Der Saarbrücker Übersetzungsservice STS - Computergestütztes Übersetzen für die Fachinformation

von Heinz-Dirk *Luckhardt* und Harald H. *Zimmermann,* Saarbrücken

Nachrichten für Dokumentation (NFD) 39, 351-356 (1988)

Maschinelle Übersetzung; Übersetzung, rechnergestützt; Informationswissenschaft; MARIS

Zusammenfassung

Der im Projekt MARIS (Multilinguale Anwendung von Referenz-Informations-Systemen) an der Fachrichtung Informationswissenschaft der Universität des Saarlandes entwickelte Service für computergestützte Übersetzung (STS) wird vorgestellt. Hierbei werden maschinelle und intellektuelle Übersetzung in einer gemeinsamen Systemumgebung (Übersetzerarbeitsplatz) verknüpft. MARIS setzt Verfahren und Systeme der maschinellen Übersetzung bei der Ubersetzung (Deutsch > Englisch) von Titeln, Deskriptoren und Abstracts aus deutschen Datenbanken praktisch ein. Bisher wurden ca. 2 Mio. Wörter übersetzt, vorwiegend für die Datenbankanwendung. MARIS wird vom Bundesministerium für Forschung und Technologie gefördert.

Summary

The Saarbrücken Translation Service STS - Computer-Aided Translation for Specialised Information Centers

The paper presents the Saarbrücken Computer-Aided Translation Service (STS) being developed in the projekt MARIS (Multilingual Application of Reference-Oriented Information Systems) at the Information Science Department of the University of Saarbrücken. Intellectual and machine translation (esp. German to English) are combined in a joint system surrounding (translator's workstation). MARIS applies methods and (sub)systems developed for machine translation to titles, abstracts, and descriptors from German databases. About 2 million words have been translated yet. The MARIS project is funded by the Federal Ministry of Science and Technology.

0 Einleitung

Seit 1985 werden im Projekt MARIS (Multilinguale Anwendung von Referenz-InformationsSystemen), das vom Bundesministerium für Forschung und Technologie (BMFT) gefördert wird, Arbeiten zur Konzipierung und Entwicklung des computergestützten Saarbrücker Übersetzungsservice STS durchgeführt. Das Projekt ist in den wissenschaftlichen Teilen am Lehrstuhl für Informationswissenschaft der Universität des Saarlandes, Saarbrücken, angesiedelt. Die technisch-praktischen Arbeiten werden am Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung an der Universität des Saarlandes (IAI) durchgeführt. Die wesentlichen Teile, Funktionen und Anwendungen werden im folgenden vorgestellt.

1 Übersicht über das Verfahren

1.1 Ausgangslage zu Beginn des Projekts

Im folgenden werden kurz die Problemstellungen und die Entwicklungen im Projekt MARIS umrissen.

Ausgangsmaterial für die computergestützte Ubersetzung Deutsch > Englisch sind deutschsprachige Titel und Deskriptoren in Datenbanken (seit *Anfang* 1988 sind auch Abstracts einbezogen). Die Datenbestände liegen maschinenlesbar vor; zum Teil existieren bereits Übersetzungen ins Englische bzw. aus dem Englischen ins Deutsche (und in andere Sprachen); ein (maschinenlesbares) kontrolliertes Vokabular (Thesaurus), mit dessen Hilfe der Inhalt zusätzlich intellektuell erschlossen wird, ist nur teilweise mehrsprachig verfügbar.

Bei fachspezifischen Übersetzungen stellt die Terminologiearbeit den Übersetzer vor große Probleme. Die Übersetzung in der Fachinformation hat zudem z. Z. nicht den hohen Stellenwert: die Kosten für die Übersetzung sollen den (erheblichen!) Aufwand bei der Erschließung der Dokumente nicht wesentlich erhöhen *(vgl. Zimmermann* 1987, 1987a, 1987c).

Inzwischen sind technische Hilfsmittel entwickelt worden, die eine rationellere Durchführung der Übersetzungsaufgaben ermöglichen. Diese Hilfsmittel unterstützen vor allem die fremdsprachige Textgenerierung (Übersetzung, Postedition) und führen im terminologischen Bereich zu größerer Konsistenz. Der Einsatz bzw. die Anpassung lohnen sich jedoch erst ab einem gewissen Mindestvolumen an Übersetzungen.

