Bröckelt der Turm zu Babel?

Harald H. Zimmermann

In: Das Inforum Nr. 18 4/84, S. 12

Es gibt kaum ein scheinbar unumstößlicheres, unverrückbareres Bild in der Bibel als dasjenige vom Turm zu Babel. Ist diese Geschichte von der Entstehung der Sprachvielfalt (1. Mose 11) auch nur ein Erklärungsversuch, so belastet die "Sprachverwirrung" doch erheblich die menschliche Interaktion und Kommunikation, besonders im internationalen Bereich. Ins Profane, Moderne übersetzt - bezogen auf die Ziele des Wissens- und Informationstransfers - heißt dies: das Problem der Überwindung der Sprachbarrieren in der Fachinformation ist bis heute nicht gelöst. Wird das Computerzeitalter auch diesen Wunschtraum der Menschheit nach einer idealen, schrankenlosen Kommunikationsgesellschaft erfüllen, oder, wem dies zu hoch gegriffen scheint: werden wir uns dank einer (Micro-)Computersoftware, die (gesprochene) Sprache so weit versteht und übersetzt, wie dies heute ein menschlicher Übersetzer oder Dolmetscher tut, dabei "brauchbar" verständigen können, auch wenn die Gesprächspartner unterschiedliche (natürliche) Sprachen verwenden?

Wir wollen hier nicht die Frage erörtern, ob die Entwicklung maschineller Verfahren in irgendeiner Weise noch menschenwürdig erscheint, ob die Beschäftigung mit derartigen Fragen vielmehr nicht einen prometheischen Kampf gegen eine gottgewollte Ordnung darstellt. Wir wollen uns konkret fragen, welche Anforderungen an die Entwicklung von Instrumenten zur Sprachübersetzung zu stellen sind - und nicht zuletzt: wo wir heute dabei stehen.

Wenn man unter diesem Vorzeichen beispielsweise in die Schubladen der Experten für Künstliche Intelligenz (KI) schaut, so findet sich eine Reihe von Theorien und Modellen, hier und da auch so genannte "Mini-Welten", in denen exemplarisch und ansatzweise verschiedene (z.T. frappierende) Lösungen zum sog. "Sprachverstehen" vorgestellt bzw. angeboten werden. (Ein interessantes Beispiel ist das in Hamburg entwickelte Modell HAM-ANS.) Doch trotz z.T. jahrzehntelanger Forschungen von Sprachwissenschaftlern, Computerspezialisten, Psychologen usf. ist ein Durchbruch zu praktikablen, hoch entwickelten Verfahren, die das Problem zumindest äquivalent zu den möglichen Leistungen menschlicher Übersetzer, ohne eine entsprechende menschliche Interaktion lösen, nicht in Sicht. Die Fachwelt hofft nun auf die Ergebnisse der Forschungen der "5th Generation Computer" in Japan; im Rahmen einer "europäischen" Antwort auf dieses japanische Forschungsprogramm, dem europäischen ESPRIT- Konzept, soll wie in Japan das Problem des Sprachverstehens durch Computer mitbehandelt werden. Doch was an praktikablen Lösungen (oder besser: Lösungsansätzen) herauskommen wird, ist heute noch völlig ungewiss. Oder umgekehrt: nach aller bisherigen Erfahrung wird man auf die 6. oder 7. oder n-te Computergeneration warten müssen, ehe (vielleicht) derartige Wunschträume Wirklichkeit werden.

Woran liegt das? Nun: einmal an der Komplexität und der ungeheuren Vielfalt der natürlichen Sprache. Zu (fast) jeder sprachlichen Regel gibt es eine Ausnahme, die ein maschinelles Sprachanalyse- oder -verstehenssystem aufschwemmt, es unübersichtlich und unhandlich macht. Dies gilt analog für die Evaluierung der Ergebnisse, da die Bewertung fast mehr Aufwand verursacht als die Erstellung der Verfahren selbst. Zum anderen stellt sich ein "internes" Transfer- und Ergebnis- Sicherungsproblem: die Forschungs- und Arbeitsgruppen, die sich mit derartigen Themen befassen, sind (zu) klein, die damit beschäftigten Wissenschaftler eher

an prinzipiellen Lösungen interessiert als an "Knochenarbeit". Die verwendeten Technologien wandeln sich zudem so schnell, dass Anpassungen an neue Rahmen (Soft- und Hardware) häufig mehr Zeit verschlingen als die Entwicklungsarbeit(en) selbst. Dem Financier grundlegender problemspezifischer Entwicklungen (sei es nun ein Ministerium oder auch die Industrie) geht - meist unter einem ökonomischen Erfolgszwang stehend - frühzeitig der Atem aus. Arbeitsgruppen zerfallen - man denke in Deutschland an das LIMAS- Projekt, an die PLIDIS-Entwicklungen, an die CONDOR-Gruppe - und nur mühevoll können erreichte Teilergebnisse halbwegs gesichert werden.

