

Syntax und Parsing mit CFGs

- Die Syntax natürlicher Sprachen ist nicht-regulär. Sie kann durch endliche Automaten nicht vollständig beschrieben werden.
- Ein geeignetes (vorsichtiger: ein geeigneteres) Darstellungsformat für die Syntax natürlicher Sprachen sind kontextfreie Grammatiken (CFG). Sprachen, die mit CFGs beschrieben werden können, heißen „kontextfreie Sprachen“.
- Kontextfreie Sprachen werden von Kellerautomaten erkannt bzw. generiert.
- Verfahren oder Systeme, die die Regeln einer Grammatik in einen syntaktischen Analyseprozess umsetzen, heißen Parser.
- Der Top-Down-Parser und der Bottom-Up-Parser, die in der Vorlesung vorgestellt wurden, sind Kellerautomaten, die die Regeln einer CFG zur syntaktischen Analyse benutzen.

Ein Algorithmus für den Top-Down-Parser

- Die Parsing-„Verfahren“, die in der letzten Vorlesung vorgestellt wurden, sind nicht-deterministisch - und deshalb keine Verfahren, die gewährleisten, dass eine mögliche Analyse auch gefunden wird.
- Im folgenden wird ein Algorithmus für das Top-Down-Parsing vorgeführt, der, wie der Tiefensuch-Algorithmus für Zustandsdiagramme, eine Agenda verwendet, in der die alternativen Resultate für die verschiedenen möglichen Analyseschritte abgelegt werden.
- Unterschied: Zusätzlich zur Position in der Eingabekette muss der Stackinhalt des Kellerautomaten notiert werden (der Zustand des Automaten braucht in diesem Beispiel nicht notiert zu werden, weil er nach der Initialisierung des Stack mit „S“ immer 1 ist).

Beispielgrammatik

- Die folgende Grammatik wird für die Beispiele in den folgenden Folien zugrundegelegt:

$S \rightarrow NP VI$

$NP \rightarrow ART NN$

$S \rightarrow NP VT NP$

$NP \rightarrow PN$

$S \rightarrow S NP$

$NP \rightarrow NP PP$

$PP \rightarrow P NP$

$PP \rightarrow PAC NN$

$VI \rightarrow \textit{schläft, arbeitet}$

$VT \rightarrow \textit{verfolgte, studiert, liest, sah}$

$NN \rightarrow \textit{Detektiv, Gangster, Studentin, Sportwagen, Buch, Mann, Teleskop}$

$ART \rightarrow \textit{der, den, dem}$

$PN \rightarrow \textit{Peter, Maria}$

$P \rightarrow \textit{in, mit}$

$PAC \rightarrow \textit{im}$

Verarbeitung mit Top-Down-Parser

₀Der ₁ *Detektiv* ₂ *im* ₃ *Sportwagen* ₄ *verfolgte* ₅ *den* ₅ *Gangster* ₇

Agenda:

Stack:

S

Verarbeitung mit Top-Down-Parser

₀Der ₁ *Detektiv* ₂ *im* ₃ *Sportwagen* ₄ *verfolgte* ₅ *den* ₅ *Gangster* ₇

Agenda:

0, NP VI
0, NP VT NP
0, S PP

Stack:

S

Verarbeitung mit Top-Down-Parser

Der₁ Detektiv₂ im₃ Sportwagen₄ verfolgte₅ den₅ Gangster₇

Agenda:

0, NP VT NP
0, S PP

Stack:

S
NP VI

Verarbeitung mit Top-Down-Parser

Der₁ Detektiv₂ im₃ Sportwagen₄ verfolgte₅ den₅ Gangster₇

Agenda:

0, ART NN VI
0, PN VI
0, NP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI

Verarbeitung mit Top-Down-Parser

₀Der ₁ *Detektiv* ₂ *im* ₃ *Sportwagen* ₄ *verfolgte* ₅ *den* ₅ *Gangster* ₇

Agenda:

0, PN VI
0, NP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN VI

Verarbeitung mit Top-Down-Parser

₀Der ₁ *Detektiv* ₂ *im* ₃ *Sportwagen* ₄ *verfolgte* ₅ *den* ₅ *Gangster* ₇

Agenda:

0, *der* NN VI
0, PN VI
0, NP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

0, PN VI
0, NP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN VI
der NN VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

0, PN VI
0, NP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN VI
der NN VI
NN VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

1, Detektiv VI
0, PN VI
0, NP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN VI
der NN VI
NN VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

0, PN VI
0, NP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN VI
der NN VI
NN VI
Detektiv VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

0, PN VI
0, NP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN VI
der NN VI
NN VI
Detektiv VI
VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

2, arbeitet
0, PN VI
0, NP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN VI
der NN VI
NN VI
Detektiv VI
VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

0, PN VI
0, NP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN VI
der NN VI
NN VI
Detektiv VI
VI
arbeitet

Verarbeitung mit Top-Down-Parser

*₀Der ₁ *Detektiv* ₂ *im* ₃ *Sportwagen* ₄ *verfolgte* ₅ *den* ₅ *Gangster* ₇*

Agenda:

0, NP PP VI
0, NP VT NP
0, S PP

Stack:

PN VI

Verarbeitung mit Top-Down-Parser

*₀Der ₁ *Detektiv* ₂ *im* ₃ *Sportwagen* ₄ *verfolgte* ₅ *den* ₅ *Gangster* ₇*

Agenda:

0, NP PP VI
0, NP VT NP
0, S PP

Stack:

NP VI
PN VI

Verarbeitung mit Top-Down-Parser

*₀Der ₁ *Detektiv* ₂ *im* ₃ *Sportwagen* ₄ *verfolgte* ₅ *den* ₅ *Gangster* ₇*

Agenda:

0, NP VT NP
0, S PP

Stack:

NP PP VI

Verarbeitung mit Top-Down-Parser

Der₁ Detektiv₂ im₃ Sportwagen₄ verfolgte₅ den₅ Gangster₇

Agenda:

0, ART NN PP VI
0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

NP PP VI

Verarbeitung mit Top-Down-Parser

*₀Der ₁ *Detektiv* ₂ *im* ₃ *Sportwagen* ₄ *verfolgte* ₅ *den* ₅ *Gangster* ₇*

Agenda:

0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

NP PP VI
ART NN PP VI

Verarbeitung mit Top-Down-Parser

₀Der ₁ *Detektiv* ₂ *im* ₃ *Sportwagen* ₄ *verfolgte* ₅ *den* ₅ *Gangster* ₇

Agenda:

0, *der* NN PP VI
0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

NP PP VI
ART NN PP VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

NP PP VI
ART NN PP VI
der NN PP VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

NP PP VI
ART NN PP VI
der NN PP VI
NN PP VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

1, Detektiv PP VI
0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

NP PP VI
ART NN PP VI
der NN PP VI
NN PP VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

NP PP VI
ART NN PP VI
der NN PP VI
NN PP VI
Detektiv PP VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN PP VI
der NN PP VI
NN PP VI
Detektiv PP VI
PP VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

2, PAC NN VI
2, P NP VI
0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN PP VI
der NN PP VI
NN PP VI
Detektiv PP VI
PP VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

2, P NP VI
0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN PP VI
der NN PP VI
NN PP VI
Detektiv PP VI
PP VI
PAC NN VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

2, P NP VI
0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN PP VI
der NN PP VI
NN PP VI
Detektiv PP VI
PP VI
PAC NN VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

2, P NP VI
0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN PP VI
der NN PP VI
NN PP VI
Detektiv PP VI
PP VI
NN VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

2, P NP VI
0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN PP VI
der NN PP VI
NN PP VI
Detektiv PP VI
PP VI
NN VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

2, P NP VI
0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN PP VI
der NN PP VI
NN PP VI
Detektiv PP VI
PP VI
NN VI
VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

2, P NP VI
0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN PP VI
der NN PP VI
NN PP VI
Detektiv PP VI
PP VI
NN VI
VI

Verarbeitung mit Top-Down-Parser

₀Der ₁Detektiv ₂im ₃Sportwagen ₄verfolgte ₅den ₅Gangster ₇

Agenda:

2, P NP VI
0, PN PP VI
0, NP PP PP VI
0, NP VT NP
0, S PP

Stack:

S
NP VI
ART NN PP VI
der NN PP VI
NN PP VI
Detektiv PP VI
PP VI
NN VI
VI

Etliche Schritte weiter ...