Im Terminologiebereich entstehen bei „zentraler" Sammlung und Nutzung für einen Übersetzungsservice weitere Vorteile: während der bisherigen Projektlaufzeit von MARIS zeigte sich bereits an den intensiv bearbeiteten Fachgebieten „Bauwesen", „Technische Regeln" und „Umweltschutz", dass sich erhebliche Überlagerungs- bzw. Berührungspunkte ergeben (z.B. Baurecht, Verwaltungsvorschriften, Technische Regeln, Bauausführung, Normen etc.). Insgesamt steht zu erwarten, dass durch das kumulierte Sammeln der technischen Terminologien eine erhebliche Unterstützung des Übersetzungsprozesses möglich wird.

1.2 Der Saarbrücker Translationsservice (STS)

Ziel des Projekts MARIS ist die prototypische Entwicklung und der praxisorientierte Einsatz der organisatorischen und technischen Infrastruktur einer Serviceeinrichtung für Übersetzungsdienstleistungen im Fachinformationsbereich. Die wissenschaftliche Konzeption und modellhafte Realisierung werden dabei mittels praktischer Aufträge überprüft (vgl. *Zimmermann/Kroupa/ Luckhardt 1987).* Orientiert an den Problemstellungen im Fachinformationsbereich wird ein *mehrstufiges* Konzept entwickelt und erprobt.

Der effiziente Einsatz maschineller Übersetzungshilfen, sieht man einmal von reinen Textverarbeitungsprogrammen ab, ist erst bei Vorliegen eines Mindestumfangs an maschinenlesbarer Fachterminologie (d. h. einer hohen „Trefferquote" bei der Identifikation in lexikalischen Datenbasen bzw. maschinellen Übersetzungssystemen) sinnvoll. Da dies für die meisten Fachgebiete z. Z. noch nicht gegeben ist, sieht der stufenweise Aufbau von STS in der ersten Stufe (STS-I) zunächst den verstärkten Ausbau entsprechender Terminologie in einem neu hinzukommenden

Fachgebiet vor. Ausgehend von bereits übersetzten Texten wird eine Terminologiedatenbank aufgebaut, die die retrospektive Generierung von Terminologie unterstützt: gleichzeitig kann sie den Übersetzern bei Problemfällen als elektronisches Wörterbuch dienen.

Im Anschluss an STS-I kann jeweils (d.h. je neues Fachgebiet) zunächst eine Variante STS-II einbezogen werden. Zu den Wörtern der zu übersetzenden Texte werden automatisch zielsprachliche Entsprechungen aus den Computerwörterbüchern herausgezogen und den Übersetzern angeboten. Auf der Stufe STS-III kommt die maschinelle Übersetzungskomponente zum Einsatz.

Die mit dem Jahr *1987* abgeschlossene Konzeptions- und Implementierungsphase hat inzwischen zu einer Integration der entwickelten Konzepte geführt. Daraus resultieren zwei alternative Systemanwendungen:

- CAT-H: intellektuelle Übersetzung mit Computerunterstützung

- CAT-C: Computerübersetzung mit Prä- und Postedition

Beide Verfahren können vom Übersetzer bzw. Organisator alternativ eingesetzt und nach Bedarf auf besondere Textsorten angepasst werden (vgl. dazu auch Kap. *3).*

2 Vertragspartner

In Abstimmung mit dem Projektträger (GID/GMD) erfolgen die Vertragsabschlüsse phasen- und schrittweise, da in jedem Fall gewährleistet sein muß, dass die geschlossenen Verträge auch vollständig erfüllt werden können. Dabei wurden den Übersetzungen (von Titeln) für die Datenbank ICONDA des Informationszentrums Raum und Bau, für die Datenbank DITR des Deutschen Informationszentrums für Technische Regeln und die Datenbank SOLIS des Informationszentrums Sozialwissenschaften Priorität eingeräumt.

Die Verträge mit den einzelnen Vertragspartnern werden zudem nicht über die gesamte Projektlaufzeit abgeschlossen, sondern gestaffelt nach Umfang und Bearbeitungszeitraum (in der Regel mit Laufzeit von einem Jahr).

Bislang wurden folgende Vereinbarungen geschlossen: (Stand der Übersetzungsleistungen jeweils zum *30.* Juni *1988)*

(1) Informationszentrum RAUM und BAU (IRB), Stuttgart

Verträge über: 110 000 Titel + 5000 Abstracts
Übersetzt sind: 104 000 Titel + 1000 Abstracts

(2) Deutsches Informationszentrum für technische Regeln (DITR), Berlin

1 Vertrag über: 20 000 Titel und 20 000 Termini
Übersetzt sind: 15 000 Titel

(3) Informationszentrum Sozialwissenschaften, Bonn

2 Verträge über: 60 000 Titel
Übersetzt sind: 38 000 Titel

(4) Deutsches Patentamt, München

1 Vertrag über: 130 000 Termini (Übersetzung abgeschlossen)

Es ist ein weiterer Vertrag über die Übersetzung der in der 5. Auflage der internationalen Patentklassifikation (IPC) neu auftretenden Begriffe geplant.

