Algorithmen und Formale Sprachen

- **Algorithmen und formale Sprachen**
- **Formale Sprachen und Algorithmen**
- **Formale Sprachen und formale Algorithmen**

(formale (Sprachen und Algorithmen))
((formale Sprachen) und Algorithmen)

(((formale) Sprachen) und (Algorithmen))
(und ((Sprachen (formale)) (Algorithmen)))

Eine formale Grammatik beschreibt eine Sprache, indem sie exakt festlegt, welche Wörter in welcher Reihenfolge einen Satz dieser Sprache bilden. Außerdem weist sie diesem Satz eine bestimmte grammatische (syntaktische) Struktur zu. Mehrdeutige Sätze besitzen mehrere Strukturen.

Ist Deutsch eine formale Sprache??

Ist es sinnvoll, sämtliche deutschen Sätze mit derselben formalen Grammatik zu beschreiben?

Zur Bestimmung der syntaktischen Struktur eines Satzes hinsichtlich einer bestimmten formalen Grammatik gibt es verschiedene Verfahren (Parsingalgorithmen), die so präzise formuliert sind, dass sie vom Computer ausführbar sind.

Weitere Algorithmen beschreiben die Verarbeitung einfacher grafischer Darstellungen (Baum, Netz)

Formale Sprachen

Aus Sicht der Informatik ist eine Sprache eine Menge von Zeichenfolgen über einem Alphabet.

<u>Beispiele:</u>

- **Morsezeichen**
- **{3, 33, 298, 38956}**
- **Menge aller Primzahlen**
- **Menge aller Namen der EinwohnerInnen von Duisburg**
- **Menge aller syntaktisch korrekten VisualBasic-Programme**
- **Menge aller deutschen Sätze**

Manche Sprachen (Mengen) sind endlich, andere unendlich groß.
Jedes einzelne Element der Menge (Wort der Sprache) hat aber nur eine endliche Länge.

Um eine Sprache benutzen zu können, benötigt man eine Berechnungsvorschrift, die angibt, welche Ausdrücke zu einer Sprache gehören und welche nicht.

Bei endlichen Sprachen kann diese Berechnungsvorschrift darin bestehen, alle Elemente aufzuzählen. Für Sprachen mit unendlich vielen Elementen muß es endlich viele Regeln geben, mit denen sich die Ausdrücke dieser Sprache erzeugen lassen.

Worin könnten solche Regeln für eine natürliche Sprache (z.B. Deutsch) bestehen?

Zur Erstellung einer deskriptiven Grammatik betrachtet man eine Menge (Korpus) von Elementen der zu beschreibenden Sprache:

Die Wirtschaft sucht ausländische Computerexperten.
Die IT-Branche boomt.
Kritiker fordern stärkere Ausbildungsaktivitäten.
Der Bundeskanzler hält eine programmatische Rede.
Er plant großzügigere Regelungen.
Die zuständigen Ministerien diskutieren verschiedene Modelle.

Eine sehr einfache traditionelle sprachwissenschaftliche Grammatik würde diese Sätze folgendermaßen beschreiben:

Ein Satz besteht aus einem Subjekt und einem Prädikat.
Das Prädikat besteht aus einem Verb sowie eventuell einem Objekt.
Subjekt und Objekt bestehen jeweils aus einem Substantiv, vor dem eventuell ein Artikel und/oder ein Adjektiv steht, oder aus einem Personalpronomen.

Die Wörter eines Satzes werden also durch ihre Wortarten beschrieben und stufenweise zu immer umfassenderen Konstituenten zusammengefaßt. Formal lauten diese sog. Phrasenstrukturregeln:

S → SUBJEKT PRÄDIKAT
SUBJEKT → SUBSTANTIV
SUBJEKT → ARTIKEL SUBSTANTIV
SUBJEKT → ADJEKTIV SUBSTANTIV
SUBJEKT → ARTIKEL ADJEKTIV SUBSTANTIV
SUBJEKT → PERSONALPRONOMEN
PRÄDIKAT → VERB
PRÄDIKAT → VERB OBJEKT

OBJEKT → SUBSTANTIV
OBJEKT → ARTIKEL SUBSTANTIV
OBJEKT → ADJEKTIV SUBSTANTIV
OBJEKT → ARTIKEL ADJEKTIV SUBSTANTIV
OBJEKT → PERSONALPRONOMEN

Phrasenstrukturregeln beschreiben die Bestandteile (Konstituenten) eines Satzes sowie deren Reihenfolge (Präzedenz).