*₀Der ₁ *Detektiv* ₂ *im* ₃ *Sportwagen* ₄ *verfolgte* ₅ *den* ₅ *Gangster* ₇*

Agenda:

0, S PP

Stack:

NP VT NP

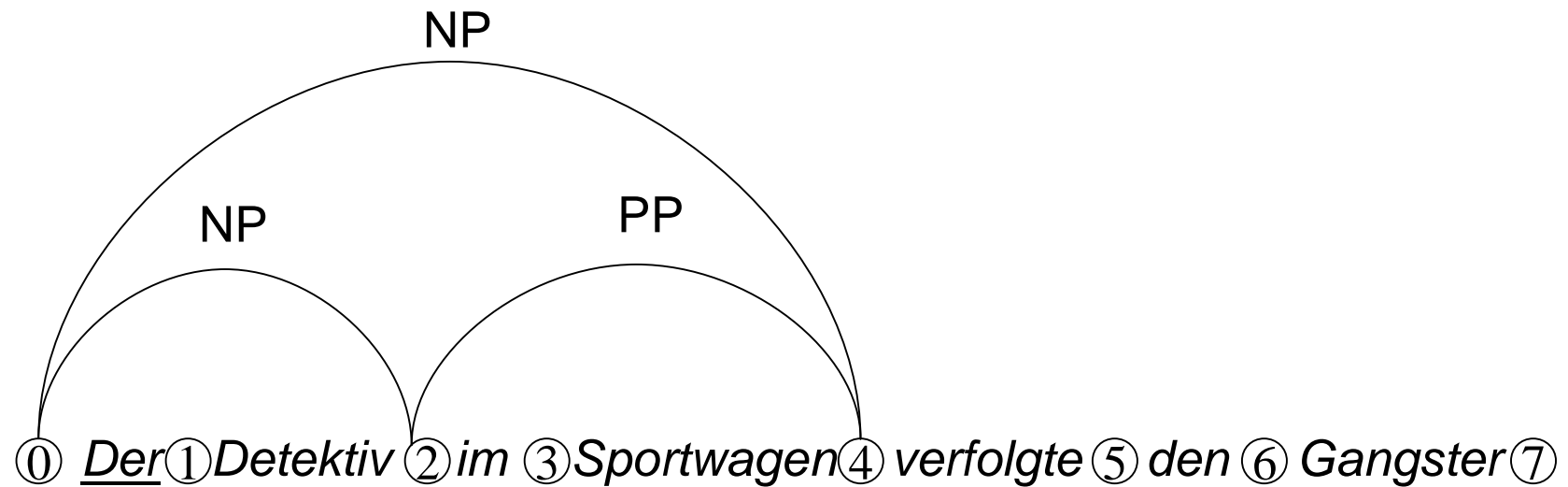
Ein Effizienzproblem

- Der Parser hat beim Abarbeiten des Verbs „festgestellt“, dass nicht die Regel $S \rightarrow NP VI$, sondern $S \rightarrow NP VT NP$ verwendet werden muss. Bis er zur alternativen Analyse gelangt, muss er aber die komplette NP-Analyse - mit allen Irrwegen - noch einmal identisch durchlaufen.
- Lösungsidee: Gesicherte Teilanalysen werden zwischen-gespeichert. Die Repräsentation, die für die Speicherung verwendet wird, nennt man Chart, das Parsen mithilfe einer Chart „Chart-Parsing“.

