



## Vorlesung: Einführung in die Computerlinguistik

Hans Uszkoreit

© 2004 Hans Uszkoreit

### Übersicht des ersten Teils



- ☐ Aufgaben und Einordnung des Faches
- ☐ Motivationen für die Modellierung menschlicher Sprache
- ☐ Computerlinguistik als eine moderne Sprachwissenschaft
- ☐ Repräsentationen und Verarbeitungskomponenten

© 2004 Hans Uszkoreit



- 26.10. HU: Einführung, Vorstellung der Teilgebiete
- 2.11. HU: Sprach-, Ingenieur- und Kognitionswissenschaft
- 9.11. HU: Teilbereiche der CL
- 16.11. SP: Spracherkennung / Synthese.
- 23.11. HU: Morphologie 1
- 30.11. HU: Morphologie 2
- 7.11. HU: Syntax / Parsing 1
- 14.11. HU: Syntax / Parsing 2
- 21.11. SP: Semantik.
- 11.1. SP: Einf. Statistische Verarbeitung
- 18.1. SP: PCFGs und Parsing mit PCFGs
- 25.1. HU: Ueberblick Anwendungen - Hybride Systeme?
- 1.2. HU: Anwendung: Information Management (IR, IE, etc.)
- 8.2. HU: Anwendung: Cross-lingual Applications -- Maschinelle Übersetzung
- 15.2. SP: Annotation und Ressourcen

## EINSTIEG



Faszination

Wissenschaft

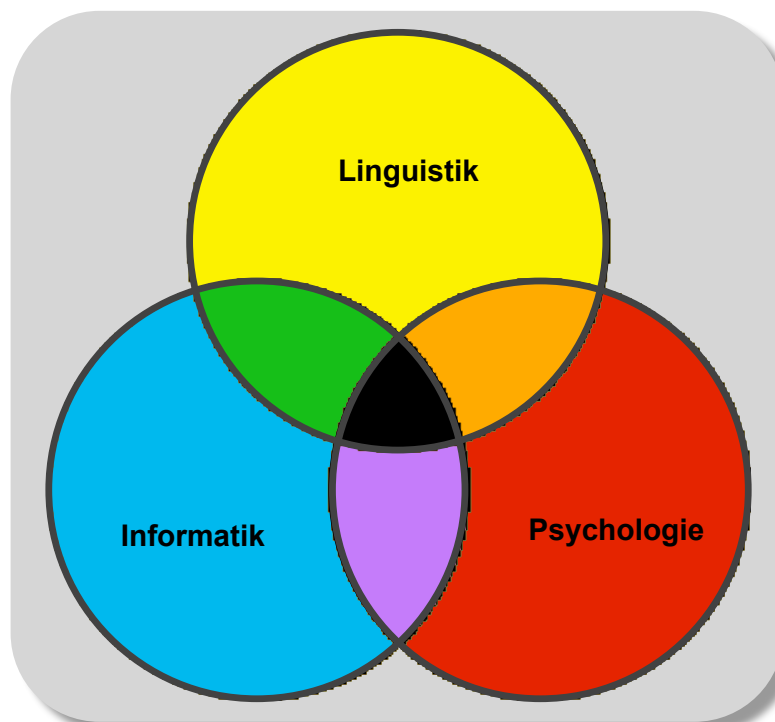
Technologie

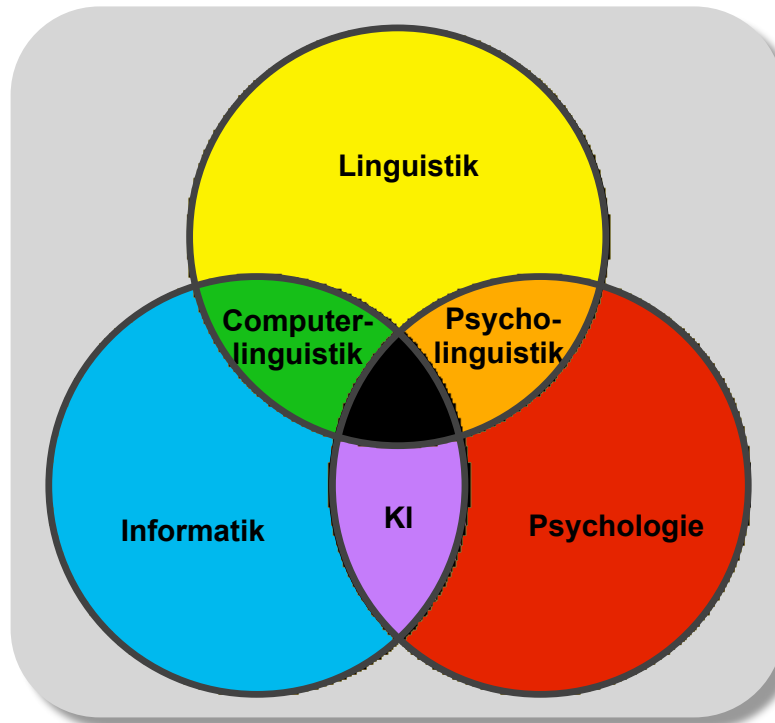


Mehr noch als Denken ist die Sprache eine Fähigkeit, die nur der Mensch besitzt.

Es ist ein Wunder, wie wir in Sekundenschnelle komplexe Gedanken in einem Satz ausdrücken können.

Es ist nicht weniger erstaunlich, wie das Kind in nur wenigen Jahren zehntausende von Wörtern und eine komplexe Grammatik lernt.





## Die Disziplin



### Computerlinguistik im weiteren Sinne

ist ein zwischen Linguistik und Informatik liegendes interdisziplinäres Forschungsgebiet, das sich mit der maschinellen Verarbeitung natürlicher Sprachen beschäftigt.

### Computerlinguistik im engeren Sinne

ist ein Teilgebiet der modernen Linguistik, das berechenbare Modelle menschlicher Sprache entwirft, implementiert und untersucht.

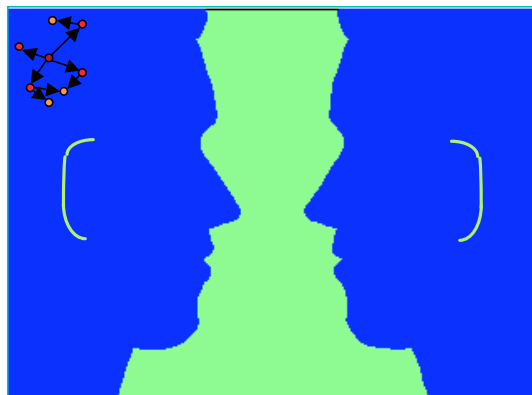


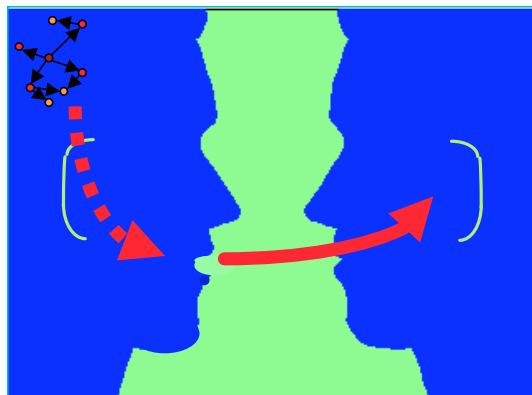
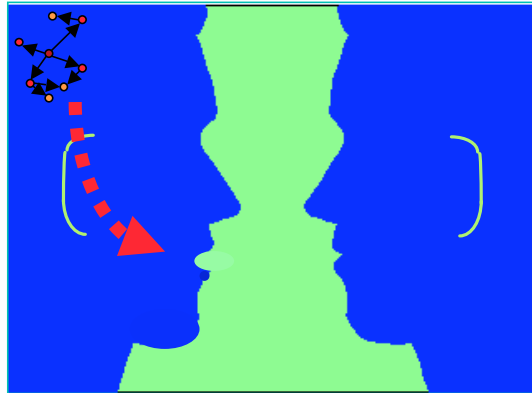
### Theoretische Computerlinguistik

