

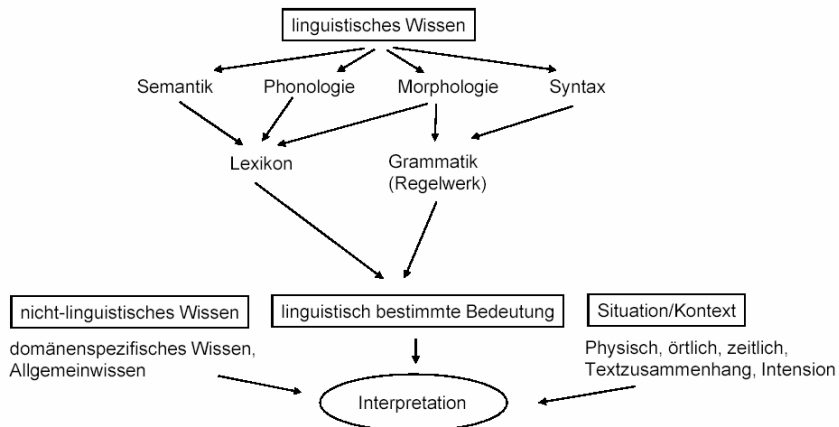
# Spezielle Themen der KI

## NLP

### Natural Language Processing

### Sprachverstehen

## NLP - Verarbeitungsstufen des Sprachverstehen



## **Phonologie** – Lautlehre oder Sprachgebilde-Lautlehre untersucht

- Systeme von **Phonemen** als kleinsten bedeutungsunterscheidenden Elementen von Sprachen
- Laute als Einheiten im System einer Sprache
- Phoneminventar durch Benutzung von Minimalpaaren,
  - Wörter, die sich nur in einem Laut unterscheiden. Dieser Laut hat dann Phonemstatus.
  - Beispiel: *Kutter* und *Futter* -> /k/ und /f/
- Notierung der Phoneme durch Internationales Fonetisches Alphabet in /x/.

## **Phonetik** - Sprechakt-Lautlehre

- Lehre der von Menschen hervorgebrachten Lauten, den **Phonen**.
- Detaillierte Beschreibung von Lauten unabhängig von Systemüberlegungen.
- Kleinste Phonologische Repräsentation einer Äußerung:
  - Sequenz von Einheiten, gekennzeichnet durch Eigenschaften (Merkmale).
  - Eigenschaften können entweder vorhanden (+) oder nicht vorhanden (-) sein.
  - Phonem ist ein Bündel von Merkmalen. (Chomsky und Halle)
- Notierung der Phone durch Internationales Fonetisches Alphabet in [x].

## Beispiele (Wikipedia):

- *ch* im Deutschen wird häufig mit Phonem /x/ notiert, Artikulation
  - im Bereich des harten Gaumens (palatal): *ich* [ɪç]
  - hinten im Mund: (*ach* [ax]).
  - Unterschied ist nicht bedeutungsunterscheidend: auch
    - Keine Bedeutungsänderung für *ich*, wenn statt des Ich-Lautes in *ich* ein Ach-Laut gesprochen wird, wie es manche Nichtmuttersprachler tun.
  - Der *Ich-Laut* und der *Ach-Laut* sind so genannte Allophone (fonetisch realisierte Varianten eines einzigen Phonems).
  - Vor Variantenidentifikation zwei verschiedene Phone:
    - [ç] und [x].
- s-Laute im Deutschen
  - Bedeutungsunterscheidend ist Unterschied zwischen stimmhaftem und stimmlosem s.
  - Zwei Phoneme (/s/ und /z/) unterschiedlicher Schreibweise: (ß oder ss und s):
    - *reißen* und *reisen* unterscheiden sich lediglich in der Stimmhaftigkeit des s-Lautes
    - /s/ versus /z/.

# Morphologie und Lexikon

## **Morphologie** – Lehre von den Wortformen

- beschäftigt sich mit der Erforschung der kleinsten bedeutungstragenden Elemente einer Sprache, den **Morphemen**
- Klassifiziert Wörter nach Wortarten und Wortformen (Verb, Substantiv, Adjektiv, Artikel, Pronomen...) und
- der Struktur der Wortformen (Fall, Geschlecht bei Nomen, Person, Numerus, Tempus...):
  - Wortbildung (Gesetzmäßigkeiten zur Bildung neuer komplexer Wörter) und
  - Flexion (Änderung der Gestalt eines Wortes zum Ausdruck seiner grammatischen Funktion innerhalb eines Satzgefüges)

## Aufgaben der Morphologie

- Zurückführung von Worten auf Grundform und Kategorisierung
  - Automatische Wortformerkennung neben dem Lexikon
- Morphologische Phänomene lassen sich formal mit regulären Ausdrücken beschreiben.

