Reguläre Grammatiken/Sprachen und endliche Automaten

Bei regulären Grammatiken ist die Form der Grammatikregeln am stärksten eingeschränkt. Trotzdem lassen sich bereits weite Teile einer natürlichen Sprache damit beschreiben, wenn auch auf sehr umständliche Weise. Gleichzeitig erlaubt diese Einschränkung eine einfache syntaktische Analyse.

Grundgedanke:

> **Zerlegung einer Konstituente in einen Anfang (Terminalzeichen) und eine dazu passende Fortsetzung (Nichtterminalzeichen)**

Am einfachsten ist dies, falls eine Konstituente mit obligatorischen Elementen beginnt:

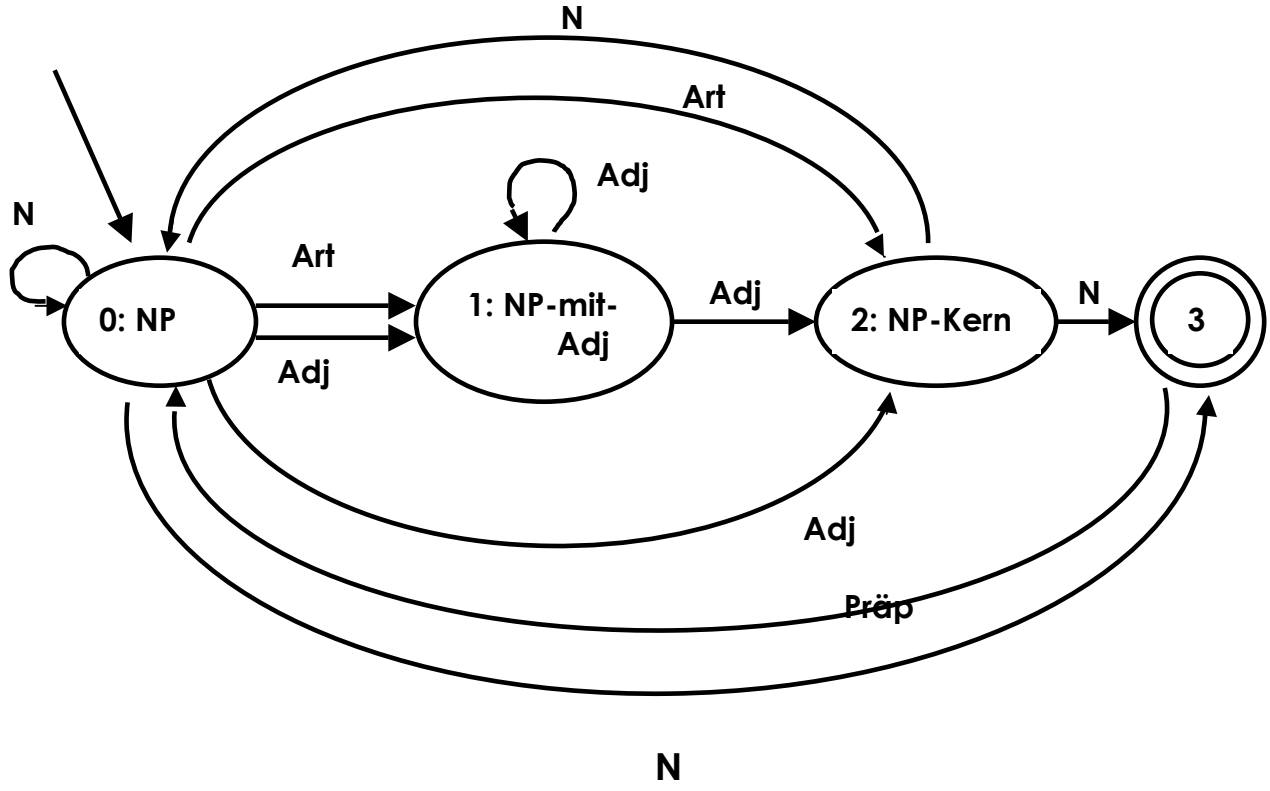
S → Pron VP
VP → V
VP → V NP

Beginnt eine Konstituente mit fakultativen Elementen, müssen alle Möglichkeiten aufgezählt werden:

NP → Art NP-mit-Adj	
NP → Art NP-Kern	
NP-mit-Adj → Adj NP-Kern	**NP → Adj NP-Kern**
NP-mit-Adj → Adj NP-mit-Adj	**NP → Adj NP-mit-Adj**
NP-Kern → N	**NP → N**
NP-Kern → N NP	**NP → N NP**
NP-Kern → N PP	**NP → N PP**
PP → Präp NP	

BNF-Form: **NP → [Art] {Adj} N {NP} {PP}**

graphische Darstellung als sog. endlicher Automat



Tabellarische Darstellung:

alter Zustand	Eingabezeichen	neuer Zustand
0	Art	1
0	Art	2
0	Adj	1
0	Adj	2
0	N	3
0	N	0
1	Adj	1
1	Adj	2
2	N	3
2	N	0
3	Präp	0

Startzustand: 0; Endzustand: 3

Definition: Endlicher Automat:

- **Menge von Zuständen,**
  darunter Startzustand und ein oder mehrere Endzustände
- **Eingabealphabet**
- **Überführungsfunktion, die für jeden Zustand und jedes Eingabezeichen den oder die Zielzustände angibt**

Ein endlicher Automat veranschaulicht den Ablauf der Analyse eines Ausdrucks:

Die Analyse beginnt im Startzustand.

Anschließend wird für jedes Zeichen des zu analysierenden Wortes in einen neuen Zustand gewechselt (oder derselbe Zustand beibehalten); das Eingabezeichen und der bisherige Zustand legen zusammen fest, welches der Zielzustand ist.

Am Ende des Wortes muß sich der Automat in einem Endzustand befinden; ansonsten gehört es nicht zur Sprache des Automaten. Nicht dazu gehören demnach etwa "Art Art N" oder "Art Adj".

Die Zustände eines Automaten drücken in der Regel Informationen über die Struktur der Ausdrücke aus (ebenso wie die Nichtterminalzeichen einer Grammatik).

Beispiel NP-Analyse:

Zustand 0: Anfang einer NP
Zustand 1: Artikel gelesen
Zustand 2: Adjektiv gelesen
usw.

Zusammenhang endlicher Automat - reguläre Sprache:

Jede reguläre (= Typ 3-) Sprache wird durch einen endlichen Automaten erkannt und umgekehrt.

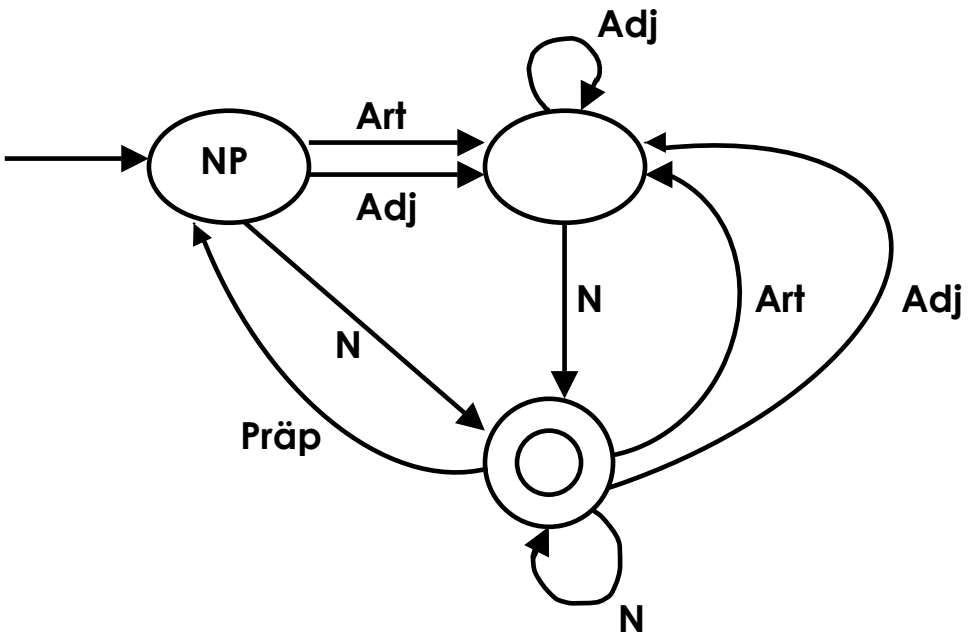
Jedem Zustandsübergang $\delta(z_1,a)=z_2$ entspricht eine Regel $z_1 \rightarrow a\ z_2$ (sowie, falls z_2 ein Endzustand ist, zusätzlich $z_1 \rightarrow a$).