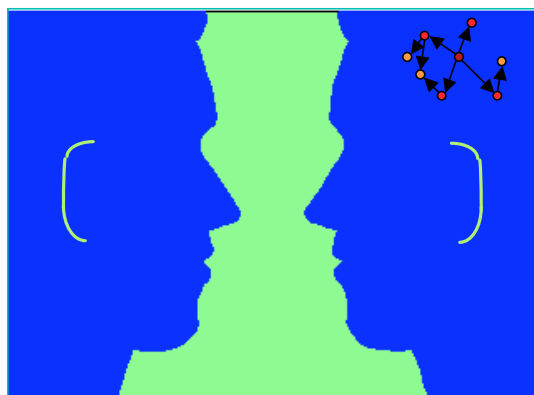
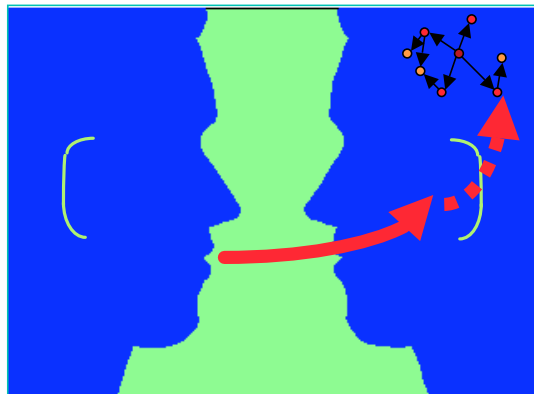
entwirft, implementiert und untersucht die Modelle mit dem Ziel, zum Verständnis, zur Verifikation und zur Verbesserung der zugrundeliegenden linguistischen und psychologischen Theorien beizutragen.

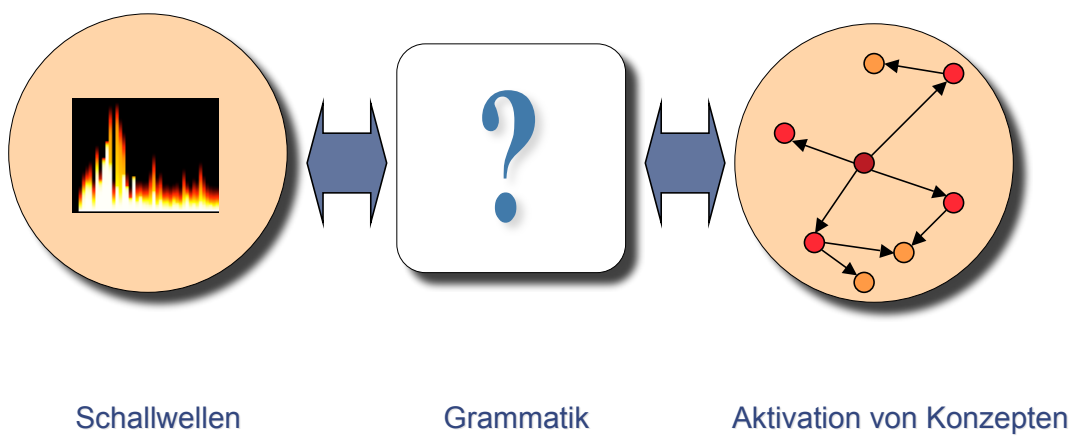
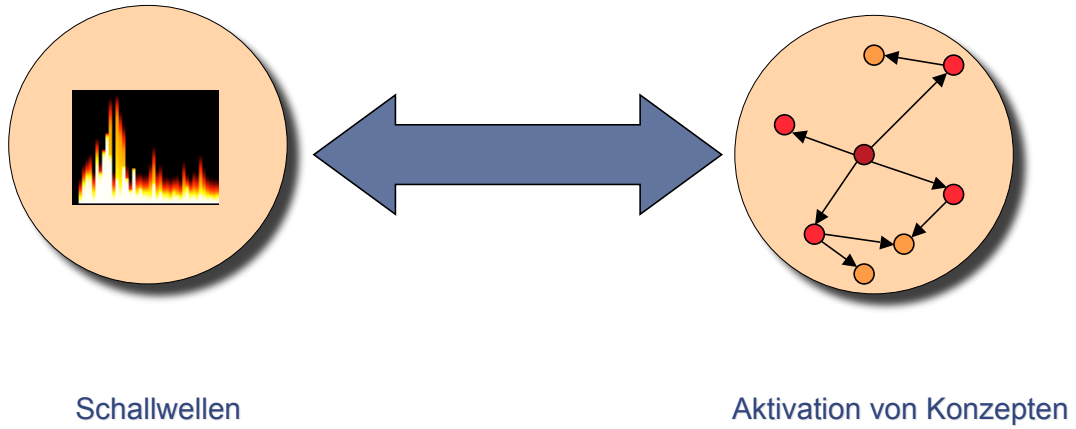
### Angewandte Computerlinguistik

entwirft, implementiert und untersucht die Modelle mit dem Ziel, Softwareanwendungen zu ermöglichen, die über eine (eingeschränkte) Beherrschung menschlicher Sprache verfügen.

