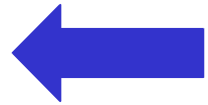
Elementare Probleme und Lösungen in der Sprachverarbeitung

Cristina Vertan

Walther v. Hahn

Inhalt

- Grenzen der Voice-XML unterstützten Grammatiken
- Grammatiken für natürliche Sprache
- Lexikonaufbau
- Elementare Parsing-Strategien
- Anaphorische Ausdrücke in Dialogsystemen
- Aufgaben für das Praktikum



Grenzen der Voice-XML unterstützten Grammatik -1-

Welche Farbe hat Ihr Wagen ?

Mein Wagen ist rot

Die Farbe meines Wagens ist rot

Mein Wagen ist blau

Die Farbe meines Wagens ist blau

Mein Wagen hat eine blaue Farbe

```
FarbWahl
(?FarbeIst : x){<colour $x|}
Haben
  [(Mein Wagen ist)(Die Farbe meines
Wagens ist)]
Farbe [ rot {return("rot")}
          blau {return ("blau")}
        ]
```

```
Farbe

(?(Mein Wagen hat eine) [blaue rote] (Farbe) )
```

Grenzen der Voice-XML unterstützten Grammatik -2-

Man kann mit solchen einfachen Grammatiken nicht Sätze modellieren wie:

- Mein Problem ist am Vergaser
- Mein Problem ist an den Rädern

.

Komplexere Aussagen wie

- Ich komme mit dem Wagenheber nicht zurecht
- Beim Losfahren höre ich manchmal seltsame Geräusche
- Wie fülle ich Kühlmittel nach ?

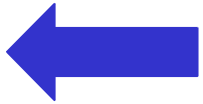
werden in einem VoiceXML System gar nicht unterstüzt.

Wenn der Sprecher solche Sätze benutzt, wird ein “default”-Prompt “Ich habe nicht verstanden, bitte benutzen Sie ....” zurückgegeben

Grenzen der Voice-XML-Systeme

- Der Benutzer wird gehindert, natürliche Sprache zu benutzen
- Das ist besonders unangenehm bei Informationssystemen, bei denen der Benutzer nicht konkrete Entscheidungen treffen will, sondern ein reales Gespräch führen will.
- Wenn man möglichst viele Aussagen erlauben will, muß man für fast jede von ihnen eine Grammatik schreiben, was unrealistisch und teuer ist.
- Durch die Beschränkung der Sprache kann es sogar passieren daß das beabsichtigte Thema teilweise falsch dargestellt wird.

Inhalt

- Grenzen der Voice-XML unterstützten Grammatiken
- Grammatiken für natürliche Sprache 
- Lexikonaufbau
- Elementare Parsing-Strategien
- Anaphorische Ausdrücke in Dialogsystemen
- Aufgaben für das Praktikum

Definitionen

- Grammatik: Beschreibung der Regeln zur korrekten Kombination der lexikalischen Elemente einer Sprache
- Es gibt verschiedene Grammatikformalismen (Konstituentenstruktur, Dependenzstruktur, LFG, HPSG, usw...)
- Konstituenten sind Wörter oder Folgen von Wörtern die als Einheit größere Einheiten bilden, wie etwa:
  - NP (Objekte, Plätze, Konzepte, Ereignisse)
  - VP (Verben oder Verben+Objekt)
  - AdjP (Adjektive oder Adjektive+Gradpartikeln)
  - PP (Präposition+Konstituente)
  - usw....(Das ist nur eine Liste mit den meistbenutzten Konstitutentenarten)

Konstituentenstrukturgrammatik

- Meistens wird die Struktur eines Satzes mit Hilfe eines Baums dargestellt
  - Wörter sind Blätter (Terminale)
  - Wortklassen sind Präterminale
- Für die maschinelle Bearbeitung kann aber eine geklammerte Darstellung einfacher zu verarbeiten sein.

