In: Buder, Rehfeld, Seeger (1990): Grundlagen der praktischen Information und Dokumentation. 264-274

B 10 Linguistisch-technische Aspekte der maschinellen Übersetzung

Harald H. Zimmermann

B 10.1 Einleitung

Betrachtet man die technische Entwicklung in der Informationsindustrie, kommt man nicht umhin festzustellen, dass der Computer in nahezu jeden Bereich menschlicher Tätigkeiten Einzug hält. Auf dem Gebiet der Wort- oder Textverarbeitung wird die Mensch-Maschine-Kommunikation - auf PC-Ebene oder an Workstations anstelle von Schreibmaschinen - der Standard der nahen Zukunft sein, d.h. vom Ende der achtziger Jahre an.

Auf der anderen Seite wird es wegen der Komplexität der natürlichen Sprachen nicht gelingen, das Problem der maschinellen Übersetzung natürlicher Sprache im allgemeinen - ganz zu schweigen von der Erkennung und Übersetzung gesprochener Sprache - im Sinne einer FAHQT (fully automatic high-quality translation, d.h. einer vollautomatischen hochqualitativen Übersetzung) zu lösen. So gibt es also *Grenzen* (mehr oder weniger beim wissenschaftlichen Herangehen an das Problem) und Möglichkeiten der Einbindung technischer Einrichtungen in den Prozess der (intellektuellen) Übersetzung oder gar des Textverstehens durch Maschinen. Ziel dieses Beitrags ist es, die Möglichkeiten (Stand Okt. 1989) des *praktischen Einsatzes* des Computers im Bereich der maschinellen und maschinengestützten Übersetzung zu erkunden. Wenn es auch einigermaßen verlockend wäre, *grundlegende Fragen* der Übersetzbarkeit von Texten (und des Textverstehens) zu behandeln, so wird doch das Gewicht auf der technischen Übersetzung liegen, d.h. auf der Übersetzung technischer und gemeinsprachlicher Texte.

B 10.2 Systematische Aspekte

Selbst wenn es von einigem Interesse ist, wie maschinelle Übersetzung möglich gemacht wird, so spielen doch die linguistischen Aspekte (insb. die Grammatikmodelle und die möglichen bzw. benutzten Strategien) eine untergeordnete Rolle. Man darf also annehmen, dass es eine Art „black box" gibt, in die ein Text bzw. Wörter einer natürlichen Sprache eingegeben werden und aus der mit oder ohne menschliche Unterstützung eine oder alternative Übersetzungen herauskommen. Die Übersetzung kann aus Sicht des Benutzers gut, brauchbar oder schlecht sein. Bezogen auf den Nutzen und die Nutzung der maschinellen oder maschinengestützten Übersetzung lassen sich hauptsächlich zwei Benutzergruppen unterscheiden:

(1) der so genannte *Endnutzer*, d.h. ein Experte. der sich über einen Artikel informieren möchte, der in einer ihm mehr oder weniger unbekannten natürlichen Sprache verfasst ist; eine Person, die einem Freund/einer Freundin einen Brief in einer fremden Sprache schreiben möchte;. . .und

(2) ein professioneller Benutzer, insb. ein *Übersetzer*, der den Computer bei seiner Arbeit als Werkzeug benutzen möchte.

Unter diesem Aspekt kann man Projekte und Grundlagenforschung zur maschinellen Übersetzung und zum Sprachverstehen außer acht lassen und sich auf praktische Werkzeuge bzw. Sys-

teme konzentrieren. Zu Beginn ist es wichtig, zwischen den beiden Hauptstrategien zu unterscheiden: maschineller Übersetzung (MT) und maschinengestützter Übersetzung (CAT).

- Ein maschinelles Übersetzungssystem (MT-System: Machine Translation) liegt nur dann vor, wenn der Übersetzungsprozess - ausgehend von einem maschinenlesbaren Quelltext - ohne menschliche Hilfe abläuft und zu einem Zieltext führt, dessen Qualität für Informationszwecke mindestens „gut genug" ist.

- Ein maschinengestütztes Übersetzungssystem (CAT-System: Computer Aided Translation) liegt dann vor, wenn der Mensch eingreifen muss, damit eine „gute" Übersetzung des Quelltextes (sei er maschinenlesbar oder nicht) erreicht wird.

Es ist offensichtlich, dass ein MT-System zusätzlich als *Komponente* eines CAT-Systems eingesetzt werden kann: ein Text kann vor Beginn der MT „angepasst" werden („pre-editing"), oder der von der Maschine übersetzte Text kann von Übersetzern post-ediert werden, damit eine höhere Qualität erzielt wird. Es sind bereits zahlreiche MT- und CAT-Systeme auf dem Markt, die von ihren Entwicklern jeweils als das „richtige praktische Werkzeug" dargestellt werden, und es fällt schwer, ohne präzise Kriterien eine Entscheidung zu treffen. Da es beim Einsatz von MT-Systemen wesentlich auf den Umfang der Computerwörterbücher ankommt, spielt in beiden Alternativen die Erweiterung dieser Wörterbücher (Wörterbuch-Updating) eine wichtige Rolle:

- ein Konzept sieht vor, dass - besonders im Falle sehr umfangreicher Wörterbucheinträge - der Benutzer, z.B. der Übersetzer, *nicht* selbst am Updating mitwirkt (vgl. z.B. das SYSTRAN-Konzept), so dass die Systemverwaltung Spezialisten für die Verbesserung der Lexikondaten einsetzen muss.