Bei der Analyse eines Satzes mit Hilfe des o.g. Automaten (S. 12) entsteht das Problem, dass von demselben Knoten mehrere Kanten mit derselben Markierung ausgehen: der Automat ist nichtdeterministisch.

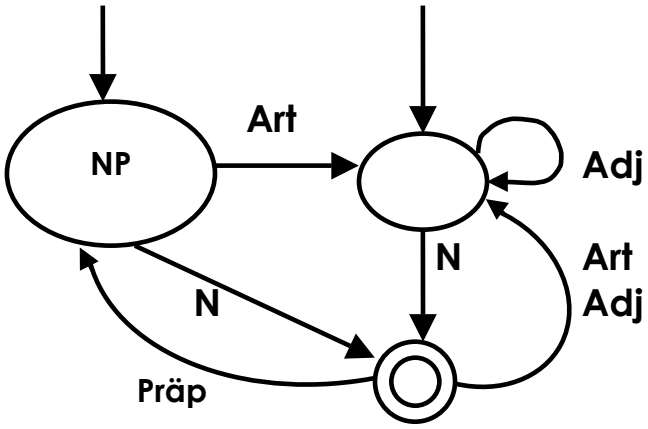
Jeder solche nichtdeterministische Automat läßt sich in einen äquivalenten deterministischen Automaten überführen.

Äquivalenz bedeutet: Für den nichtdeterministischen Automaten gilt ebenso wie für den deterministischen, daß er genau diejenigen Wörter erkennt, für die es einen Pfad von einem Startzustand zu einem Endzustand gibt.

<u>Deterministischer Automat:</u>



Nichtdeterministische endliche Automaten können beliebig viele Startzustände haben.

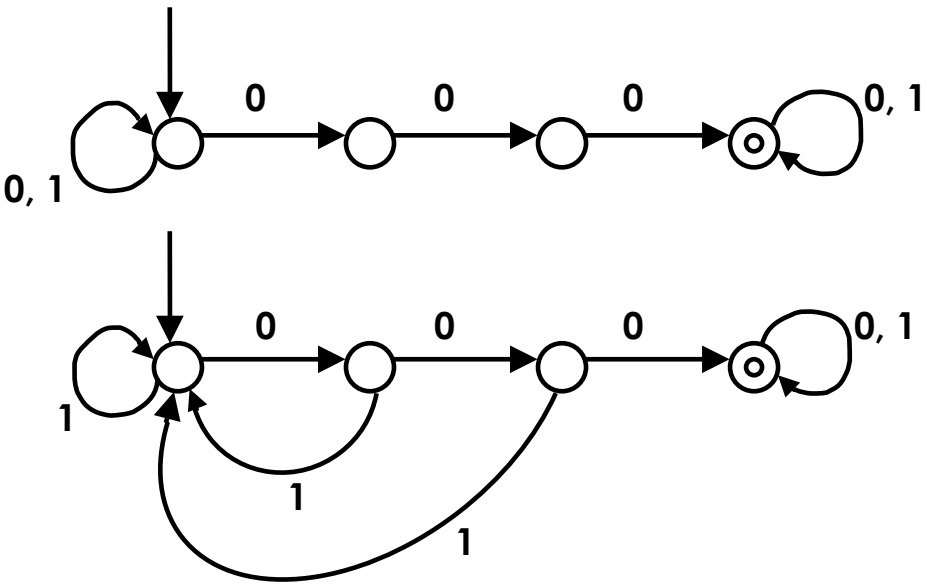


Für viele Sprachen ist es einfacher, einen nichtdeterministischen statt eines deterministischen Automaten anzugeben. Allerdings ist es schwieriger, einen gegebenen Ausdruck durch einen nichtdeterministischen Automaten zu analysieren:

Wenn es mehrere Startzustände gibt oder wenn es zu einem Eingabesymbol mehrere Zustandsübergänge gibt, muß der Automat "raten", bei welchem Startzustand die Analyse beginnen soll, bzw. welches die richtige Fortsetzung ist.