(5) Umweltbundesamt, Berlin

1 Vertrag über: 64 000 Titel (plus Deskriptoren)
Übersetzt sind: 44 000 Titel

(6) Fachinformationszentrum Technik, Frankfurt

1 Vertrag über: 20 000 Termini (Übersetzung abgeschlossen)

3 Stand der Arbeiten

Beim Stand der Arbeiten werden die einzelnen Ergebnisse, die in Arbeit befindlichen Aufgaben sowie weitere Planungen in Übersicht dargestellt.

3.1 Intellektuelle Übersetzung mit Computerunterstützung (CAT-H)

3.1.1 Übersetzerpool

Zur Durchführung der intellektuellen Übersetzung wurde zu Beginn des Projekts ein Übersetzerpool (vgl. *Sharma 1987)* eingerichtet, dessen Mitglieder mit der zu übersetzenden Textsorte und den Terminologien der verschiedenen Fachgebiete vertraut gemacht werden. Dem Pool sind derzeit zwischen 15 und 20 Übersetzer(innen) angeschlossen. Insgesamt gehörten ihm bisher über *30* Übersetzer(innen) an; die Fluktuation ist also recht stark. Der Pool ist in Fachgruppen aufgeteilt, die den derzeitigen Anwendern (IRB, IZ, UBA, DPA, DITR) zugeordnet sind und von jeweils einem Projektmitarbeiter betreut werden. Aus Mitgliedern des Übersetzerpools rekrutieren sich auch die Posteditoren für die maschinelle Übersetzung.

3.1.2 Technische Abwicklung der intellektuellen Übersetzung

Die Abwicklung der intellektuellen Übersetzung CAT-H vollzieht sich nach dem folgenden Verfahren:

- Datenaustausch zwischen Auftraggeber und Auftragnehmer per Magnetband
- Datenumsetzung auf STS-Format
- Datenaufbereitung für die Übersetzer auf Computer und Papier

- automatische Lemmatisierung und Übersetzung der Lemmata
- intellektuelle Übersetzung am PC unter Zuhilfenahme der automatisch übersetzten Termini
- Nachkorrektur, Abrechnung, Datensicherung und Übersendung der übersetzten Titel an den Auftraggeber.

An konkreten Projektarbeiten fielen im Projektzeitraum an:

- laufende Umsetzung von Bändern der Anwender auf STS-Format - laufende Überspielung von Übersetzungen in die Originaldateien und Versenden der Daten
- Anpassungen der Umsetzungsprogramme an geänderte Inputformate
- Entwicklung von Prüfprozeduren für formale Fehler
- Erstellung von Datenbankaufbereitungs-, Statistik-, Terminologieerschließungsprogrammen
- technische Betreuung der Übersetzungen - Datensicherung.

3.2 Terminologie

3.2.1 Auswahl von Übersetzungsäquivalenten

Mit dem Anwachsen des Terminologiepools stellt sich mehr und mehr das Problem der Auswahl zwischen mehreren verschiedenen Übersetzungsäquivalenten (= zielsprachlichen Entsprechungen, vgl. *Luckhardt 1987),* die im Laufe des Projekts teils aus konkreten Übersetzungen, teils aus maschinenlesbar vorliegenden Sammlungen gewonnen wurden. Bei der Titelübersetzung werden von Übersetzern Äquivalente in Abhängigkeit vom Fachgebiet, von Anwendervorschriften und/oder vom Kontext vergeben, wobei ggf. alle drei Gesichtspunkte ineinandergreifen. Die Möglichkeit des Abwägens der verschiedenen Kriterien hat der Übersetzer dem Computer voraus, da diese intellektuelle Leistung kaum formalisiert ist. Insbesondere ergeben sich die folgenden Probleme:

(a) Fachgebiet

Die Titel eines Auftraggebers können nicht ohne weiteres einem fest umrissenen Fachgebiet zugeordnet werden. Ein Beispiel dafür stellen die Daten des Umweltbundesamtes dar, die den Fachgebieten Umweltschutz, Chemie, Biologie, Land- und Forstwirtschaft, Recht, Wirtschaft, Raumordnung, Bauwesen etc. zuzuordnen sind, wobei u. U. *in einem* Titel mehrere von ihnen vertreten sind.