- der andere Weg besteht darin, es dem professionellen Benutzer zu überlassen, das Systemwörterbuch zu ergänzen oder Spezialvokabulare aufzunehmen (wie es das LOGOS-Konzept vorsieht).

Es gibt weitere Systemmerkmale, die im Entscheidungsprozess eine wichtige Rolle spielen: die Verfügbarkeit von Sprachpaaren; die Möglichkeit, besondere Texttypen (z.B. Sitzungsprotokolle, Briefe,. . .) zu verarbeiten etc.

10.3 Bewertungskriterien
10.3.1 Qualität

Obwohl man normalerweise das System als „black box" ansehen kann (oder muss), gibt es doch Unterschiede in der Qualität der „reinen" maschinellen Übersetzungsergebnisse („Rohübersetzung"). Es ist nicht leicht, ein genaues Maß anzugeben, aber es soll hier auf einige wichtige Kriterien hingewiesen werden (zu den Einzelheiten vgl. das Konzept von Van Slype (Lit. 04.) und die Beschreibung der beiden SYSTRAN-Evaluationen von Van Slype (Lit. 03.)).

Die Hauptkriterien dabei sind:

(1) Verlässlichkeit und Wiedergabetreue, d.h.: in welchem Maße (schlecht, ausreichend, gut) sind Inhalt/Bedeutung des Originals wiedergegeben.

(2) Verständlichkeit und Lesbarkeit, d.h.: in welchem Maße ist der Benutzer in der Lage, den übersetzten Text zu lesen und zu verstehen.

Dies sieht einfach genug aus, aber die Probleme stecken im Detail: So produziert ein MT-System in syntaktischer und stilistischer Hinsicht normalerweise schlechtere Übersetzungen, als es der Mensch vermag, wohingegen die Übersetzung des Systems auf lexikalischer Ebene, insb. in bezug auf die Erkennung der korrekten Fachbegriffe, sogar präziser und konsistenter sein kann als eine intellektuelle Übersetzung.

10.3.2 Anwendungsumgebung

MT oder CAT müssen in einer konkreten Anwendung im Rahmen der Einbettung in eine technische Umgebung gesehen werden. So hängt eine Entscheidung normalerweise nicht nur von der Qualität ab, sondern auch von der Möglichkeit der Integration in ein komplexes Text- oder Wortverarbeitungssystem. In dieser Hinsicht müssen die folgenden Gesichtspunkte beachtet werden:

B 10.3.2.1 Integration in (Literatur- oder Text-) Datenbanken

Heute sind Datenbanken - technisch gesehen - weltweit über Paketvermittlungsnetze und sogar Satellitenkommunikation zugänglich. Dadurch wird die Überwindung von Sprachbarrieren, z.B. der zwischen Englisch und Japanisch, aber auch und gerade auf dem *multilingualen europäischen Markt*, ein äußerst erstrebenswertes Ziel. Experimentelle Bemühungen, MT-Systeme in Informationsprozesse einzubinden, werden in Japan unternommen: als Beispiel mag die INSPEC-Datenbank dienen, die im Original in Englisch vorliegt und auf die über japanische Deskriptoren zugegriffen werden kann. Die Deskriptoren und später die (englischen) Titel werden während des Dialogs ins Japanische übersetzt (vgl. Lit. 02.). Auf ähnliche Art und Weise wird ein anderes MT-System im Batch-Betrieb (mit Postedition) für die deutsch-englische Übersetzung von Titeln aus deutschen Datenbanken eingesetzt (vgl. Lit. 05.).

Es ist offensichtlich, dass die Titel und Abstracts auch von Übersetzern übersetzt werden könnten. Es gibt jedoch Argumente, die für MT und CAT sprechen: Die zu übersetzenden Texte liegen maschinenlesbar vor, so dass eine ideale Grundlage für den Einsatz von Computern existiert; das Fachgebiet der Titel/Texte ist normalerweise „physikalisch" markiert, so dass eine Klassifikation oder sogar Thesaurusfunktionen insbesondere für den lexikalischen Transfer (Vereindeutigung von Begriffen) benutzt werden können; das verwendete Vokabular muss äußerst präzise und konsistent sein, was mit Hilfe des Computers erreicht werden kann.

B 10.3.2.2 Automatische Indexierung

Die Indexierung von (Voll-)Texten kann ein wichtiger Nebeneffekt der Verwendung von MT oder CAT sein. Für den lexikalischen Transfer müssen Wortformen in Grundformen überführt werden; Wortzusammensetzungen und -ableitungen müssen erkannt werden; Wortklasseninformationen, Beziehungen zwischen Begriffen werden für die Vereindeutigung benutzt. So können MT-Output bzw. Zwischenergebnisse für die Dokumentarchivierung und das Information Retrieval verwendet werden.