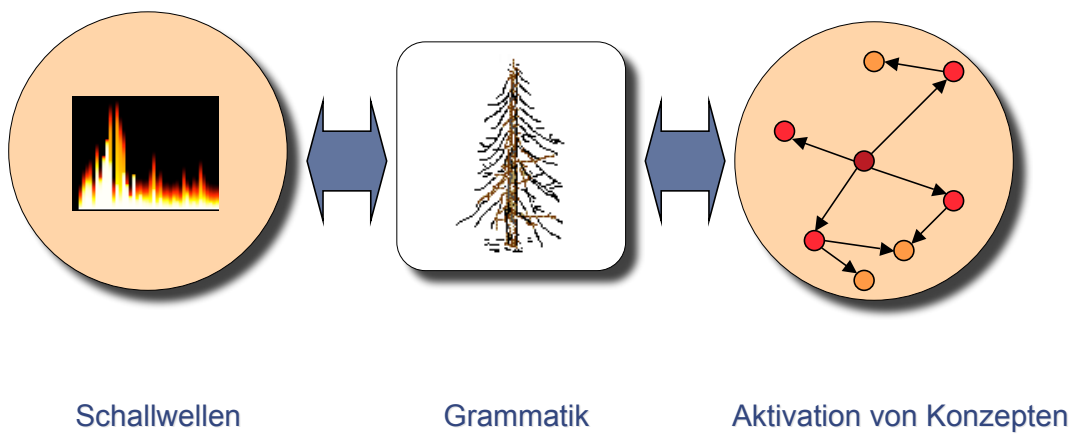
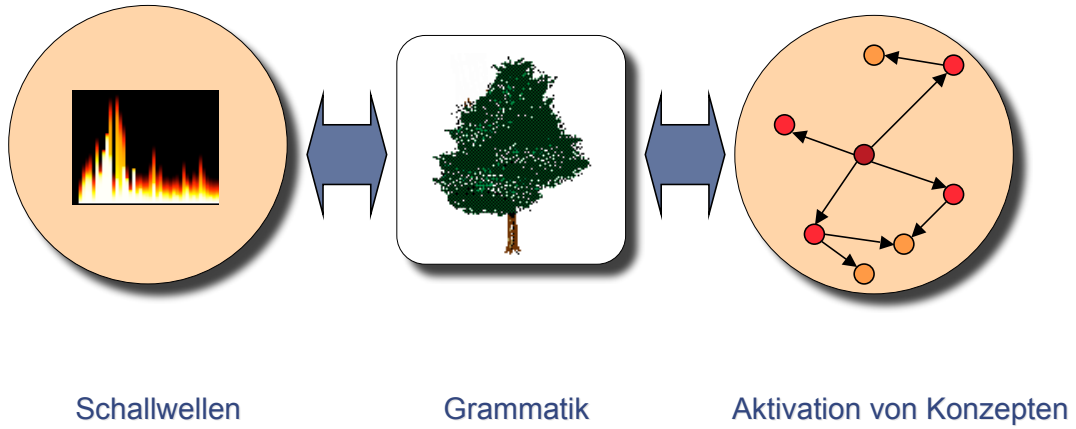


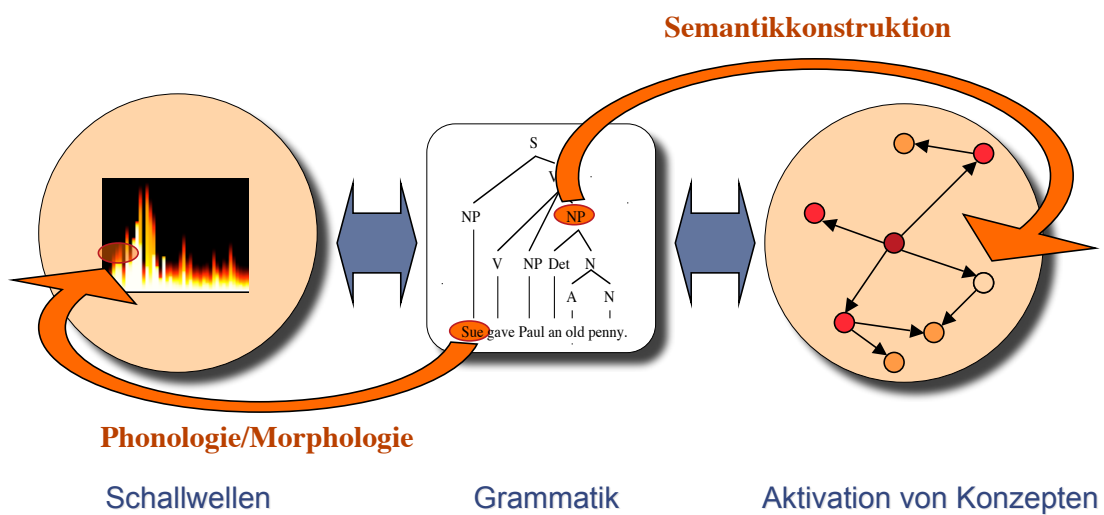
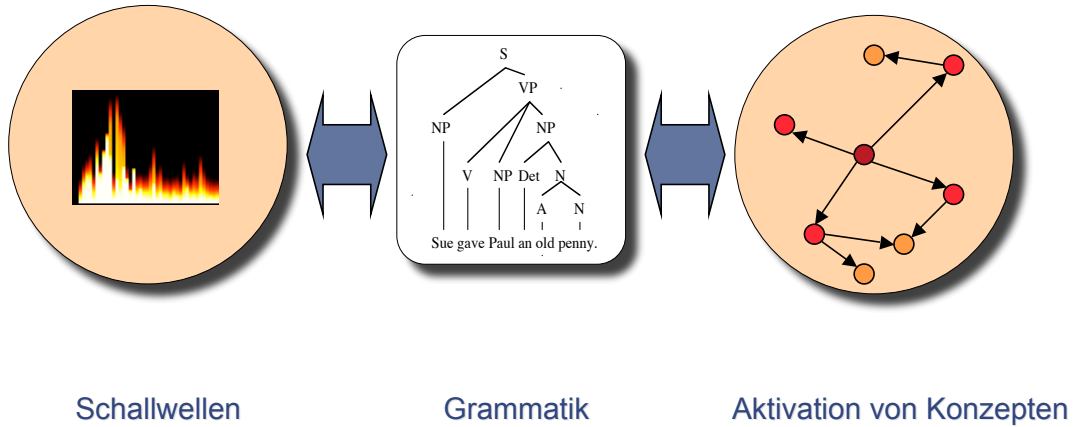


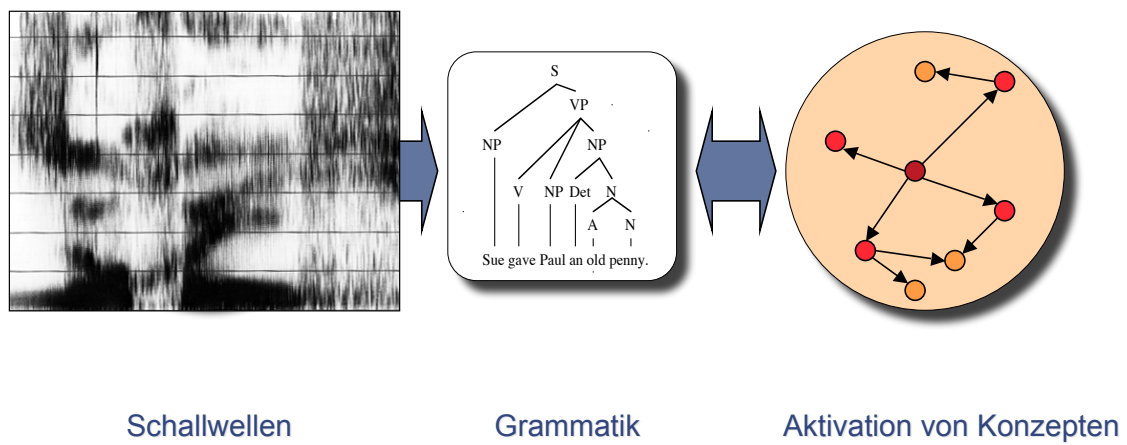
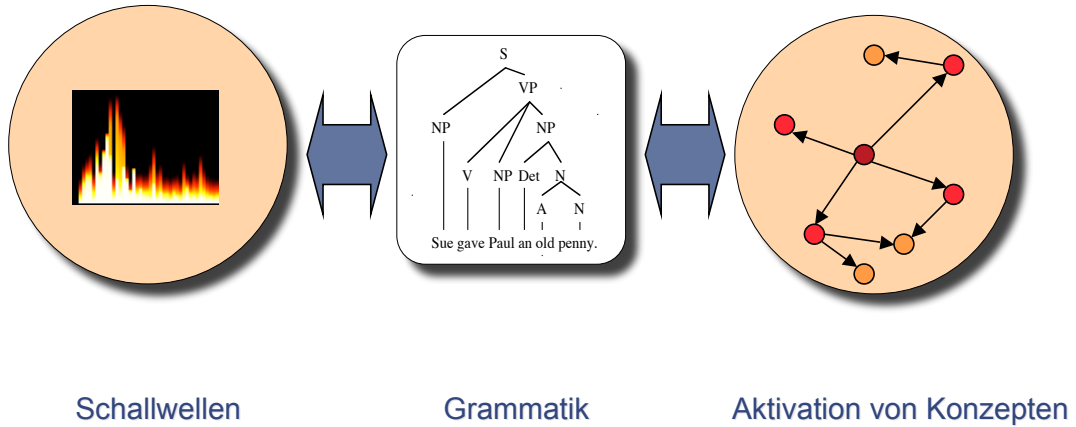