## **Lexikon** – Auflistung der bekannten Wörter

- möglichst vollständiger Wortschatz einer Sprache
- Wortformen und Wortart
- Wörterbuch - Bedeutung der Worte und Beziehungen untereinander

## Modelle für Lexika

- Liste aller möglichen Worte
- Kombiniert mit der morphologischen Analyse (Regeln)
- automatische Systeme zu Nutzung und Auswertung bestehender Lexika

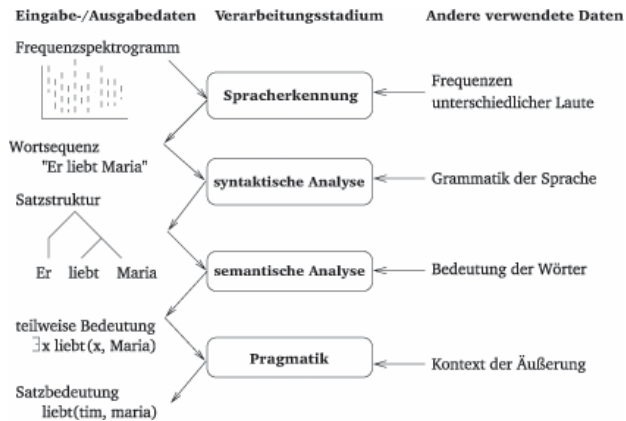
# NLP - Verarbeitungsstufen des Sprachverstehen

## Stadien des natürlichsprachlichen Verständnisses

- **Spracherkennung:**
  - Analyse der unbearbeiteten Sprachsignale.
  - Bestimmung der gesprochenen Worte und ihrer Abfolge.
- **Syntaktische Analyse:**
  - Die Abfolge der Worte wird mit Grammatikwissen analysiert.
  - Die Satzstruktur wird bestimmt.
- **Semantische Analyse:**
  - Bedeutungsdarstellung durch Strukturinformation und Wortbedeutung.
- **Pragmatische Analyse:**
  - Kontextinformation vervollständigt die Bedeutungsdarstellung aus der semantischen Analyse.

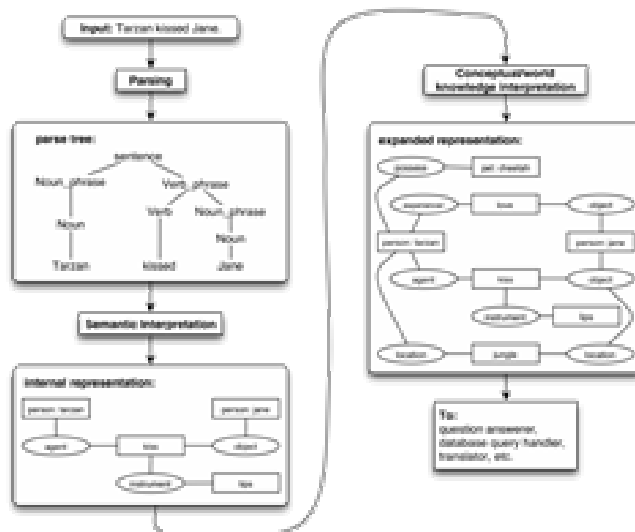
# NLP - Verarbeitungsstufen des Sprachverstehen

## Stadien des natürlichsprachlichen Verständnisses



52

# NLP - Verarbeitungsstufen des Sprachverstehen



53

# Verarbeitung natürlicher Sprache

## Spracherkennung

- Aufteilung des Signals in verschiedene Frequenzen
- Messung der Energie für diese Frequenzen → Frequenzspektrum (FS)
- Zuordnung von elementaren Lauten – Phonen- zu Mustern im FS
- Suche dieser Muster im Signal
  - Evt. Geschwindigkeitsanpassung durch dynamic time warping (DTW)
- Statistische Modellierung:
  - Wahrscheinlichkeit der Phone
  - Übergangswahrscheinlichkeiten zwischen den Phonen
  - Häufigkeit der Worte
  - Folge der Worte
  - ...
- Benutzung von HMMs zur statistischen Modellierung