B 10.3.2.3 Text- und Wortverarbeitung

Ohne Zweifel spielt die Textverarbeitung in jeder Übersetzungsumgebung eine bedeutende Rolle. Sogar „normale" Übersetzer gehen mehr und mehr zur Textverarbeitung auf PC über, und es ist nur ein kleiner Schritt bis zur Integration von (eigenen) Glossaren oder Wortlisten, die über so genannte „Windows" auf dem Bildschirm sichtbar gemacht werden, anstelle der Benutzung von Karteikästen. Natürlich werden auch andere Funktionen wie Rechtschreib-, Grammatik- oder Stilhilfen in zunehmendem Maße in solche Prozesse integriert.

Der „Quelltext" liegt als Ergebnis einer Textverarbeitung maschinenlesbar vor. Das MT- oder CAT-System muss jedoch auf die verschiedenen Textverarbeitungssysteme angepasst werden (Wang OIS ist beispielsweise mit LOGOS und SYSTRAN kombinierbar, WordPerfect mit SYSTRAN). Wenn solche Werkzeuge verfügbar sind, kann zudem die *Postedition* von MT-Ergebnissen durch besondere Editoren unterstützt werden.

Ein Problem ist in diesem Zusammenhang die Verbindung solcher Werkzeuge mit lokal verfügbaren MT-Systemen (vgl. z. B. LOGOS) oder die Verbindung mit einem zentralen Übersetzungsdienst (vgl. z.B. das Konzept des SYSTRAN-Einsatzes in der Europäischen Gemeinschaft oder auch die Nutzung des MINITEL-Systems in Frankreich für maschinelle Übersetzung mit SYSTRAN).

Zweifellos wird der Einsatz von MT und CAT insbesondere in Kombination mit Textverarbeitung und Online-Zugang zunehmen. Die Frage ist im Augenblick noch, ob die existierenden Werkzeuge (was die Qualität anbelangt) mächtig genug sind, um vom Benutzer angenommen zu werden. Besonders die Ergebnisse der MT-Experimente der Fa. Gachot S. A. über MINITEL und über den PC-Zugang mit SYSTRAN werden hier einen wichtigen Beitrag leisten.

B 10.3.2.4 Elektronisches Publizieren

Zu übersetzende Texte (insb. technische Texte wie *Wartungsanleitungen* und *Handbücher*) werden immer öfter mit Hilfe des elektronischen Publizierens (auch: Desktop Publishing) aufbereitet - incl. Abbildungen, Zeichnungen und Tabellen. Firmen, die Aufträge an Übersetzer in ihrer Firma oder nach außerhalb vergeben, wollen sich eine Aufbereitung (oder das Setzen) fertiger Übersetzungen ersparen, zumal wenn in mehr als eine Sprache übersetzt wird.

So müssen große Anstrengungen unternommen werden (vielleicht auf beiden Seiten: von den Herstellern von Desktop-Publishing-Systemen und den Entwicklern von MT- oder CAT-Software), um die Übersetzungshilfen so zu integrieren, dass die Dokumentstruktur nicht zerstört oder auch nur verändert wird. Natürlich gibt es Probleme bei der Zeilen-, Absatz- oder Seitenabstimmung (wegen der unterschiedlichen Längen von Original- und übersetzten Texten), und auch die Umstellung von Phrasen/Wörtern infolge unterschiedlicher Wortstellung in den einzelnen Sprachen wirft das Problem der korrekten Einfügung typographischer Zeichen (Fettdruck, Unterstreichungen etc.) in den Zieltext auf. Wenn diese „technischen" Aspekte des Übersetzungsprozesses nicht dem Übersetzer/Posteditor überlassen werden sollen, muss eine hochstandardisierte Textbeschreibung in das Electronic Publishing integriert werden. Dies stellt eine große Herausforderung an die existierenden und die kommenden MT- und CAT-Systeme dar. (Siehe Kap. B 11 und 12)

10.3.3 Benutzerfreundlichkeit

Was die professionellen Übersetzerarbeitsplätze anbelangt, so spielt die *Benutzerfreundlichkeit* auf jeder Ebene der MT und der CAT eine große Rolle. Nicht wünschenswert wäre es, wenn die Benutzung von MT oder CAT dazu führte, dass der Mensch nur noch „Sklavenarbeit" leistet und Tag für Tag die trivialen Fehler im Maschinenoutput korrigiert. Diese Gefahr besteht derzeit, da die verfügbaren Systeme nicht sehr flexibel und anpassungsfähig sind.

Zukünftige Entwicklungen in MT und CAT müssen sich daher darauf konzentrieren, dem Nutzer mehr und direktere Feedback-Möglichkeiten zu geben. Das manchmal in Datenbanksystemen benutzte „Privatdateikonzept", nach dem sich ein Benutzer in einer Datenbank einen „privaten Bereich" anlegen kann, könnte als Beispiel dienen: zumindest auf Wörterbuchebene müsste dem Benutzer die Möglichkeit gegeben werden, - auf der Grundlage existierender Daten - „eigene" Wörterbücher (physikalisch oder logisch) zu kreieren.