### Maschinelle Sprachverarbeitung

Analyse und Generierung von natürlicher Sprache mit dem Computer. Englisch: Natural Language Processing (NLP).

### Sprachtechnologie(n)

Übergriff für die Technologien sprachbeherrschender Systeme.  
Ingenieurwissenschaftliches Forschungsgebiet, in dem die Sprachtechnologien entwickelt werden.

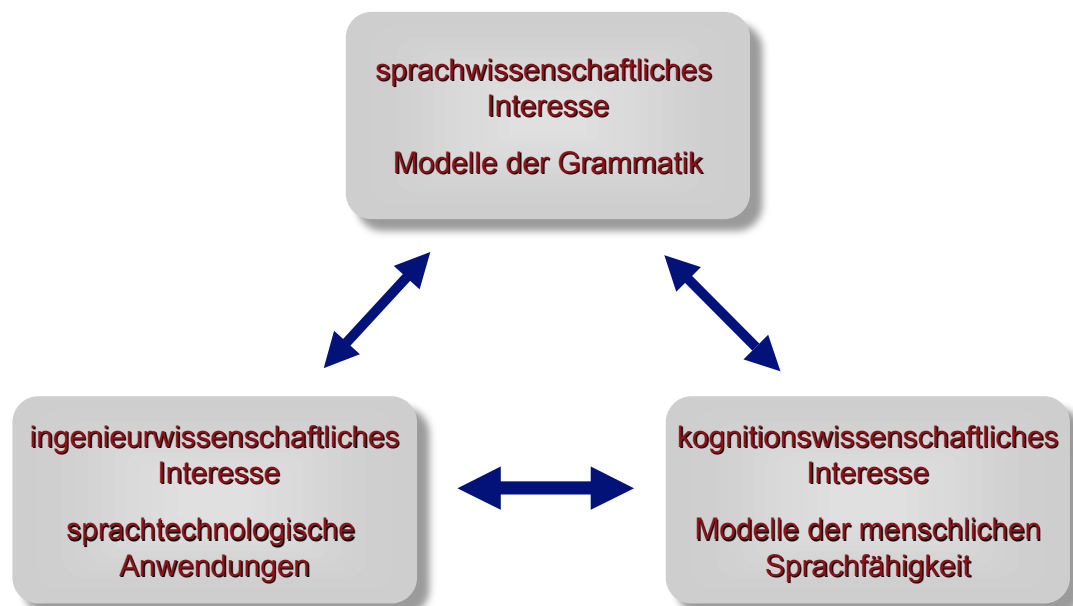
### Linguistische Datenverarbeitung (LDV)

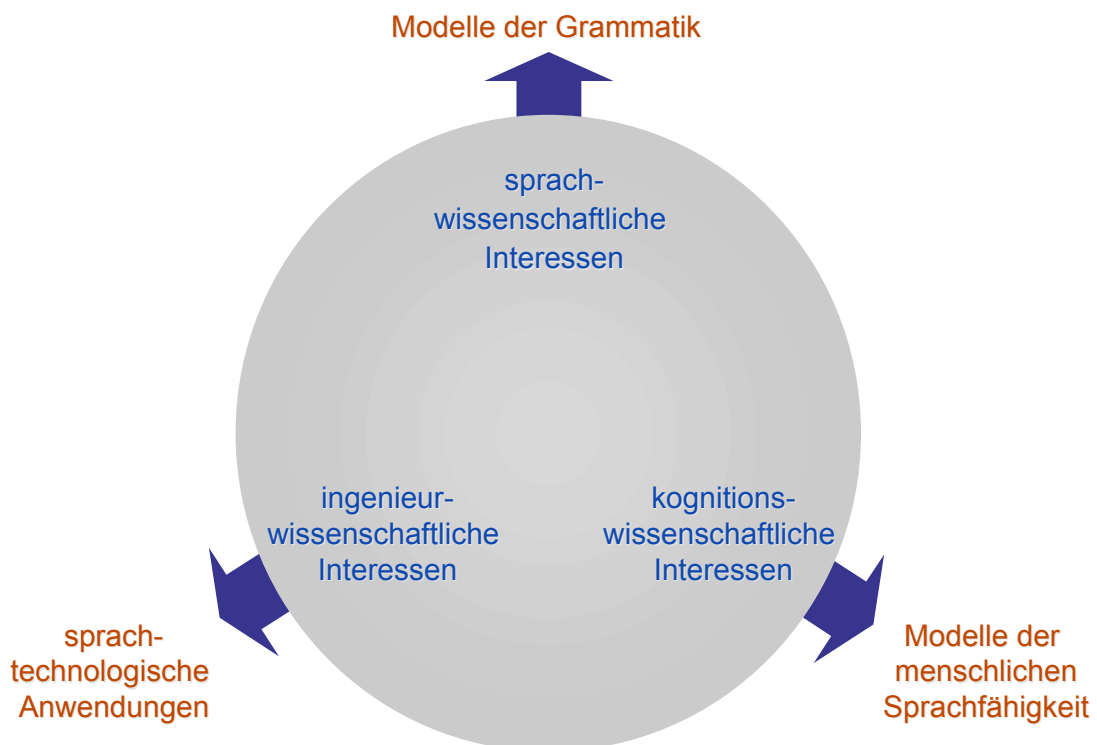
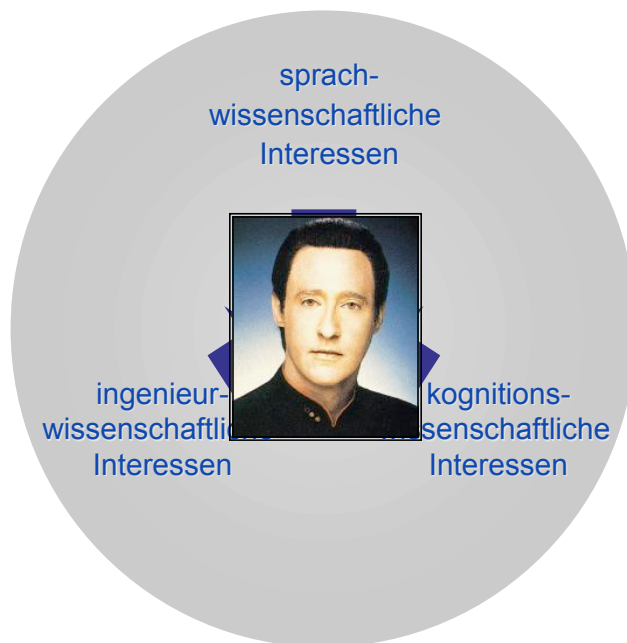
Traditionell ein Teilgebiet der elektronischen Datenverarbeitung, das sich sowohl mit der Anwendung von Methoden der Datenverarbeitung für die linguistische Forschung als auch mit maschineller Sprachverarbeitung beschäftigt. Die LDV versteht sich heute als ein Gebiet, das die Computerlinguistik einschließt.

