Harald H. Zimmermann

Stand und Perspektiven der Sprachtechnologie mit dem Beispiel der Maschinellen Übersetzung

LEND.DOC 2002-09-08

Übersicht

Nach einigen generellen Bemerkungen zum Gegenstandsbereich ‚Sprachtechnologie' wird ein kurzer Rückblick zu den Anfängen sprachtechnologischer Forschung und Entwicklung v.a. in Deutschland gegeben, die die ‚Generation', zu der Winfried Lenders und ich gehören, als angehende Wissenschaftler erlebt hat. Ein Stück dieses Weges sind wir beide übrigens als Herausgeber der Zeitschrift ‚Sprache und Datenverarbeitung' gemeinsam gegangen. Der Themenbereich der ‚Maschinellen Übersetzung' hat uns bis heute begleitet, ohne dass hier bislang ein Durchbruch – v.a. mit Blick auf eine qualitativ hoch stehende Übersetzung für ‚beliebige' Texte – erreicht wurde. Es gibt dafür sicherlich mehr als einen Grund. Vor allem sind die immensen Kosten zu nennen, die mit der Entwicklung eines Übersetzungssystems verbunden sind, aber auch die letztlich unzureichenden oder zumindest unbefriedigenden Konzepte und Verfahren, deren Anwendung immer wieder damit begründet wurde, einen hinreichend hohen ‚Prozentsatz' an vorkommenden Sätzen / Texten (etwa im Bereich der Dokumentation) zureichend übersetzen zu können. Im letzten Abschnitt werden einige Anregungen zu zukünftigen Entwicklungen vorgestellt, die angesichts der wachsenden weltweiten Kommuniationsmöglichkeiten zumindest die Chance eröffnen, den Anteil des Einsatzes der Maschinellen Übersetzung deutlich zu steigern.

Rahmensetzung

In *naturwissenschaftlich* basierten Technik-Bereichen, wie etwa der Kraftfahrzeugtechnik, der Luft- und Raumfahrttechnik oder der Meerestechnik, lassen sich Theorien wie Entwicklungen in aller Regel auf – meist physikalischen – Gesetzmäßigkeiten aufbauen, die mit entsprechenden Messungen ermittelt bzw. mit Messgeräten exakt überprüft werden können.

Sobald es um den *Einfluss der Technik auf den Menschen* als 'Objekt' oder Betroffenen geht, werden empirische Erhebungen und Aussagen schon schwieriger. Beispiele sind die Medizintechnik und die Gentechnik, aber auch die Kerntechnik und die Umwelttechnik, hier z.B. mit der Problematik der Festlegung von Grenzwerten bei Schadstoff-Emissionen.

Die *Sprachtechnik* – als *Werkzeug* zur Verarbeitung natürlicher Sprache – ist grundsätzlich in einem Dilemma: Die Grundlagen des Funktionierens *menschlicher* Sprache – als Fähigkeit der (sinnvollen) Sprache-Erzeugung und insbesondere des Sprache-*Verstehens* – sind noch weitgehend unerschlossen. Obwohl man mit entsprechenden psychologischen Experimenten und auch aufgrund von Auswirkungen von Hirnschädigungen auf Sprach-Erzeugung und Sprache-Verstehen bestimmte Teile des Gehirns als für die Sprache 'zuständig' identifizieren kann, weiß man noch zu wenig über den Prozess bzw. die 'Speicherung' sprachlicher Daten oder Regeln, auch wenig über den Zusammenhang von 'Wissen' oder 'Glauben' und 'Meinen' (also dem *Bild*, das sich jemand von der 'Welt' macht) und dem zugehörigen *sprachlichen* Teil, also dem, was letztendlich in sprachlichen Äußerungen gegenüber der Umwelt zutage tritt (Wörter, Mimik, Gestik ...).

Wenn man also schon über die *Grundlagen* der Sprache(n) und des Sprachverstehens mehr oder weniger nur spekulieren kann – wissenschaftlich(er) ausgedrückt: sich nur mehr oder weniger exakte *Modelle* davon machen kann –, ist es schwer, *Standards* oder Normen zu entwickeln, die von einer größeren Wissenschaftsgemeinschaft akzeptiert werden. Zwar hatte schon Noam Chomsky in den 60er Jahren ein Kriterium, das allgemein für die Bewertung von Modellen gilt, auch in die Sprachwissenschaft eingeführt. Es lautet – frei übersetzt –: Ein *Mo-*

dell A, das ein Sprachsystem (etwa *gemessen* an den möglichen *korrekten* Sprachäußerungen einer bestimmten Sprache) *umfassender* beschreibt als ein *Modell B*, sei diesem vorzuziehen. Ein *Modell A'*, das bei gleicher *Beschreibungsstärke* ein Sprachsystem besser *erklärt* als ein *Modell B'*, sei wiederum diesem vorzuziehen (*Erklärungsstärke*). Erst wenn zwei Modelle (bzw. Verfahren) in diesen Bereichen die gleiche 'Stärke' haben, sei dasjenige vorzuziehen, das weniger Regeln (Aufwand) dazu benötigt.

Dazu ein kleines Beispiel: Wenn zwei (Grammatik-)Systeme die Sätze 'Peter singt ein Lied' bzw. 'Ein Lied wird von Peter gesungen' 'generieren' und 'analysieren' können, ist dasjenige überlegen, das den *Zusammenhang* zwischen beiden Sätzen 'erkennt' (Aktiv – Passiv). Wenn ein System (z.B. anhand sog. 'semantischer' Merkmale) 'verhindert', dass der (in der normalen 'Welt' – nicht z.B. in der Märchenwelt' –) ungrammatische Satz 'Der Topf singt ein Lied' erzeugt wird (bzw. von dem System als ungrammatisch 'erkannt' wird), ist es einem System überlegen, das dies nicht unterscheiden kann bzw. einen solchen Satz erzeugt.