Lange Zeit bedeutete die (mangelnde) Geschwindigkeit des Computers und die relativ geringe Speicherkapazität ein Handicap für die diesbezügliche Forschung und Entwicklung. Dies scheint sich jetzt langsam zu bessern, jedenfalls wird den technischen Rahmenbedingungen zunehmend weniger Bedeutung zukommen. Umgekehrt wächst der Bedarf an derartigen Verfahren oder zumindest an Teillösungen mit dem Ansteigen der Nutzung von Tele- und Bürokommunikation, aber auch mit der zunehmenden Internationalisierung der Fachinformation. Es bringt nämlich wenig, wenn man heute aus Japan auf deutschsprachige Informationsbanken zugreifen kann (und umgekehrt), ohne dass man dabei den Titel oder das Abstract (und später - beim On-Demand-Publishing - auch die Textfassung) lesen bzw. verstehen kann. Natürlich könnte sich die Wissenschaft - und tut es ja zum Teil - auf eine 'Wissenschaftssprache' (z.B. Englisch) verständigen, aber jedermann leuchtet ein, dass die Verfügbarkeit textueller Information in der jeweiligen Muttersprache eine weitaus interessantere Lösung darstellen würde - vorausgesetzt, sie lässt sich ökonomisch und auch qualitativ ausreichend gut herstellen.

Angesichts der generellen Problematik, aber auch angesichts der mangelnden Verfügbarkeit von Mitteln (die - um einen drastischen Vergleich zu wagen - letztlich in ihrer Gesamtheit den Investitionen für ein Raumfahrtprogramm entsprechen dürften) muss zur Überwindung der Sprachbarrieren in der Fachinformation eine Politik der "kleinen Schritte" entwickelt werden. Für eine derartige Strategie sind unter Kosten/Nutzen-Gesichtspunkten vor allem zwei Kriterien wichtig:

- die jeweiligen Zwischenschritte müssen praxisrelevante Ergebnisse bringen;
- die Zwischenschritte müssen nach Möglichkeit auf den vorausgehenden Etappen (Verfahren, Daten) aufbauen.

Vielleicht könnte man noch ein drittes Element hinzufügen:

- vorhandene Ressourcen sollten so weit wie möglich mitgenutzt werden.

Zu diesen Ressourcen für eine maschinelle Sprachdatenverarbeitung in der Fachinformation gehören v.a. bestehende Lexika, Thesauri und Enzyklopädien. Bei aller Problematik einer unmittelbaren Übernahme bestehender "gedruckter" Sprachdatensammlungen - sie sind meist wenig formalisiert, z.T. ad hoc auf den "menschlichen" Benutzer zugeschnitten: derartige Inventare enthalten so viel "kondensiertes" sprachliches Wissen, dass sie zumindest als Steinbruch, vielfach auch mit ihren sprachlichen Kodierungen für den Aufbau eines systematischen Wortinventars herangezogen werden können.

Was sind das aber für Zwischenschritte? Hier lassen sich verschiedene Wege gehen. Einmal kann man die Software-Entwicklungen zunächst auf bestimmte Fachgebiete/Branchen und innerhalb dieser Bereiche auf bestimmte Textsorten konzentrieren. Zugleich kann man - vorausgesetzt, dies lässt sich praktisch verwirklichen - nur Teilstrukturen der natürlichen

Sprache behandeln. Ein derartiger Weg wird z.B. in der Textilinformation mit TITUS beschritten. Hierbei gelangt man frühzeitig zu maschinellen Übersetzungen (die z.T. noch einer knappen Nachredaktion bedürfen), v.a. verfügt man über ein interessantes terminologisches Kontrollinstrument. Analog zu TITUS kann dann anschließend versucht werden, die verfügbaren Sprachstrukturen schrittweise zu erweitern.

In Kanada hat man vor einigen Jahren den Versuch gemacht, eine sehr stark restringierte Textsorte zu bearbeiten. Es ging darum, Wettermeldungen maschinell zu übersetzen (System METEO). Ein scheinbar trivialer Fall, doch der Teufel steckt hier im Detail - heute ist dieses zunächst viel versprechende Projekt bereits Makulatur.

Ein weiterer Ansatz, der bei den Systemen SYSTRAN (auf Großrechnerebene) und LOGOS (auf Minirechner-Basis) unternommen wird, ist die Unterstützung des menschlichen Übersetzers bei der Übersetzungsarbeit. Hier ist bereits das Vorhandensein eines umfassenden Fachwörterbuchs eine wichtige Grundlage; zudem muss der Übersetzer über eine Schnittstelle verfügen, die einerseits ein bequemes Ergänzen von Wörterbuchdaten ermöglicht, andererseits zugleich die Edition der maschinellen Roh-Übersetzungen (mehr schafft der Computer bei "normalen" Texten nicht) über Textsystem-Funktionen zulässt. Hier heißt die Gleichung ganz einfach: ist ein derartiges maschinelles Verfahren preiswerter (und sichert es daneben eine größere Konsistenz?). Diese Frage ist heute vielleicht noch nicht ganz so eindeutig beantwortet, zeigt aber eine richtige Entwicklungsrichtung an.