In der Regel sind die Dokumente (Titel/Abstracts) der Datenbankanbieter *klassifiziert,* d.h. durch entsprechende Notation in Fach- oder Themengebiete eingeordnet. Dies bedeutet *an sich* eine wertvolle Unterstützung bei der Disambiguierung (dies gilt auch für die maschinelle Übersetzung (MÜ)). Ein Problem stellen jedoch die unterschiedlichen Fachgebietsklassifikationen der verschiedenen Anwender dar. Jeder Anwender stellt eine eigene Klassifikation nach den für ihn wesentlichen Kriterien auf und markiert die Klassen an den Stellen besonders detailliert, an denen er es für sinnvoll erachtet. Dies steht den Anforderungen einer *allgemeinen Klassifikation* entgegen, wie sie für die Ordnung von Fachgebieten innerhalb eines maschinellen Übersetzungswörter-

buchs für verschiedene Anwender/Fachgebiete (eines Terminologiepools) sinnvoll erscheint (vgl. das Beispiel in Luckhardt *1987c).*

(b) Anwender

Die Nutzer von MÜ-Systemen (oder auch allgemeiner: die Auftraggeber von Übersetzungen) legen in der Regel großen Wert darauf, dass die in ihrem Hause übliche Terminologie (Inhouse-Terminologie) verwendet wird. So hat das Auswahlkriterium „Benutzerpriorität" Vorrang vor anderen, muss dabei aber mit dem Kriterium „Fachgebiet" in Einklang gebracht werden. Der Auswahlalgorithmus muß also z.B. die folgenden Pfade verfolgen:

- stimmen Benutzercode des Textes und eines vorliegenden Lexikoneintrags überein?
- wenn ja: stimmen auch die Fachgebietscodes überein?
- wenn nein: stimmen wenigstens die Fachgebietscodes überein? - etc.

(c) Kontext

Titel werden in STS losgelöst von den dazugehörigen Texten übersetzt. „Kontext" bedeutet im vorliegenden Falle also in der Regel „knappe, meist nur nominale Strukturen, allenfalls eine einfache Verbform (fin. Verb, Infinitiv, Partizip)". Doch auch hier hat der Übersetzer derzeit dank seines Fach- oder Weltwissens, seines Assoziationsvermögens, seiner Phantasie etc. dem Computer - der z. Z. allein auf linguistisch-formaler Ebene entscheidet - intellektuelle Leistungen voraus (vgl. Beispiele in *Luckhardt 1987c),* die zudem schwer „formalisierbar" oder typisierbar sind. Wenn es der Künstlichen Intelligenz gelingt, auf dem Gebiet der Formalisierung kognitiver Prozesse voranzukommen, kann auch für die automatische Auswahl von Übersetzungsäquivalenten eine neue Qualität erreicht werden. Doch wird dies bis zur Erfüllung des Anspruchs, beliebige Texte/Daten aus beliebigen Fachgebieten mit uneingeschränkter Sprachform zu übersetzen, noch ein weiter - heute nicht präzise zeitlich vorhersehbarer - Weg sein.

3.2.2 Der STS-Terminologiepool

Der STS-Terminologiepool umfasst derzeit (März *1988)* ca. *180 000* deutsch-englische Übersetzungsäquivalente, die gewonnen wurden durch:

- Einspielen fremder bzw. anwendereigener Terminologiesammlungen;
- Extraktion von Äquivalenten aus fertigen Übersetzungen nach dem halbautomatischen STS-CTX-Verfahren;
- intellektuelle Extraktion aus vorliegenden Übersetzungen;
- titelunabhängige Übersetzung von z.B. Schlagwortverzeichnissen bzw. Thesauri.

