Computergestützte Übersetzung in der Anwendung: Das Projekt MARIS und das Saarbrücker Translationssystem STS

Heinz-Dirk Luckhardt Institut für Angewandte Informationsforschung (IAI) Martin-Luther-Str.14 D-6600 Saarbrücken

Kurzfassung

Die Autoren stellen in ihrem Beitrag den im Projekt MARIS an der Fachrichtung Informationswissenschaft der Universität des Saarlandes entwickelten Service rur computergestützte Übersetzung (STS) vor. Es werden in dem vom BMFT geförderten Projekt maschinelle und intellektuelle Übersetzung in einer gemeinsamen Systemumgebung (Übersetzerarbeitsplatz) miteinander verknüpft. MARIS setzt die entwickelten Verfahren und Systeme bei der Übersetzung (Deutsch-Englisch) von Titeln, Deskriptoren und Abstracts aus deutschen Datenbanken praktisch ein. Bisher wurden ca. 2 Mio. Wörter übersetzt.

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

1 Die STS-Verfahren zur computergestützten Uber¨setzung

Die im folgenden beschriebenen Forschungs- und Entwicklungsarbeiten werden im Rahmen des Projekts MARIS (Multilinguale Anwendung von Referenz- Informationssystemen, Laufzeit: 5/85 - 12/88) an der Fachrichtung Informationswissenschaft der Universität des Saarlandes und am Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung an der Universität des Saarlandes (IAI) durchgeführt. MARIS wird vom Bundesministerium für Forschung und Technologie (Förderkennzeichen 1013209 2) gefördert. Ziel des Projekts ist die exemplarische technische und organisatorische Entwicklung eines Systems für computergestützte Übersetzungen im Bereich Fachinformation, zu dem als Rahmen das Saarbrücken Translationssystem als Service (STS) eingerichtet wurde (vgl. [Zimmermann & Kroupa & Luckhardt 1987]).

Von Projektbeginn an wurden deutschsprachige Titel aus Datenbanken ins Englische übersetzt und wieder in die Datenbanken integriert. Das Zusammenwirken von intellektueller und maschineller Übersetzung sowie die Terminologiearbeit sollen im folgenden beschrieben werden.

1.1 Ausgangslage

Ausgangsmaterial für die computergestützte Übersetzung Deutsch-Englisch sind deutschsprachige Titel und Deskriptoren in Datenbanken (seit Anfang 1988 auch Abstracts). Diese Datenbestände liegen maschinenlesbar vor, zum Teil existieren bereits Übersetzungen ins Englische bzw. aus dem Englischen ins Deutsche (und in andere Sprachen). Zu den Datenbanken gibt es in der Regel kontrollierte Vokabulare (Thesauri), mit deren Hilfe der Inhalt zusätzlich intellektuell erschlossen wird und die bislang nur teilweise mehrsprachig verfügbar sind.

LDV-Forum Bd. 5. Jg. 1987/88. Nr. 4. September 1988: S. 12 - 20

Bei fachspezifischen Übersetzungen stellt insbesondere die Terminologiearbeit den Übersetzer vor große Probleme. Demgegenüber sollen die Kosten für die Übersetzung den erheblichen Aufwand bei der Erschließung der Dokumente nicht ungebührlich erhöhen.

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

Im Terminologiebereich können zudem bei "zentraler" Sammlung und Nutzung für einen Übersetzungsservice weitere Vorteile entstehen: während der bisherigen Projektlaufzeit
von MARIS zeigte sich bereits an den Fachgebieten "Bauwesen" und "Technische Regeln", aber auch für "Bauwesen" und "Umweltschutz", daß sich erhebliche 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 Computergestützte intellektuelle Übersetzung (CAT-H)

Es werden im folgenden zwei Varianten der computergestützten Übersetzung (computeraided translation; CAT), d.h. der Unterstützung der Arbeit des Übersetzers durch computergestützte Verfahren v.a. zu Textverarbeitung und zu Terminologieverwendung, unterschieden:

(a) CAT-H: Eine vorwiegend intellektuell ge fertigte (Human-) Übersetzung;

(b) CAT-C: Eine vom Menschen durch Präedition, Interaktion und/oder Postedition gesteuerte bzw. korrigierte Übersetzung durch den Computer.