Nebenbei gesagt können beide Seiten - der Hersteller des Datenbanksystems und der Nutzer - von einem solchen Konzept profitieren: das Fachvokabular des Systems wird erweitert und der Benutzer hat selbst Einfluss auf die Auswahl der Übersetzungen (aber auch große. Verantwortung).

Was für das Lexikon richtig ist, gilt auch für die strukturellen Komponenten (z.B. den Einfluss auf die Satzlänge, stilistische Komponenten etc.). Bestehende Systeme müssen flexibler werden und künftige sollten solche Komponenten von Anfang an einbeziehen.

B 10.3.4 Kosten und Nutzen

Wie überall in der Wirtschaft wird eine Entscheidung für oder gegen Werkzeuge wie MT oder CAT nach Kosten/Nutzen-Analysen getroffen. Im vorliegenden Fall spielen nicht allein die reinen Kosten eine Rolle, denn „Zeit ist Geld" und eine Übersetzung sofort zu bekommen ist einen hohen Preis wert. Letzten Endes jedoch wird die Entscheidung aus wirtschaftlichen Gründen gefällt, wobei soziale und menschliche Gründe mitspielen werden.

Da (wenigstens dem Autor) keine ausreichenden Daten über die Entwicklungskosten von MT- und CAT-Systemen vorliegen, soll das Augenmerk im Folgenden den Kosten und Nutzen auf Seiten des Benutzers gelten.

Es scheint im Augenblick so zu sein, dass die Übersetzungsleistung eines Übersetzers (in Seiten pro Tag gemessen) durch Interaktion mit dem Computer und/oder Postediting beträchtlich erhöht werden kann. Wenn man annimmt, dass z.B. die Pflege und technische Unterhaltung eines Systems wie SYSTRAN incl. ComputerHardware rund 300.000 $ im Jahr kostet und dass im Jahr 300.000 Seiten technisch gesehen übersetzt werden können, sind die Kosten der Rohübersetzung - die Wörterbuchpflege und die Vor- und Nachbereitung der MT-Ergebnisse nicht eingeschlossen - fast zu vernachlässigen (1 $ pro Seite). Die Kosten des gesamten Prozesses (Übersetzung mit Mensch-Maschine-Interaktion) hängen von der gewünschten Qualität ab. Für eine „good-enough", d.h. eine *informative Übersetzung* (z.B. von Arbeitspapieren), ist eine so genannte schnelle Postedition (rapid postediting) ausreichend, bei der ein Übersetzer ca. 20 Seiten pro Tag produziert (anstelle von 6 - 8 Seiten ohne MT).

Um eine der professionellen intellektuellen Übersetzung vergleichbare hohe Qualität zu erreichen, muss mehr Zeit für die Nachkorrektur angesetzt werden. Es scheint jedoch, als beginne sich der Einsatz vom MT oder CAT in dem Sinne zu rechnen, dass - wenn das Vokabular des MT-Systems an die Bedürfnisse des Benutzers angepasst ist - die Kosten der Übersetzung deutlich unter denen der intellektuellen Übersetzung liegen, selbst wenn man berücksichtigt, dass die Verwendung von Textverarbeitungssystemen im intellektuellen Übersetzungsprozess bereits ca. 20 % Zeit spart.

B 10.4 Prinzipielle linguistische und strategische Probleme

Die Frage der Morphologie (d.h. die Probleme der Flexion, Ableitung und Wortzusammensetzung) kann in der MT als gelöst gelten, zumindest für praktische Zwecke in einer Anwendungsumgebung, auch wenn die Übersetzung korrekt zerlegter, abgeleiteter oder zusammengesetzter Wörter nicht immer automatisiert werden kann. Dies ist nicht der Fall für Lösungen auf syntaktischer oder semantischer Ebene. Selbst wenn man annimmt, dass ein Problem wie die Vereindeutigung syntaktischer Homographen wie *plays* in *he plays* und *the plays* durch strenge und voll formalisierte Analysesysteme gelöst werden kann, führt doch die Komplexität der Strukturen der natürlichen Sprache zu einer Explosion der Rechenzeit, wenn man versucht, jede mögliche (Teil-) Struktur zu berücksichtigen und zu verarbeiten. Daher versuchen die kommerziellen Systeme, den Prozess der Identifikation (oder Disambiguierung) durch besondere deterministische oder probabilistische Regeln abzukürzen. Im Ergebnis laufen sie 10.000 oder 1.000 mal schneller als vollkommen linguistisch orientierte Systeme, aber ihre Ergebnisse erreichen evtl. nicht die gleiche Qualitätsstufe.

Heute spielt Computerrechenzeit zwar nicht mehr die gleiche Rolle wie noch vor einigen Jahren, aber in der maschinellen Übersetzung ist sie bis heute nicht vollkommen zu vernachlässigen. Das gleiche gilt für die Lösung der Probleme der Homonymie, d.h. im Bereich der Semantik. Einerseits bestehen Beschränkungen bei der Verarbeitung der *Text*struktur (im Vergleich zur *Satz*struktur).