3.2.2.1 Terminologiegewinnung

Fremde bzw. anwendereigene Terminologiesammlungen

Wenn beim Anwender fremdsprachige Terminologie vorliegt, hat diese Vorrang bei den anwenderbezogenen Übersetzungen. So wurden im Falle des Informationszentrums Raum und Bau (IRB) und des Deutschen Informationszentrums für technische Regeln (DITR) hauseigene

deutsch-englische Terminologiesammlungen in das entsprechende STS-Computerlexikon überspielt. Diese Sammlungen sind allerdings nicht ohne Anpassungen zu übernehmen (vgl. Beispiele in *Luckhardt 1987c).*

Extraktion aus vorliegenden intellektuellen Übersetzungen

Die Daten der Anwender werden automatisch auf das Eingabeformat für ein automatisches Texterschließungs- und Indexierungssystem (CTX, vgl. *Kroupa 1982)* gebracht, mit dem die Grundformen aus den Eingabedaten ermittelt werden. Das Vergleichsprogramm speichert diejenigen Grundformen in einer besonderen Datei, für die es kein zielsprachliches Äquivalent findet. Wenn die intellektuellen Übersetzungen fertig sind, ordnen Übersetzer/Terminologen den dem Pool noch „unbekannten" Begriffen zielsprachliche Äquivalente aus den entsprechenden Titelübersetzungen zu.

Titelunabhängige Terminiübersetzung

Seit Mai 1987 wird im Projekt das deutsche Stich- und Schlagwortverzeichnis des Deutschen Patentamts zur internationalen Patentklassifikation IPC (ca. 130 000 Begriffe mit Unterstichwörtern und erklärenden Kontexten) ins Englische übersetzt. Nach Fertigstellung dieser Arbeit wird ein umfassender fachsprachlicher Wortschatz zur Verfügung stehen, der weitgehend die Grundbegriffe der gesamten Technik abdeckt (und zudem mit IPC-Kennungen markiert ist).

Das Fehlen eines umfassenderen Kontextes schafft hier neue Bedingungen. Allerdings gibt die Beschreibung der IPC-Klasse, die jedem Stichwort zugeordnet ist, dem Übersetzer/Terminologen einen wichtigen Hinweis auf die vorliegende Lesart eines (ggf. mehrdeutigen) Begriffs und somit auf das auszuwählende zielsprachliche Äquivalent. So kann der Begriff „Schnecke" im Bereich *Backwarenherstellung* mit „worm" und im Bereich *Gartenbau* mit „snail" eindeutig übersetzt werden.

3.2.2.2 Fachgebietsklassifikation

Es erschien wenig sinnvoll, die Disambiguierung in STS unmittelbar auf den Anwenderklassifikationen aufzubauen. Dies würde vor allem in der Entwicklungsphase zu einer Reihe von Problemen führen, u.a. in bezug auf das Copyright. Für STS ist daher eine „neutrale", möglichst allgemein gehaltene Klassifikation gewählt worden. Sie erlaubt eine Grobauswahl zwischen Übersetzungsäquivalenten. Auf eine genauere Klassifizierung wurde für STS verzichtet, da die Abgrenzungsprobleme besonders bei Anwendung der maschinellen Übersetzung mit der Feinheit der Klassifikation zunehmen.

3.2.2.3 Nutzung des STS-Terminologiepools DEENWO

Unter einem Terminologiepool wird in MARIS die für die maschinelle Übersetzung zugreifbare Terminologie für alle Fachgebiete/Anwender verstanden. Daneben existiert ggf. eine dBase-Datenbank (z.B. für Sozialwissenschaften bzw. Raum und Bau) für die intellektuelle Übersetzung, die von Zeit zu Zeit mit dem Terminologiepool für das maschinelle Übersetzungssystem kompatibilisiert wird.

Auf den Terminologiepool greifen verschiedene Verfahren zu:

1. die automatische Übersetzung von Deskriptoren aus Datenbanken,
2. die Wortübersetzung in der Teilkomponente „Transfer" der maschinellen Übersetzung von Titeln und Abstracts,
3. die automatische Indexierung CTX und anschließende Übersetzung der erzeugten Deskriptoren.

Im allgemeinen reicht die Angabe des Auftraggebers (als Merkmalkennung) zur Auswahl der korrekten zielsprachlichen Äquivalente aus. Man versieht die zu übersetzenden Deskriptoren bzw. Dokumente mit dem Code des Auftraggebers, und der Suchalgorithmus kann die mit dem gleichen Code versehenen zielsprachlichen Äquivalente zuordnen.

Die Fachgebietsmarkierung kommt dann ins Spiel, wenn entweder kein Übersetzungsäquivalent mit dem passenden Anwendercode vorhanden ist oder der Pool für einen Anwender mehrere Äquivalente anbietet. Es ist offensichtlich, dass auch Pooleinträge mit Anwendercode für Texte anderer Anwender nutzbar sein müssen. Wenn der Begriff „Umweltverschmutzung" für das Umweltbundesamt mit „environmental pollution" übersetzt wird, soll diese Übersetzung auch für andere Anwender verfügbar sein, ohne dass sie dupliziert und/oder mit anderen Anwendercodes versehen wird. Wenn mehrere Äquivalente zur Verfügung stehen, wird dasjenige gewählt, dessen Fachgebietscode mit dem des Dokuments/Textes übereinstimmt.