Da von Projektbeginn an Übersetzungen für deutsche Fachinformationszentren anzufertigen waren, wurde zunächst ein herkömmliches (Human-) Übersetzungsverfahren mit Computerunterstützung entwickelt und eingesetzt, gleichzeitig aber die Integration eines maschinellen Übersetzungsverfahrens in einen Übersetzungsservice konzipiert und implementiert:

(a) Das als Kernsystem für das maschinelle Übersetzungsverfahren verwendete maschinelle Übersetzungssystem SUSY wurde von der universitätseigenen Siemensanlage (Betriebssystem BS2000) auf die projekt eigene Nixdorfanlage Targon /35 (Betriebssystem UNIX) migriert.

(b) Da es sich bei SUSY um ein forschungsorientiertes System handelte, mußte eine produktionsorientierte Systemumgebung geschaffen werden.

(c) Da keines der zu bearbeitenden Fachgebiete durch das SUSY- Übersetzungswörterbuch abgedeckt war (Stand zu Projekt beginn: Testwörterbuch mit 7.000 Einträgen), mußten Verfahren zur (teil-) automatischen Lexikonerweiterung entwickelt werden (Stand des Übersetzungswörterbuchs, d.h. des Terminologiepools (4/88): 200.000 Einträge).

Die Organisation der intellektuellen Übersetzung in einem computergestützten Service - wie sie im Rahmen des Projekts MARIS für maschinenlesbare Daten entwickelt wurde läßt sich wie folgt beschreiben:

Datenübergabe

Die Auftraggeber liefern die zu übersetzenden Dokumente per Magnetband oder Diskette an. Ein Dokument ist in der Regel ein Datenbankeintrag mit einem Identifikationscode, einem Titel, einem Abstract, einem Fachgebietsschlüssel und evtl. intellektuell vergebenen Deskriptoren.

Datenumsetzung

Die Daten werden auf das STS-Format umgesetzt. Die zu bearbeitenden Dokumente werden dabei in eine Form gebracht, die eine angemessene Weiterverarbeitung erlaubt. Für den Vorgang nicht benötigte Teile der Dokumente werden ausgeblendet: Es wird ein Bereich für das Eintragen der Übersetzungen bereitgestellt. Jede Lieferung (zwischen 500 und

5000 Dokumenten) wird in einzelne "Pakete" zu 50 oder 200 Dokumenten aufgegliedert. Jedes Paket wird ausgedruckt.

Paketverteilung und -übersetzung

Jeder Übersetzer erhält eine Anzahl von Paketen zur Übersetzung. Entsprechend den implementierten Verfahren kann diese auf zweierlei Arten geschehen:

(a) Die Übersetzungen werden vom Übersetzer auf Papier vorgeschrieben und in einem zweiten Schritt im Dialog mit der Siemensanlage mit dem zeilenorientierten Editor erfaßt. (Dieses Verfahren wird nicht mehr eingesetzt).

(b) Nachdem das Projekt inzwischen über eine ausreichende Anzahl an PCs verfügt, werden die Pakete auf Disketten gelegt und mit einem bequemen Textprozessor bearbeitet.

Terminologieunterstützung

Die Übersetzer erhalten zu den Grundformen der Originaltexte alle Übersetzungsäquivalente aus dem Terminologiepool. Diese werden durch automatische Lemmatisierung der Texte und automatische Wörterbuchsuche nach dem ALT-Verfahren (automatic lemmatisation and translation) ermittelt. Die Übersetzer werden dabei auf bevorzugte Varianten (z.B. in-house-Terminologie der Auftraggeber) hingewiesen.

Datenrücksendung

N ach einer abschließenden Kontrolle werden die übersetzten Pakete wieder zu einer großen Datei zusammengefügt und an den Auftraggeber zurückgeschickt.