Mit der Umsetzung derartiger Konzepte kann man in der Sprachtechnik schon ziemlich weit kommen. Allerdings gibt es ein großes – vielleicht entscheidendes – Problem: 'Bedeutungen' – also das, was im menschlichen Gehirn ein-eindeutig differenziert ‚verfügbar' ist – und die Möglichkeit einer ebenso ein-eindeutig differenzierten sprachlichen *Äußerung* sind nicht deckungsgleich: die Sprache als *System* ist – sei es aus sprachhistorischen, sei es aus sprachökonomischen Gründen – nicht in der Lage, diese *begrifflichen* Ausdifferenzierungen auch in den *Bezeichnungen* (d.h. dem, was in der zwischenmenschlichen Kommunikation über Laute oder auch Texte vermittelt wird) ein-eindeutig umzusetzen. Einzelne Bezeichnungen, ja manchmal auch ganze Sätze, können daher – für sich genommen – *mehrdeutig* sein.

Der Mensch ist meistens in der Lage, diese 'lokale' Mehrdeutigkeit (einer Bezeichnung, etwa des Wortes 'Schloss') aufzulösen. Dazu reicht *sprachliches* Wissen (d.h. die Kenntnis der Regeln, wie Sprache funktioniert) oft nicht aus, vielmehr ist ‚Weltwissen' einbezogen, d.h. die Verbindung des ‚Gesagten' mit dem (inhaltlich) ‚Gemeinten'. Vielfach ist dieser 'Mechanismus' so internalisiert, dass der Betreffende diese Mehrdeutigkeit nicht einmal bemerkt und erst durch ungewöhnliche (im Grunde 'fehlerhafte') Formulierungen darauf aufmerksam wird. Der Komiker Heinz Erhard hat beispielsweise dieses Phänomen benutzt, wenn er sagte: "Ich heiße Heinz Erhard und Sie herzlich willkommen". Hätte er nur gesagt "Ich heiße Sie herzlich willkommen", wäre die Mehrdeutigkeit von 'heißen' kaum jemandem aufgefallen. Niemand wird auch bei dem 'Verstehen' der folgenden Sätze Probleme haben: 'Der Kellner brachte die Suppe heiß herein' bzw. 'Der Kellner brachte die Suppe schnell herein': Im ersten Satz ist (im ‚Normalfall') die 'heiße Suppe' gemeint, im zweiten Satz wohl in erster Linie die Tatsache, dass es nach der Bestellung nicht lange gedauert hat, bis die Suppe kam (der Kellner kann sie durchaus normalen Schrittes gebracht haben).

Wir Menschen sind dabei nicht nur in der Lage, in den meisten Fällen diese Mehrdeutigkeiten (und viele andere mehr) *im Kontext unseres Wissens* zu *vereindeutigen* bzw. zu *erkennen* (manches Mal ist die Mehrdeutigkeit ja gewollt und manchmal muss man ‚zwischen den Zeilen lesen'), sondern wir haben auch die Fähigkeit, fehlerhafte Äußerungen zu *'korrigieren'*. Dabei kann es sich um Fehler eher oberflächlicher Natur handeln, z.B. Tippfehler oder Kommafehler in Texten, aber auch um sinnentstellende Fehler. Der Tippfehler 'Tippfeler' wird leicht 'als solcher' erkannt, bei dem Satz "Die Mehrwortsteuer wurde um 1 Prozent erhöht" kann man hoffen, dass der Staat nicht auch noch auf das Sprechen von mehreren Wörtern eine Steuer erhebt, und das Wort 'Mehrwortsteuer' getrost als Schreibfehler für 'Mehrwertsteuer' ansehen (wenn man den Fehler beim ersten Lesen nicht sogar 'überlesen' hat). Auch mehr oder weniger 'verhunzte' Sätze wie "Err Standd aufdrer Trepe" lassen sich noch einigermaßen zuverlässig rekonstruieren. Die Äußerung "Gib mir kein Geld zurück" ist demgegenüber zwar

etwas ungewöhnlich (wahrscheinlicher ist "Gib mir mein Geld zurück"), aber situativ durchaus möglich.

Was hier am Beispiel der *geschriebenen Sprache* verdeutlicht wurde, gilt in noch größerem Maße für die *gesprochene Sprache.* Hier wird man – bis zu gewissen Grenzen – mit Lärm-Umgebungen, mit Wort-Verschleifungen, mit fremdsprachigen Akzenten, mit regionalen Aussprache-Besonderheiten 'fertig', immer *bemüht*, den Gesprächspartner 'zu verstehen'.

All diese Fähigkeiten erwirbt man – zumindest für die Muttersprache – im Laufe der ersten Lebensjahre, wobei – und dies ist besonders wichtig – *alle Sinnesorgane* mitwirken: es handelt sich also um einen sehr komplexen, langandauernden *Lernprozess*. Was davon (etwa strategisch gesehen) schon angeboren ist oder ebenfalls ‚erlernt' wird, kann hier dahingestellt bleiben. Dass sich beim einzelnen Menschen schon bald so etwas wie ein ‚Sprachmodell' entwickelt, kann man schon daran erkennen, dass sehr früh versucht wird, Wörter systematisch zu bilden (bei Kleinkindern z.B. ‚gegeht' statt ‚gegangen', d.h. Anwendung einer Art ‚Partizip-Bildungs-Regel').

Auch wenn sich jeder Mensch letztendlich sein eigenes 'Bild' von der Umwelt aufbaut, so verfügt er doch im Laufe der Zeit über ein (Sprach-)System, das ihm erlaubt, sich mit anderen Menschen unter Verwendung von Gesprochenem oder Geschriebenem auszutauschen, d.h. sein Wissen (Glauben, Meinen ...) weiterzugeben und externes Wissen auf diesem Wege auch einzuwerben. Dass der Mensch bei 'unbekannten' Wörtern auch etwas 'spekuliert', gehört dazu. Wenn ein kleines Kind beispielsweise einen 'Wolkenkratzer' für eine Art Flugzeug hält, wird man ihm eher sinnvolles Schlussfolgern denn Dummheit unterstellen. Ja diese Fähigkeit des Verstehens geht noch weiter: Da bei jeder Äußerung mehr oder weniger auch das jeweils subjektiv erworbene ‚Weltbild' mitschwingt und man sich der damit verbundenen ‚Differenzen' bewusst ist (dies gilt in besonderem Maße für das Verstehen von Äußerungen in einer anderen Sprache als der Muttersprache), wird davon so gut es geht ‚abstrahiert' (sozusagen ‚interpoliert') mit dem Ziel, eine möglichst übereinstimmende Abbildung der externen Aussagen mit dem eigenen (internen) System zu erreichen.

