Harald H. Zimmermann, Saarbrücken:

Sprache und Sprachtechnik
- Alfred Hoppe zum Geburtstag -

In: International Classification 18/4 (1991), 196-199

Einführung

Wer sich mit Sprachproblemen beschäftigt, der leistet vielfach Sisyphusarbeit: Kaum glaubt man einen Schritt weiter zu sein, tauchen weitere Phänomene auf, die das ganze Gebäude wenn nicht in Frage stellen, so doch in seiner Begrenztheit aufzeigen.

Ich bin in meinem Leben drei vorzüglichen Wissenschaftlern begegnet, die jeder für seinen Teil und in abgesteckten Bereichen versucht haben, Sprache und Sprachformalismen zu beschreiben und dabei auch für praktische Zwecke nutzbar zu machen. Es sind dies - wenn man es etwas plakativ ausdrücken darf - der Lexikologe und Lexikograph Gerhard Wahrig, der Syntaktiker und Philologe Hans Eggers und der Semantiker und Kybernetiker Alfred Hoppe.

Alle haben sie sich nicht mit abstrakten Theorien und der Entwicklung formaler Beschreibungssysteme begnügt, sondern verbanden Theorie mit Empirie, mit harter, systematischer Arbeit am Material. Und alle waren bzw. sind sich bewusst, dass die sprachliche Kompetenz des Menschen mehr ist, als man heute selbst auf die modernsten Maschinen abbilden kann. Lassen Sie mich stellvertretend A. Hoppe zitieren. Im Kapitel "Sprache und Maschine" seines neuesten Buches, "Theorie der semantischen Syntax: Feste Fügungen" lesen wir einleitend:

> "Will man das Verstehen einer Sprache und damit das Denken einer Maschine anvertrauen, ist das Verfahren (...) mitbestimmt von der Konstruktion und der Arbeitsweise dieser Maschine. Schon die erste semantische Sprosse der Leiter ist für sie nicht erreichbar. Der Sprachkompetente hingegen steigt dank seiner sprachlichen Kompetenz mühelos von Sprosse zu Sprosse der Leiter auf und ab." (S.124)

Ich darf mich - in aller Bescheidenheit und mit großem Respekt vor diesen Wissenschaftlern - mit meinen folgenden Überlegungen dazu gesellen. Hierbei nehme ich - entsprechend meiner Lehr-, Forschungs- und Entwicklungstätigkeit im Bereich der Informations- und Sprachtechnik - in erster Linie einen *Ingenieurs-Standpunkt* ein. Dieser Ingenieur steht vor der allgemeinen Frage, ob es denn und wenn ja: in welchen Grenzen Möglichkeiten gibt, "Sprache" (hier noch ganz vage gebraucht) in sprachtechnische Prozesse so zu integrieren, dass es möglich ist, menschliche Sprachäußerungen (wie man sie an der Oberfläche - geschrieben oder gesprochen - vorfindet) maschinell auszuwerten bzw. zu verarbeiten.

Ausklammern möchte ich im Folgenden den umfassenden, letztendlich aber "unintelligenten" Bereich der physikalischen Speicherung und des Transports von Sprachäußerungen, etwa die digitale Sprachübermittlung, digitale Sprachspeichersysteme.

Welche ingenieurmäßigen Aufgaben, bei denen Sprachverarbeitung eine Rolle spielt, sind anzuführen? Ohne dies hier weiter systematisieren zu wollen, nenne ich einige Bereiche, zu denen es seit Jahren anwendungsorientierte Entwicklungen gibt:

(1) die elektronische Textverarbeitung,
(2) die Mensch-Maschine-Kommunikation (Frage-Antwort-Systeme),
(3) das maschinelle "Verstehen" gesprochener Sprache,
(4) das automatische Indexieren (bis hin zur Inhaltsanalyse),
(5) die maschinelle und maschinengestützte Sprachübersetzung.

Ich bestreite dabei nicht die von A. Hoppe mehrfach dargestellte Wechselwirkung verschiedenster sprachlicher Faktoren, insbesondere nicht die zentrale Bedeutung der Semantik (Hoppe-scher Prägung) für den sprachlichen Verstehensprozess. Dies gilt insbesondere für den generativen Teil, d.h. für den Fall, dass die Maschine die Erzeugung von Sprachäußerungen mehr oder weniger selbsttätig vornimmt.