Syntax-Beispiel - (Baum)

S → NP VP

NP → DET N

PP → PREP NP

VP → V PP

DET → das

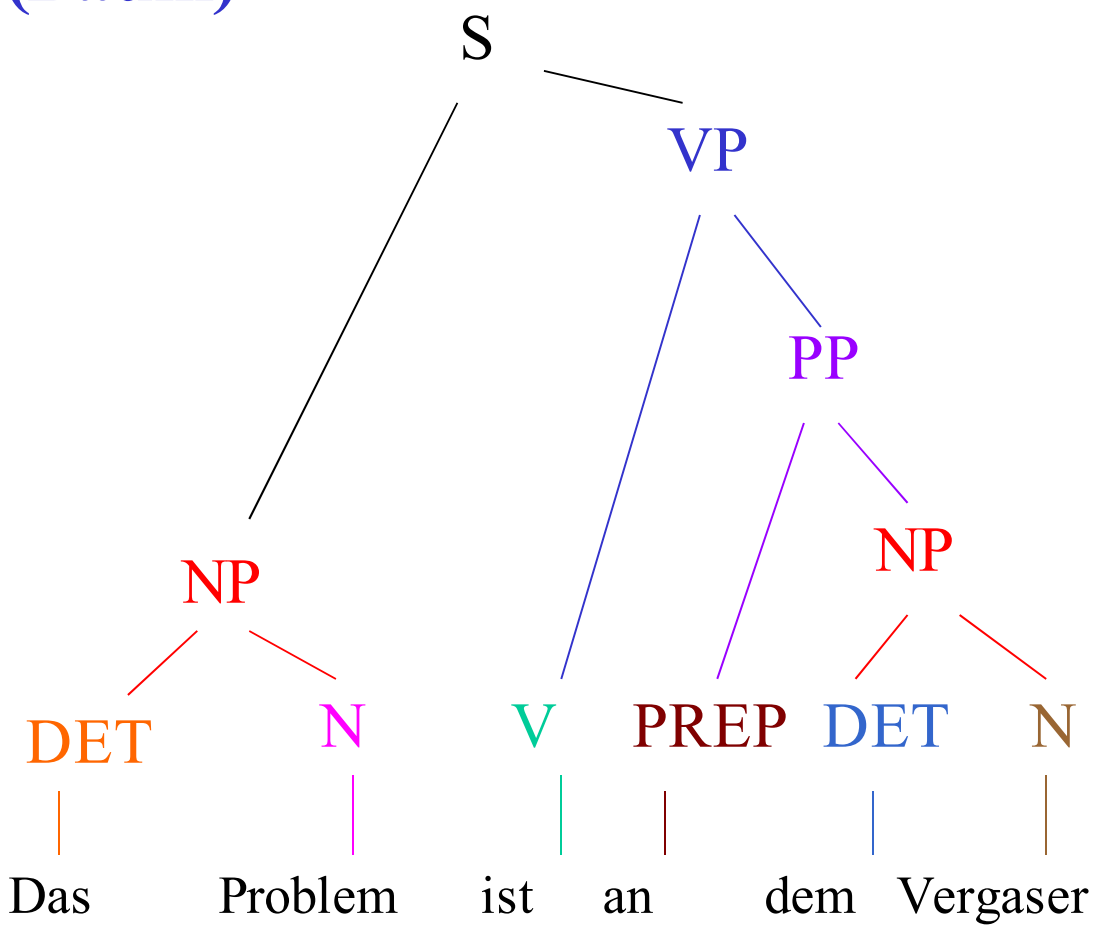
DET → dem

PREP → an

N → Problem

N → Vergaser

V → ist

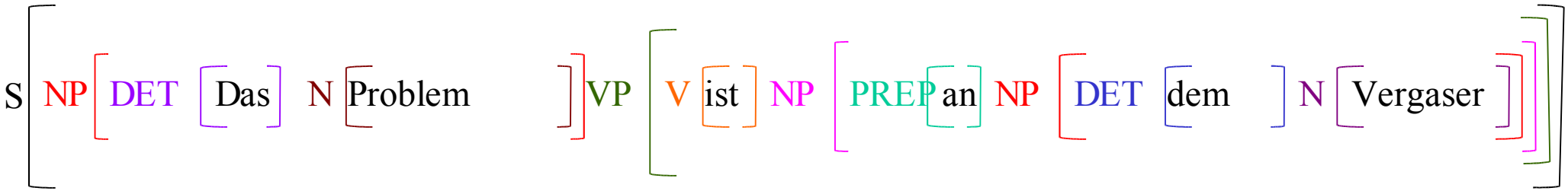


Syntax-Beispiel - (geklammerte Ausdrücke)

S → NP VP
NP → DET N
PP → PREP NP

VP → V PP
DET → das
DET → dem

PREP → an
N → Problem
N → Vergaser
V → ist

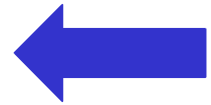


Konstituentenstrukturgrammatik: Formale vs. Natürliche Sprachen

- Nicht alle Ausdrücke, die der Grammatik entsprechen, sind in natürlicher Sprache korrekt,
- z.B. ein Satz wie *Das Problem ist an der Vergaser* wird von der gegebenen Grammatik auch als korrekt erkannt.
- In natürlicher Sprache gelten komplexe Regeln, die mit einfachen Konstituentenregeln nicht darstellbar sind.
- Normalerweise erweitert man die Regeln mit Merkmalen, z.B.

NP [Genus, Numerus, Kasus] $\rightarrow$Det [Genus, Numerus, Kasus]
N [Genus, Numerus, Kasus]

Inhalt

- Grenzen der Voice-XML unterstützten Grammatiken
- Grammatiken für natürliche Sprache
- Lexikonaufbau 
- Elementare Parsing-Strategien
- Anaphorische Ausdrücke in Dialogsystemen
- Aufgaben für das Praktikum

Notwendigkeit eines Lexikons

- Die Grammatik wird kompakter (alle Regeln, die Terminale enthalten, werden ausschaltet)
- Trennung von Daten und Regeln (dieselbe Grammatik kann auch für einen anderen Wortschatz benutzt werden)
- In das Lexikon kann man Merkmale hineinschreiben, die für die Grammatik nicht wesentlich sind, aber für die Anwendung wichtig sind (z.B. semantische Merkmale)

Aufbau eines Lexikons -1-

- Man muß zunächst entscheiden ob für die Anwendung ein Vollformen-Lexikon (alle flektierte Formen) oder ein Stammformen-Lexikon (nur die Lemmas) notwendig ist.
- Beide Formen haben Vor- und Nachteile:
- ein Vollformenlexikon erfordert keine morphologischen Prozesse, wird aber bei realistischen Systemen schnell sehr groß und viele Infomationen sind redundant.
- Ein Stammlexikon ist sinnvoll bei Sprachen mit sehr regelmäßiger Flexion und wenn der Wortschatz, den die Anwendung braucht, sehr groß ist.