Wenn nun die ‚Umgebung' (auch im praktischen Sinn: das, was in der ‚Umgebung' eines einzelnen Menschen von Bedeutung ist) einen so großen Einfluss auf das persönliche Sprachsystem hat (spezifische dialektale Ausprägungen, Wortschatz, Sprachstil, Themenbereiche, ‚Kulturen' ...), so ergeben sich innerhalb einer (eher virtuell gedachten) Sprachgemeinschaft (etwa der deutschen Sprache) Unterschiede je nach Region oder gesellschaftlicher Gruppenzugehörigkeit, die zu Un- oder Missverständnis führen können. So dürfte es es kaum einen Angehörigen der deutschen Sprachgemeinschaft geben, der den Satz „Der Boofke alfanzt mit dem Schwiemel" komplett ‚versteht' (es sei denn, er ist weit gereist). Überträgt man die regional spezifischen Wörter ins ‚Allgemeindeutsch', so wird er allgemeiner verständlich: „Der Narr treibt seine Possen mit dem Trunkenbold".

Ist vor diesem Problem-Hintergrund und angesichts der Komplexität natürlicher Sprachen so etwas wie 'Sprachtechnik' überhaupt machbar? Vor wenigen Jahren hätte mancher an der *prinzipellen* Machbarkeit noch stark gezweifelt, auch wenn immer wieder in den Medien ein 'Durchbruch' etwa bei der *maschinellen Text-Übersetzung* oder bei der *automatischen Erkennung fließend gesprochener Sprache* als kurz bevorstehend angekündigt wurde und wird.

Mit diesen beiden Themen – der *Sprachübersetzung* und der *Spracherkennung* i.S. einer Erkennung gesprochener Sprache – sind natürlich die *Highlights* der Sprachtechnik angesprochen. Ein drittes Höchstziel ist die automatische 'Verdichtung' komplexer Texte bzw. Darstel-

lungen auf *Zusammenfassungen* (das sog. "Abstracting'). Selbst wenn man eine oder mehrere fremde Sprachen einigermaßen fließend beherrscht, selbst wenn man ein Befürworter der (zusätzlichen) Benutzung einer weltweiten Verkehrssprache (heute Englisch) ist, so stellt man sich doch vor, dass ein Werkzeug, mit dessen Hilfe ein Text von einer natürlichen Sprache in eine andere maschinell übersetzt wird, der weltweiten Kommunikation dienlich ist, wobei – gerade im vielsprachigen Europa – sozusagen als ‚Nebeneffekt' ein Beitrag dazu geleistet werden kann / könnte, die Vielfalt der Kulturen zu bewahren (zumindest als theoretische Annahme oder auch Motivation).

Insbesondere der Erkennung *gesprochener* Sprache (etwa im Sinne eines Verstehens von *Anweisungen*, z.B. die Teilnehmer-Anwahl beim Telefonieren im Auto; die Umsetzung gesprochener Sprache in Schriftsprache) gilt dabei als ein besonderer Meilenstein. Zur *Kommunikation mit dem Computer*, aber auch bei jeder textuellen Kommunikation, benötigt man heute noch eine *Tastatur* bzw. eine Hand-Eingabe. Auch wenn es wohl auf lange Sicht Situationen gibt, in denen man sich der Tastatur bedienen muss oder sollte, wäre eine Kommunikation über gesprochene Sprache, entweder mit direkter 'Wirkung' oder über eine automatische Textgenerierung, eine allgemein nützliche Lösung, die zudem dazu beitragen kann (wiederum zumindest theoretisch), demjenigen eine Chance zu geben, der aufgrund besonderer Umstände (Behinderung) die Fertigkeit des Schreibens nicht hinreichend beherrscht.

Natürlich hat jemand, der ein Wort lesen oder schreiben kann, damit noch nicht notwendig gezeigt, dass er auch den *Sinn* verstanden hat. Dies lässt sich bei Grundschulkindern sehr gut zeigen. Entsprechendes gilt für den Sinnzusammenhang bei sprachlichen Äußerungen. Wenn jemand in einer entsprechenden Situation sagt ‚es zieht', dann ist – wie schon oben angesprochen – ‚eigentlich' die Aufforderung gemeint, ein Fenster oder eine Tür zu schließen. In diese Kategorie der Probleme gehört auch das Zwischen-den-Zeilen-Lesen.

Aus all dem lässt sich ableiten, dass es nicht leicht sein wird, sprachverarbeitende Systeme zu entwickeln, die *alle* diese Probleme beherrschen oder zumindest durch eine Art ‚Lernmechanismus' an die gegebenen Kommunikationssituationen adaptiert werden können.

Wegbereiter der elektronischen Sprachverarbeitung und -forschung in Deutschland

Zu Beginn der 60-er Jahre wurden erstmals in Deutschland auch Computer für Forschungszwecke in der Erforschung der natürlichen Sprache eingesetzt. So entwickelte *Hans Eggers*, soeben als Professor für Mediävistik (!) an die Universität des Saarlandes berufen, Anfang der 60er Jahre des 20. Jahrhunderts die Absicht, eine neue deutsche Grammatik (im Syntax-Bereich) auf der Basis umfassender *statistischer* Erhebungen zu erstellen: Zur damaligen Zeit waren die traditionellen Grammatiken *präskriptiv*, d.h. sie gaben vor, wie man schreiben *sollte*, und nicht *deskriptiv*, d.h. sie zeigten nicht auf, wie tatsächlich (üblicherweise) geschrieben wurde. Grundlage der Untersuchungen waren Auszählungen aus zwei Bereichen: ausgewählte Texte aus Rowohlts Deutscher Enzyklopädie und ausgewählte Artikel von Autoren der Frankfurter Allgemeinen Zeitung (F.A.Z.), wobei angenommen wurde, dass damit zumindest das 'Hochdeutsche' einigermaßen 'getroffen' wurde (wohl wissend, dass dies nicht 'repräsentativ' im statistischen Sinne sein konnte). Eggers war einer der Ersten, die dabei auch ‚Elektronenrechner' einsetzten: die Daten (Wortklassen und Satzstrukturen der ausgewählten Sätze) wurden kodiert auf Lochkarten übertragen, die Auswertungen erfolgten zunächst über Hollerithmaschinen (Sortierer und Tabellierer), später wurden im sog. 'Deutschen Rechenzentrum' in Darmstadt auch Magnetbänder eingesetzt und Großrechner zum Sortieren genutzt: Um 100.000 Sätze (Kodierungen) zu sortieren, liefen die Magnetbänder damals mehrere Stunden:

heute würde man dazu wenige Minuten oder gar nur einige Sekunden auf dem PC benötigen). Die Ergebnisse hatten unmittelbare Auswirkungen auf die Gestaltung der modernen Grammatiken. Ein Beispiel sind die Aussagen zur *Verbalklammer* bei komplexen Sätzen (Beispiel: "Fritz *kam* gestern mit dem Zug, der drei Stunden Verspätung hatte, in Stuttgart *an*"), wo die sog. 'Ausklammerung' inzwischen keine Ausnahme, sondern fast schon die Regel geworden ist: Fritz *kam* gestern in Stuttgart *an* mit dem Zug, der drei Stunden Verspätung hatte.". Nach der Publikation dieser spezifischen (statistischen) Untersuchung durch *Rainer Rath* auf der Basis des Eggers'schen Materials hat sich beispielsweise der Grammatik-Duden entsprechend 'umgestellt'.

Eggers blieb aber bei diesen statistischen Untersuchungen nicht stehen: Bekannt geworden ist er durch die Forschungen zur 'Elektronischen Syntaxanalyse', bei der Ende der 60er Jahre des letzten Jahrhunderts in Deutschland erstmals ein Prototyp zur syntaktischen Analyse beliebiger deutscher Sätze (Rechner: eine Philipps Electrologica X1, programmiert in der *Assembler*-Sprache 'Kieler Code') entstanden ist. Aus einem Auftrag der Deutschen Forschungsgemeinschaft (DFG) zur Bewertung des (damaligen) SYSTRAN-Systems zur maschinellen Übersetzung Russisch -> Deutsch (von Peter Toma, s.u.) ging ein eigenes MT-Projekt Russisch -> Deutsch (als *Forschungsprojekt*) hervor, das dann (unter Verwendung von FORTRAN) auf einem Rechner Control Data 3300 realisiert wurde. Beide Entwicklungen wurden schließlich von Eggers in den auf über 12 Jahre angelegten Sonderforschungsbereich 'Elektronische Sprachforschung' (SFB 100) der DFG mit eingebracht; hier entstand (zunächst auf einer Telefunken TR 440) das Saarbrücker Übersetzungssystem SUSY, das übrigens heute noch (nach einer Portierung auf eine UNIX-Anlage) im Internet 'getestet' werden kann. Der deutsche Analyseteil von SUSY – er ist u.a. gekennzeichnet durch eine sog. 'Lemmatisierung' – d.h. Grundformenermittlung – unter 'syntaktischer' Disambiguierung (kontextbasierte Ermittlung der vorliegenden Wortklassen) – wurde von mir später an der Universität Regensburg (1974 – 1980) in anwendungsorientierte Projekte zur Analyse und Indexierung juristischer Texte (Projekt 'JUDO' – mit Urteilstexten als Datenmaterial) überführt und später (Anfang der 80er Jahre) an der Universität des Saarlandes auch zur Indexierung von Patenttexten (Projekt TRANSIT) genutzt.

Im Rahmen der Tests des SYSTRAN-Systems (Russisch -> Deutsch) wurde in Deutschland auch der 'Vater' dieses maschinellen Übersetzungssystems, *Peter Toma*, bekannt. Toma hatte eine Vorversion, die 'GAT' (Georgetown Automatic Translation) mit der Sprachrichtung Russisch -> Englisch, die schon Ende der 50er Jahre des 20. Jahrhunderts vorlag, überarbeitet und von einer IBM 7090 auf eine IBM 360/50 portiert; die DFG hatte Toma mit der Erstellung einer Variante Russisch -> Deutsch beauftragt und die Saarbrücker Arbeitsgruppe unter Leitung von Hans Eggers mit einer *systematischen Dokumentation* des Systems betraut. Toma, ungarischer Herkunft, beeindruckte durch seine unermüdliche Schaffenskraft (man konnte den Rechner im Rechenzentrum der Universität Bonn, auf dem SYSTRAN installiert war, nur nachts benutzen: eine damals für 'elektronische Sprachforscher' übliche Zeit, da die Rechner tagsüber den – meist naturwissenschaftlichen – Aufgaben zugeordnet waren), aber auch durch seine vielfältigen fundierten Sprachkenntnisse (neben Russisch u.a. Englisch, Deutsch, Spanisch ...) und seine leidenschaftliche Begeisterung für die computertechnischen Problemlösungen, auch in den vielen 'Details'. Das *damalige* System *SYSTRAN* – noch ohne jegliche Semantik-Komponente – hielt allerdings nicht das, was man sich in Deutschland davon versprochen hatte.

Eggers war in seinen Forschungen vielseitig und größtenteils 'konservativ' im besten fachlichen Sinne. Seine 'Deutsche Sprachgeschichte' in drei (Taschenbuch-)Bänden ist hier zu nennen, aber auch ein anderes Werk hinterließ einen starken Eindruck: Das von ihm bearbeitete

(letztmals 1967 erschienene) Wörterbuch "Deutscher Wortschatz – ein Wegweiser zum treffenden Ausdruck" (kurz – unter Einbeziehung des vorherigen Bearbeiters Hugo Wehrle – auch 'Wehrle-Eggers' genannt). Bezeichnenderweise ist dieses nach dem englischen Vorbild, dem 'Roget Thesaurus', gestaltete Werk, bei dem das Wortmaterial nach Sach- oder Themengruppen 'hierarchisch' geordnet ist, noch völlig ohne Computerhilfe entstanden: Eggers hatte seine ganze Familie eingespannt, um das gesamte Material (über 100.000 Wörter) zu 'verzetteln' und für das Register alphabetisch zu ordnen (vgl. Vorwort zur 13. Auflage, S. VI). Das Werk verzeichnet übrigens mehr als 10.000 komplexe Verben und verbale Wendungen des Deutschen, deren (lexikalische) Einbindung und Identifizierung in ein maschinelles Übersetzungssystem in diesem Bereich die Übersetzungsqualität erkennbar steigern könnte.