1.3 Computergestützte (maschinelle) Übersetzung (CAT-C)

Mit der Migration von -SUSY auf die Nixdorf-Anlage wurde 1986 die Voraussetzung für die Einrichtung des computergestützten Saarbrücker Translationsservice STS geschaffen. Weitere Voraussetzungen bzgl. der Präedition, der Auswahl von Übersetzungsäquivalenten und der Postedition waren zu erfüllen:

1.3.1 Präedition

Bevor die Daten maschinell übersetzt werden konnten, mußten drei Probleme gelöst werden: Beseitigung von Rechtschreibfehlern, Auflösung der Mehrdeutigkeit Abkürzungspunkt/Satzendepunkt, Selektion fremdsprachiger Titel.

Die sprachlich-formale Qualität der angelieferten Daten läßt z. T. zu wünschen übrig. Für den menschlichen Übersetzer bedeuten Rechtschreibfehler in der Regel keine unüberwindlichen Hindernisse. Die MÜist jedoch auf vollkommene Korrektheit des Eingabetextes angewiesen. Alle zu bearbeitenden Titel werden demzufolge mit der automatischen Rechtschreibhilfe PRIMUS überprüft.

Bezüglich der *Punkte,* die in den u. U. aus mehreren Teilen bestehenden Titeln vorkommen, ergibt sich das Problem der Entscheidung "Satzendepunkt oder Abkürzungspunkt?". Derzeit in der Erprobung befindliche automatische Verfahren werden zur gegebenen Zeit in die Systemumgebung integriert. Bis dahin führen die Übersetzer die Vereindeutigung in einem halbautomatischen Verfahrensschritt durch.

Bei einigen Anwendern kommen in den angelieferten Daten außer den deutschen auch fremdsprachige Titel vor (außer in Französisch u.a. auch in Italienisch, Spanisch, Finnisch etc.), die von der Übersetzung ausgeschlossen werden müssen. Auch dies geschieht z. Zt. intellektuell, für später sind Sprachidentifikationsverfahren vorgesehen.

1.3.2 Auswahl von Übersetzungsäquivalenten

Mit dem Anwachsen des Terminologiepools stellt sich mehr und mehr das Problem der Auswahl zwischen mehreren verschiedenen Übersetzungsäquivalenten (vgl. [Luckhardt 1987]), die im Laufe des Projekts teils aus konkreten Übersetzungen, teils aus maschinenlesbar vorliegenden Sammlungen gewonnen wurden (vgl. 2.). 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. in Fach- oder Themengebiete eingeordnet. Dies bedeutet *an sich* eine wertvolle Unterstützung bei der Disambiguierung. 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 den.en er es für sinnvoll erachtet. Dies steht den Anforderungen einer *allgemeinen Klassifikation* entgegen, wie sie für die Ordnung von Fachgebieten innerhalb eines maschinellen Übersetzungswörterbuchs für verschiedene Anwender/Fachgebiete (eines Terminologiepools) sinnvoll erscheint. Dazu das folgende Beispiel:

Das Umweltbundesamt hat die folgende Grobklassifikation gewählt (die "Umweltgebiete" d.h. die "Säulen" der Klassifikation):

LU LUFT

WA WASSER

BO BODEN

NL NATUR UND LANDSCHAFT

LF LAND- UND FORSTWIRTSCHAFT / NAHRUNGSMITTEL

AB ABFALL

LE LÄRM/ERSCHÜTTERUNGEN

CH UMWELTCHEMIKALIEN/SCHADSTOFFE

SR STRAHLUNG

EN ENERGIE UND ROHSTOFFE

UW UMWELTÖKONOMIE

UA ALLGEMEINE UND UBERGREIFENDE UMWELTFRAGEN

Diese Gebiete sind in zwei Hierarchiestufen weiter klassifiziert (d.h. facettiert), z.B.: "Boden" in:

BO 50 technische Bodenschutzmaßnahmen

BO 60 planerische Bodenschutzmaßnahmen

BO 70 Theorie, Grundlagen und allg. Fragen

BO 71 Bodenkunde und Geologie

BO 72 Bodenbiologie

Das heißt: Die Fachgebiete selbst (wie Biologie, Geologie etc.) tauchen auf einer niedrigen Hierarchieebene auf, und zwar an verschiedenen Stellen der Klassifikation. Das Umweltbundesamt hat einen anderen Querschnitt durch Wissenschaft und Technik angesetzt, als es z.B. das Informationszentrum RAUM und BAU tut. Dermaßen verschiedenartige Klassifizierungen sind nur schwer miteinander bzw. mit einer allgemeinen Klassifizierung zu kompatibilisieren.