54

# Syntaktische Analyse

- Satzbildung durch Wortkomposition
- Ausgangspunkt für Bedeutungsanalyse
  - (1) *Der Hund fraß den Knochen*
  - (2) *Der Knochen wurde vom Hund gefressen*
- Struktur aus Syntaxregeln hilft der Bedeutungsfindung im Gegensatz zu:
  - Es ist immer das 2. Substantiv, das gefressen wird.
- Zerlegung der Bedeutungsfindung auf Basis von syntaktischen Zerlegungen (etwa von Phrasen)
  - (3) *sp[Der Hase mit den langen Ohren] erfreute sich an sp[einem großen grünen Salatblatt]*
- Bedeutungsfindung durch Komposition der Teilbedeutungen
- Komposition abhängig von gewählter Clusterung
  - (4) *Tim sah Maria mit dem Fernglas*
  - (5) *Ich sah den Kölner Dom auf dem Flug nach Frankfurt*

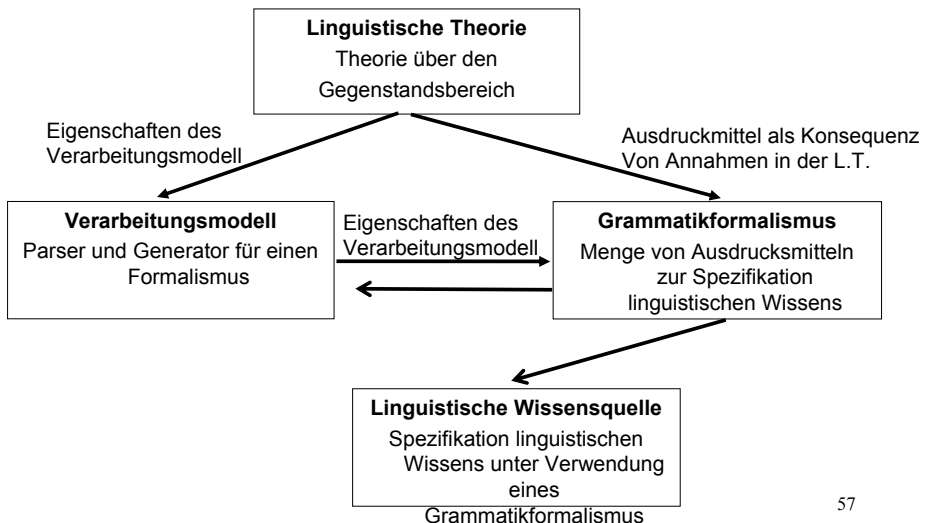
55

# Beschreibungsformalismus

**Definition:** Ein Grammatikformalismus ist ein Formalismus zur Repräsentation sprachlichen Wissens auf den verschiedenen linguistischen Beschreibungsebenen.

- Kriterien zur Ausgestaltung eines Grammatikformalismus unter Berücksichtigung der Beziehung von Grammatikformalismus zu linguistischer Theorie:
  - **Ausdrucksstärke:** Welche Phänomenbereiche können beschrieben werden?
  - **Algorithmisierbarkeit:** Welche Verarbeitungsmodelle werden unterstützt? Welche Algorithmisierbarkeit ist möglich?
  - **Linguistische Motiviertheit:** Umfang direkt zu verwendender linguistisch abgesicherter Ausdrucksmittel? Inwieweit sind Ausdrucksmittel direkt an spezifische linguistische Theorie gebunden?

# Beschreibungsformalismus





# Grammatik

- Grammatik als formales Beschreibungssystem für Syntaxregeln zur Definition der Wortorganisation
- Menge von Ersetzungsregeln für eine Kette von Symbolen (=Satz)
- Symbole eingeordnet in syntaktische Kategorien
  - Nomen, Verb, Adjektiv, ...
  - Nominalphrase, Verbalphrase
- Satz ist regulär, wenn ausgehend von einem Startsymbol der Satz mit Hilfe der Regeln erzeugt werden kann (Parsen)
- Was soll die Grammatik tun?
  - Unterscheidung zwischen „korrekten“ und „inkorrekten“ Sätzen
    - Der Junge wirft den Ball. (+)
    - Der wirft Junge Ball den. (-)
  - Zuweisung einer „sinnvollen“ Struktur
    - (Der Junge) (wirft den Ball)
    - (Der) (Junge wirft) (den Ball)
- Eigenschaften der Grammatik: Kompakt, modular
- Syntaxanalytiker (Parser) überprüft Satz auf Grammatikkonformität



# Grammatik

- Linguistische Strukturen und Prozesse
  - Kompositionalität
    - Interpretation einer Struktur lässt sich aus Teilen und ihrer Zusammensetzung ableiten.
    - Interpretation jedes Teils kann lokal erfolgen.
    - Bedingt analytische (paradigmatische Betrachtung).
  - Deklarativität
    - Trennung zwischen Wissen und Verarbeitung.
    - Verarbeitung kann nach rein strukturellen Gesichtspunkten erfolgen
    - Unterstützung von analytischen und generativen Prozessen.
- Konkatenation
  - Lineare Aneinanderreihung einzelner Elemente.
  - Häufigste Operation über linguistischen Daten.
  - Bei natürlichen Sprachen finden sich auch Tilgung und Ersetzung.