3.2.3 Terminologisches Material

Die STS-Systemlexika haben derzeit (31. Mai 1988) den folgenden Inhalt:

	Einträge
dt. morpho-synt. Analyselexikon:	143 546
dt. Kompositalexikon:	158 900
dt. semantisches Lexikon:	75 853
dt./engl. Transferlexikon:	200 000
engl. Syntheselexikon:	2 896

Dazu kommen die folgenden Datenbanken für Übersetzer:

1) dBASEIII (IZ, IRB, DITR)	17 000
2) GOLEM (IRB und DITR)	13 000

3.3 Maschinelle Übersetzung (CAT-C)

Das MÜ-System STS ist lauffähig. Verwendet wird eine von Siemens-BS2000 auf Nixdorf TARGON/35 (Unix) migrierte Version des sog. „Saarbrücker Übersetzungssystems" SUSY. STS wurde während der MARIS/STS-Präsentation am 5. Februar 1987 zum ersten Mal in der Öffentlichkeit vorgestellt. Das System wurde in den folgenden Monaten anhand von Titeln des IRB auf seine Praxistauglichkeit untersucht und wird laufend - auch technisch - fortentwickelt. Inzwischen erstreckt sich der Einsatz auf die Daten aller Auftraggeber. Im folgender werden Konzeption und Stand der CAT-C-Variante des STS im Detail beschrieben.

3.3.1 Datenumsetzung

Die Anwenderdaten werden derzeit auf einer Siemens-Anlage (Universität des Saarlandes) umgesetzt und in einzelne Pakete aufgeteilt, danach auf die projekteigene Nixdorf-Anlage (im IAI) überspielt.

3.3.2 Halbautomatische Präedition

In STS musste zunächst eine halbautomatische Präeditionskomponente integriert werden, da an den Anwenderdaten einige Besonderheiten festgestellt wurden, die die intellektuelle Übersetzung kaum behinderten, die jedoch die MÜ blockiert hätten.

3.3.2.1 Auswahl der zu übersetzenden Titel

Aus dem Format bzw. der Kennzeichnung der gelieferten Daten lässt sich nicht in allen Fällen mit hundertprozentiger Sicherheit automatisch feststellen, welche Titel zu übersetzen sind. Die Dokumente enthalten neben den zu übersetzenden deutschen Titeln auch englische oder anderssprachige Titel mit oder ohne deutsche Übersetzung. Es gilt also die Dokumente (intellektuell) zu identifizieren, die über eine deutsche Titelformulierung verfügen, aber über keine englische.

3.3.2.2 Satzendeerkennung

Zahlreiche Titel bestehen aus mehreren Sätzen, so dass bei mehr als einem Punkt pro Titel das Problem der Erkennung möglicher Satzgrenzen auftritt. Da keine hundertprozentig sicheren Algorithmen für die Automatisierung dieses Problems existieren, die MÜ aber auf korrekte Satzgrenzen angewiesen ist, werden Satzgrenzen zumeist halbautomatisch (d.h. mit intellektueller Überprüfung) gesetzt.

3.3.2.3 Rechtschreibfehlererkennung, Schreibvarianten

Die Anwenderdaten sind in der Regel mit Problemen behaftet, die vor der Bearbeitung durch die Maschine bereinigt werden müssen: Rechtschreibfehler, Schreibvarianten, ökonomische Schreibweisen etc. Diese Probleme sind in den meisten Fällen von geringer Bedeutung für die intellektuelle Übersetzung, da hier Faktoren wie Assoziationsfähigkeit, Kreativität, Intuition, Weltwissen etc. dem Menschen das Textverständnis erleichtern bzw. erst möglich machen.

Rechtschreibfehler werden von der SOFTEX-PRIMUS-Rechtschreibhilfe erkannt, die in den Präeditionsprozeß integriert wurde. Ökonomische Schreibweisen wie „ein- und zweistöckig" werden durch die automatische morphologische Analysekomponente bearbeitet, ebenso wie Schreibvarianten von der Art „Anzeigegerät/Anzeigengerät". Letztere können allerdings nicht automatisch einander zugeordnet werden, so dass beide getrennt bearbeitet und terminologisch erfasst werden.