Beispiel:

Menge aller Zeichenfolgen über {0,1}, die die Zeichenfolge "000" enthalten; der Automat "rät", wann diese Zeichenfolge beginnt



Notationskonventionen:

1. Hängt man Wörter verschiedener regulärer Sprachen aneinander, ergibt sich wieder eine reguläre Sprache. (= "Die regulären Sprachen sind abgeschlossen unter Konkatenation.")

➔

Weiß man, dass ein bestimmter Satzteil (z.B. eine Nominalphrase) durch eine reguläre Grammatik beschrieben und damit durch einen endlichen Automaten erkannt werden kann, so kann man diesen Satzteil als Etikett einer einzigen Kante verwenden.

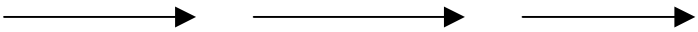
Allerdings ist es nicht ganz einfach, die einem solchen Automaten zugrundeliegende reguläre Grammatik zu rekonstruieren. Insbesondere werden für die beiden NPs teilweise unterschiedliche Regeln benötigt.

Außerdem ist Vorsicht bei der Anwendung unmittelbarer Rekursion geboten (dasselbe Nichtterminalzeichen auf beiden Seiten der Regel).

2. Die Menge aller Wörter, die nicht zu einer regulären Sprache gehören, bilden ebenfalls eine reguläre Sprache. (= "Die regulären Sprachen sind abgeschlossen unter Komplementbildung.")

➔

Kanten dürfen auch mit denjenigen Kategorien etikettiert werden, mit denen sie nicht durchlaufen werden dürfen



Probleme:

- **Eine reguläre Sprache läßt nur eine einzige Gestalt für die Strukturbäume zu ("rechtsverzweigende Binärbäume"), also nicht die unter linguistischen Aspekten gewünschten Strukturen.**
  (Beispiel: beliebige Folge von NP und PP nach einer NP)

- **Treten dieselben Konstituenten mehrfach auf (z.B. eine NP vor dem Verb und eine NP nach dem Verb), müssen sie trotzdem mehrfach in die Grammatik aufgenommen werden, weil sie unterschiedliche Fortsetzungen besitzen.**

- **Nicht alle Elemente einer natürlichen Sprache lassen sich durch eine reguläre Grammatik beschreiben.**

Grenzen:

Wann ist eine Sprache nicht mehr regulär?

Da ein endlicher Automat nur endlich viele (*n*) Zustände hat und bei jedem neu eingelesenen Zeichen ein Zustandsübergang erfolgt, lassen sich Ausdrücke, die länger sind als *n*, nur akzeptieren, wenn der Automat Schleifen enthält. Diese Schleifen kann man beliebig oft durchlaufen. Wenn es ein Wort aus der vom Automaten erkannten Sprache gibt, das länger als *n* ist, kann man es also "aufpumpen", so daß beliebig lange Wörter entstehen. Alle längeren Wörter lassen sich so auf kürzere zurückführen. Sprachen, bei denen dies nicht möglich ist, sind nicht mehr regulär.

Beispiele:

a) Die Sprache L={ $(ab)^m$, $m \geq 1$} ist regulär.

b) Die Sprache L={ $a^m b^m$, $m \geq 1$} ist nicht regulär.

c) Die koordinative Verknüpfung von Nominalphrasen ist regulär.

Eine Nominalphrase und noch eine, die dritte und die vierte können durch Schleifen dargestellt werden.

d) Die rekursive Einbettung von Relativsätzen ist nicht regulär.

Dieses Problem, das in einem deutschen Satz, der mehrere Nebensätze, die ineinander geschachtelt sind, enthält, auftritt, erfordert eine kontextfreie Lösung.