Ich bin mir auch bewusst, dass vor allem die syntakto-semantischen Merkmale, wie sie in der "Kommunikativen Grammatik", d.h. der Theorie einer semantisch dominierten Syntax von Hoppe systematisch erarbeitet wurden, hier eine zentrale Rolle spielen. Ähnliche Überlegungen finden sich in den Arbeiten zu Französisch von M. und G. Gross.

Der Ansatzpunkt des Ingenieurs ist demgegenüber etwas anders. Sein Problem ist es, für einen bestimmten Bereich praktisch wirksame Problemlösungen für seine "Kundschaft" zu bieten. Die Frage, die sich letztendlich stellt, ist (nur): gibt es in diesem Bereich überhaupt Lösungen, bei denen die Semantik - oder weiter gefasst - das sprachliche und weltbezogene Wissen - nur partiell genutzt werden kann? Mit Recht sagt nämlich Hoppe zum Abschluss des erwähnten Kapitels von der Informationstechnologie, dass sie "nicht vergessen (darf), dass ihre Kundschaft auch denkt und spricht".

Lexikon und Morphologie

Das (elektronische) Lexikon wird im folgenden als der Wissensspeicher der Maschine angesehen. Das für die Sprachverarbeitung benötigte Wissen gelangt (heute) über (menschliche) Sprachexperten in diesen Speicher. Es spielt dabei im vorliegenden Zusammenhang eine untergeordnete Rolle, wie dieses Lexikon technisch organisiert ist. Allerdings bieten sich relationale Datenbanksysteme für die Speicherung und konsistente Datenpflege an, verbunden mit Expertensystem-Teilen, die v.a. Regeln zur Ableitung von Merkmalen beinhalten. Auch im Hoppe'schen Konzept spielt das Lexikon eine wichtige Rolle, da hier u.a. die semantischen Rollen der Wörter verzeichnet sind (vgl. "Theorie" S. 108).

Verfolgt man im lexikalischen Bereich ein offenes Konzept, beispielsweise mit der Möglichkeit, beliebige Merkmale hinzuzufügen bzw. ggf. auch zu modifizieren, so ist man m.E. für alle denkbaren Anwendungen gerüstet. Dies bedeutet aber nicht zugleich, dass von vornherein alle möglichen Anwendungen bedacht werden müssen oder können. Dies ist nämlich in der Praxis eine Kosten- und Marketingfrage. Der Sprachingenieur wird sich in erster Linie danach orientieren, inwieweit mit vertretbarem Aufwand spezifische Lösungen erreicht werden können. Aus den o.a. Bereichen greife ich drei Beispiele heraus:

(a) Automatische Silbentrennung und Rechtschreibkorrektur bei der Textverarbeitung
(b) Lexikalische Synonym- und Übersetzungshilfen
(c) Automatische (wortorientierte) Indexierung

In allen Fällen kommt es zunächst darauf an, dass der Kunde vom Volumen her nicht "enttäuscht" wird. So hat ein Ansatz, wie er z.B. von Knuth bei der Silbentrennung fürs Englische verfolgt wurde (System TEX): "Nur bekannte Wortformen trennen" im Deutschen (als einer stark komponierenden Sprache) praktisch keinen Wert. Eine automatische Silbentrennung und Rechtschreibhilfe greift im Deutschen erst, wenn über 100.000 Wortstämme gespeichert sind und darüber hinaus eine morphologische Flexionsanalyse und Derivations- und Dekompositionsverfahren eingesetzt werden. Nur dann bleibt das System nicht bei sprachlich korrekten Wörtern stehen, weil sie systemseitig unbekannt sind. Das Problem der möglichst vollständigen morphologischen Identifikation muss zunächst gelöst werden, ehe man sich mit Restriktionen dazu befassen kann. Interessanterweise führen die gängigen gedruckten Wörterbücher im Deutschen keine Fugenmerkmale auf, die jedoch für die Erkennung fehlerhafter Wortzusammensetzungen von Bedeutung sind.