In (West-)Deutschland gab es in den 60er Jahren des 20. Jahrhunderts 'nur' einen ernsthaften Mitbewerber zu dem Saarbrücker Forschungsschwerpunkt: die Bonner Arbeitsgruppe LIMAS mit ihrem Leiter, Dr. Alfred Hoppe. LIMAS wurde v.a. durch das Bonner Verteidigungsministerium finanziell gefördert, so dass die fachliche wissenschaftliche Kommunikation – was die Details der Entwicklungen anging – immer etwas erschwert war. Auch Hoppe hatte für seine sprachanalytischen Konzepte eine repräsentative Materialbasis erstellt: Das maschinenlesbare LIMAS-Korpus umfasste 1000 Artikel aus den verschiedensten Themenbereichen mit je 1000 laufenden Wörtern, d.h. rund eine Million laufende Wörter; es wurde inkl. des Registers auch Dritten in Microfiche-Form zur Verfügung gestellt. Weitaus interessanter war jedoch das Grammatikmodell, die sog. 'Kommunikative Grammatik', das Hoppe allerdings erst später (nach seiner Pensionierung und dem Auslaufen der LIMAS-Forschungen) unter dem Titel 'Grundzüge der Kommunikativen Grammatik' in zwei Bänden (1. Teil: 1981; 2. Teil: 1991) veröffentlichte. War Eggers der ‚Syntaktiker’, so kann Hoppe als der 'Semantiker' unter den deutschen elektronischen Sprachforschern eingestuft werden; seine Ideen und Konzepte der 'semantischen Syntax' sind m.E. (vielleicht auch aufgrund seiner vom gewöhnlichen Fachvokabular völlig abweichenden Nomenklatur) und ihrer Komplexität bis heute nicht in die Praxis der Sprachverarbeitung eingedrungen; das LIMAS-Verfahren selbst hat (wohl auch aufgrund seiner Komplexität) nie hinreichend in einen Demonstrator zur maschinellen Übersetzung überführt werden können.

Ein anderer 'Pionier' der ersten Stunde war ohne Zweifel *Gerhard Wahrig*. Er ist einem breiten Publikum v.a. durch seine gedruckten Wörterbücher bekannt, etwa den 'Einbänder'-Wahrig' (heute bei Bertelsmann) oder den Brockhaus-Wahrig in 6 Bänden. Er war *Lexikograph* – also Produzent von (gedruckten) Lexika – und *Lexikologe*, d.h. wissenschaftlich über Wörterbücher arbeitend, zuletzt als Professor an der Universität Mainz, ehe er nach kurzer schwerer Krankheit noch vor der Fertigstellung des Brockhaus-Wahrig verstarb. Ich habe ihn im Zusammenhang mit seiner Frage kennen gelernt, das lexikalische Material maschinenlesbar aufzubereiten und für Lösungen für die maschinelle Pflege zu erstellen. Bis zu diesem Schritt waren alle Wörterbücher konventionell entstanden: Stichwörter wurden auf Zetteln erfasst, diese sortiert und dann (meist per Lochstreifen) in Satzsysteme eingegeben. Nachbesserungen waren fast unmöglich, Systematisierungen (Sortierungen nach verschiedenen Gesichtspunkten) undenkbar. Wahrigs erstes 'elektronisch' erstelltes (deutsches) Wörterbuch war das sog. *dtv-Wörterbuch*: hier konnte er sein theoretisches Konzept, bei den Definitionen Zirkel zu vermeiden und im Definitionsteil nur Grundbegriffe zu verwenden, weitgehend durchhalten; der über 220.000 Stichwörter umfassende *Brockhaus-Wahrig* (1980 ff.) wurde (interessanterweise mit Hilfe einer Variante des Information-Retrieval-Systems STAIRS) *vollständig* elektronisch erstellt und gepflegt.

Von Eggers und Wahrig konnte man lernen, dass Lexikonarbeit – auch wenn sie maschinell gestützt wird – *umfangreiche intellektuelle Vorleistungen* erfordert. Hoppe hatte – wenn auch

weitgehend aufgrund theoretischer Überlegungen und anhand von Beispielen – gezeigt, dass ein *einfaches syntaktisches Parsing* nicht hinreicht, um Schriftsprache zu übersetzen. Andererseits wurde es immer wichtiger, den Computer als Hilfsmittel bei der Übersetzungsarbeit einzusetzen. Die *Kommission der Europäischen Gemeinschaft* hatte daher als Erste damit begonnen, ein Übersetzungssystem schrittweise für die praktischen Übersetzungsarbeiten einzusetzen. Zuständig wurde die Generaldirektion XIII in Luxemburg, der Verantwortliche war *Loll Rolling.* Die Wahl fiel auf das inzwischen von Peter Toma – auch mit langjährigen Aufträgen der US-amerikanischen Regierung – deutlich weiterentwickelte amerikanische System SYSTRAN. Zur Begleitung dieser Entwicklungen wurde ein Sachverständigenkreis einberufen, dem auch ich angehörte. Nach mehreren Jahren der Weiterentwicklung – im Vordergrund stand zunächst das Sprachpaar Französisch / Englisch – wurde schließlich die *'Kommissions-Variante'* von SYSTRAN zum Zwecke der Nutzung und des Ausbaus auf andere EG-Sprachen vom Toma'schen Unternehmen an die EG-Kommission lizenziert. Toma selbst verkaufte später seine Firma (zumindest den für Europa relevanten Sprachen-Teil) an den französischen Unternehmer Jean Gachot, dessen Hauptgeschäft zum damaligen Zeitpunkt eine Armaturenfabrik war.