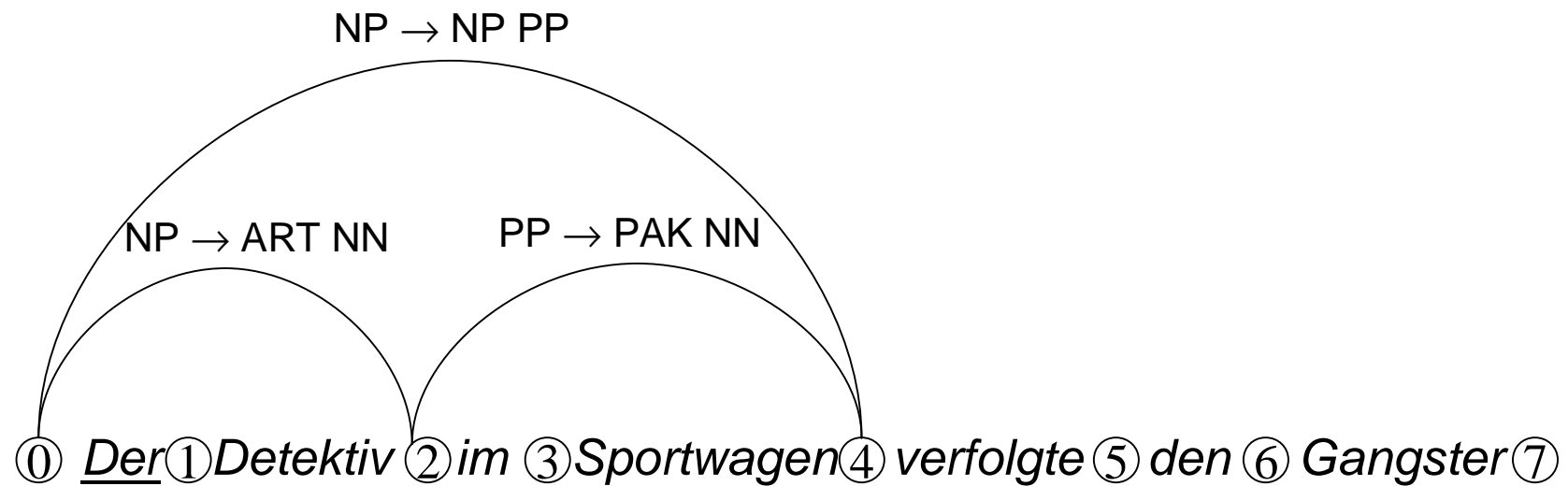
Die Chart, Beispiel

① Der ① *Detektiv* ② *im* ③ *Sportwagen* ④ *verfolgte* ⑤ *den* ⑥ *Gangster* ⑦

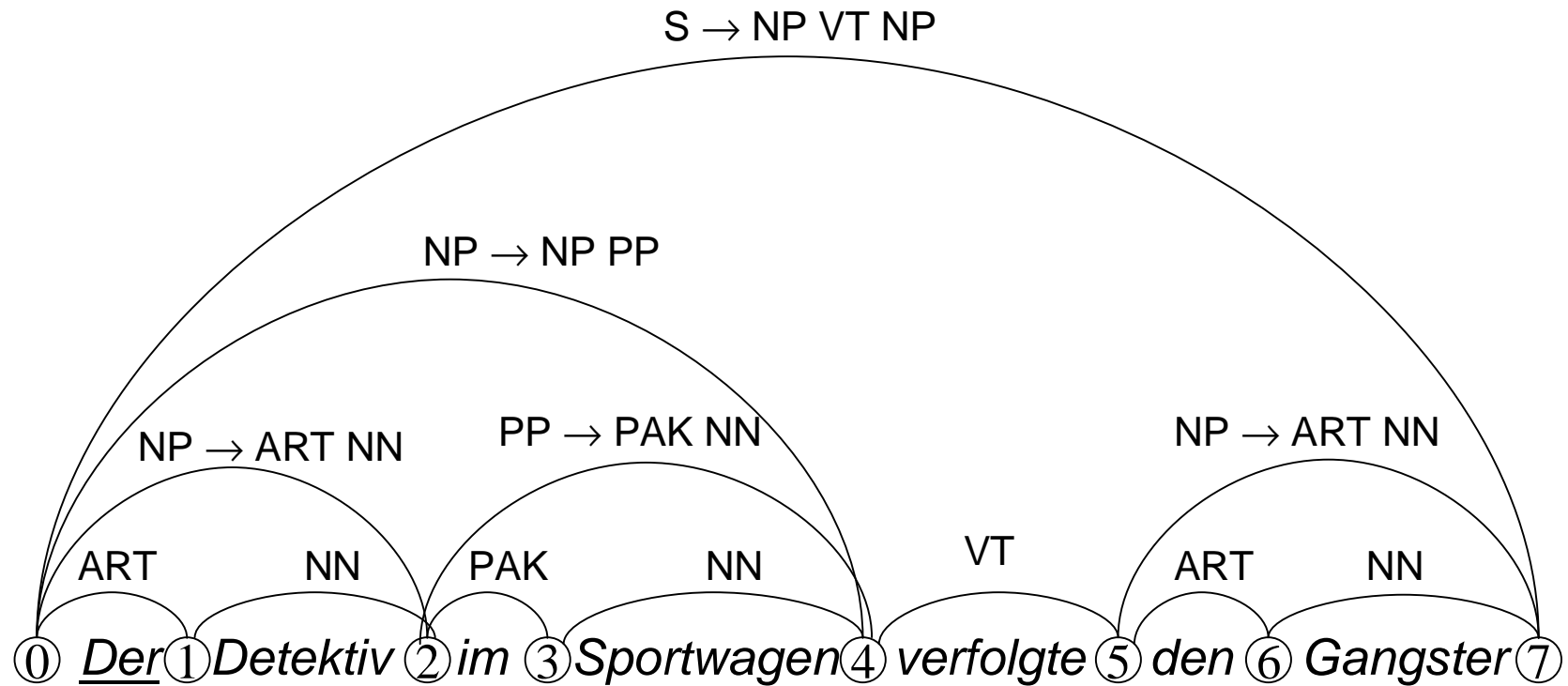
Die Chart



Die Chart



Die Chart



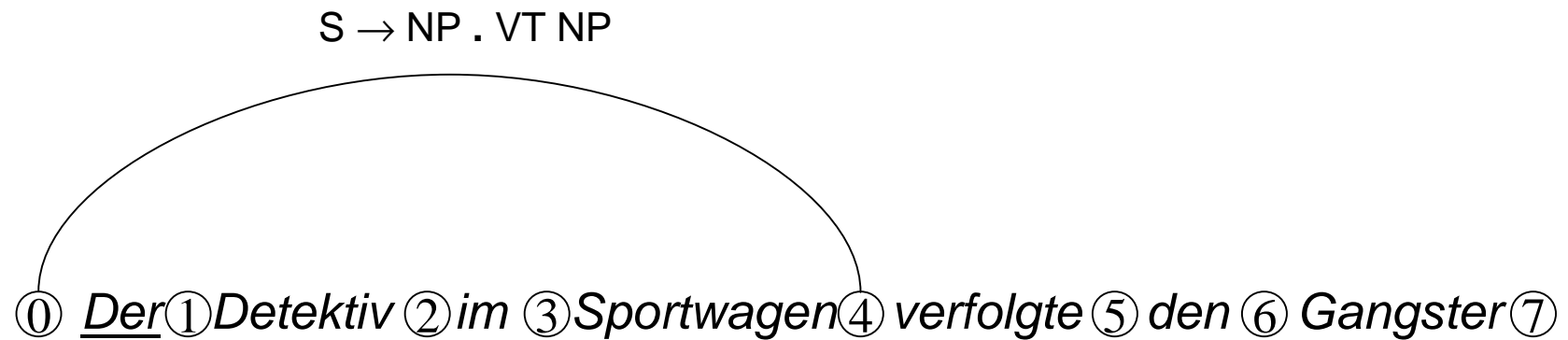
Die Chart

- Die Eingabekette wird als initialer Graph dargestellt: Knoten durchnummeriert, Kanten mit Wörtern beschriftet.
- Für eine erfolgreich abgearbeitete Regel wird eine neue Kante eingeführt, die die gefundene Konstituente überspannt. Die Kante wird mit der Regel beschriftet (Kategorie der Konstituente und ihrer Teilausdrücke).