Die Möglichkeit, auch Augenblicksbildungen bei der Komposition als sprachlich korrekt zuzulassen (Kanzlerreise, Buchüberreichung ...), führt allerdings notwendig zu Überidentifikationen, wenn nicht weitere Kriterien (welche?) zur Blockierung genutzt werden können: Die Schreibfehler "Waldkauf" für "Waldlauf" "Maustür" für "Haustür" gehören hierher (und erscheinen in gewissen Grenzen zumindest identifizierbar), was aber ist mit dem Satz "gib mir seinen Brief zurück" (statt "... meinen Brief ...")? Untersuchungen zur Fehleridentifikation auf Wortbasis haben gezeigt, dass rd. 95 % aller Fehler solcher Natur sind, dass sie verlässlich erkannt werden, die restlichen 5 % zu finden bleibt (heute) dem Menschen überlassen ...

Anders sieht es bei der automatischen Silbentrennung aus: Mit guten Verfahren lassen sich im Deutschen praktisch bereits auf Wortebene Qualitäten erreichen, die jeder intellektuellen Trennung standhalten. "Schwächen" gibt es v.a. an Nahtstellen, wo etwa das Suffix "er" und das Präfix "er" kollidieren (Druck-er-zeugnis) und bei (seltenen) Fällen unterschiedlicher Zerlegungsmöglichkeiten (bekanntestes Beispiel: Wach-stube / Wachs-tube).

Im Falle der Bereitstellung elektronischer (lexikalischer) Übersetzungshilfen wird zunächst das gedruckte Buch durch das elektronische Lexikon ersetzt: Die Strategie ist hier vergleichbar mit der Nutzung gedruckter Wörterbücher. Der Vorteil der elektronischen Verfahren ist an zwei Punkten zu sehen: Man muss als Nutzer das Alphabet (fast) nicht mehr kennen und hat die Hilfe mehr oder weniger auf Knopfdruck während des Schreibens zur Verfügung.

Natürlich wird man zur Differenzierung von Bedeutungen Hilfen geben und Merkmale setzen. Es ist sehr interessant, einmal zu prüfen, inwieweit die Hoppe-sehen formalen Klassifizierungen als äußere Grundlage herangezogen werden können. Ich mache hier bewusst eine Unterscheidung: Auf der Systemseite kann durchaus das formale Merkmal stehen; nach aller Erfahrung wird man aber annehmen müssen, dass der Laien-Benutzer hiermit wenig anfangen kann. Es gilt also, eine Brücke zu bauen von der "Systemsicht" auf die "Anwendersicht", ein in der Informationstechnik / Datenbanktechnik durchaus übliches Verfahren). Als Möglichkeiten bieten sich an: die Ersetzung der Merkmale durch prototypische Beispiele (GETR/DONS - "Vater"; GEZL/DONM - "Auto") oder aber die automatische Generierung eines Beispiels (schenken_1: Fritz schenkt Paul ein Auto) usf.

Die Verbesserung der Suche in (bibliographischen wie textuellen) Datenbanken bzw. deren tiefere Erschließung mit sprachtechnischen Methoden ist ein besonderes Desiderat. Es sind dabei zwei

Schwerpunkte zu erkennen: Es ist auf längere Sicht nicht wirtschaftlich machbar, die "großen" Datenbanken (etwa DPA, JURIS, PATDPA in Deutschland; CHEMICAL ABSTRACTS international) schon beim Aufbau sprachlich tiefergehend zu erschließen. Gegenwärtig werden praktisch alle Textdatenbanken auf Wortformenbasis (im Freitextbereich) recherchiert, und dem Benutzer werden z.T. regelrechte Verrenkungen (bei der sog. Trunkierung) zugemutet. Eine große Hilfe ist sicherlich die Bereitstellung automatischer Trunkierungshilfen, bei denen das System automatisch die möglichen Stämme und - wenn wortklassenübergreifend - auch Pseudostämme bereitstellt. Hierzu reicht zunächst ein Reduktionsalgorithmus aus, der gegenüber dem Identifikationsverfahren, wie es bei der Rechtschreibkontrolle verwendet wird, noch Verweise zu (Basis-)Stämmen aufführt. Angesichts der Unzuverlässigkeit des Originalmaterials fallen mögliche Überindexierungen (Beispiel: Schraubenmutter / - muttern? / mütter?) kaum ins Gewicht, im Gegenteil: Eine Vorabdifferenzierung (etwa Bank = Geldinstitut oder = Sitzgelegenheit) würde von der Datenbank nicht honoriert, da dort entsprechende Unterscheidungen fehlen.