Wenn man weiß, welches Volumen und welche 'Qualität' zum Zeitpunkt der ersten Verträge mit der EG das SYSTRAN-System hatte, dann kann man den Mut, den Rolling damals aufbrachte, letztendlich auch seine Ausdauer und Zuversicht nur bewundern. Auch wenn man bis heute nicht sagen kann, dass SYSTRAN die intellektuelle Übersetzung überflüssig gemacht hat, so scheint sich doch inzwischen die Erkenntnis durchgesetzt zu haben, dass sich der Einsatz maschineller Verfahren (wirtschaftlich gesehen) 'rechnet', sofern ihre Stärken genutzt und ihre Schwächen vermieden oder zumindest 'umgangen' werden. Das System der EU-Kommission wird jedenfalls – eingebunden in das EU-interne Kommunikationssystem – zunehmend zur *Informativ*-Übersetzung genutzt, wobei z.B. inzwischen auch das zunächst getrennt zum Zwecke der ‚Humanübersetzung' entwickelte Kommissions-Lexikon EURODICAUTOM adaptiert wurde.

Die Überlegung, durch die Entwicklung eines europäischen Übersetzungssystems einerseits den europäischen Forschungen in diesem Bereich ein hohes praktisches Ziel zu setzen und andererseits neuere sprachwissenschaftlich-computerlinguistische Ansätze zu erproben, gehen ebenfalls in großen Teilen auf Rolling zurück. So wurden von ihm – ohne dass die SYSTRAN-Entwicklungen der Kommission eingestellt wurden – die Forschungen und Entwicklungen zu einem *europäischen Übersetzungssystem EUROTRA* ins Leben gerufen, in dessen Finanzierung sich die beteiligten (damaligen) EG-Länder und die EG-Kommission teilten. EUROTRA hat allerdings zumindest die Erwartungen der Kommission – ein *praktisch brauchbares Übersetzungssystem für alle EG-Sprachen* zu entwickeln – nicht erfüllt: Vielleicht waren letztendlich einfach zu viele Köche am Werk, vielleicht wollte man auch vom konzeptionellen Ansatz her zu viel und verlor die praktische Seite aus den Augen. EUROTRA hat aber zumindest eines bewirkt: Die EG-Mitgliedsstaaten haben ihre *Forschungszentren* im Bereich der *Computerlinguistik* (z.T. auch im 'Wettbewerb' zu den EUROTRA-Zentren) deutlich ausgebaut, so dass man inzwischen in Europa zwar nicht eine blühende Sprachindustrie, wohl aber starke (und weitgehend kooperierende) computerlinguistische Forschungszentren mit entsprechenden universitären Ausbildungsmöglichkeiten vorfindet.

Der französische Unternehmer *Jean Gachot* übertrug nach dem Erwerb der kommerziellen Rechte an SYSTRAN von Peter Toma (bis auf den japanischen Marktbereich) seinem Sohn *Denis Gachot* die Unternehmungen in den USA (La Jolla, Calif.) sowie die gesamten Entwicklungsarbeiten; er selbst begann damit, SYSTRAN in Europa zu *vermarkten*, wobei er damit naturgemäß in Frankreich anfing. Ähnlich wie Rolling war er von den prinzipiellen

Möglichkeiten der maschinellen Übersetzung überzeugt; seine Strategie bestand aber nicht nur darin, die Übersetzungsleistung des Systems zu steigern (sein Qualitätsanspruch, auf eine einfache Formel gebracht, lautete: 95% der Qualität der menschlichen Übersetzungsleistung sind eine Voraussetzung zur praktischen Nutzung), sondern er begann, den 'Rohdiamanten' zu schleifen und zu fassen: Über das MINITEL (die französische Variante des Bildschirmtext) konnte man schon bald kleinere Texte übersetzen lassen (ein Teil der Einnahmen aus der Telefongebühr ging an die GACHOT S.A. / SYSTRAN S.A.), über Mailboxen konnte man per Modem im Abonnement Texte verschiedener Formate an den SYSTRAN-Zentralrechner übermitteln, zurück kam eine 'Rohübersetzung' im Format des Quelltextes. Interessant auch die Möglichkeit, in einer Art ‚Chat'-Version (so würde man dies im Internet-System bezeichnen) mit einem gleichzeitig im MINITEL anwesenden Gesprächspartner schriftlich *in unterschiedlichen Sprachen zu kommunizieren*, wobei die Dialog-Passagen jeweils 'zwischendurch' vom SYSTRAN-System in die Sprache des Partners übersetzt werden. Den größeren Unternehmen, die SYSTRAN lizenzierten, wurden Hilfen zum computergestützten Aufbau der notwendigen Fachwörterbücher angeboten. Schließlich gelang es, neben der Mainframe-Lösung eine PC-Version auf den Markt zu bringen. Eine neuere Anwendung von SYSTRAN ist TRANSLATE, ein kostenloser Service zum Übersetzung von Internet-Sites in Verbindung mit der Suchmaschine AltaVista.

Den Vorteil, den SYSTRAN aufgrund seiner *extrem hohen Übersetzungsgeschwindigkeit* besitzt, versuchen andere Systeme durch Steigerung der Übersetzungsqualität und verbesserte Möglichkeiten seitens des Nutzers, das System an die eigenen Bedürfnisse zu adaptieren, wettzumachen. Ein Übersetzungssystem, das in diesem Zusammenhang erwähnt werden soll, ist *LOGOS*. Was Peter Toma für SYSTRAN war, ist *Budd Scott* für LOGOS. Auch hier stand in den 70er Jahren das amerikanische Verteidigungsministerium bei den ersten Entwicklungen Pate, auch hier gibt es einen Pionier, der den ‚Traum' verwirklichen wollte, durch maschinelle Unterstützung die Sprachbarrieren zwischen den Menschen wenigstens ein Stück weit abzubauen. Anders als Jean Gachot ist Budd Scott ein Sprachingenieur und Entwickler; LOGOS ist 'sein' System, 'seine' Ideen zur Sprachanalyse sind überall zu spüren; er kann jeden Arbeitsschritt, fast jede Funktion, des LOGOS-MT-Systems beschreiben und erklären.