In den meisten Fällen ist die Operationsbasis eines MT-Systems die Satzumgebung, d.h. dass Informationen oder Lösungen aus vorangegangenen Sätzen verloren sind und so gut wie nichts über die Textebene bekannt ist. Dies führt zu vielen Fehlern, insbesondere bei der pronominalen Referenz, aber auch bei der Artikelinsertion und der Vereindeutigung von Homonymen. Die existierenden Systeme gehen das Problem der semantischen Mehrdeutigkeiten mit Hilfe semantischer Codes an (die auf einer allgemeineren Ebene auch bei der Vereindeutigung syntaktischer Strukturen eine Rolle spielen), insbesondere mit Fachgebietsmarkierungen, die zur Auswahl des „richtigen" Worts (oder der Wortfolge) in Abhängigkeit von den vom Benutzer vergebenen Fachgebietsparametern verwendet werden. Sie versuchen auch, dieses Problem durch Identifizierung von Wortsequenzen oder Redewendungen im Wörterbuch zu lösen (vgl. z.B. *es regnet Bindfäden*: diese Wendung kann nicht analysiert werden. Sie muss im Lexikon durch *it rains cats and dogs* ersetzt werden).

Seit Chomsky hat der systematische strukturell-semantische Zugang zur Sprachanalyse, zum Sprachverstehen und zur Übersetzung Fortschritte gemacht. So gibt es im Forschungsbereich verschiedene moderne formalisierte Grammatiktypen und Parser. Insbesondere in Japan (vgl. z.B. das MU-System) und Europa (vgl. z.B. die Anstrengungen der Europäischen Gemeinschaft und ihrer Mitgliedstaaten mit dem Europäischen Übersetzungssystem EUROTRA) wird die MT-

Forschung vorangetrieben. Es scheint jedoch, als brauche man mehr als eine computerlinguistische Entwicklung: Linguisten, Computerfachleute, Informatiker und Benutzer müssen in Großprojekten zusammenarbeiten, um das Ziel der praktischen Einsetzbarkeit zu erreichen.

B 10.5 Beispiele

Um einen Eindruck vom Entwicklungsstand der so genannten „produktiven" (nicht unbedingt kommerziellen) Systeme zu vermitteln und um die Anwendbarkeit der genannten Kriterien aufzuzeigen, werden zwei Systeme beschrieben: die MT-Systeme SYSTRAN und SUSY/STS.

B 10.5.1 SYSTRAN

SYSTRAN (die Rechte liegen bei der Fa. Gachot S. A., Soisy, Frankreich) hat - in seiner neuesten Version 3.7 - die folgenden Merkmale:

- Übersetzung von Volltexten. Selbst wenn die Strukturen nicht stimmen oder Wörter falsch geschrieben sind oder nicht im Computerlexikon gefunden werden, wird eine Übersetzung produziert.

- Die Übersetzungsgeschwindigkeit beträgt (in Abhängigkeit von der Rechnerkapazität) bis zu 350.000 Wörter in der Stunde. Damit ist es das schnellste System auf dem Markt.

- Das System ist sprachenpaarorientiert. Übersetzungen können für die Sprachenpaare Englisch > Französisch, Englisch >Italienisch, Französisch > Englisch, Russisch > Englisch (US-Airforce), Englisch > Japanisch (SYSTRAN Japan), Englisch > Arabisch erzeugt werden; entwickelt werden u.a. Englisch - Deutsch, Französisch > Deutsch, Deutsch > Englisch und Deutsch > Französisch. Die Qualität hängt einerseits von der Verfügbarkeit fachgebietsorientierter Lexika ab. Bei der Europäischen Gemeinschaft sind große Anstrengungen unternommen worden, um die SYSTRAN-Wörterbücher zu entwickeln. Für die englisch-französische Übersetzung steht jetzt ein Wörterbuch mit 150.000 Einträgen zur Verfügung. Die gleiche Qualität kann für die Übersetzung Deutsch-Französisch, die sich noch in der Anlaufphase befindet, nicht erzielt werden.

- Der Einsatz von SYSTRAN erfordert technische Spezialisten (und Systemverwalter). So können nur Firmen, die sich solche Spezialisten leisten können (wie die EG oder die US-Airforce), eine SYSTRAN-Version auf ihrem eigenen Computer halten (wenn es ein IBM- oder IBM-kompatibler Mainframe ist). Es gibt jedoch eine interessante Alternative: das System ist über Telekommunikationsnetzwerke zu nutzen, z.B. über Paketvermittlung (in Deutschland: vom PC aus via DATEX-P) oder - als sehr futuristische Variante - über BTX. In Frankreich ist eine BTX-Anwendung (die die französische BTX-Version namens TELETEL in Verbindung mit einem Telefon über einen Monitor (MINITEL) benutzt) bereits verfügbar (und wird sogar von Schülern genutzt).

