In: G. Nickel; A. Raasch (Hrsg., 1974): IRAL-Sonderband: Kongressbericht der 5. Jahrestagung der GAL, Heidelberg, 172-179

Harald H. Zimmermann

EIN KONZEPT ZUR SYNTAKTISCHEN OBERFLÄCHENANALYSE

Die automatische Lemmatisierung (AL) von Wortformen eines Textes, also die Zuordnung einer Flexionsform zu ihrem Lemma[1], setzt neben einem maschinellen Lexikon eine automatische Analyse voraus. Diese Analyse hat - im Zusammenhang mit dem Ziel der AL die Aufgabe, (lexikalisch) mehrdeutige Wortformen aufgrund des vorliegenden Kontextes zu vereindeutigen. Folgende Ansprüche werden aufgrund des derzeitigen Konzepts der AL an die Analyse gestellt:

1. Syntaktische Mehrdeutigkeiten sollen beseitigt werden. Beispielsweise ist LIEBE in ICH SEHE LIEBE KINDER als Adjektiv ('LIEB'), in ICH LIEBE SCHOKOLADE als Verb ('LIEBEN') und ER KENNT KEINE LIEBE als Substantiv ('LIEBE') zu klassifizieren[2].

2. Mehrwortige Flexionsformen sollen ihrem Lemma zugeordnet werden. Beispielsweise sind die zusammengesetzten Zeiten (IST GEKOMMEN, - 'KOMMEN'; WAR GEZWUNGEN WORDEN - 'ZWINGEN') solche Flexionsformen, aber auch feste Syntagmen wie 'BLINDER PASSAGIER', 'IN ANSPRUCH NEHMEN', bei denen als Lemmaname mehrere Wortformen verwendet werden, gehören hierher.

3. Semantische Mehrdeutigkeiten sollen nur insoweit aufgelöst werden, als die lexikalischen Informationen dazu ausreichen. Beispielsweise lässt sich das Substantiv SCHLOSS mittels der verfügbaren morphologischen, syntaktischen und semantischen Beschreibungsmerkmale nicht mehr differenzieren in (Königs-)$Schloss_1$ und (Tür-)$Schloss_2$. Eine Reduktion dieser Mehrdeutigkeit kann also auch durch die Analyse, die nur auf den lexikalischen Informationen - kontextbezogen - operiert, nicht erfolgen. Andererseits kann überall dort, wo morphologische, syntaktische oder semantische Informationen zur Differenzierung ausreichen, eine Mehrdeutigkeit reduziert werden: '$SCHLAGEN_1$' in DIE NACHTIGALL SCHLAEGT lässt sich von '$SCHLAGEN_2$' in DIE UHR SCHLAEGT aufgrund der semantischen Kongruenz der Subkategorisierungsmerkmale[3] in dem vorliegenden Kontext unterscheiden; bei DER MANN SCHLAEGT SEIN KIND NIE lässt sich über syntaktische und semantische Merkmale das Lemma '$SCHLAGEN_3$' ermitteln. Dennoch ist es in dieser ersten Phase nicht das Ziel der AL, eine Differenzierung aller semantisch mehrdeutigen Lemmata in allen Fällen zu "erreichen: dazu reichen die lexikalischen Informationen nicht aus.

4. Unbekannte Flexionsformen sollen weitgehend automatisch syntaktisch klassifiziert und lemmatisiert werden. Zu diesem Teilziel ist auch die Ermittlung von - lexikalisch nicht erfassten - Eigennamen zu rechnen.

Die zugrundeliegende grammatische Beschreibung der Lexikoneinträge, vor allem der Hauptwortklassen Adjektiv, Verb und Substantiv steckt weitgehend den Rahmen für die Tiefe der Analyse ab. Es ist also häufig nicht zu erwarten, dass die Ergebnisse der Analyse wesentlich über eine Oberflächenbeschreibung von Texten (oder besser: Sätzen) hinausgehen werden. Transformationelle Elemente, wie Nominalisierungstransformationen und Tilgungstransformationen im Rahmen der Junktion, sind bei der Analyse beispielsweise ausgeklammert.

Eine weitergehende Analyse von Texten setzt m.E. allerdings eine Oberflächenanalyse voraus. Ich betrachte daher die im Folgenden beschriebene Analysephase als einen ersten Schritt zu tieferen Analysen (auf syntaktisch-transformationeller oder semantisch-pragmatischer Ebene). Derartige Analysen setzen eine Erweiterung und Verfeinerung der Lexikonkomponente voraus.