Diese Einschränkung gilt nicht, wenn beispielsweise ein Beratungssystem dazu entwickelt wird, um einem Datenbanknutzer geeignete Termini für die Suche in einer Datenbank vorzuschlagen: Hierbei können die Hoppe-sehen Kategorisierungen (etwa zur Bedeutungsdifferenzierung) eine wichtige Rolle spielen.

Syntax und Semanto-Syntax

Spätestens seit Fillmore ist der "bedeutungsneutralen" Syntaxanalyse Chomsky-scher Prägung international eine Absage erteilt worden. Die frühen systematischen Arbeiten von A. Hoppe waren bereits in eine ähnliche Richtung gegangen, wie überhaupt in der Weisgerber-Nachfolge eine ganzheitliche Betrachtung von Sprachphänomenen im Vordergrund stand.

Ich möchte an dieser Stelle nicht weiter auf das Stufenmodell von Hoppe eingehen; hierzu kann auf die Lektüre der Neuerscheinung verwiesen werden. Eines ist jedoch anzumerken: Alle heutigen Verfahren und Ansätze v.a. im Bereich der maschinellen Übersetzung (auch die klassischen Systeme SYSTRAN und METAL, nicht zuletzt EUROTRA), bauen inzwischen auf der Erkenntnis auf, dass Sprachanalyse und -synthese eine semantische Komponente brauchen. Viele der bestehenden Unterschiede liegen eher im analyse- und synthesestrategischen Bereich: Während Hoppe der Semantik einen "Steuerungsfaktor" zuspricht ("Theorie" S. 127), ist sie bei EUROTRA eine "Ebene" neben der Oberflächensyntax, kommt ihr bei SYSTRAN v.a. die (auch von Hoppe als wesentlich angesprochene) Disambiguierungsfunktion zu.

Der Hoppe-Graph und das sog. Wechselwirkwerk-Netzwerksystem stellen in meinen Augen eine interessante Repräsentationsform dar, lösen aber beispielsweise nicht das Problem des Parsing (allenfalls lässt sich ein Sprachgenerator bauen), da man in den sprachlichen Ausdrucksformen ein Gewirr von (Oberflächen-)Ambiguitäten findet, das erst einmal entflochten werden muss. Ohne Zweifel ist dabei die Idee interessant, das Begriffssystem selbst (genauer gesagt: die den Ausdrucksformen = "Begriffswörtern` zugeordneten Systemteile) als Basis der Analyse zu nehmen und nicht - wie allgemein üblich - die syntaktische Struktur (vgl. "Theorie" S. 114).

Dass es Interdependenzen zwischen den einzelnen "Ebenen" gibt (sofern überhaupt eine solche Analogie am Platze ist), zeigen schon die einfachsten Beispiele. Die Sätze "Er sehte das Frau im Garten" und "weil er kam zu spät, er nicht mehr wurde eingelassen" bleiben verständlich: bei dem formal korrekten Satz "Der Boofke alfanzt mit dem Schwiemel" kann nicht notwendig ent-

schieden werden, ob er semantisch korrekt ist (er ist es, wenn man seine "Übersetzung" betrachtet: "der Narr treibt seine Possen mit dem Trunkenbold"). Bei dem Satz "Diese Maus frisst die Katze" "versagt" (theoretisch) auch die semantische Syntax.

Dennoch bleibt festzuhalten: Ohne die (konsequente) Anwendung einer semantisch orientierten Syntax müssen *höherwertige* Systeme der maschinellen Sprachverarbeitung, insbesondere zur maschinellen Übersetzung, fehlschlagen. Ihre besondere Leistungsfähigkeit zeigt sich v.a. bei der sog. "Monosemierung", d.h. der Vereindeutigung formal-syntaktisch mehrdeutiger Strukturen (potentieller Alternativen) bzw. - was mindestens ebenso wichtig ist: der Disambiguierung von Wortbedeutungen. Zur "Ehrenrettung" bestehender Übersetzungsverfahren ist jedoch anzumerken: Es hat fast ein Lebensalter gedauert, in einigen Teilbereichen die Grundlagen zu legen für eine Systematik der Semantischen Syntax, die zudem sich in der Praxis erst noch bewähren muss. In jedem Falle wird ein solches Verfahren erst wirksam, wenn es satzübergreifend (bis hin zum Absatz oder ganzen Text als Kontextebene) realisiert und dabei u.a. auch das Problem der pronominalen Referenz bedacht (und mit-gelöst) wird. Eine derartige semantische Analyse auf Satzkontextebene - dies zeigten Saarbrücker Forschungen in den 80er Jahren - lassen wegen der Ambiguität der Pronomen ein solches Verfahren scheitern (d.h. es wird zu wenig monosemiert).