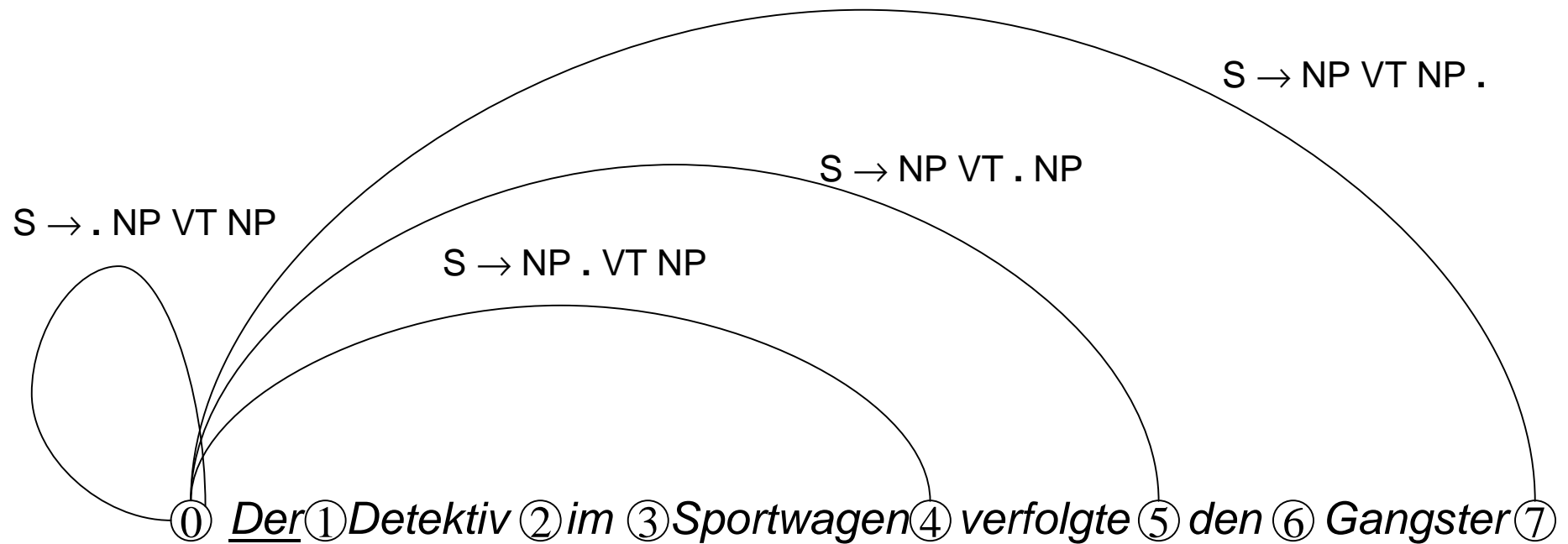
Chart-Parsing

- Wie können wir das Chart-Konzept in einen Parser integrieren - d.h., wie können wir beim Parsing-Prozess auf zwischengespeicherte Information zurückgreifen?
- Hierzu benötigen wir „aktive Kanten“.

Aktive Kanten



Aktive Kanten



Aktive Kanten

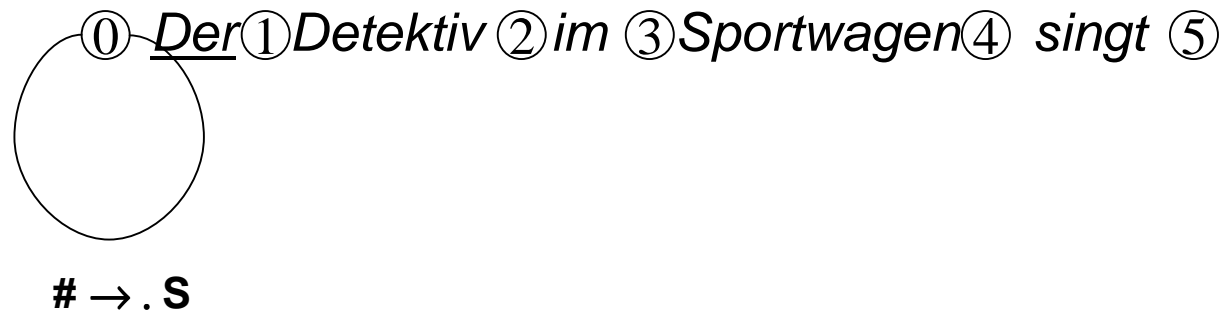
- $A \rightarrow \beta . \gamma [i,j]$ wird gelesen als: Zwischen i und j wurde eine Konstituente(nfolge) β abgearbeitet; wenn nach j eine Konstituente(nfolge) γ folgt, ergibt dies zusammen eine Konstituente vom Typ A .
- Beispiel: $S \rightarrow NP . VT NP [0,4]$ besagt: Zwischen 0 und 4 wurde eine NP erkannt, wenn nach 4 ein VT und eine NP folgen, ergibt das Ganze ein S.
- Extremfall 1, β leeres Wort: $S \rightarrow . NP VT NP [0,0]$
„An Position 0 könnte ein S beginnen, bestehend aus NP, VT und NP“
- Extremfall 2, γ leeres Wort: $S \rightarrow NP VT NP . [0,7]$:
Komplette Kante: Zwischen 0 und 7 wurde ein S, bestehend aus NP, VT und NP analysiert.