Bei der automatischen Analyse sind zwei unterschiedliche Vorgehensweisen bekannt, die Konsequenzen für die gesamten Verfahrensschritte bringen:

(1) Grammatisches Regelsystem und Analyse-Algorithmus werden strikt voneinander getrennt.
(2) Der Algorithmus beinhaltet die Grammatik.

Im ersten Fall werden (zumeist kontextfreie) Phrasenstrukturregeln durch einen Algorithmus in bestimmter Weise interpretiert. Die Regeln werden dabei in einer festgelegten Reihenfolge auf die konkreten Elemente angewendet, um Texten/Sätzen eine im Regelsystem vorgesehene Struktur zuzuordnen. Bekannte Vorgehensweisen sind die Bottom-up- und die Top-down-Analyse, Analyseverfahren oder Modifikationen davon, die auch bei Programmiersprachen verwendet werden.

Will man eine natürliche Sprache wie das Deutsche kontextfrei beschreiben, so verlangt dies eine Vielzahl von allerdings relativ leicht verständlichen Regeln und Subregeln. Diese Beschreibung kann ohne Programmierkenntnisse umgesetzt werden, wenn Compiler für diese Beschreibung zur Verfügung stehen.

Wenn Regelsystem und Algorithmus eine Einheit bilden, können die Regeln in einer beliebigen Beschreibungssprache von Linguisten vorformuliert werden. Sie werden dann unmittelbar in den Algorithmus, das Analyseprogramm, integriert. Dies bedeutet nicht, dass jede "Regel" in eine Computerinstruktion umgesetzt werden müsste, vielmehr können hier Tabellen bei der Ermittlung von Strukturen herangezogen werden. Diese Integration der Analysegrammatik in den Algorithmus ist nicht zuletzt deswegen umstritten, weil eine Dokumentation der zugrundeliegenden Grammatik erschwert ist. Zudem muss ein Linguist über entsprechende programmtechnische Kenntnisse verfügen, um sich einen ausreichenden Überblick zu bewahren.

In Saarbrücken hat dieser zweite Weg bereits eine gewisse Tradition: das in den 60er-Jahren entwickelte Verfahren zur automatischen syntaktischen Analyse deutscher Sätze (vgl. /5/) war in dieser Weise organisiert, einem automatischen Übersetzungsverfahren Russisch-Deutsch liegt ebenfalls ein entsprechender Parser zugrunde (vgl. /6/).

Die Erfahrungen, die dabei gemacht wurden, lassen es sinnvoll erscheinen, einen modifizierten Algorithmus auf der Grundlage dieser beiden Analyseverfahren zu erstellen. Ihm wird sowohl eine deutsche als auch eine russische Grammatik zugrunde liegen, wobei die strukturellen Gemeinsamkeiten dieser Sprachen ausgenutzt und auch Besonderheiten berücksichtigt werden.

Das integrierte Analyseverfahren lässt sich in folgende wesentliche Abschnitte gliedern:

1. Verknüpfung Text - Wörterbuch
2. Homographenreduktion im Rahmen. eines Textabschnittes (Satzes)
3. Aufbau eines Analysekellers
4. Analyse je Kellerteil

5. Ergebnisanzeige und Weiterverarbeitung (Lemmatisierung)

1. Verknüpfung Text - Wörterbuch

Das zugrundegelegte Verfahren. zur Wörterbuchsuche ist bereits beschrieben [4]. Einzelheiten und Ergänzungen werden veröffentlicht werden, sobald die Erfahrungen mit dieser Vorgehensweise ausgewertet sind.