Sprach- und Weltwissen

Die Frage, wo "sprachliches" Wissen (= "sprachsystembezogenes Wissen") endet und wo "Weltwissen", d.h. "sachbezogenes Wissen" beginnt, ist - aus der Sicht des Sprachingenieurs - ein "philosophisches" Thema. Ich möchte auch hier ein Beispiel geben (und zugleich auf das o.a. Beispiel von der Katze und der Maus verweisen): Wenn in einem Frage-Antwortsystem (das Beispiel stammt aus PLIDIS, einer früheren Entwicklung am Institut für Deutsche Sprache, Mannheim) die Frage auftaucht: "Um wieviele Punkte stieg VW gestern?", sind zur Beantwortung (mindestens) folgende Daten notwendig:

- aktuelles Datum (Tagesdatum)
- VW = VW-Aktie
- Wert der Aktie vom Vortag und vom Vor-Vortag
- steigen = Wertsteigerung / Wertveränderung (die Aktie könnte ja auch gefallen sein)
- Punkt = "ganzzahliger numerischer Wert..."

Folgende Operationen (Regeln) müssen mindestens angewendet werden:

- gestern = aktuelles Datum -1 (Datumswissen!) ...
- Abruf der Werte aus Datenbank
- mathematische Vergleichsoperation

Wenn man sich in einer kleinen "Welt" bewegt (Börsenauskunft, Wetterbericht, Veranstaltungsinformation: man kann sich beliebige derartige "Welten" vorstellen), gewinnen Funktionen an Bedeutung, die in allgemeinsprachlichen Systemen allenfalls rudimentär aufscheinen, aber ebenfalls starke Wirkungen auf das "Verstehen" haben (ja für die Antworten unbedingt nötig sind).

Fazit und Ausblick

Die Sprachtechnik muss sich - will sie den Kunden / Nutzern wirkliche Erleichterungen verschaffen oder Hilfen geben - aller verfügbaren Mittel bedienen, die von Sprachwissenschaftlern, Psychologen, Fachexperten, Informatikern eingebracht werden. Höherwertige Systeme (etwa zur maschinellen Übersetzung) benötigen (allerdings) eine starke semantisch basierte Syntax.

Die Arbeiten von Alfred Hoppe sind nicht erst seit heute, sondern schon seit den Forschungen der LIMAS-Gruppe ein wichtiges Element dieser Entwicklungen. Sie sind aufgrund ihres hohen Formalisierungsgrades zudem besonders geeignet, in Anwendungssystemen - etwa zur Textanalyse und zur maschinellen Übersetzung - Eingang zu finden. Mit den beiden Teilen der "Semantischen Syntax' (1981/1991) steht jetzt vielfältig nutzbares Material zur Verfügung.

Dennoch sollte man nicht unterschätzen, dass von der theoretisch fundierten Darstellung - trotz der relativ breiten materiellen Grundlegung - bis zur praktischen Anwendung ein weiter (und zudem kostspieliger) Weg ist. Die größten Chancen, die Arbeiten praktisch anzuwenden, finden sich m.E. im Bereich der maschinellen Übersetzung und - wie von Hoppe gelegentlich selbst eingebracht - im (sich gerade neu entwickelnden) Bereich der sog. Teachware, d.h. des computergestützten Lernens.

Wir sind dankbar, dass jetzt nicht nur das Gedankengebäude von Alfred Hoppe vorliegt, sondern eine breite Materialbasis für gezielte Anwendungen geschaffen ist. So weit es in meinen Kräften steht, werde ich dazu beitragen, daß seine Konzepte und Modelle mittelbar und unmittelbar wirksam werden. Dem Geburtstagskind wünsche ich von Herzen weitere Schaffenskraft und Gesundheit.

Saarbrücken, Fassung: September 1991