3.3.3 Maschinelle Übersetzung

Die so aufbereiteten Titel werden vom STS-Kernsystem auf der TARGON/35 von Nixdorf automatisch übersetzt. Hierzu waren einige Anpassungsschritte vor allem bei der Datenauswahl

erforderlich. In die Transferkomponente wurde zudem ein automatisches Auswahlverfahren integriert, das Übersetzungsvarianten anhand der Benutzererkennung selektiert. Insgesamt hat sich die MÜ-Komponente in ihrer Funktion als ausreichend robust erwiesen, d.h. für jeden eingegebenen Titel wird eine Übersetzung geliefert.

3.3.4 Postedition

Der Output der MÜ wird auf PC überspielt und mit Hilfe des Textprozessors WORDSTAR 2000 postediert. Diese Systemkomponente wurde 1987 verbessert und um eine Komponente erweitert, die dem Posteditor für bestimmte quellsprachliche Begriffe mehrere zielsprachliche anbietet, für die eine automatische Selektion nicht möglich ist. Hier wurden zwei Varianten getestet, eine, in der die Varianten in den Text integriert werden, und eine, in der die Varianten an den übersetzten Titel angefügt werden.

3.4 Technologische Arbeiten

Der Schwerpunkt der technologischen Arbeiten lag auf der Erstellung und Modifikation der Systemumgebung für Übersetzung und Postedition, der Datenumsetzung und der CTX-Migration.

Technische Ausstattung: Zur Durchführung der Übersetzungsarbeiten steht z. Z. am IAI folgende technische Ausstattung zur Verfügung:

1 Minirechner Nixdorf TARGON/35 (Unix) mit 400 mB-Platte
2 PCs Victor VPC, 15 mB-Platte
6 PCs Tandon PCA, 40 mB-Platte
1 PC Tandon PCA, 30 mB-Platte
1 Drucker, IBM-Wheel-Printer

Zur Kommunikation des Unix-Rechners mit anderen Rechnern wurden verschiedene Programme getestet, modifiziert bzw. neu programmiert.

3.5 Wissenschaftliche Begleitung, weitere konzeptionelle Entwicklungen

3.5.1 Schnittstelle Mensch-Maschine

Dieser Aspekt ist in allen Projektphasen von großer Bedeutung. In der ersten Hälfte des Projektzeitraums ging es im wesentlichen um die Schnittstelle zwischen Übersetzer(in) und dem Siemens-Betriebssystem BS 2000. Hier galt es, Methoden und Prozeduren zu entwickeln, um den Übersetzer(inne)n die Erfassung und Korrektur der übersetzten Daten zu erleichtern.

In der zweiten Projekthälfte trat die Schnittstelle Posteditor-Maschine in den Vordergrund. In der ersten MÜ-System-Version handelte es sich um den Textprozessor WORDSTAR 2000, dem im Laufe des Jahres 1987 eine zweite Editor-Version an die Seite gestellt wurde, eine Eigenentwicklung auf der Grundlage des Turbo-Editors.

Das größte Problem bei der Postedition von Computerübersetzungen stellt jedoch nicht die geeignete Textverarbeitungssoftware dar, sondern die korrekte Auswahl von Übersetzungsvarianten (bei Mehrdeutigkeit des ausgangssprachlichen Worts oder bei Vorliegen mehrerer zielsprachli-

cher Entsprechungen für einen ausgangssprachlichen Begriff). Und hier hat sich gezeigt, dass zur Erzielung einer hochqualitativen Übersetzung trotz des umfangreichen Terminologiepools und des großen Angebots an Übersetzungsvarianten immer noch intellektuelle Recherchen der Posteditoren während des Posteditionsvorgangs oder - wenn der maschinelle Output auf Papier vorliegt - während der Posteditions-Vorphase (Vorbereitung auf Papier) notwendig sind.

3.5.2 Automatische Auswahl von Äquivalenten

Die in 3.5.1 dargestellten Fragen machen die Lösung des Problems der Auswahl von Übersetzungsäquivalenten immer dringlicher. Hier muß eine stärkere Hinwendung zu semantischen Verfahren erfolgen, die die bisher eingesetzten syntaktischen Verfahren ergänzen.