### Sprachdatenverarbeitung

Verarbeitung von sprachlichen Daten mit dem Computer. Schließt ein: mono- und multilinguale Textverarbeitung, elektronische Wörterbücher, Konkordanzen, Terminologiebanken, maschinelle und maschinengestützte Übersetzung.

## Motivationen







- ❑ Die Linguistik ist eine "moderne", synchron orientierte, auf die interne Struktur der Sprache bezogene Wissenschaft, die sprachliche Regularitäten auf allen Beschreibungsebenen untersucht und ihre Ergebnisse in explizierter (formalisierter) Beschreibungssprache und in integrierten Modellen darlegt.
- ❑ (H. Bußmann *"Lexikon der Sprachwissenschaft"*)

## Teilgebiete der Linguistik



- ❑ Nach Beschreibungsebenen
  - Phonetik
  - Phonologie
  - Morphologie
  - Syntax
  - Semantik
  - Pragmatik/Text/Diskurs
- ❑ Andere Teildisziplinen
  - Psycholinguistik
  - Neurolinguistik
  - Historische Linguistik
  - Sozio- und Ethnolinguistik,
  - Dialektologie
  - Mathematische Linguistik



### SPRACHLICHES WISSEN

Was sind die Inhalte und Strukturen dieses unbewußten Wissens?

### SPRACHVERARBEITUNG

Wie produzieren und verstehen wir sprachliche Äußerungen?

### SPRACHERWERB

Wie lernt das Kind seine Muttersprache?

### SPRACHWANDEL

Wie entstehen Sprachen, Dialekte, Soziolekte?

## Kompetenz und Performanz

---



### ☐ Sprachliche Kompetenz:

- ☐ die endliche strukturierte Wissensbasis, die es den Sprechern einer Sprache ermöglicht, die wohlgeformten Äußerungen der Sprache zu generieren und zu interpretieren.

### ☐ Sprachliche Performanz:

die Generierung oder Interpretation realer Äußerungen, bzw. die Gesamtheit der Prozesse, die beteiligt sind, wenn der Mensch auf der Basis der sprachlichen Kompetenz reale Äußerungen generiert und interpretiert.



**Ein Kompetenzmodell sollte beinhalten:**

**Regeln, Prinzipien, Beschränkungen auf jeder Beschreibungsebene, die in ihrem Zusammenwirken genau die wohlgeformten Sätze der Sprache charakterisieren.**

**Es bietet für jede Sprache eine formalisierte endliche Definition einer unendlichen Menge von Paaren <Satz, Bedeutung>.**

**(Dazu gehören: Grammatik, Lexikon, morphologische Regeln, semantische Regeln.)**



**Ein Performanzmodell sollte erklären:**

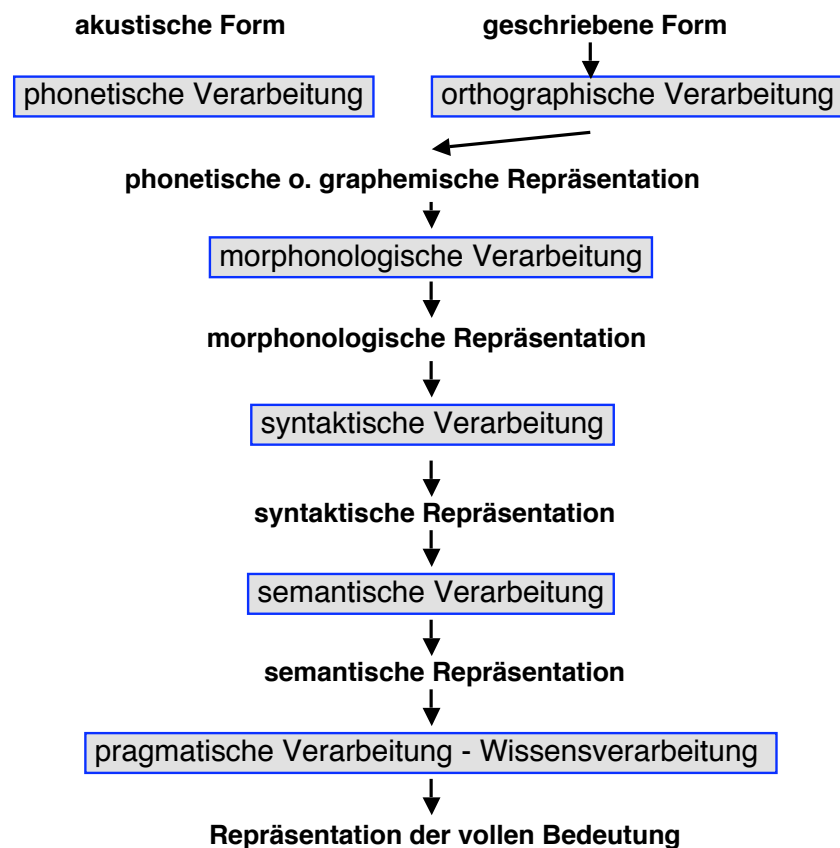
- ☐ warum viele ungrammatische Sätze erzeugt werden
  - ➔ z.B. Sprechfehler, Grammatikfehler
- ☐ warum viele ungrammatische Sätze verstanden werden
  - ➔ z.B. in der Kommunikation mit Kindern oder Ausländern
- ☐ warum viele grammatische Sätze nicht erzeugt werden
  - ➔ z.B. durch Präferenzen in der Generierung
- ☐ warum viele grammatische Sätze nicht verstanden werden
  - ➔ z.B. Holzwegsätze
- ☐ wie die Verarbeitung zeitlich strukturiert ist
  - ➔ z.B. Effizienz, Abfolge der Verarbeitungsschritte
- ☐ welchen Aufwand die Verarbeitungsschritte erfordern
  - ➔ z.B. Abhängigkeiten von anderen kognitiven Belastungen