2. Homographenreduktion

Nach der Wörterbuchsuche wird der vorbereitete Text in Analysesegmente zerlegt. Um später in der Lage zu sein, auch satzübergreifende Zusammenhänge zu erfassen, wird der Satz (= Wortfolge zwischen Punkt/ Fragezeichen/ Ausrufezeichen) nicht grundsätzlich die Basis der Segmentierung bilden, sondern eine Folge zwischen beliebigen Grenzzeichen (etwa Absatz) gewählt werden können. Über diesem Textabschnitt operiert nun ein Verfahren zur Reduktion syntaktischer Mehrdeutigkeiten. Es ist orientiert an der Vorgehensweise der "Elektronischen Syntaxanalyse" von 1969, allerdings mit anderer Zielsetzung: Es kommt nicht darauf an, alle Mehrdeutigkeiten (auf Wortklassenebene) zu reduzieren, sondern nur solche, bei denen eine Reduktion mit einfachen Mitteln (u.a. Stellungsregeln) absolut zuverlässig erreichbar ist. Daneben spielt der Häufigkeitsgesichtspunkt eine Rolle: Das Verfahren ist auf solche Fälle eingeschränkt, in denen sich der prozentuale Anteil noch verbleibender Mehrdeutigkeiten im Textabschnitt deutlich einschränken lässt. Betroffen sind hier vor allem funktionale Mehrdeutigkeiten wie Relativwort/ Demonstrativwort, Präposition/ Verbzusatz und finites Verb/ Infinitiv. Obwohl hier natürlich linguistische Gesichtspunkte die Grundlage der Reduktion bilden, ist dieser Analyseabschnitt für die Ergebnisse der Analyse prinzipiell ohne Belang: er dient allein der Verkürzung der Analyse-Zeit und der Einschränkung des Speicherbedarfs.

3. Aufbau des Analysekellers

Die nach der Homographenreduktion nach verbleibenden Mehrdeutigkeiten auf Wortklassenebene steuern im Folgenden den Aufbau des Analysekellers. Die Analyse selbst operiert auf scheinbar eindeutigen Informationen und versucht, diesen eine Strukturbeschreibung zuzuordnen. Lässt sich eine derartige Zuordnung nicht erreichen, wird angenommen, dass es sich um eine Informationsfolge handelt, die nicht durch die zugrundeliegende Grammatik beschrieben ist. Man kann dabei von der Annahme ausgehen, dass eine lexikalisch mögliche, kontextuell aber falsche Information vorlag, und wird daraufhin eine Lexikonalternative zugrunde legen. Mit dieser Alternative wird nun der Analysevorgang erneut ablaufen, bis alle Alternativen entsprechend die Analyse durchlaufen haben, unabhängig davon, ob unter einer bestimmten Konfiguration zwischenzeitlich eine vollständige Strukturbeschreibung vorgenommen werden konnte oder nicht. Auf diese Weise ist sichergestellt, dass im Rahmen der Grammatik noch mehrdeutige Texteinheiten in allen möglichen Strukturvarianten beschrieben sind.

Jede nach der Homographenreduktion noch verbleibende Mehrdeutigkeit vervielfacht auch die Basisinformationen für die nachfolgende Analyse. Diese Vervielfachungen werden unter Verwendung der entsprechenden Alternativen vor der Analyse in einem Speicher abgelagert (gekellert), dem sog. "Textkeller". In dieser Phase (teilweise kann dies bereits bei der Wörterbuchsuche

und der Homographenreduktion geschehen) sind weitere Differenzierungen des Analyseinventars durchzuführen. Dies betrifft vor allem:

a) die Ermittlung solcher "festen Syntagmen", die zu wortklassenspezifischen Veränderungen führen oder die weitere Strukturierung des Textes beeinflussen, also Wendungen wie DEN GANZEN TAG (Temporaladverb), UM DIE ECKE BRINGEN (Verb) u.a., nicht notwendig aber Fälle wie "BLINDER PASSAGIER" vgl. /2/);

b) die Feststellung und Zuordnung von Verbzusätzen. Hier kann sich beispielsweise die Rektion von der des Simplex-Verbs unterscheiden. Über eine spezielle Wörterbuchsuche ist hier während der Analyse ein entsprechender Informationstausch durchzuführen;

c) die Aufhebung der Subsatzdiskontinuität. Die unter a) und b) erwähnten Probleme sind bereits teilweise nur über die Ermittlung des Zusammenhangs diskontinuierlich stehender Elemente lösbar. Zugleich kann die weitere Strukturanalyse dadurch vereinfacht werden, dass die Diskontinuität - etwa auf Subsatzebene - formal aufgehoben wird. Die Ermittlung möglicher (nicht unbedingt tatsächlicher) zusammenhängender Elemente der Wortfolgen erscheint aufgrund bestimmter Hinweiselemente (vor allem der Konjunktionen und der verbalen Wortformen) befriedigend lösbar, vor allem wenn man derartige Zuordnungen - ähnlich der Homographenreduktion - als vorläufig betrachtet und die endgültige Verifizierung über die strukturelle Beschreibung abwartet. Mit der dadurch notwendigen Ermittlung und Beschreibung der Verbalgruppe ist zugleich eine wesentliche Grundlage für die weitere Analyse geschaffen, da mit der Ernennung des finiten Verbs (Numeruskongruenz) und des Rektionsträgers entscheidende Beschränkungen für den möglichen Kontext festliegen.