Benötigt werden außerdem Lexikoneinträge, d.h. Regeln, die jedem Wort seine Wortart zuordnen:

SUBSTANTIV → ***Wirtschaft***
SUBSTANTIV → ***IT-Branche***
VERB → ***sucht***
PERSONALPRONOMEN → ***er***
usw.

Die o.g. Grammatik lässt sich vereinfachen, indem man feststellt, dass Subjekt und Objekt offenbar dieselbe Form haben:

S → **SUBJEKT PRÄDIKAT**
SUBJEKT → **NOMINALPHRASE**
PRÄDIKAT → **VERB**
PRÄDIKAT → **VERB OBJEKT**
OBJEKT → **NOMINALPHRASE**
NOMINALPHRASE → **SUBSTANTIV**
NOMINALPHRASE → **ARTIKEL SUBSTANTIV**
NOMINALPHRASE → **ADJEKTIV SUBSTANTIV**
NOMINALPHRASE → **ARTIKEL ADJEKTIV SUBSTANTIV**
NOMINALPHRASE → **PERSONALPRONOMEN**

Beide Grammatiken beschreiben dieselbe Sprache.

Dieselbe formale Sprache kann durch unterschiedliche formale Grammatiken beschrieben werden.

Die grammatische Struktur eines Satzes muss nicht eindeutig sein. Bei der grammatischen Analyse eines Satzes ist daher zu beachten, dass ein Satz mehrere Interpretationen haben kann.

Beispiel für eine Grammatik, die derselben Wortfolge mehrere Strukturen zuweist:

S → SUBJEKT PRÄDIKAT
SUBJEKT → PERSONALPRONOMEN
PRÄDIKAT → VERB OBJEKT
PRÄDIKAT → VERB ADVERB OBJEKT
OBJEKT → ADJEKTIV SUBSTANTIV
OBJEKT → ADVERB ADJEKTIV SUBSTANTIV

Er erfindet schnell konstruierte Sätze.
Sie ißt häufig aufgewärmtes Sauerkraut.

Er verkauft frisch zubereitete Pizzen.

Eine formale Grammatik besteht also aus 4 Elementen (einem sog. Quadrupel):

- **einer Menge von Terminalzeichen**

  (d.h. denjenigen Zeichen, aus denen die Kette besteht, die von der Grammatik beschrieben werden soll)

  Strenggenommen sind bei einer Grammatik für eine Natürliche Sprache die Wörter die Terminalzeichen.
  Es ist jedoch sinnvoller, als Terminalzeichen die Menge aller Wortarten zu verwenden („Präterminalzeichen“) und zusätzliche lexikalische Regeln zu verwenden, die die Zuordnung von Terminalzeichen zu Wörtern steuern.

- **einer Menge von Variablen**

  (syntaktischen Kategorien, Nichtterminalzeichen)

- **einer Startvariable**

- **einer Menge von Regeln, die eine Kette von Terminal- bzw. Nichtterminalzeichen in eine andere solche Kette überführen**

Sinnvolle Beschränkung:

Keine Regel sollte Terminalzeichen durch andere Terminalzeichen ersetzen.

Die Menge aller Ketten von Terminalzeichen, die sich aufgrund der Regeln aus der Startvariablen ableiten lassen, bildet die von der Grammatik erzeugte Sprache.

Leider erzeugen die Grammatiken auf S. 3/4 auch Sätze, die nicht zu der gewünschten Sprache gehören:

Die IT-Branche boomt stärkere Ausbildungsaktivitäten.

Kritiker fordern der Bundeskanzler.