# Chomskys Hierarchie der Grammatiken

Regeln zum Ausdruck der korrekten grammatikalischen Struktur einer Sprache für Analyse und Synthese

Format der hierarchischen Chomsky Grammatiken

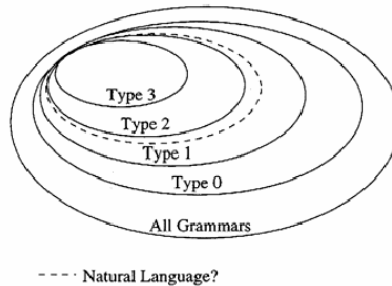
$A B C \dots \rightarrow E F G \dots$

Typ 0: Transformation Grammar  
Regeln der Form: *beliebig*  $\rightarrow$  *beliebig*  
Mächtigkeit: allgemeine Turing Maschine

Typ 1: kontextsensitive Grammatik  
Regeln der Form:  $A B C \rightarrow A D C$   
Mächtigkeit: Linear Bound Automata

Typ 2: kontextfreie Grammatik  
Regeln der Form:  $A \rightarrow B C D \dots$   
Mächtigkeit: Push Down Stack Automata

Typ 3: Regulär or Rechtslinear  
Regeln der Form:  $A \rightarrow x B$  und  $A \rightarrow x$  wobei x terminal  
Mächtigkeit: Finite State Automata



# Grammatikformalismen

Phrasenstrukturgrammatiken Beispiel:

(7)  $NP \rightarrow Det Adj N$

- (7) beschreibt einfache Sätze wie
  - „der alte Mann“
  - „ein junger Mann“
  - „dem großen Haus“
- Menge von Regeln wie (7) bilden eine Phrasenstrukturgrammatik (PSG).
- PSG sind kontextfreie Grammatiken.

**PSG:**

Grammatik	$G = (\Sigma, \Phi, P, S)$
Terminale	$\Sigma = \{alte, dem, der, ein, großen, Haus, junger, Mann, Vater, \dots\}$
Nonterminale	$\Phi = \{Adj, Det, N, NP, \dots\}$
Regeln	$P = \left\{ \begin{array}{l} S \rightarrow NP VP \mid \dots, NP \rightarrow Det N \mid Det Adj N \mid \dots \\ Det \rightarrow dem \mid der \mid ein \mid \dots, Adj \rightarrow alte \mid junger \mid \dots, \\ N \rightarrow Haus \mid Mann \mid Vater \mid \dots \end{array} \right\}$
Startsymbol	$S$



## Beispiel für eine einfache deutsche Grammatik

Grammatik in DCG (definite clause grammar oder definiten Klauselgrammatik):

```

satz      -> substantiv_phrase, verb_phrase
substantiv_phrase -> passendes_substantiv
substantiv_phrase -> bestimmungswort, substantiv
verb_phrase -> verb, substantiv_phrase
passendes_substantiv -> [Maria]
passendes_substantiv -> [Tim]
substantiv -> [Löwen]
substantiv -> [Keks]
verb      -> [aß]
verb      -> [küsste]
bestimmungswort -> [den]
    
```

62

## Ableitungsbaum

Ableitungsbaum für „*Tim aß den Löwen*“



63



# Grammatikformalismen

**Eine einfache Grammatik mit zahlenmäßiger Übereinstimmung zwischen Subjekt und Verb:**

```
satz          -> substantiv_phrase(Num),  
                verb_phrase(Num)  
substantiv_phrase(Num) -> passendes_substantiv(Num)  
substantiv_phrase(Num) -> bestimmungswort (Num),  
                substantiv(Num)  
verb_phrase(Num)     -> verb(Num), substantiv_phrase (_)  
passendes_substantiv(s) -> [maria]  
substantiv(s)        -> [[löwe]  
substantiv(p)        -> [[löwen]  
best(s)              -> [den]  
best(p)              -> [die]  
verb(s)              -> [isst]  
verb(p)              -> [essen]
```