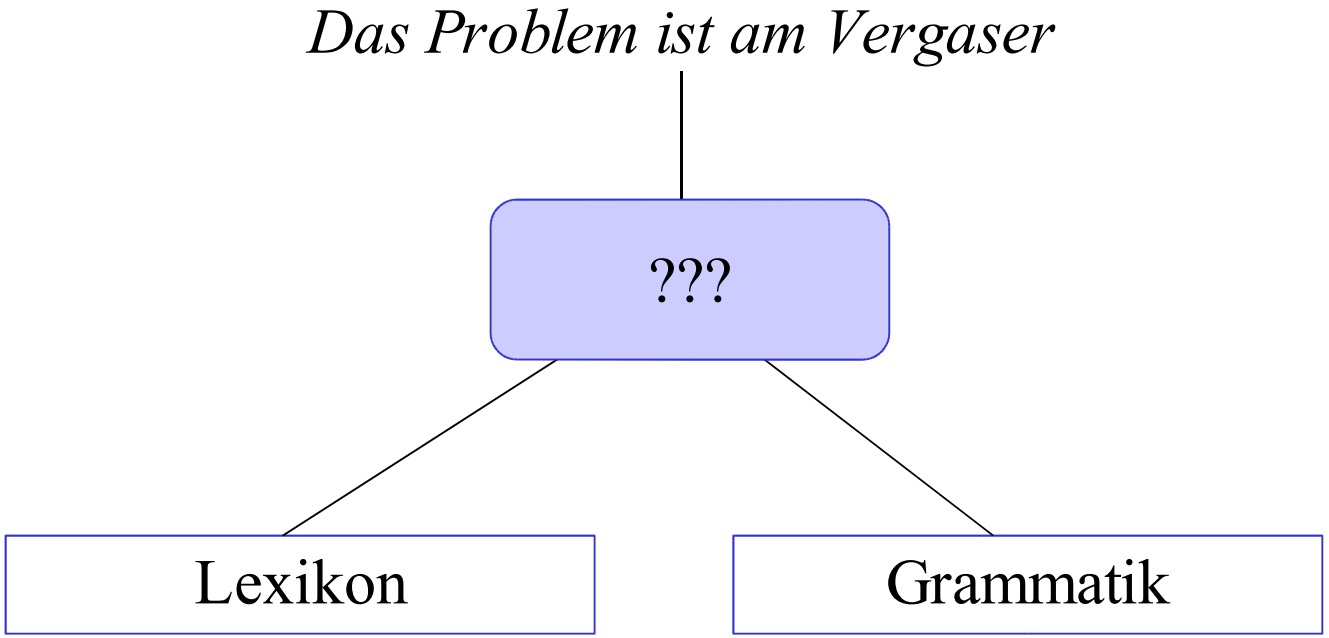
Aufbau eines Lexikons -2-

- Formal: Normalerweise werden heute die Lexika in einem XML-Format kodiert
- Inhaltlich: Üblicherweise enthält ein Lexikon je Eintrag mindestens
  - Wortklasse (Nomen, Adjektiv, Verb, …)
  - Flexionsklasse, falls Stammlexikon (stark, schwach, ...),
  - morphologische Merkmale, falls Vollformenlexikon, (Genus, Numerus, Kasus),
  - Subkategorisierungsmerkmale (transitiv/intransitiv),
  - Semantische Merkmale (z.B. belebt/nicht belebt; selbst reparierbar/nicht selbst reparierbar, …).

Beispieleintrag in einem Vollformen-Lexikon

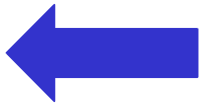
<entry name="Motor">
    <genus> maskulinum </genus>
    <kasus> Nominativ</kasus>
    <numerus> singular</numerus>
    <sem> nicht_selbst_reparierbar </sem>
</entry>

Wie nutzt man Grammatik und Lexikon für die Analyse?



Man braucht einen Prozeß, der automatisch einen Satz analysiert und entscheidet, ob der Satz korrekt ist (der Grammatik entspricht)

Inhalt

- Grenzen der Voice-XML unterstützten Grammatiken
- Grammatiken für natürliche Sprache
- Lexikonaufbau
- Elementare Parsing-Strategien 
- Anaphorische Ausdrücke in Dialogsystemen
- Aufgaben für das Praktikum

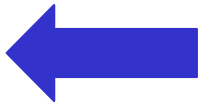
Parsingstrategien für Konstituentenstrukturgrammatiken -1-

- Grundsätzlich benutzt man die von formalen Sprachen bekannten Strategien
  - bottom -up: Der Prozeß beginnt mit den Wörtern und versucht, die Grammatikregeln so lange anzuwenden, bis das Startsymbol erreicht wird.
  - top-down: Der Prozeß beginnt mit dem Startsymbol und versucht die Grammatikregeln so lange anzuwenden, bis der Satz generiert ist.
- In beiden Fällen wird der Prozeß beendet. Wenn der Satz nicht in der Grammatik enthalten ist, wird abgebrochen und eine Fehlermeldung angezeigt,.