Der Bundeskanzler hält programmatische Rede.

Er diskutieren großzügigere Regelungen.

Die zuständigen Ministerien diskutieren verschiedenen Modelle.

Die Wirtschaft sucht eine programmatische Rede.

Der Bundeskanzler boomt.

Kritiker fordern die IT-Branche.

Die letzteren Sätze sind nur aus semantischen Gründen inkorrekt. Im folgenden betrachten wir nur die syntaktische Korrektheit.

Neben Konstituenz- und Präzedenzvorschriften muss eine Grammatik für das Deutsche also auch weitere syntaktische Phänomene abdecken:

- **Numeruskongruenz zwischen Subjekt und Verb**
- **Kasus-Numerus-Genus-(KNG-)Kongruenz innerhalb einer Nominalphrase; Abhängigkeit des Adjektivs vom Artikel**
- **Verbvalenz (Anzahl und Form der Objekte; Subjekt im Nominativ)**

Die Erzeugung der inkorrekten Sätze läßt sich auf verschiedene Weisen vermeiden:

- **durch Einführung neuer Kategorien:**

  SATZ → SUBJEKT-SG PRÄDIKAT-SG
  SUBJEKT-SG → ARTIKEL-SG SUBSTANTIV-SG
  SUBSTANTIV-SG → ***Wirtschaft***
  SUBSTANTIV-PL → ***Ministerien***
  usw.

- **durch Angabe des Kontextes:**

  PRÄDIKAT → VERB ARTIKEL ADJEKTIV SUBSTANTIV
  PRÄDIKAT → VERB ADJEKTIV SUBSTANTIV
  ARTIKEL ADJEKTIV → ARTIKEL ***verschiedenen***
  VERB ADJEKTIV → VERB ***verschiedene***

Üblicherweise verwendet man Grammatiken, deren linke Seite aus einem einzigen Nichtterminalzeichen besteht. Bei mehreren Zeichen auf der linken Seite ist meist keine Baumdarstellung der syntaktischen Struktur mehr möglich.

SATZ → SUBJEKT VERB OBJEKT
SUBJEKT → ARTIKEL SUBSTANTIV
ARTIKEL SUBSTANTIV → EIGENNAME
EIGENNAME → ***Duisburg***
VERB OBJEKT → ***blüht***

Alternative: Regeln, die das sog. leere Wort ableiten. Dies ist zwar zulässig, führt aber bei der Syntaxanalyse zu großen Schwierigkeiten.

VERB → ***blüht***
OBJEKT → ε

Chomsky-Hierarchie der formalen Sprachen

Durch zunehmende Einschränkung der Gestalt der Grammatikregeln legt man verschiedene Typen von Grammatiken fest. Je eingeschränkter die Regeln, desto einfacher sind die Analyseverfahren, desto eingeschränkter aber auch die entsprechenden Sprachen.

Typ 0 (allgemeine Phrasenstrukturgrammatiken):

keine Einschränkung der Regeln

Typ 1 (kontextsensitive Grammatiken):

Die rechte Seite einer Regel darf nicht kürzer sein als die linke.
(Das heißt insbesondere, daß keine Regeln zulässig sind, die das leere Wort ableiten.)

Typ 2 (kontextfreie Grammatiken):

Kontextsensitive Grammatiken, bei denen auf der linken Seite nur ein einzelnes Nichtterminalzeichen steht.

Typ 3 (reguläre Grammatiken):

Kontextfreie Grammatiken, bei denen auf der rechten Seite nur ein einzelnes Terminalzeichen steht, eventuell gefolgt von einem einzelnen Nichtterminalzeichen (bzw. umgekehrt).

Backus-Naur-Form (BNF):

Mehrere Regeln mit derselben linken Seite kann man verkürzt in der sog. Backus-Naur-Form (BNF) schreiben:

A → a | b | c

statt

A → a
A → b
A → c

A → a [b] c

statt

A → a c
A → a b c

A → a {b} c

statt

A → a c
A → a B c
B → b
B → b B