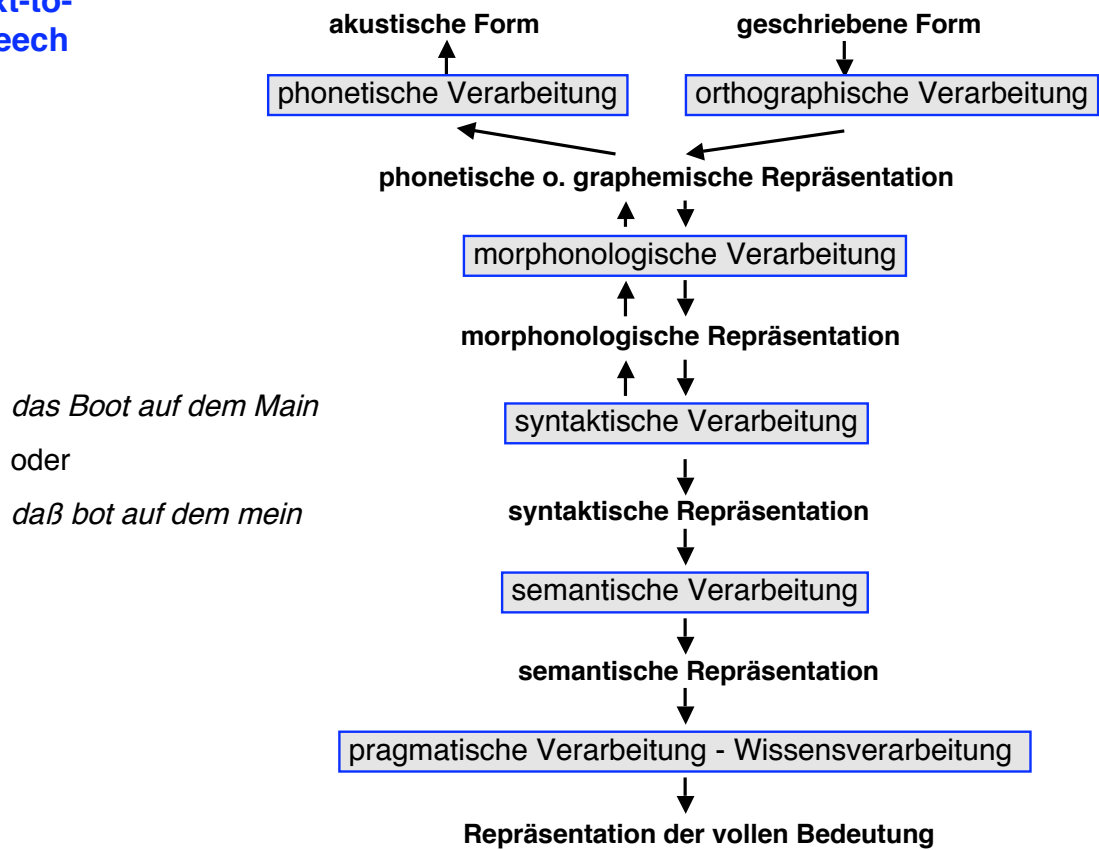


- efficiency**      Fähigkeit, Lösungen mit geringem Zeit- und Speicherbedarf zu liefern
- accuracy**      Fähigkeit, linguistisch korrekte Lösungen zu liefern
- robustness**      Fähigkeit, mit allen möglichen Eingaben fertigzuwerden
- coverage**      größtmögliche Abdeckung der Grammatik
- specificity**      Fähigkeit, die intendierte Analyse zu selektieren

