SYNTAKTISCHE ANALYSE DEUTSCHER SÄTZE IM EUROTRA-FORMALISMUS

Konfigurationale und relationale Struktur

Paul Schmidt
IAI/EUROTRA-D
Martin-Luther-Straße 14
D-6600 Saarbrücken

Der stratisfaktionelle Ansatz des EUROTRA-Projektes zur maschinellen Übersetzung wird beschrieben: Zwischen Quell- und Zieltext werden verschiedene linguistisch motivierte Repräsentationsebenen angenommen, wodurch die Übersetzung in eine Abfolge kleinerer Übersetzungsteilschritte unterteilt wird. Die einzelnen Ebenen, die morphologische Struktur (EMS), die konfigurale Struktur (ECS), die relationale Struktur (ERS) und die Interface Struktur (IS) werden vorgestellt und das Konzept der Repräsentation und der Beziehungen zwischen benachbarten Ebenen werden kurz erläutert. Es wird an Beispielen gezeigt, wie die Regeln der deutschen Syntax auf der konfiguralen Ebene formuliert werden und wie ECS-Repräsentationen mittels *t*-Regeln auf die relationale Ebene abgebildet werden.

Probleme des Ansatzes, die in der Automie der ECS und in der Beschränktheit der *t*-Regeln liegen, werden angesprochen. Für den beschriebenen Ausschnitt der EUROTRA-D-Sprachbeschreibung liegt eine Implementierung vor.

1. Grundprinzipien

In diesem Beitrag werden die Grundprinzipien sowie einige Beispiele für syntaktische Analyseregeln eines Fragments des Deutschen gegeben. Dieses Fragment ist die Grundlage für die Implementierung der deutschen Syntaxkomponente in EUROTRA (vergleiche *Hallet/Reyle 1987*).

Die Repräsentationssprache ist der sogenannte C,A,T-Formalismus *(Arnold et al. 1986)*. Er arbeitet mit den drei Elementen

Constructor
Atom
T-Rule

und ist eine formale, linguistisch motivierte Stratifikationstheorie des Übersetzungsvorganges. In der EUROTRA-Theorie werden zwei Grundannahmen gemacht:

1) Die Übersetzungsrelation wird nicht zwischen Quell- und Zieltext definiert, sondern muß in einzelne einfachere Übersetzungsschritte aufgeteilt werden.
2) Diese kleineren Übersetzungsschritte werden zwischen linguistisch motivierten Ebenen vorgenommen.

Aus diesen Annahmen folgen drei Grundeigenschaften des EUROTRA-Formalismus:

1) Das Übersetzungssystem hat stratifikationellen Charakter: die Übersetzungsrelation besteht zwischen $T_1 \ldots R_1 \ldots R_n \ldots T_2$, wobei jede Ebene R eine künstliche Repräsentationssprache ist.
   Zur Zeit gilt: R_i = (EMS, ECS, ERS, IS)
   EMS = EUROTRA Morphological Structure
   ECS = EUROTRA Configurational Structure
   ERS = EUROTRA Relational Structure
   IS = Interface structure

2) Die Repräsentationen sind durch Grammatiken bestimmt, die Generatoren genannt werden. Diese enthalten zwei Arten von Regeln:
   b-rules bauen Repräsentationen auf
   a-rules drücken Generalisierungen über Attribute aus.
   a-rules werden auf von b-rules erzeugte Strukturen angewendet. Diese "Anwendung" bedeutet, daß eine Unifikation der a-rule mit der erstellten Struktur versucht wird. Wenn diese gelingt, wird die Struktur durch das Ergebnis der Unifikation ersetzt, im anderen Fall gibt es zwei Möglichkeiten:
   - die Struktur überlebt ohne Änderung (etwa nach einer sogenannten "gentle"-a-rule-Anwendung)
   - die Struktur wird gelöscht (nach einer "strikten" a-rule)

3) Die Beziehung zwischen zwei Ebenen wird durch einen "Translator" determiniert, der aus einer Menge von "t-rules" besteht. Diese t-rules haben zwei Charakteristika:
   - sie sind ein-eindeutig, d.h. sie besitzen keine interne Strategie; ein Objekt der eine Ebene wird in ein einziges der anderen Ebene übersetzt
   - sie sind kompositionell, d.h. die Übersetzung eines strukturierten Objekts ist eine Funktion der Übersetzung seiner Teile.

2. Formale Probleme

Die wichtigsten Probleme der EUROTRA-Theorie, die zur Zeit bestehen, sind folgende:

- die Autonomie der ECS, die zu einer erheblichen Übergenerierung führt
- die relative Indadäquatheit der t-rules für die Behandlung von Sprachen mit freier Wortstellung.