Der Earley-Algorithmus

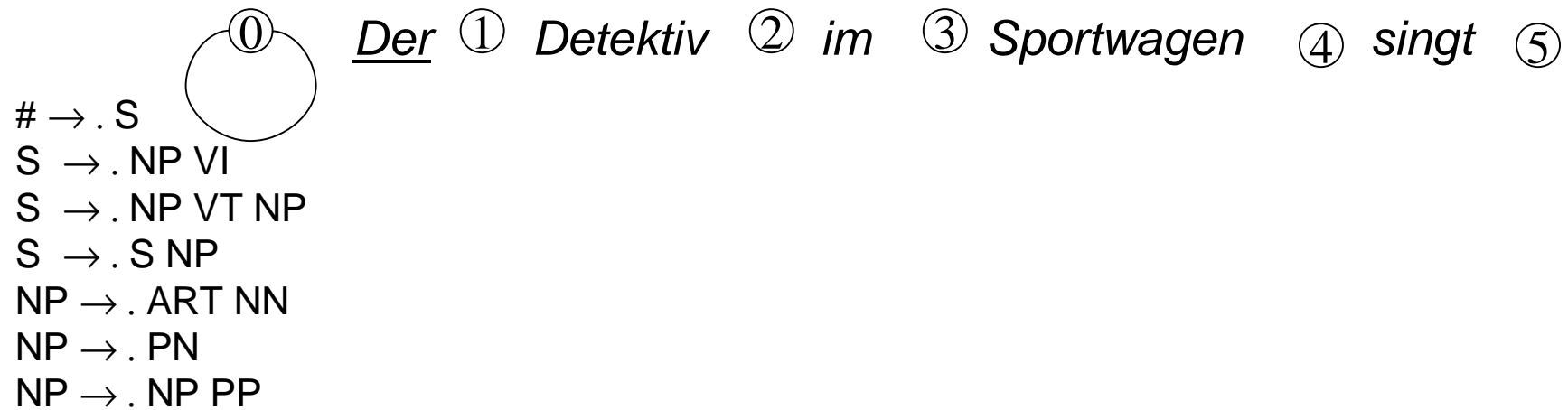
- Initialisierung:
 - Trage Eingabekette der Länge n in eine Chart ein, deren Knoten von 0 bis n nummeriert sind.
 - Trage neue aktive Kante $\# \rightarrow . S [0,0]$ ein.

Der Earley-Algorithmus

- Initialisierung:
 - Trage Eingabekette der Länge n in eine Chart ein, deren Knoten von 0 bis n nummeriert sind.
 - Trage neue aktive Kante $\# \rightarrow . S$ [0,0] ein.