Was Alfred Hoppe theoretisch vorstellte, hat Budd Scott wenigstens in Teilen praktisch umgesetzt: In komplexen (z.T. auch komplizierten) Schritten, die von Scott heute gerne mit den (modernen) Verfahren in *neuronalen Netzen* verglichen werden, 'reduzieren' sich die quellsprachigen Satzoberflächen zu immer abstrakteren Strukturen mit entsprechenden Attributen, ehe dann aus dieser ‚tiefen' Repräsentationsebene sozusagen in einer Art Umkehrprozess wieder die zielsprachigen Sätze entstehen. Syntaktische und semantische Analysen, z.T. auch die morphologischen Prozesse sind miteinander verwoben; es gibt nicht – wie etwa bei SUSY oder SYSTRAN – Analysephasen, die genau eine entsprechend spezifische Aufgabe erfüllen. Semantische Merkmale können beispielsweise zur syntaktischen Disambiguierung an nahezu jeder Stelle anhand entsprechender (kontextbasierter) Regeln herangezogen werden. Scott hatte übrigens schon sehr früh die Idee, *thematisch* gegliederte Lexika (wie sie der Roget-Thesaurus darstellt) zur *Disambiguierung von Wortbedeutungen* mit heranzuziehen (dieser Teil ist aber bis heute in LOGOS nicht implementiert).

Aus meiner persönlichen Erfahrung heraus können die bisherigen konkreten Entwicklungen nur bruchstückhaft wiedergegeben werden. Alleine in Europa gibt es viele Forschungs- und Entwicklungsstätten, die auch im vorliegenden Zusammenhang noch dazustellen gewesen wären; ich nenne stellvertretend die Aktivitäten in Grenoble (um den verstorbenen Bernard Vauquois und um Christian Boitet) und in Pisa (um Antonio Zampolli), die alle nicht erst seit heute an diesen und verwandten Fragestellungen arbeiten. Natürlich dürfen die Arbeiten der

großen Theoretiker, allen voran Noam Chomsky, aber auch von Lucien Tesnière, James Fillmore oder Yorrick Wilks, mit Blick auf konzeptionelle Überlegungen nicht unerwähnt bleiben, andererseits ist es – wie das letztlich diesbezüglich gescheiterte Projekt EUROTRA zeigt – ein weiter Weg von der Theorie hin zur Praxis.

Man macht es sich zu einfach, wenn man die mangelnde Nutzung etwa der bestehenden marktorientierten maschinellen Übersetzungssysteme auf Schwachstellen im systematisch-linguistischen Bereich zurückführt. Natürlich gab und gibt es auch 'primitive' Verfahren, die eine Übersetzung alleine auf der Wort-für-Wort-Ebene durchführen, und nicht immer sind die durchgeführten Arbeitsschritte transparent bzw. führen zu dem 'gewünschten' Erfolg. In aller Regel war es *nicht* die mangelnde Rechenleistung früherer Computergenerationen oder sind nicht unzureichende linguistisch-strategische Konzepte der Grund dafür, dass sich Übersetzungssysteme am Markt nur so mühsam 'verkaufen'.

Mein persönliches Fazit aus all diesen Erfahrungen und Kontakten war und ist, dass man auch als Wissenschaftler in diesem Bereich nicht bei den theoretischen Erkenntnissen stehen bleiben darf. So wie Eggers oder Wahrig ihre theoretischen Überlegungen in praktische Produkte (wenn auch 'nur' auf Wörterbuch-Ebene) umgesetzt haben, so sind wir und die nachfolgende Generation gefordert, ggf. auch mit entsprechender *Ausdauer* (d.h. einem *langen Atem*) schrittweise Lösungen zu entwickeln, die ihre praktische Nutzung unmittelbar deutlich werden lassen. Vielleicht war es falsch, in einem 'großen' Schritt gleich das Problem der maschinellen *Übersetzung beliebiger Texte* anzugehen, so wie es falsch wäre, bei der Erkennung gesprochener Sprache gleich das Ziel zu verfolgen, beliebige fließend gesprochene Sprache beliebiger Sprecher in beliebigen Themen zu erkennen.

Stand und Perspektiven der Maschinellen Übersetzung

Unter *Maschineller Übersetzung* (MT) wird im Folgenden die vollautomatische Übersetzung eines Textes in natürlicher Sprache in eine andere natürliche Sprache verstanden. Unter *Human-Übersetzung* (HT) wird die intellektuelle Übersetzung eines Textes mit oder ohne maschinelle lexikalische Hilfen mit oder ohne Textverarbeitung verstanden. Unter *computerunterstützter Übersetzung* (CAT) wird eine Übersetzung verstanden, die entweder auf einer maschinellen Vorübersetzung mit nachfolgender intellektueller Nachbereitung (Postedition) oder aber auf einer *interaktiven* maschinellen Übersetzung aufbaut, bei der der Mensch während des maschinellen Übersetzungsprozesses zur Klärung von Problemen eingreift. Unter ICAT wird eine spezielle Variante von CAT verstanden, bei der ein Nutzer ohne (hinreichende) Kenntnis der Zielsprache bei einer Übersetzung aus seiner Muttersprache so unterstützt wird, dass das zielsprachige Äquivalent relativ fehlerfrei ist.

Ausgeklammert wird im Folgenden die Einbeziehung von *Rahmenfunktionen* wie Translation Memory, Textfilter und Terminologie-Erschließung, obwohl gerade hiervon die HT sehr stark profitiert, insbesondere was der Kosten- und Zeitaufwand angeht.

Die *prinzipielle Machbarkeit von CAT* i.S. einer *Kostenreduktion und / oder Zeitersparnis* gegenüber der *HT* ist inzwischen erwiesen. Die Einschränkung 'prinzipiell' soll andeuten, dass die Aussage nur bei Erfüllung von Rahmenbedingungen wie 'eingeschränktes Fachgebiet', 'umfangreiches maschinelles Lexikon' und 'relativ einfache Satzstrukturen' gilt. Für die Einsetzbarkeit von *MT* zur Herstellung eines *endnutzungsfähigen Produkts*, d.h. die *unmittelbare* Verwendung einer maschinellen Übersetzung etwa als *brauchbare Verstehenshilfe*, ist der *Nachweis* bislang *nicht eindeutig* geführt.

Bei der *HT* konnte der *Durchsatz*, gemessen in Seiten / Tag, durch maschinelle *Hilfen* wie Term-Banks und Textverarbeitung, *etwa um 1/3 gesteigert* werden. *CAT* ist heute – verglichen mit *HT* – zudem nur auf wenige Sprachpaare und Sprachrichtungen – wenn auch die höherfrequenten – anwendbar, so dass ein global agierendes Unternehmen mit Lokalisierungsbedarf in vielen Ländern allenfalls nur in Teilen von dieser Funktion profitiert.