(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 aber natürlich 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. (vgL auch [Luckhardt 1987]).

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 dank seines Weltwissens, seines Assoziationsvermögens, seiner Phantasie etc. dem Computer - der z.Zt. allein auf linguistisch-formaler Ebene entscheidet schwer formalisierbare intellektuelle Leistungen voraus. Ein Beispiel:

In einem zweisprachigen Glossar ist" Altstadt" mit "old parts of a town" übersetzt. Dieses Äquivalent wird in ein deutschenglisches Computerlexikon übernommen.
Damit würde aber der Computer die Phrase "Altstadt von Saarbrücken" mit "old parts of a town of Saarbrücken" übersetzen. Ein anderes Beispiel:

"ENTSORGUNG" =} "WASTE DISPOSAL"

"ENTSORGUNG VON ABFALL" =} "WASTE DISPOSAL OF WASTE" ?

Solche Probleme lassen sich zwar auch im Rahmen des vorliegenden Systems lösen, jedoch nur mit einem erheblichen Aufwand an Regelformulierung und Lexikonkodierung.

1.3.3 Postedition

Die maschinell übersetzten Titel werden von Übersetzern an PCs postediert. In einer ersten Implementation des Posteditionsverfahrens wurden die MÜ- Ergebnisse auf Disketten gelegt und vom Posteditor mit dem Textprozessor WordStar2000 direkt bearbeitet. Dabei entstanden zwei Probleme:

- Nicht übersetzte Wörter mußten während des Posteditionsvorgangs recherchiert werden, so daß der PC unnötig lange blockiert wurde;
- Übersetzungsvorschläge des Computers mußten darauf überprüft werden, ob die Inhouse-Terminologie des Auftraggebers berücksichtigt war.

In einer zweiten Implementationsstufe wurden diese Probleme inzwischen wie folgt gelöst:

- Nicht übersetzte Wörter bzw. Übersetzungsvorschläge des Computers für Komposita und abgeleitete Wörter werden gesondert ausgegeben und können vom Posteditor in einer Recherchesitzung vor der Posteditionsphase übersetzt bzw. überprüft werden;
- bei der automatischen Auswahl von Übersetzungsäquivalenten wird jetzt der Benutzercode ausgewertet, so daß der Posteditor sicher sein kann, daß im maschinellen Output die Termini des Anwenders berücksichtigt sind.

2 Der STS-Terminologiepool

Der STS- Terminologiepool umfaßt derzeit (April 1988) ca. 200.000 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 vorliegen den Übersetzungen;
- titelunabhängige Übersetzung von z.B. Schlagwortverzeichnissen bzw. Thesauri.

2.1 Terminologiegewinnung

2.1.1 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) haus eigene deutsch- englische Terminologiesammlungen in das entsprechende STS-Computerlexikon überspielt. Diese Sammlungen sind allerdings nicht ohne Anpassungen zu übernehmen. Einige Probleme sollen kurz angedeutet werden:

Derartige Einträge sind nicht immer als Übersetzungsäquivalente zu verstehen und zu

benutzen. Dazu einige Beispiele aus den Quellen:

MESSWESEN > MEASUREMENT, TESTING, AND INSTRUMENTS

UMRECHNUNG > CONVERSION (UNITS OF MEASUREMENT)

ORTUNGSGERÄT > DETECTORS

KIPPEN > BUCKLING

Zum einen handelt es sich oft um Begriffe, die (z.B. durch Klammerausdrücke) genauer spezifiziert sind, weil sie sonst als *Deskriptoren* (und das ist ja ihr eigentlicher Zweck) nicht aussagekräftig genug sind. Zum anderen werden die englischen Äquivalente entsprechend dem anglo-amerikanischen Usus nach Möglichkeit im Plural wiedergegeben. Dies ist jedoch für ein MU-Lexikon problematisch, weil hier *Grundformen* bzw. Stämme erwartet werden. Es kann aber erst über ein Deflektionsverfahren entschieden werden, ob ein auslautendes *-s* eine Pluralen dung ist oder zum Wortstamm gehört. Aus einem Verb der Quellsprache ("kippen") ist ein Verb der Zielsprache abzuleiten ("buckle") und nicht ein Gerundium ("buckling"). (Daß die Substantivierung - hier: das "Kippen" - gemeint ist, ist eine andere Frage, die wiederum aus der unterschiedlichen Problemstellung abgeleitet ist).

