoretischen Semantik. Wenn man auch über die Wertungen streiten kann, die das Buch in Bezug auf andere Grammatiktheorien enthält, so sind doch die vorgetragenen Grundgedanken plausibel, und die Konsequenz, mit der dieses Konzept durchgehalten wird, bewundernswert.

Eine letzte Leitidee, die hier zu würdigen bleibt, wird durch den Anspruch des Buches markiert, neben der Präsentation einer computerorientierten Kommunikations- und Sprachtheorie sich auch als Lehrbuch zu eignen. Dieser Anspruch wird erfüllt. Denn es finden sich, wie schon dargestellt werden konnte, neben den theoretischen ‚Hintergründen'auch viele einfache praktische Beispiele und Erklärungen auch kompliziert scheinender Sachverhalte (z.B. die Ableitung der drei elementaren Grammatikformalismen). Der Lehrbuchcharakter des Buches wird auch durch die Übungen herausgestellt, die sich an jedem Ende der 24 Kapitel finden und die weniger dem schulischen Abprüfen des ‚gelernten' Stoffes dienen, als vielmehr zu weiterem Nachdenken anregen sollen. Allerdings ist Haussers Buch eher in die Kategorie der anspruchsvollen Lehrbücher einzuordnen. Denn es enthält sehr viele Querbezüge (z.B. zur Semiotik, zu Karl Bühler, Tarski, Frege etc.) und Hinweise auf Hintergründe, die der Anfänger erst nach ergänzender Lektüre verstehen und einordnen kann. Dies betrifft z.B. die Ursprünge der PSG in der Automatentheorie und in der Theorie formaler Sprachen bei E. Post 1936, die Hintergründe der Theorie Chomskys (vgl. S. 142) und den oben schon genannten Komplex der modelltheoretischen Semiotik mit all ihren logik- und philosophiegeschichtlichen Facetten. Der Autor spannt damit einen ‚Background' auf, der dem wirklich Interessierten hervorragende Anregungen zu weiterem Nachdenken und Nachlesen gibt.

Wenn das Buch also in vielerlei Hinsicht auch einfachere Grundlagen der Computerlinguistik verständlich darstellt, so liegt doch vor allem eine anspruchvolle Gesamtdarstellung vor, die zu Recht die Bezeichnung *foundations* führt. Es wäre freilich zu wünschen, dass der kognitive Agent, den dieses Gesamtkonzept modelliert, auch in seinen komplexeren Teilen (der semantischen und pragmatischen Interpretation) als Pilotsystem realisiert werden könnte.

Bericht über die 11. Jahrestagung der GLDV

Multilinguale Corpora: Codierung, Strukturierung, Analyse

vom 8.–10. Juli 1999 in Frankfurt

Nachdem vor zwei Jahren, vom 21.-24. Oktober 1997, mit der „6th International Conference about the Use of Computers in Historical and Comparative Linguistics“ (6. Internationale Konferenz über den Gebrauch von Computern in der Historischen und Vergleichenden Sprachwissenschaft) bereits eine arbeitskreisinterne Tagung an der Universität Frankfurt stattgefunden hatte, war es eine besondere Ehre für unser Fach, nunmehr die 11. Jahres-tagung in Frankfurt zu organisieren.

Es hatten sich etwa 90 Teilnehmer, darunter 45 Vortragende, zur Tagung angemeldet. Viele davon waren Mitglieder der GLDV, viele Studierende und Hochschulangehörige, auch aus dem Ausland (den weitesten Weg hatte Tom Lai aus Hong Kong). Es kamen aber auch Vertreter aus dem kommerziellen Bereich, wo an linguistischen, übersetzungstechnischen oder Kodierungsproblemen usw. gearbeitet wird.

Der vielbeachtete Festvortrag „The Encoding of Language Corpora: The TEI Recommendations in Principle and Practice“ wurde am Abend des Donnerstags (8.7.) von Lou Burnard vom Oxford Text Archive gehalten. Die Vorträge in den Arbeitskreisen verteilten sich während der drei Tage auf 2 große und 3 kleinere Hörsäle und gliederten sich in folgende Sektionen:

Codierung und Anordnung (3 Vorträge), Statistische Corpusanalyse (4), Texttechnologie (6), Lexikographie (6), Sprachspezifische Vorträge (5), Analyse und Generierung (5), Multi-linguale Corpora (7 Vorträge) sowie 4 Präsentationen. In einer eigens organisierten Sektion „Translation Memory Systems“ wurden in 9 Vorträgen die neuesten Entwicklungen auf dem Gebiet der automatischen Sprachübersetzung vorgestellt.

Leider hatten wir auch 4 kurzfristige Absagen, dies gab den Teilnehmern jedoch Zeit für Diskussionen und eingehendere Beschäftigung mit den Präsentationen.

Es versteht sich von selbst, daß eine solche Tagung mit den vergleichsweise geringen Mitteln, die uns zur Verfügung standen, nur schwer zu realisieren war; herzlich gedankt sei deshalb unseren Sponsoren, die uns unterstützt haben: die BHF-Bank mit einem großzügigen finanziellen Zuschuß, die Frankfurter Sparkasse, die Hessische Landesbank, und die Firmen Compaq und Misco sowie der Verlag Mouton deGruyter mit Sachspenden, wie Tagungsmappen, Blocks, Bleistiften, Kugelschreibern usw.

Dank der Kooperationsbereitschaft der Autorinnen und Autoren konnte in erstaunlich kurzer Zeit – nicht einmal ein halbes Jahr später – der Tagungsband[1] erscheinen, indem Beiträger und Herausgeber in ständigem e-mail-Kontakt standen; so konnten auch ursprünglich nicht vorgesehene Korrekturen eingearbeitet werden. Bei einem Band dieses Umfangs (knapp 400 Seiten) ist natürlich eine hundertprozentige Fehlerfreiheit unmöglich; es mag als Tribut an die erforderliche Aktualität der wissenschaftlichen Beiträge gelten.

Peter Olivier
Vergleichende Sprachwissenschaft der
Johann Wolfgang Goethe-Universität Frankfurt am Main
e-mail: olivier@em.uni-frankfurt.de

[1] Multilinguale Corpora. Codierung, Strukturierung, Analyse. 11. Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung. Hrsg. von Jost Gippert in Verbindung mit Peter Olivier. Praha: Enigma Corp., 1999. ISBN 80-86126-04-8. Preis: US$ 48,-. Siehe: http://titus.uni-frankfurt.de/curric/gldv99/papers.htm (Tagungsband) sowie http://titus.uni-frankfurt.de/curric/gldv99.htm (Index).