- Die SYSTRAN-Wörterbuchpflege muss bis jetzt von Systemexperten vorgenommen werden. Das Hauptproblem liegt nicht in der Kodierung selbst (die sehr komplex ist, aber mittels einer benutzerfreundlichen Schnittstelle bewältigt werden kann und wird), sondern in der Konsistenz der Lexikondatenbank. Die Wörterbücher enthalten Fachgebietscodes,

aber diese Komponente muss weiterentwickelt werden, um eine flexiblere Nutzer- und Nutzungsorientierung zu erfahren.

- SYSTRAN ist wenig portabel. Das bedeutet: die Sprache ist IBM-Assembler, auch wenn die linguistischen Regeln normalerweise in einer besonderen Makrosprache geschrieben sind. Das System selbst benötigt - wie oben erwähnt - einen (IBM- oder Siemens- oder Amdahl-) Mainframecomputer oder Computer gleicher Größe. Eine Softwareumstellung (evtl. auf UNIX) ist allerdings geplant.

- Die Qualität der (Roh-) Übersetzung schwankt, und zwar je nach Sprachenpaar und Lexikonumfang. Für Englisch > Französisch können etwa die folgenden Prozentzahlen genannt werden: Morphologische Erkennung: etwa 100 %; syntaktische Strukturen: etwa 90 %; semantische Disambiguierung: zwischen 80 und 90 %, in Abhängigkeit von den sogenannten „limited semantics"-Regeln.

- SYSTRAN kann mit verschiedenen Textverarbeitungsumgebungen kombiniert werden. Eine davon ist Wang OIS (die bei der EG eingesetzt wird); es können aber auch (IBM-kompatible) PCs mit Editoren wie WordPerfect oder MS-WORD verwendet werden. Es gibt Anwendungen, bei denen mit Beleglesern WordPerfect-Textdateien erstellt werden, die an ein SYSTRAN-Servicezentrum (z. B. Gachot S. A. in Soisy bei Paris) geschickt werden. Man erhält dann den übersetzten Text über Postleitung und ein Softwaretool zurück, das die Postedition des Textes über einen geteilten Bildschirm erlaubt (mit dem Original in dem einen Fenster und der Übersetzung im anderen).

- Die Nutzung des Übersetzungssystems selbst ist batch-orientiert. Während des Übersetzungsprozesses selbst gibt es keine Möglichkeit der Interaktion (trotz der Tatsache, dass der Systemadministrator in einem Zwischenschritt Übersetzungen für unbekannte Wörter einfügen kann). Andererseits behandelt der Benutzer das System als „black box"; er benötigt keinerlei Kenntnis des Systems.

- Es gibt keine genauen Informationen über die Entwicklungskosten des Systems. Schätzungen bewegen sich zwischen 20 und 50 Millionen Dollar. Die EG-Version allein kostete bislang etwa 4 bis 6 Millionen Dollar. Die laufenden Kosten für die Unterhaltung einer Komplett-Version (mit mehreren Sprachen, inkl. Hardware, ohne Lizenzgebühren) belaufen sich meiner Meinung nach auf 300.000 Dollar im Jahr. Wenn also (bei der EG) etwa 300.000 Seiten im Jahr übersetzt werden können, betragen die Systemkosten 1 $ pro Seite (Textverarbeitung, Postedition und Computerrechenzeit nicht eingeschlossen). Die Übersetzungskosten bei Benutzung des französischen MINITEL-Systems (BTX) liegen bei 0,10 $ pro „Fenster" (das sind bis zu 10 Zeilen). In dieser Version ist die automatische (Roh-) Übersetzung innerhalb von 20- 30 Sekunden verfügbar. Auf dem deutschen Markt wird die Übersetzung mit SYSTRAN z.Z. mit 0,09 DM/Wort (Standardpreis) angeboten.

- Dies alles lässt den Schluss zu, dass SYSTRAN endbenutzerorientiert ist, d.h. dass sein Markt in erster Linie die „Informationsgesellschaft" ist, die keine 100%ige Qualität benötigt, sondern eine Übersetzung, die ausreicht, um einen Text in einer dem Benutzer fremden oder so gut wie fremden Sprache zu verstehen. Es besteht jedoch kein Zweifel daran, das SYSTRAN auch gute Chancen hat, in einer professionellen Übersetzungsumgebung

als superschnelles Werkzeug und als eine Alternative im Bereich der computergestützten Übersetzung eingesetzt zu werden.

B 10.5.2 SUSY/STS

Das MT-System SUSY wurde im Rahmen eines großen Forschungsprojekts zur Computerlinguistik an der Universität des Saarlandes in Saarbrücken entwickelt und wird nun im Saarbrücker Translationsservice STS zur Übersetzung von Datenbanktexten (Titel/Abstracts) als Produktionssystem benutzt. Dieser Service wurde am Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung (IAI) an der Universität des Saarlandes eingerichtet. Hauptmerkmale von SUSY-STS sind:

- Übersetzung von Volltexten (auf Satzebene) ist möglich. In dieser speziellen Anwendungsumgebung liegt der Schwerpunkt jedoch auf der Übersetzung von Titeln und Abstracts.