2.1.2 Extraktion aus vorliegenden intell. 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. Diese können dann mit dem Terminologiepool verglichen 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.

2.1.3 Titelunabhängige Terminiübersetzung

Seit Mai 1987 wird im laufenden Projekt MARIS das deutsche Stich- und Schlagwortverzeichnis des Deutschen Patentamts zur internationalen Patentklassifikation IPC (ca. 130.000 Begriffe) ins Englische übersetzt. Nach Fertigstellung dieser Arbeit wird ein umfassender fachsprachlicher Wortschatz zur Verfügung stehen, der weitgehend die Grundbegriffe der gesamten Technik abdeckt.

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.

Ein Problem stellen echte Synonyme dar, die z.B. ins Spiel kommen, wenn in der in gedruckter Form mit herangezogenen englischen Version der IPC Begriffe an verschiedenen Stellen verschieden benannt werden, z.B. "baler" oder "baling preß" für "Ballenpresse" . Hier kann keine eindeutige Entscheidung zugunsten des einen oder des anderen Aquivalents getroffen werden.

Ein weiteres Problem wurde schon oben angedeutet: Die Einträge in den Thesauri/Stichwortverzeichnissen der MARIS-Anwender stellen *Deskriptoren* dar, wie sie für das Information Retrieval vergeben werden. So kommen, z.B., im Thesaurus der DOMA-Datenbank (Fachinformationszentrum Technik: Maschinenbau) die folgenden Einträge vor:

Abfall (Fertigung)
Abfall (Kerntechnik)
Abfall (Land wirtschaft)
Abfall (Müll)
Abfall (Papier)
Abfall hochaktiv
Abfall mittel aktiv
Abfall schwach aktiv

Die Differenzierungen sind in diesen Fällen für den Aufbau des Terminologiepools irrele-

vant, da es sich in allen Fällen um "refuse" ("waste") handelt. Klammerausdrücke in der Quelle können allerdings auch der terminologierelevanten Bedeutungsdifferenzierung dienen, wie in den folgenden Beispielen:

Anhänger	⇒	supporter
Anhänger (Fahrzeug)	⇒	trailer
Anker (Schiff)	⇒	anchor
Anker (el. Maschine)	⇒	armature

2.2 Fachgebietsklassifikation

Die Problematik von Fachgebietsklassifizierungen ist bereits zur Sprache gekommen. Es erschien wenig sinnvoll, die Disambiguierung in STS unmittelbar auf den Anwenderklassifikationen aufzubauen. Dies würde v.a. 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 aus den oben (vgl. 1.2.2) genannten und aus einigen weiteren Gründen verzichtet. Am Beispiel der Internationalen Patentklassifikation IPC soll das erläutert werden.

Die IPC ist in Sektionen, Untersektionen, Klassen, Unterklassen, Haupt- und Untergruppen unterteilt, z.B.:

Sektion **A**: täglicher Lebensbedarf
Untersektion "Nahrungsmittel und Tabak"
Klasse **A21**: Backen; eßbare Teigwaren
Unterklasse **A21D**: Behandeln von Mehl oder Teig, Backverfahren; Bäckereierzeugnisse; deren Haltbarmachung
Hauptgruppe **A21D 15/00**: Konservieren von Bäckereierzeugnissen
Untergruppe **A21D 15/02**: Konservierung von Bäckereierzeugnissen durch Kühlen