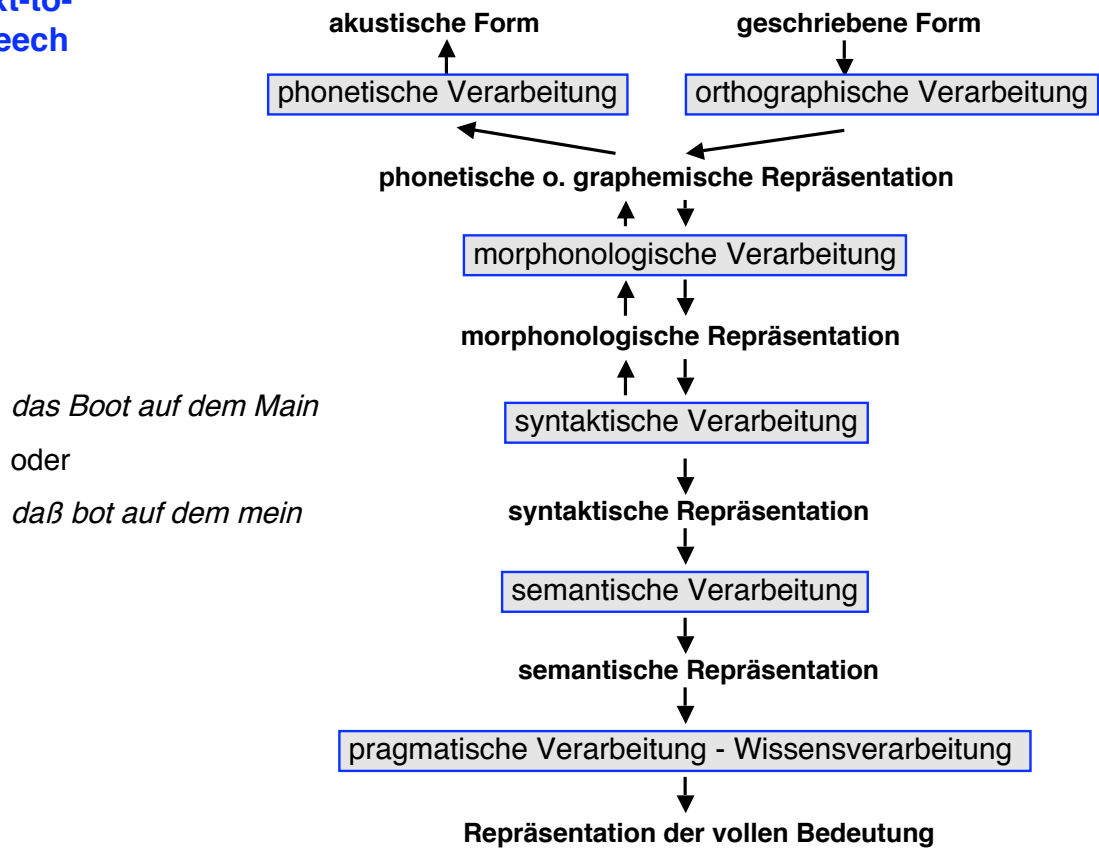
### Textverstehen

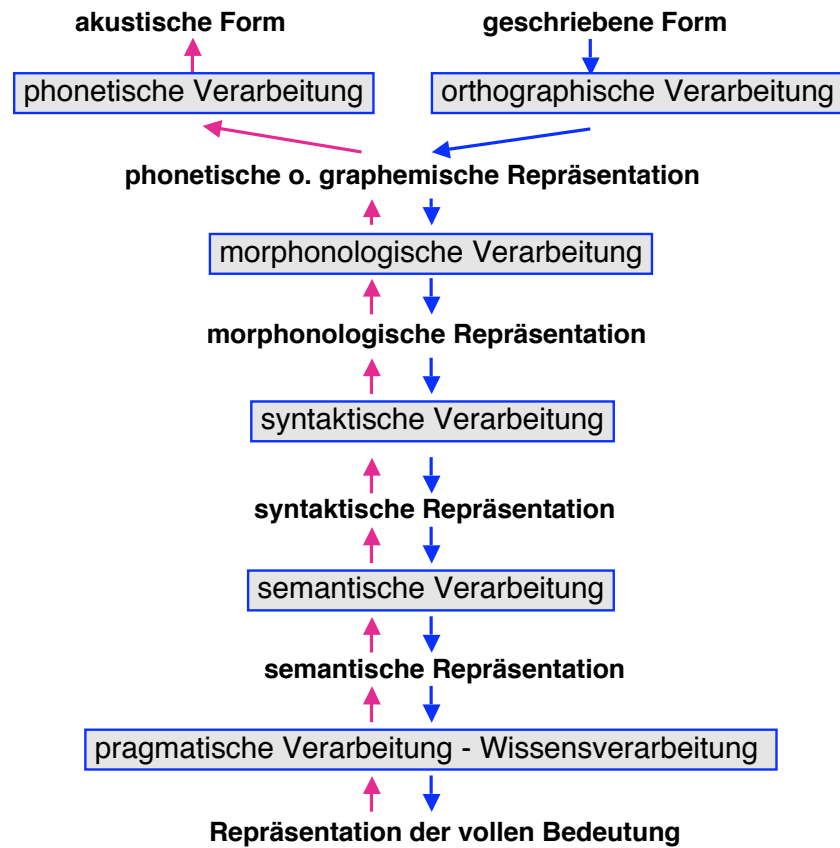


## Text-to-Speech



## Text-to-Speech





## Beispiele



- ❑ Grammatikfehler und Sprechfehler:
  - ❑ Das Verfassen der Kinderbücher und der Reiseberichte haben dem Autor viel Ruhm eingebracht.
  - ❑ Die Poxen zum Backen...
- ❑ Holzwegsätze:
  - ❑ The canoe floated down the river sank.
  - ❑ Er bezichtigte den Vater des Schreibens unkundiger Kinder.
  - ❑ Peter beschuldigte sie der Geheimniskrämerei ähnlichen Verhaltens.



phonetische Ambiguität (Homophone)

**Miene - Mine**

orthographische Ambiguität (Homographen)

**übersetzen - übersetzen**

lexikalische Ambiguität (Homonyme)

**Ball - Ball**

morphologische Ambiguität

**Staubecken - Staubecken**

**Hauptpostsekretär**



### **syntaktische Ambiguität**

Peter fuhr seinen Freund sturzbetrunknen nach Hause.

Visiting relatives can be boring.

Ich traf den Sohn des Nachbarn mit dem Gewehr.

### **kompositionell-semantische Ambiguität**

Die zwei Mitarbeiter müssen vier Sprachen beherrschen.

### **pragmatische Ambiguität**

Könnten Sie die Aufgabe lösen.



phonetische Ambiguität (Homophone)

**Miene - Mine**

orthographische Ambiguität (Homographen)

**übersetzen - übersetzen**

lexikalische Ambiguität (Homonyme)

**Ball - Ball**

morphologische Ambiguität

**Staubecken - Staubecken**

**Hauptpostsekretär**

## Lexical Ambiguity



Certain Readings are less preferred

*Auf dem Tisch lag ein **Heft**.*

*Ich habe einen **Stift** gefunden.  
gesucht.*

*Auf der Werkbank lag ein **Heft**.*

*Ich habe einen jungen **Stift***

The preference can be influenced by concept.

*The goal keeper opened the **Ball**. vs. The President opened the ball*

*The astronomer married a **star**. vs. The movie director married a star.*



- o **syntaktische Ambiguität**

- o Peter fuhr seinen Freund sturzbetrunk nach Hause.
- o Visiting relatives can be boring.
- o Ich traf den Sohn des Nachbarn mit dem Gewehr.

- o **kompositionell-semantische Ambiguität**

- o Die zwei Mitarbeiter müssen vier Sprachen beherrschen.

- o **pragmatische Ambiguität**

- o Könnten Sie die Aufgabe lösen.

## Ambiguität beim Parsing



In fast allen realen Situationen sind Sätze hochgradig ambig.

### Beispiel:

Grammatik: deutsche LFG-Grammatik von Christian Rohrer

Parser: XLE Parser von XEROX PARC (Kaplan, Maxwell, Shemtov,...)

Korpus: Teilmenge des NEGRA Korpus Frankfurter Rundschau (Saarbrücken)

ØSatzlänge: ca. 16 Wörter

ØAmbiguität: >3000 Lesarten pro Satz  
(durch heuristische Präferenzen reduziert auf 7 Lesarten)



*„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“*



*„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“*

Der Satz weist lexikalische (L), syntaktische (S) und anaphorische (A) Ambiguitäten auf, die uns nicht auffallen.



*„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“*

Der Satz weist lexikalische (L), syntaktische (S) und anaphorische (A) Ambiguitäten auf, die uns nicht auffallen.

Wieviele Lesarten besitzt dieser Satz?

**258.048**



*„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“*

Das berechnet sich so:

- L *Früher* kann sowohl eigenständiges Adverb als auch Komparativ von *früh* sein (2);
- L die Verbform *stellten* ist ambig zwischen Präteritum und Konjunktiv (2);
- S die Nominalphrase *die Frauen* kann sowohl Subjekt als auch Objekt des Satzes sein (2);
- S *am Wochenende* kann die Insel, die Frauen oder das Verb modifizieren (3);
- S *mit Blumenmotiven* kann sich auf die Kopftücher beziehen, ein Instrument der Herstellung sein oder ein Adjunkt im Sinne von *gemeinsam mit Blumenmotiven* (3);
- L *her* hat auch eine direktionale Bedeutung (2);





*„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“*

Und weiter:

- S** der Relativsatz könnte jede der vier Nominalphrasen im Plural modifizieren (4);
- S** sowohl *die* als auch *ihre Männer* kann Subjekt des Relativsatzes sein (2);
- A** das Possessivpronomen *ihre* kann auf jede der Nominalphrasen referieren (4);
- L** *Montagen* hat eine zweite Lesart als Nominalisierung von *montieren* (2);
- S** *die Hauptinsel* kann im Genitiv zu der vorangegangenen NP gehören oder im Dativ die Käuferin bezeichnen (2);
- S** die drei Präpositionalphrasen des Relativsatzes können sich in insgesamt sieben Kombinationen mit den jeweils vorhergehenden NPs oder mit dem Verb verbinden (7);
- L** *verkauften* zeigt wieder die Ambiguität zwischen Präteritum und Konjunktiv auf (2).



*„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.“*

Durch Multiplikation ergibt sich die Gesamtambiguität:

$$2 \times 2 \times 2 \times 3 \times 3 \times 2 \times 4 \times 2 \times 4 \times 2 \times 2 \times 7 \times 2 = \underline{258.048}$$



- ❑ warum viele ungrammatische Sätze erzeugt werden
  - ➡ z.B. Sprechfehler, Grammatikfehler
- ❑ warum viele ungrammatische Sätze verstanden werden
  - ➡ z.B. in der der Kommunikation mit Kindern oder Ausländern
- ❑ warum viele grammatische Sätze nicht erzeugt werden
  - ➡ z.B. durch Präferenzen in der Generierung
- ❑ warum viele grammatische Sätze nicht verstanden werden
  - ➡ z.B. Holzwegsätze
- ❑ wie die Verarbeitung zeitlich strukturiert ist
  - ➡ z.B. Effizienz, Abfolge der Verarbeitungsschritte
- ❑ welchen Aufwand die Verarbeitungsschritte erfordern
  - ➡ z.B. Abhängigkeiten von anderen kognitiven Belastungen



**Der Wissenschaftler schrieb zwei Bücher über den Ursprung der menschlichen Sprache, die in vielen Fernsehsendungen diskutiert wurden, ab.**



Der Wissenschaftler **schrieb** zwei **Bücher** über den Ursprung der menschlichen Sprache, **die in vielen Fernsehsendungen diskutiert wurden, ab.**

## Exkurs: Probleme des neuronalen Ansatzes



Wenn ein Teil der sprachlichen Kompetenz angeboren ist, dann kann der Spracherwerb nicht alleine durch neuronale Lernverfahren modelliert werden.