Earley-Algorithmus: Beispiel



Der Earley-Algorithmus

- Predictor:
 - Für aktive Kanten $A \rightarrow \beta. B \gamma [i,j]$, B nicht-lexikalische Kategorie, trage neue aktive Kanten $B \rightarrow .\delta [j,j]$ für jede Regel der Grammatik $B \rightarrow \delta$ ein.
- Scanner:
 - Für aktive Kanten $A \rightarrow \beta. A \gamma [i,j]$, A lexikalische Kategorie, $A \rightarrow a$ Regel, und a Eingabewort zwischen i und j , trage $A \rightarrow a. [j,j+1]$ ein.
- Completer:
 - Für komplette Kanten $B \rightarrow \delta. [j,k]$, aktive Kanten $A \rightarrow \beta. B \gamma [i,j]$, trage $A \rightarrow \beta B. \gamma [i,k]$ ein.

Earley-Algorithmus: Beispiel

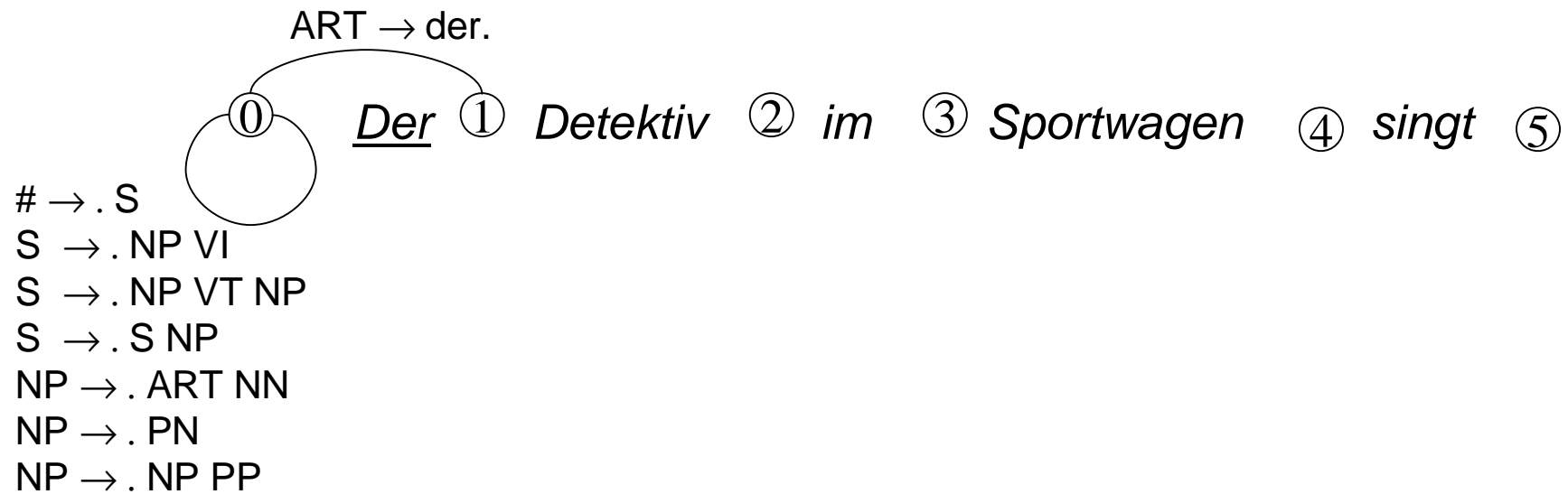


Chart-Parsing mit dem Earley-Algorithmus

- Ein Eingabesatz wird als grammatisch akzeptiert, wenn eine komplette S-Kante gefunden wird, die die gesamte Eingabe überspannt.
- Aus der Chart ist auch die syntaktische Struktur / der Analysebaum für den Eingabesatz ablesbar.
- Für syntaktisch mehrdeutige Eingabesätze enthält die Chart sämtliche Analysen in kompakter Form.
- Wenn keine korrekte Analyse gefunden werden kann, enthält die Chart trotzdem Information über Teilstrukturen.

Chart-Parsing mit dem Earley-Algorithmus

- Der Earley-Algorithmus ist ein
 - Top-Down-Parser, der
 - mit Breitensuche
 - von links nach rechts durch die Eingabe geht.
- Er hat einen entscheidenden Effizienzvorteil gegenüber dem naiven Top-Down-Parser, weil er nicht direkt auf der Eingabe, sondern auf einer Chart arbeitet, einer Datenstruktur, die zur Speicherung von Zwischenresultaten dynamisch erweitert werden kann.