Es ist offensichtlich, daß nicht die gesamte Klassifikation für die maschinelle Übersetzung übernommen werden könnte, wenn man bedenkt, daß die IPC 8 Bände mit bis zu 300 Seiten Umfang pro Band umfaßt. Die meisten Begriffe des Lexikons müßten u.U. Hunderten von Untergruppen zugeordnet werden. Selbst die Hauptgruppen und Unterklassen geben eine zu feine Klassifizierung ab, da z.B. bestimmte Maschinenteile intellektuell verschiedenen Unterklassen zugewiesen werden müßten. Die Kodierung neuer Lexikoneinträge würde unverhältnismäßig viel Zeit kosten, und der Nutzen wäre überdies zweifelhaft, da es keine Texte gibt (außer den Texten im Patentwesen), die entsprechend markiert sind, so daß ein Vergleich der Markierung des Textes und der Lexikoneinträge zur eindeutigen Zuweisung von zielsprachlichen Äquivalenten führen könnte.

Oberhalb der Unterklassen sind in der IPC-Hierarchie die "Klassen" angesiedelt. Selbst hier gibt es noch unerwünschte Überschneidungen, z.B. zwischen A22 (u.a. Fisch- und Fleischverarbeitung) und A23 (u.a. Fisch- und Fleischkonservierung); A21 (u.a. Backöfen), A47 (u.a. häusliche Backausrüstung), F23 (Feuerungen) und F24 (u.a. häusliche Heiz- oder Kochherde, die ganz oder teilweise als Backöfen ausgebildet sind).

Wohlverstanden: dies ist auch hier keine Kritik an der IPC, die "vor allem als ein wirksames Recherchenwerkzeug für das Wiederauffinden von Patentdokumenten durch die Patentämter und sonstigen Anwender, zur Feststellung der Neuheit und zur Beurteilung der Erfindungshöhe ... von Patentanmeldungen" und einigen weiteren Zwecken (vgl. [IPC (1979)], Band 9, S.7) dient. Für die allgemeine Klassifizierung von Lexikoneinträgen und Texten zum Zwecke der maschinellen Übersetzung scheint jedoch vorerst nur die zweitoberste Hierarchieebene geeignet, nämlich die der "Untersektionen", die in einigen Fällen in Klassen aufgeteilt werden kann, z.B. die Untersektion "Transportieren" in die "Klassen" *Eisenbahnen, Gleislose Landfahrzeuge, Schiffe, Luftfahrzeuge, Transportverfahren und Sattlerei/Polsterei.* Dies schließt übrigens nicht aus, daß es in Einzelfällen sinnvoll sein kann, zur Disambiguierung eines spezifischen Wortes auch eine feinere Klassifikation mit heranzuziehen, soweit dies in einem solchen Falle zur Verbesserung der Übersetzungsergebnisse führt.

Auf jeden Fall bietet *jede Art von Klassifikation* "Angriffsflächen" für den genauen Betrachter. Es kommt jedoch immer auf die Verwendung der Klassifikation an. Ein menschlicher Anwender (z.B. ein Prüfer des Patentamts), wird dank seiner Intelligenz, Assozia-

tionsfähigkeit, vor allem aber seines Weltwissens mit einer Klassifikation eher zurechtkommen, als die Maschine, die auf genau umrissene Klassen angewiesen ist, solange es nicht möglich ist, *Klassenübergänge* zu formalisieren.

Die STS-Klassifikation ist sachgebietsorientiert und hierarchisch aufgebaut. Nicht alle Fachgebiete, sondern nur solche, für die heute schon Daten vorliegen, sind vertreten (vgl. [Luckhardt 1987a]).

2.3 Nutzung des Terminologiepools DEENWO

Unter Terminologiepool wird in MARIS die für die maschinelle Übersetzung zugreifbare Terminologie für alle Fachgebiete/Anwender verstanden. Daneben existieren zwei dbase-Datenbanken (Sozialwissenschaften bzw. Raum und Bau) für die intellektuelle Übersetzung, die von Zeit zu Zeit mit dem Terminologiepool kompatibilisiert werden.