- Wie SYSTRAN ist SUSY-STS ein robustes „All-Round"-System: Ein Text wird übersetzt, auch wenn er unbekannte Wörter enthält. In solchen Fällen bleibt das Originalwort im Zieltext stehen.

- SUSY-STS ist ein „multilinguales" MT-System. Nur die Transferkomponente (die eigentliche „Verbindung" zwischen zwei Sprachen) ist zweisprachig, während Analyse und Synthese unabhängig von der jeweiligen Ziel- bzw. Quellsprache ablaufen. SUSY-STS-Anwendungen sind möglich für Deutsch > Englisch (wird in STS-Verfahren praktisch genützt), Englisch > Deutsch und Russisch > Deutsch, ansatzweise auch für Französisch > Deutsch, Deutsch > Französisch und Esperanto > Deutsch. Vor allem wird Deutsch > Englisch übersetzt, wobei ein deutsches Analysewörterbuch mit 150.000 Einträgen, ein Lexikon mit deutschen Komposita (150.000 Einträge) und ein deutsch-englisches Übersetzungslexikon mit 300.000 Einträgen zur Verfügung stehen.

- Wie SYSTRAN muss auch SUSY-STS von gut eingearbeiteten Fachkräften bedient werden. Daher wurde ein Übersetzungsservicekonzept entwickelt. Übersetzungen werden nur im Übersetzungszentrum am IAI in Saarbrücken erstellt. Die Auftraggeber schicken ihre Daten per Magnetband oder Diskette ans IAI und erhalten die übersetzten Daten auf dem gleichen Wege zurück. Hauptanwendung ist, wie erwähnt, die Übersetzung deutscher Titel aus Datenbanken ins Englische. Die zweisprachige Terminologie wird während der Vorbereitung der Daten für die verschiedenen Fachgebiete aufbereitet. Auftraggeber sind u.a.: Das Informationszentrum RAUM und BAU, wo die zweisprachige Datenbank ICONDA produziert wird; das Deutsche Patentamt, dessen deutsches Stich- und Schlagwortverzeichnis übersetzt wurde; das Deutsche Informationszentrum für Technische Regeln, für das die Titel deutscher Industrienormen ins Englische übersetzt werden, die anschließend in die entsprechende Datenbank eingebracht werden; das Informationszentrum Sozialwissenschaften, für das die Titel einer Literaturdatenbank übersetzt werden. Alles in allem sind vom STS bis heute 200.000 Titel übersetzt worden.

- Das System ist unter UNIX verfügbar (und in diesem Sinne auch portabel). Die Programmiersprache ist FORTRAN, einige Routinen sind in C geschrieben. Eine andere Version läuft unter BS2000 auf Siemensrechnern.

- Die Qualität der maschinellen (Roh-) Übersetzung Deutsch > Englisch ist zufriedenstellend: etwa 99 % der Wörter der Quelltexte werden morphologisch erkannt; was die Erkennung der syntaktischen Strukturen und die semantische Disambiguierung anbelangt, so liegen keine zuverlässigen Statistiken vor. Die Einbeziehung von Fachgebietscodes ist in Entwicklung.

- Der Service setzt in der Regel menschliche Posteditoren ein, um eine hohe Übersetzungsqualität zu erzielen. Als erster Schritt im Übersetzungsprozess erfolgt die Identifizierung und Korrektur fehlerhafter Wörter mit Hilfe einer Rechtschreibhilfe. Dann werden unbekannte Wörter (hauptsächlich im Übersetzungswörterbuch) identifiziert und kodiert. Im nächsten Schritt wird der Text maschinell übersetzt und durch die Übersetzer (Posteditoren) am Bildschirm nachredigiert. Nach einer Nachprüfungsphase werden die Daten auf Band oder Diskette zurückgeschickt. Es kann ein automatisches Indexierungssystem integriert werden, d.h. der Analysebaustein von SUSY-STS wird benutzt, um aus den Textwörtern Grundformen zu ermitteln und Komposita oder Ableitungen in ihre Bestandteile zu zerlegen. Dem Posteditor werden neben dem Text noch Wortalternativen angeboten, die er „per Knopfdruck" in den Zieltext einfügen kann. Wie SYSTRAN ist SUSY-STS batchorientiert. Die Wörterbuchpflege ist jedoch dialogorientiert, so dass ein entsprechend geschulter Benutzer neue Wörter in die entsprechenden Lexika eintragen kann. Die Postedition erfolgt separat auf PC oder per Terminal direkt mit dem Hostrechner, einer Nixdorf TARGON/35.

- Die Entwicklungskosten des SUSY-Systems betrugen etwa 5 Millionen Dollar (ohne die Grundlagenforschung an der Universität des Saarlandes). Die Anpassung an die Erfordernisse des STS kostete etwa 200.000 Dollar. Die Systemkosten betragen pro Jahr etwa 150.000 Dollar incl. des Bedienungspersonals (ohne die Übersetzer). Der Service arbeitet kostendeckend, wenn im Jahr etwa 4 Millionen Wörter (à 0,12 Dollar) übersetzt werden. STS wird z.Z. mit Fördermitteln des Bundesministers für Forschung und Technologie (BMFT) getestet.