Durch Verbesserung der *Rahmenbedingungen* – weniger durch Verbesserung der Übersetzungsqualität – kann die Einsatzfähigkeit eines MT-Systems im Rahmen der CAT noch deutlich gesteigert werden. Dazu gehören:

- die *automatische Erkennung von Fachgebieten* (und deren Nutzung),
- die Verwendung nutzerseitiger (d.h. extern vorliegender) Glossare und deren Verwendung / computergestützte Codierung 'on the spot' im Zusammenhang mit einer konkreten Übersetzung,
- die Erkennung und entsprechende Behandlung von *Eigennamen* und *Abkürzungen* aus dem Kontext heraus,
- die *satzübergreifende und kontextbasierte Disambiguierung* von Polysemen (durch Techniken, die nicht notwendig Bestandteil des 'internen' MT-Maschine sind),
- die automatische *Rückkopplung von Posteditionen* zur 'automatischen' Verbesserung des terminologischen Inventars und der kontextbasierten Disambiguierungsverfahren.

Auch der Einsatz von *MT ohne Nachbereitung* kann von diesen Funktionen profitieren, so dass sich die Brauchbarkeit der Rohübersetzung als Verstehenshilfe verbessert und ein sinnvoller Einsatz (etwa im Internet) wahrscheinlich wird.

Aus Gründen der *Akzeptanz* ist es jedoch erforderlich, (heuristische) Funktionen zu entwickeln, die eine *automatische Bewertung eines Textes* mit Bezug zur *erwarteten Übersetzungsqualität der Rohübersetzung* erlauben, so dass entweder das Ergebnis dem Nutzer mitgegeben oder aber auch zur 'Verweigerung' einer MT führt.

Kurzfristige Ziele (bis 2005 / 2010):

Die Entwicklungen sollten sich auf die Realisierung weiterer Sprachpaare als Quell- und Zielsprache konzentrieren, um (größeren) Unternehmen, die heute bereits ein MT-System i.S. der letztendlichen Kostenreduktion von HT-Übersetzungen (gegenüber ‚reinen HT-Verfahren) ‚profitabel' einsetzen (könn(t)en), eine größere Bandbreite an Lösungen zu bieten. Inwieweit dabei ein Redesign einer MT-Maschine einhergeht, etwa um linguistisch basierten Verfahren zu verbessern, müsste im Einzelfall entscheiden werden.

Es ist eine größere *Flexibilität* in Bezug auf die *Textsorten* einzubringen, um die Robustheit gegenüber einem (i.S. des grammatischen Regelwerks einer natürlichen Sprache) 'fehlerhaften' Input zu steigern (Einbindung einer automatischen Rechtschreibkorrektur, ggf. Verzicht auf die Satzstrukturierung unter Nutzung der – oft fehlerhaft gesetzten - Satzzeichen ...)

Das verfügbare lexikalische Inventar ist drastisch zu steigern, um in den gängigen Fachgebieten nicht zu große Lücken aufzuweisen.

Längerfristige Ziele (bis 2015 /2020):

Es sind *Standards* zu schaffen, deren Beachtung aus *Nutzersicht* die Möglichkeit eröffnet, ohne größere technische Schwierigkeiten entweder von einem am Markt verfügbaren System zu einem anderen umzusteigen oder aber mehrere Systeme für Übersetzungen in unterschiedlichen Sprachen und / oder Textsorten zu benutzen.

ICAT (die Interaktiv-Übersetzung von der jeweiligen Muttersprache in beliebige Zielsprache) erlaubt es dem Nutzer (= nicht geschulten ‚Übersetzer'), qualitativ hinreichend gute Übersetzungen seines muttersprachlichen Textes in (eine) ihm nicht oder nicht hinreichend bekannte Fremdsprache(n) zu realisieren.

Die MT mit einer Rohübersetzung (Informativübersetzung) erschließt einen großen (bisher nicht genutzten) Markt, dadurch dass hinreichend valide Übersetzungen zu extrem günstigen Preisen angeboten werden (etwa 1 $ / Seite).

Erforderliche Maßnahmen:

- Stärkere Zusammenarbeit der 'führenden' MT-Systemproduzenten mit dem Ziel, entsprechende Standards zu definieren (*Open MT Systems*).
- Erhebliche Investitionen im lexikalischen Bereich.
- Zügige Ausweitung des Inventars an verfügbaren Quell- und Zielsprachen.

Wegen der erforderlichen Investitionen erscheint es sinnvoll, *strategische Allianzen* dergestalt anzugehen, dass im lexikalischen Bereich stärker kooperiert und bei der Ausweitung der Sprachen Prioritäten abgestimmt werden.

Erfolgswahrscheinlichkeit

Die maschinelle Übersetzung ist und bleibt ein wichtiges Desiderat in einer polyglotten Weltgesellschaft. Eine Entwicklung in Richtung auf die alleinige Anwendung einer Universalsprache wie Esperanto oder den alleinigen Gebrauch von Englisch würde der kulturellen Vielfalt nicht gerecht, die eben auch mit der jeweiligen Sprache verbunden ist. Auch wenn man aus kommunikativen Gründen das Konzept mitträgt, dass man neben seiner Muttersprache mindestens die Welt-Verkehrssprache Englisch beherrschen sollte, bleibt noch genügend Raum für die maschinelle Übersetzung, zumindest mit den Sprachrichtungen X -> Englisch und Englisch -> X.

Zu wünschen ist nicht allein ein unternehmerischer Mut – den haben Personen wie Peter Toma, Jean Gachot oder Budd Scott genügend gezeigt –, sondern v.a. die Bereitschaft zu einer Art weltweiter Zusammenarbeit, auch im Wettbewerb, um die notwendigen Standards zu setzen, um die Transparenz der Systeme und Lösungen zu erhöhen, um das Vertrauen der Anwender zu stärken – und nicht zuletzt auch Kreativität, um aus den ausgefahrenen Geleisen etwas herauszukommen. Aufgaben, die jungen Wissenschaftlern und Ingenieuren ein weites Arbeitsfeld bieten.