Auf den Terminologiepool greifen verschiedene Verfahren zu:

a. die automatische Übersetzung von Deskriptoren aus Datenbanken

b. die Wortübersetzung in der Teilkomponente "Transfer" der maschinellen Übersetzung von Titeln und Abstracts

c. die automatische Indexierung CTX und anschließende Übersetzung der erzeugten Deskriptoren

U.U. reicht die Angabe des Auftraggebers 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 das gewählt, dessen Fachgebietscode mit dem des Dokuments/Textes übereinstimmt.

2.4 Umfang der STS-Systemlexika

Die STS-Systemlexika haben derzeit (April 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 PC-Datenbanken für Übersetzer:

1. dBaseIII (IZ,IRB,DITR)

   17.000 Einträge

2. GOLEM (IRB und DITR)

   13.000 Einträge

3 Ausblick

Der Erfolg maschineller Übersetzungssysteme wird m.E. in Zukunft von drei Faktoren abhängen (vgl. auch [Zimmermann 1987]):

- der Qualität des MÜ-Kernsystems
- der Qualität der Systemumgebung
- der Qualität der Terminologieunterstützung

In allen drei Bereichen besteht bzgl. Forschung und Entwicklung ein erheblicher Nachholbedarf, der erwarten läßt, daß die legendäre "FAHQT" (fully automatic high quality translation) erst im nächsten Jahrhundert verfügbar sein wird. Bis dahin können die verfügbaren Systeme eingesetzt werden,

für die ins besondere in den beiden zuletzt genannten Bereichen noch sehr viel getan werden kann. Die Qualität der MÜ-Kernsysteme - d.h. ihre computerlinguistische Basis und ihre daraus resultierenden Fähigkeiten der Verarbeitung und Disambiguierung natürlicher Sprachen - kann durchaus noch verbessert werden. Hier sind jedoch besonders die in der Entwicklung befindlichen Systeme wie EUROTRA und andere computerlinguistische Forschungsprojekte gefordert. Verbesserungen in den Bereichen *Terminologieunterstützung* und *Systemumgebung* werden die bestehenden Systeme allgemeiner verwendbar und v.a. auch weiteren Anwendern verfügbar machen. Darüberhinaus werden die Ergebnisse dieser Arbeiten weitestgehend auch rur Systeme der nächsten Generation nutzbar sein.

Literaturverzeichnis

[IPC (1979)] IPC (1979): Internationale Patent klassifikation. München: Carl-Heymanns . Verlag

[Kroupa 1982] Kroupa, E.: Strategien der Dokumentrepräsentation bei CTX. Ein Verfahren zur computergestützten Texterschließung und Textwiedergewinnung. In: I. Batori, S. Krause, H.-D. Lutz (Hrsg.). Linguistische Datenverarbeitung. Sprache und Information Band 4, Tübingen, 155-161, 1982

[Luckhardt 1987] Luckhar4t, H.-D.: Probleme bei der Auswahl von Übersetzungsäquivalenten. In: [Zimmermann&Kroupa&Luckhardt, 1987]

[Luckhardt 1987a] Luckhardt, H.-D.: Terminologieerfassung und -nutzung im computergestützten Saarbrücker Translationssystem STS. Veröffentlichungen der Fachrichtung Informationswissenschaft. Saarbrücken: Universität des Saarlandes, 1987

[Luckhardt 1987b] Luckhardt, H.-D.: Der Transfer in der maschinellen Sprachübersetzung. Tübingen: Niemeyer, 1987

[Luckhardt 1988] Luckhardt, H.-D.: Generation of Sentences from a Syntactic Deep Structure with a Semantic Component. In: McDonald/ Bole (Hrsg.). Natural Language Generation Systems. Reihe Symbolic Computation. New York: Springer, 1988

[Peters 1987] Peters, J .P.: Die Mono-/Multilingualität von Datenbanken. In: [Zimmermann&Kroupa&Luckhardt, 1987], 10-22

[Zimmermann 1987] Zimmermann H.H.: 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. September 1987, 1987

[Zimmermann & Kroupa & Luckhardt 1987] Zimmermann, H. H., E. Kroupa, H.-D. Luckhardt: Das Saarbrücker Translationssystem STS - Eine Konzeption zur computergestützten Übersetzung. Veröffentlichungen der Fachrichtung Informationswissenschaft. Saarbrücken: Universität des Saarland es, 1987