- Zusammenfassend lässt sich sagen, dass SUSY-STS ein spezialisiertes CAT-System ist. Seine Geschwindigkeit (5.000 laufende Wörter pro Stunde CPU-Zeit) ist mit der von SYSTRAN nicht zu vergleichen, und eine Erweiterung um weitere Sprachpaare ist derzeit nicht vorgesehen. Es ist jedoch offensichtlich, dass MT-Systeme bei Ausnutzung ihrer Fähigkeiten und nach Anpassung an besondere Bedürfnisse in der Zukunft eine gute Rolle spielen werden.

Im Rahmen dieses Artikels können ähnliche Beschreibungen anderer existierender Systeme nicht gegeben werden. Zu nennen wären vor allem METAL (Siemens), LOGOS und GAT-Systeme wie ALPS oder TERMEX. Dazu wären auch eine gewisse Vertrautheit mit der Nutzung und gute Kenntnisse der konzeptuellen Ebene erforderlich. Ziel sollte es sein, einige konkrete *Beispiele* zu präsentieren, um einen ersten Eindruck von der Komplexität der Materie zu vermitteln. So erscheint es zu einfach, MT nur aus sprachlicher Sicht zu betrachten (in dem Sinne, dass die Quali-

tät der MT mit der der Humanübersetzung vergleichbar oder nicht vergleichbar sei) oder auf der anderen Seite MT als rein technisches Softwarewerkzeug zu behandeln: Fortschritte sind nur dann zu erzielen, wenn die Grenzen und Möglichkeiten der MT in konkreten Anwendungsumgebungen berücksichtigt werden.

B 10.6 Ausblick

Schaut man in eine ferne Zukunft, so scheint sicher zu sein, dass sich die professionelle Übersetzung im Normalfall der MT bedienen wird. Dies wird mehr oder weniger von den „geeigneten" Sprachenpaaren abhängen, von der Vollständigkeit der Maschinenwörterbücher, von der „richtigen" technischen Systemumgebung und nicht zuletzt auch von den Kosten der Systemnutzung. Daneben werden auch CAT-Systeme als Vokabel- und Terminologiehilfen - in Textverarbeitungssysteme integriert oder an sie angeschlossen - eine wichtige Rolle spielen, insbesondere als Unterstützung bei der Erstellung eines fremdsprachigen Textes durch den Autor selbst oder um einen über Telekommunikation (TELETEX) übermittelten Text zu verstehen. Im Büro wird CAT hauptsächlich mit bilingualen Wörterbüchern auf CD-ROM oder Festplatte als Stil- oder Grammatikhilfe benutzt werden.

Im Gegensatz zu Ansichten wie z. B. der von Hutchins (Lit. 01., S. 331-334) sehe ich nicht die Notwendigkeit eines besonderen *Übersetzerarbeitsplatzes*, aber es wird besondere Übersetzungshilfe-*Funktionen* geben, die - in leistungsfähige Autoren-Arbeitsplätze integriert - die Nutzung von Fortschritten im Bereich von Computerlexika und -thesauri möglich machen.

Was wird im Hinblick auf technische Übersetzungen mit dem professionellen Übersetzer geschehen? Man kann MT-Systeme als besondere „Expertensysteme" ansehen, wobei angemerkt werden muss, dass sie so gut wie nie die Qualität guter spezialisierter Übersetzer erreichen werden. Um jedoch solche Übersetzungs-Expertensysteme zu entwickeln, zu verbessern und zu unterhalten, werden „Sprachtechnologen" (linguistic knowledge engineers) gebraucht, die in der Lage sind, mit diesen Systemen. umzugehen. Was die Komplexität der natürlichen Sprache angeht, wo wird es zur Zusammenarbeit zwischen System und Übersetzer kommen. Und es besteht die gute Chance, dass - falls Übersetzungen billiger und schneller werden - die Nachfrage nach ihnen steigt und letzten Endes die Übersetzer - unter veränderten Bedingungen - nach wie vor eine wichtige Rolle spielen werden.

Literatur

01. Hutchins, W. J.: Machine Translation:. Past, Present, Future. Chichester 1986.
02. Nagao. M. ei al.: An English-Japanese Machine Translation System of ihe Titles of Scientific and Engineering Papers. In: COLING 1982, Amsterdam 1982.
03. Van Slype, G.: Deuxieme evaluation du systeme des traduction automatique SYSTRAN anglais-francais de la Commission des Communautes Europeennes. Bruxelles 1979.
04. Van Slype, G.: Conception dune methodologie generale d'evaluation de la traduction automatique. Multilingua 1 - 4, 1982, S. 221 - 237.
05. Zimmermann, H. H., Kroupa, E., Luckhardt, H.-D.: STS - Das Saarbrücker Übersetzungssystem. Veröffentlichungen der Fachrichtung Informationswissenschaft. Saarbrücken: Universität des Saarlandes 1987.