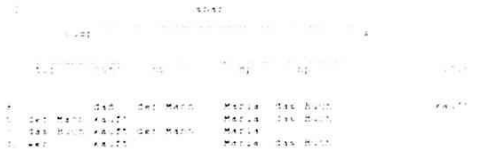
In der derzeitigen Version sieht der EUROTRA-Formalismus eine autonome ECS vor, d.h. Konstituentenstrukturen werden ohne Kontrolle durch Subkategorisierungsinformation erzeugt. Diese Art von Informationen wird erst auf ERS eingeführt, wo sie zur Reduzierung der Übergenerierung führt. Da dies keine prinzipiellen linguistischen Probleme erzeugt, ist dieses Problem nicht so bedeutend; lediglich für die Testarbeiten müssen längere Laufzeiten in Kauf genommen werden. Dagegen führt die Inad-

äquatheit der t-rule-Komponente zu prinzipiellen linguistischen Schwierigkeiten: da die t-rules nach dem Prinzip der expliziten Aufzählung arbeiten, wird diese Komponente bei Sprachen mit relativ freier Wortstellung sehr umfangreich und unübersichtlich. Alle linken Seiten der t-rule müssen aufgezählt werden, und es muß jeweils auf der rechten Seite notiert werden, an welchen Platz der Konstituent gehen soll. Dies bedeutet beispielweise, daß für die Beispiele in (1) zwei t-rules existieren müssen, eine für das topikalisierte Subjekt und eine für das topikalisierte direkte Objekt:

Dies bedeutet allgemein, daß alle möglichen Wortstellungskombinationen aufgezählt werden müssen, was eine große Anzahl von t-rules zur Folge hat. Dabei gehen alle Generalisierungen wieder verloren, die auf ECS bereits festgelegt sind; außerdem ist keine (ökonomische) Behandlung von long-distance-Phänomenen möglich. Zur Zeit werden vom EUROTRA-Zentralteam verschiedene Abänderungen am Formalismus diskutiert, die diese Probleme beheben sollen.

3. Konfigurationale Struktur (ECS)

Auf dieser Ebene selbst ist es möglich, Regeln zur deutschen Syntaxanalyse zu formulieren, die den aktuellen Forschungsstand widerspiegeln. Die Regeln folgen der einschlägigen Literatur (M. Reis, H. Haider, H. den Besten etc.): von einer kanonischen Wortstellung (Endstellung des finiten Verbs) werden alle Varianten durch "movement-rules" abgeleitet.



Der Verbalkomplex wird auf die folgende Weise behandelt:

Diese Regel erlaubt eine leichte Berechnung von Zeit und Diathese sowie eine Tilgung der Hilfsverben beim Übergang zur relationalen Ebene.

Dadurch ergibt sich der derzeitige Stand der Gesamtgrammatik wie folgt:

Den Variationsmöglichkeiten der freien Ergänzungen wird durch das optionale Auftreten von adverbialen oder prapositionalen Elementen zwischen den Komplementen Rechnung getragen. Obwohl diese Lösung nur vorläufigen Charakter hat (siehe oben), können damit die meisten Strukturen des EUROTRA-Korpus, eines Textes aus der EG-Verwaltungssprache, bearbeitet werden.

4. Relationale Struktur (ERS)

Die relationale oder Dependenz-Struktur, die in etwa den f-Strukturen der Unifikationsgrammatiken entspricht, wird durch die Eigenschaft der lexikalischen Einheiten definiert, andere Elemente zu binden. Diese Eigenschaft wird "Valenz" genannt. Die Definition dieser Ebene für das Deutsche basiert auf Arbeiten des IdS Mannheim und ist detailliert in *Schmidt 1986 und 1987* beschrieben.

Für die Beschreibung der relationalen Struktur eines Satzes wird im Prinzip eine einzige Regel benötigt:

Ein Satz besteht aus dem regierenden Verb und einem bestimmten Satz an Komplementen, die optional einmal vorhanden sind, sowie einer beliebigen Zahl von freien Ergänzungen, was wieder durch den Kleene-Star ausgedruckt wird. Welche Features im aktuellen Fall mit Werten gefüllt werden, bestimmt der Eintrag fur das Verb:

```
(6)  gov,(cat = vrb, lu = kommen,
          comp0= (cat = np, case = nom)),
          comp4= (cat = pp, prep = von)))
```

Die Werte im gov-Argument werden durch Unifikation automatisch in die comp-Features des Satzes übertragen. Sogenannte "gentle rules" führen einen Default-Wert "no" ein fur alle anderen comp-Features. Auf diese Weise wird sichergestellt, daß bei der Überführung auf die relationale Ebene nur die Syntagmen in die Slots des Verbs eingesetzt werden, die die entsprechenden Eigenschaften aufweisen.

5. Die Beziehung zwischen ECS und ERS

Wie bereits erwähnt, sind die Überführungsregeln zwischen diesen beiden Ebenen, die ja für Analyse und Synthese in gleicher Weise verwendet werden, ein komplexes Problem. Dies gilt speziell für Sprachen mit relativ freier Wortstellung, in denen keine konfigurationelle Behandlung von Komplementen möglich ist. Auch für die Verbalgruppe sind vier Arten von T-Regeln nötig, die durch zwei paarweise Merkmale charakterisiert sind:

- die Stellung des Verbs (1,2, final)
- das Vollverb bzw. Hilfsverb

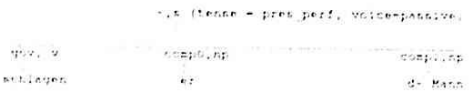
Die vier Regelarten sind also:

V/1,2 Vollverb
V/1,2 Hilfsverb
V/final Vollverb
V/final Hilfsverb

Ein Beispiel soll die Überführung von der ECS-Struktur in die ERS veranschaulichen:

(7) ECS:

(8) ERS:



Auch für diese Beschreibungsebene ist für die Strukturen des EUROTRA-Korpus eine Implementierung erfolgt.

Literatur

Die Literaturangaben für die Fachaufsätze zum Themenschwerpunkt Maschinelle Übersetzung finden sich zusammengefaßt auf Seite 23.