Wenn man kurz- und mittelfristig das Mengenproblem im Hinblick auf die Sprachbarrieren in der Fachinformation überwinden will, so kann die Nachredaktion oder Post-Edition (sie setzt einen qualifizierten Übersetzer voraus) auf Dauer nicht die Lösung sein. Daher wird an der Universität des Saarlandes - mit Unterstützung des BMFT - ein anderer Weg versucht: die Entwicklung eines so genannten maschinellen "Informativ-Textübersetzungs-Systems" (ITS). Voraussetzung - wie bei allen anderen Verfahren - sind umfangreiche maschinelle Lexika, die fachgebietsspezifische und auch textsortenrelevante Kennungen haben, um dem Problem der Mehrdeutigkeit von Benennungen zu begegnen. Nützlich sind auch thesaurusartige Begriffsvernetzungen, die zusammen mit Teilwortrelationen und Wortableitungen ein lexikalisches Netzwerk bilden, das bezogen ist auf algorithmisch umgesetzte sprachliche Regeln. Damit stellt es ein praktikables - wenn auch nicht perfektes - Instrumentarium für begriffliche Vereindeutigungen dar. Der Aufbau eines derartigen Lexikons bildet im Grunde die wesentliche Investition.

Im algorithmischen/bearbeitungsstrategischen Bereich verzichtet man demgegenüber zunächst auf komplizierte Regeln, die jeden möglichen 'Schlenker' der natürlichen Sprache mitmachen, in der Hoffnung, dass zwar nicht unbedingt ein "gutes" Deutsch oder Englisch, aber ein verstehbarer, informativer Text entsteht. "Versteht" der Computer einen Satz nicht, so zeigt er dies in einem Statusbericht an: so hat man ggf. Gelegenheit, in einer Art "Interedition" einen allzu komplexen Ausgangssatz an die System-Grammatik anzupassen. Da in der Regel sowieso bestimmte Prä-Editionen erforderlich sind, z.B. um Rechtschreibfehler zu korrigieren, die das System aufgrund eines Wörterbuchabgleichs feststellt, kann ein System-Experte die eine oder andere Problematik schon "vorausahnen" und den Text - möglichst ohne Sinnänderung - an das System formal leicht anpassen.

Das System ITS (Informativ-Textübersetzungs-System) stellt eine anwendungsorientierte Systemvariante des an der Universität des Saarlandes entwickelten Basis-Systems SUSY dar, so dass alle im Grundsystem realisierten Funktionen (v.a. die Analyse- und

Synthesealgorithmen zu den Sprachen Deutsch, Englisch, Französisch und Russisch) nebst den dazugehörigen Basis-Wörterbüchern Verwendung finden können.

Und noch ein weiterer Vorteil kann hier genannt werden: ITS ist ein modulares System, so dass wesentliche Teile identisch sind mit dem ebenfalls in Saarbrücken verfügbaren System CTX (Computergestütztes Texterschließungs-System). Automatische Indexierung und maschinelle Informativübersetzung greifen also ineinander über, so dass sowohl die einsprachige Texterschließung als auch die Textübersetzung lexikalisch, strukturell und auch rechen-ökonomisch voneinander profitieren.

Das Indexierungssystem CTX hat in mehreren, z.T. umfangreichen Anwendungen die ersten praktischen Proben bereits bestanden. So werden beispielsweise deutschsprachige Dokumente einer Reihe von Mitgliedern der Arbeitsgemeinschaft Fachinformation (AFI), z.B. Daten des Deutschen Patentamts, des Fachinformationszentrums Werkstoffe (FIZ 5) und aus dem Rechtsbereich zum Datenschutz verarbeitet. ITS ist dagegen erst in der ersten Entwicklungsphase. Aber auch bei der Informativ-übersetzung wird bereits mit handfesten Daten (wiederum des Deutschen Patentamts, aber auch des Fachinformationszentrums Technik, FIZ 16) gearbeitet.

Die Hannover-Messe 1984 bietet die Gelegenheit, Verfahrensweisen von CTX bzw. ITS wie auch die erreichten Ergebnisse am Stand der Arbeitsgemeinschaft Fachinformation (AFI) näher kennenzulernen. Es kann anschaulich - z.B. unter Zugriff auf Modell-Datenbanken - gezeigt werden, dass mit praxisrelevanten Zwischenschritten eine reelle Chance besteht, das Problem der Sprachbarrieren in der Fachinformation zu reduzieren. Bis der "Turm zu Babel" einmal wirklich überwunden sein wird, ist aber noch ein weiter Weg.