3.5.3 CAT-H und CAT-C

Das Projekt MARIS ist mit dem Vorhaben angetreten, einen computergestützten Übersetzungsservice zu konzipieren und zu implementieren. Die computergestützte Übersetzung (CAT, Computer-Aided Translation) in STS bietet sich zu Beginn des Jahres 1988 in zwei einander ergänzenden Systemkomponenten dar:

- CAT-C: Computerübersetzung mit Prä- und Postedition
- CAT-H: Humanübersetzung mit Computerunterstützung

Die Variante CAT-C wird für lexikalisch erschlossene und von der syntaktischen Struktur her geeignete Textsorten eingesetzt (z.B. Titel vom Informationszentrum Raum und Bau), CAT-H für lexikalisch unerschlossene und syntaktisch oder pragmatisch schwierige Textsorten (z.B. UBA-Titel bis zur terminologischen Sättigung des Terminologiepools oder Anweisungstexte vom Typ der Datenbank Hommel).

Anschrift der Autoren:

Diplomübersetzer Heinz-Dirk *Luckhardt* und Prof. Dr. Harald *H. Zimmermann.* Projekt MARIS, Fachrichtung 5.5 Informationswissenschaft. Universität des Saarlandes, D-6600 Saarbrücken.

Literatur

Kroupa, E. (1982): Strategien zur Dokument-Repräsentation bei CTX. Ein Verfahren zur computerunterstützten Texterschließung und Textwiedergewinnung. In: I. *Batori, J. Krause, H.-D. Lutz* (Hrsg.): Linguistische Datenverarbeitung. Sprache und Information Band 18. Tübingen: Niemeyer.

Kroupa, E., Zimmermann, H. H. (1987): Multilinguale Anwendungen der Sprachdatenverarbeitung in Referenz-Informationssystemen. In: *Wilß, W., Schmitz, K.-D.* (Hrsg.).

Kunze, J. (1986): Transfer as a Touchstone for Analysis. In: Proceedings of IAI-MT 86, 51-64.

Luckhardt, H.-D. (1987): Probleme bei der automatischen Auswahl von Übersetzungsäquivalenten. In: *Zimmermann, H. H., Kroupa, E., Luckhardt, H.-D.* (Hrsg.), S. 86.

Luckhardt, H.-D. (1987b): The STS Computer-Aided Translation System. Transfering a Research-Oriented MT System into Practice. Vortrag auf dem XIV. Weltlinguistenkongreß, Ostberlin, August 1987. Veröffentlichungen der FR Informationswissenschaft. Saarbrücken: Universität des Saarlandes, August 1987.

Luckhardt, H.-D. (1987c). Terminologieerfassung und -nutzung im computergestützten Saarbrücker Translationsservice STS. Veröffentlichungen der Fachrichtung Informationswissenschaft. Saarbrücken: Universität des Saarlandes, August 1987.

Luckhardt, H.-D. (1987d): Der Transfer in der maschinellen Sprachübersetzung. Sprache und Information Band 18. Tübingen: Niemeyer.

Nitta, Y. (1986): Idiosyncratic Gap: A Tough Problem to Structure-bound Machine Translation. In: Proceedings of COLING 86, 107-111.

Schmidt, P. (1986): Valency Theory in a Stratificational MT-System. In: Proceedings of COLING 86, 307-312.

Somers, H. L. (1986): The need for MT-oriented versions of Case and Valency in MT. In: Proceedings of COLING 86, 118-123.

Steiner, E. (1986): Generating Semantic Structures in EUROTRA-D. In: Proceedings of COLING 86, 304-306.

Wilss, W., Schmitz, K.-D. (Hrsg.) (1987): Maschinelle Übersetzung: - Methoden und Werkzeuge - Sprache und Information. Niemeyer, Tübingen.

Zimmermann, H. H. (1987): Computergestützte Übersetzung als ein Beitrag zur Überwindung von Sprachbarrieren. In: Zimmermann, H. H., *Kroupa, E., Luckhardt, H.-D.* (Hrsg.), S. 1.

Zimmermann, H. H. (1987a): Perspektiven der maschinellen Übersetzung in der Fachinformation. In: Zimmermann, H. H., *Kroupa, E., Luckhardt, H.-D.* (Hrsg.), S. 200.

Zimmermann, H. H. (1987c): Linguistic-Technical Aspects of Machine Translation. In: Proceedings of the AGARD TIP Meeting on „Barriers to Information Transfer and Approaches toward their Reduction", Washington D.C., 23.-24. 9. 87 (= AGARD-CP-430).

Zimmermann, H. H., *Kroupa, E., Luckhardt, H.-D.* (1987): Das Saarbrücker Translationssystem STS - eine Konzeption zur computergestützten Übersetzung. Saarbrücken, Universität des Saarlandes: Fachrichtung Informationswissenschaft, Januar 1987.