Parsingstrategien für Konstituentenstrukturgrammatiken -2-

- Für Grammatiken, die natürliche Sprachen beschreiben, sind diese Strategien oft zu aufwändig, der Suchraum wird schnell sehr groß
- Deswegen benutzt man oft verbesserte Methoden (z.B. Chart Parsing), bei denen man durch zusäzliche Datenstrukturen erfolgreiche Zwischenergebnisse für weitere Durchgänge speichern kann.

Inhalt

- Grenzen der Voice-XML unterstützten Grammatiken
- Grammatiken für natürliche Sprache
- Lexikonaufbau
- Elementare Parsing-Strategien
- Anaphorische Ausdrücke in Dialogsystemen 
- Aufgaben für das Praktikum

Sg.Die Anapher
Plr.Die Anaphora/
Die Anaphern

Was ist eine Anapher ?

- Anaphora sind Verweise auf Elemente (Worte, Konzepte), die in dem Dialog schon benutzt wurden.
- Meistens referenziert man auf Elemente im vorigen Satz.
- Am häufigsten sind pronominale Anaphora:

Ich denke, dass ich ein Problem am Motor habe. Er startet überhaupt nicht mehr.

- In einigen Fällen ist die automatische Identifizierung des Referenten schwierig

Beispiel: Die Bremslichter leuchten nicht. Ich habe die Sicherung herausgenommen und sie ersetzt

Traditionelle Lösungen für Anaphora

- Anaphora sind grundsätzlich schwer zu behandeln,
- Meistens wird eine Diskurs-Entitäten- Liste aufgebaut und entspechend einem Algorithmus (Centering theory), wird der wahrscheinlichste Referent gefunden.

Inhalt

- Grenzen der Voice-XML unterstützten Grammatiken
- Grammatiken für natürliche Sprache
- Lexikonaufbau
- Elementare Parsing-Strategien
- Anaphorische Ausdrücke in Dialogsystemen
- Aufgaben für das Praktikum

Aufgaben für das Praktikum -1-

- Erstellung eines Vollformen-Lexikons für das schon erzeugte Korpus.
- Sie müssen sich überlegen, welche Merkmale für das Lexikon notwendig sind.
- Erstellung einer Grammatik für die Sätze, die in der Voice-XML-Grammatik nicht darstellbar sind.
- Implementierung eines traditionellen Parsingalgorithmus (top-down oder bottom-up)
- Identifizierung von Pronomen die auf vorige Sätze referenzieren. Das System muß ein sinnvolles Prompt zeigen "Bitte benutzen Sie keine ...." oder vielleicht sogar "Meinen Sie mit "sie" die Sicherung?

Aufgaben für das Praktikum -2-

- Automatische Identifizierung zweier Sprachen:
  - Der erste Benutzerausdruck wird analysiert. Für jedes Wort wird geprüft, in welchem Lexikon es vorkommt (Englisch oder Deutsch)
  - Wenn mehr als 4 Wörter nur in demselben Lexikon vorkommen, begrüßt das System (mit einem Prompt) den Benutzer in der entsprechenden Sprache.
  - (Danach wird der Dialog nur auf Deutsch weitergeführt).
  - Wenn die Wörter sich auf die Lexika gleich verteilen, muß der Benutzer einen neuen Ausdruck formulieren oder wird gefragt, ob er Deutsch oder Englisch benutzen will.
  - Wenn sie in keinem der Lexika vorkommen, wird der Bernutzer aufgefordert, Deutsch oder Englisch zu benutzen.