4. Analyse je Textkellerteil

Die verbleibende Analysephase ist auf die endgültige Strukturbeschreibung des Textes ausgerichtet. Sie kann in zwei Teilbereiche gegliedert werden: Der erste dieser Bereiche ist subsatzbezogen, d.h. es werden alle jene Teilstrukturen beschrieben und inventarisiert, die sich im Rahmen von Subsätzen feststellen lassen, d.h. solchen "Sätzen", deren Basis ein verbaler Rektionsträger ist. Alle Matrixsätze und alle oberflächig eingebetteten (Konstituenten-) Sätze werden als Subsätze aufgefasst. Bei der Beschreibung des Subsatzinventars werden alle Nominalgruppen und die formalen Abhängigkeiten der Nominalgruppen (voneinander oder vom verbalen Rektionsträger) erfasst einschließlich der präpositionalen Attribute oder Objekte, soweit entsprechende Markierungen im Lexikon vorliegen. In dieser Phase werden auch Kasusmehrdeutigkeiten reduziert - entweder aufgrund der Kongruenzbedingungen innerhalb der Nominalgruppen oder in Abhängigkeit von der (strikten) Subkategorisierung der Verben - und das adverbiale Inventar notiert. Die subsatzübergreifenden Relationen werden schließlich in einer zweiten Phase aufgezeigt. Es handelt sich um Problemkreise wie Relativsatz- und Nebensatzbezüge. An diese oberflächige Beschreibung könnten sich in der Folge Transformationen anschließen, die eine tiefenstrukturelle Beschreibung liefern.

Literatur

/ 1 / W. Klein, R. Rath et al.: Automatische Lemmatisierung. Arbeitsberichte des Germanistischen Institutes der Universität des Saarlandes, Nr. 10, Saarbrücken 1971.

/ 2 / Autorenkollektiv: Aspekte der automatischen Lemmatisierung. Berichte 10-72: G des Sonderforschungsbereiches Elektronische Sprachforschung, Universität des Saarlandes.
/ 3 / Autorenkollektiv: Skizze einer deutschen Grammatik I. Bericht 10-73: G des Sonderforschungsbereiches Elektronische Sprachforschung, Universität des Saarlandes.
/ 4 / Referate des 1. SFB-Kolloquiums vom 20.-22. Mai 1971. Bericht 10-71: G.R.A.M. des Sonderforschungsbereiches Elektronische Sprachforschung, Universität des Saarlandes.
/ 5 / H. Eggers et al.: Elektronische Syntaxanalyse der deutschen Gegenwartssprache, Tübingen 1969
/ 6 / Autorenkollektiv: Zur Strategie einer maschinellen russischen Syntaxanalyse. Bericht 9-72: G des Sonderforschungsbereiches Elektronische Sprachforschung, Universität des Saarlandes.
/ 7 / F. L. Le Renier. Simple LR (k) Grammars. CACM 14 (1971), S. 453-460.
/ 8 / Th. A. Zoetnout: Description of the Program SLRK. A Fortran-Program to Compute from an arbitrary context-free Grammar an LR (k) Parser, Saarbrücken, Dz. 1971.
/ 9 / G. Hamann: Beschreibung des Programmes "Parsing", Universität des Saarlandes, Fachbereich Angew. Mathematik und Informatik, Saarbrücken, April 1973

ANMERKUNGEN

*SFB Elektronische Sprachforschung, Universität des Saarlandes

[1] Zur Definition von Lemma, Flexionsform, Wortform u.ä. vgl. /1/ u./2/
[2] Das Lemma wird durch den Lemmanamen (in Anführungsstrichen) repräsentiert (eine ausgewählte Flexionsform).
[3] z. B. /+ belebt/, /+ menschl. / bei der entsprechenden NP
[4] vgl. Peter Krebs in /2/.