Der Mensch erlernt viele Arten der Sprachverwendung (Performanztypen). Er scheint jedoch die einmal erworbene Kompetenz immer weiter zu verwenden.

Selbst wenn das System den Anforderungen des Erstspracherwerbs technisch gewachsen wäre, müßte die Maschine eine ähnliche Sprachsozialisation wie der Mensch durchlaufen, was aus vielen Gründen nicht möglich ist.



Der Mann, der die Katze beobachtete, staunte.

Der Mann, der die Katze, die den Vogel jagte,  
beobachtete, staunte.

Der Mann, der die Katze, die den Vogel, der laut schrie, jagte,  
beobachtete, staunte.



Der Hammer, mit dem der Handwerker, den Peter angerufen hatte, die  
Nägel einschlug, war mindestens drei Pfund schwer.



Peter hat den Wagen, der seit Tagen vor der Haustür steht, gekauft.

Peter hat den Wagen gekauft, der seit Tagen vor der Haustür steht.

Peter hat den Wagen, der vor der Haustür steht, langsam und sorgfältig lackiert.

Peter hat den Wagen langsam und sorgfältig lackiert, der vor der Haustür steht.

Der Mann hat dem Jungen, der aus der Schule kam, den Ball gegeben.

Der Mann hat dem Jungen den Ball gegeben, der aus der Schule kam.



Kleine Kinder brauchen viel Liebe

Kleine Kinder brauchen viel Liebe

Peter gab dem Jungen den Ball



weil Peter dem Jungen den Ball, der vor der Haustür lag, gab

weil Peter dem Jungen den Ball gab, der vor der Haustür lag

## Hauptansätze der CL

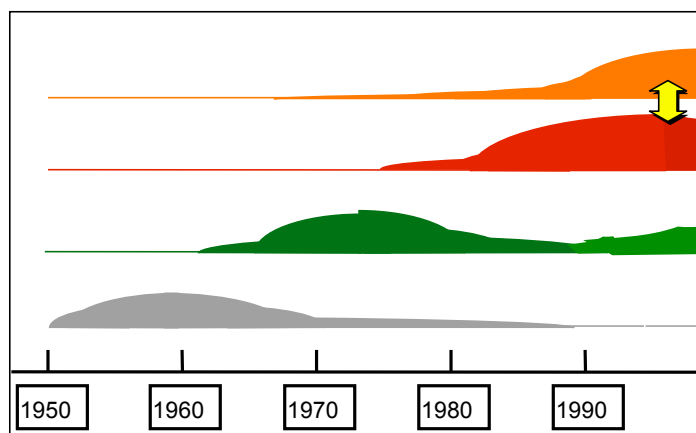


statistische und konnektio-  
nistische Methoden in der CL

deklarative linguistische  
Formalismen in der CL

spezielle Verfahren für die CL

direkte Programmierung, keine  
Trennung von Beschreibung und  
Verarbeitung





Direkte Programmierung in einer traditionellen Programmiersprache.  
Keine Trennung von Kompetenz und Performanz, also auch keine Trennung von Grammatik und Verarbeitung

- ❑ Beispiele :
  - ❑ SYSTRAN, SHRDLU, frühe SFB 100 Systeme
- ❑ Kompetenzmodellierung:
  - ❑ als Modelle theoretisch uninteressant, nicht überprüfbar, Kodierung linguistisch uninteressant, schwer erweiterbar
- ❑ Performanzmodellierung:
  - ❑ als Modelle theoretisch uninteressant, weil mit der Kompetenz vermischt, keine Ansätze zur Integration psycholinguistischer Erkenntnisse
- ❑ Anwendungspotential:
  - einige wenige Systeme sind zur Anwendungsreife gelangt (z.B. SYSTRAN), fast nicht mehr erweiterbar, für neue Entwicklungen nicht geeignet



Spezielle Verfahren und Beschreibungssprachen wurden entwickelt. Trennung von Kompetenz und Performanz, vielfach noch immer Vermischung von Wissen und Verarbeitung

- ❑ Beispiele:
  - ❑ Augmented Transition Networks (ATN), Augmented Phrase Structure Grammar (APSG), EUROTRA Framework
- ❑ Kompetenzmodellierung:
  - ❑ verschieden von den Modellen der Linguistik, als linguistische Modelle theoretisch wenig interessant, vielfach Vermischung mit prozeduralen Elementen
- ❑ Performanzmodellierung:
  - ❑ wenige aber sehr ernsthafte Versuche, einige Gesichtspunkte der Performanzmodellierung zu berücksichtigen, Einflüsse der Psycholinguistik, Hindernis ist das Fehlen plausibler Kompetenzmodelle
- ❑ Anwendungspotential:
  - fast alle der heute marktreifen Systeme gehören zu dieser Klasse (z.B. METAL, Q&A)



Deklarative Grammatikformalismen, in denen sich linguistische Grammatikmodelle und Einzelanalysen kodieren lassen. Dadurch Aufhebung der Trennung von theoretischer Linguistik und Computerlinguistik.

- ❑ Beispiele:
  - ❑ fast alle Unifikationsgrammatikmodelle, neuere semantische Formalismen
- ❑ Kompetenzmodellierung:
  - ❑ deklarative linguistisch fundierte Modelle; unabhängig von Verarbeitungsrichtung, Verarbeitungsreihenfolge und Verarbeitungsalgorithmen; logisch fundierte Semantik, transparente Modularisierung und Hierarchisierung des Wissens
- ❑ Performanzmodellierung:
  - ❑ deduktive Verarbeitung; in den fortgeschrittensten Systemen erfolgt die Verarbeitung durch Typdeduktion; bisher keine plausiblen Performanzmodelle
- ❑ Anwendungspotential:
  - noch keine marktreifen System, bisher noch mangelnde Effizienz



Statistische Verfahren in der akustischen Spracherkennung (Hidden Markov Models), und in der maschinellen Übersetzung; massiv-paralleler Ansatz zur Modellierung der neuronalen Strukturierung des menschlichen Hirns.

- ❑ Beispiele:
  - ❑ Hidden Markov Models (HMM), Parsing mit neuronalen Netzen
- ❑ Kompetenzmodellierung:
  - ❑ für die Theoriebildung uninteressant, die Kompetenz ist nicht transparent modelliert, keine Verbindung zu den Theorien der Linguistik, unzureichende Darstellung der Rekursivität
- ❑ Performanzmodellierung:
  - ❑ Lernverfahren, massive Parallelität könnte Schlüssel zum Effizienzproblem sein, Potential für die Modellierung linguistischer Präferenzen und anderer unscharfer Konzepte z.B. in der lexikalischen Semantik, Potential für holistische Ansätze
- ❑ Anwendungspotential:
  - großes Potential in der akustischen Spracherkennung und in der akustischen Sprachsynthese, für rein statistische oder neuronale Verfahren geringes Potential in der linguistischen Verarbeitung