

---

# Einführung in die Computerlinguistik

## I – Überblick

21.04.2009

Sebastian Pado

---

## Vorlesungsplan

21.04.09	Einführung/Überblick	23.04.09	Gesprochene Sprache
28.04.09	Morphologie, Automaten 1	30.04.09	entfällt
05.05.09	Morphologie, Automaten 2	07.05.09	Übung
12.05.09	Syntax 1	14.05.09	Übung
19.05.09	Syntax 2 (Parsing)	21.05.09	Übung
26.05.09	Korpuslinguistik	28.05.09	Übung
Pfingsten (vorlesungsfrei)			
09.06.09	Statistische Modellierung	11.06.09	Übung
16.06.09	Statistisches Parsing	18.06.09	Übung
23.06.09	Semantik 1	25.06.09	Übung
30.06.09	Semantik 2	02.07.09	Übung
07.07.09	Informationszugriff	09.07.09	Übung
14.07.09	MÜ	16.07.09	Übung zu Probeklausur
21.07.09	entfällt	23.07.09	Klausur

## Technisches

---

Zur Vorlesung gehören:

1. Das Vorlesungsskript (auf der Homepage des Kurses)  
<http://www.ims.uni-stuttgart.de/lehre/teaching/2009-SS/CL-Einf/>
2. Ausgewählte Texte aus drei Lehrbüchern in englischer und deutscher Sprache
  - Verbindliche Texte und Bonustexte
3. Übungsaufgaben: werden (tendenziell wöchentlich) in der Vorlesung am Dienstag ausgegeben (und auf die Homepage gestellt)
4. Übungssitzungen: dienen der Besprechung der Übungsaufgaben, der Literatur, und der Wiederholung und Vertiefung des Vorlesungsstoffes.



## Prüfungsleistung

---

Klausur über den Umfang der Vorlesung: Stoff aus dem Vorlesungsskript, den Übungen und den Lektüretexten

Klausurtermin: voraussichtlich der Übungstermin in der letzten Semesterwoche

Linguistik-HF auf Diplom: **Diese Klausur ist Teil der Orientierungsprüfung!**

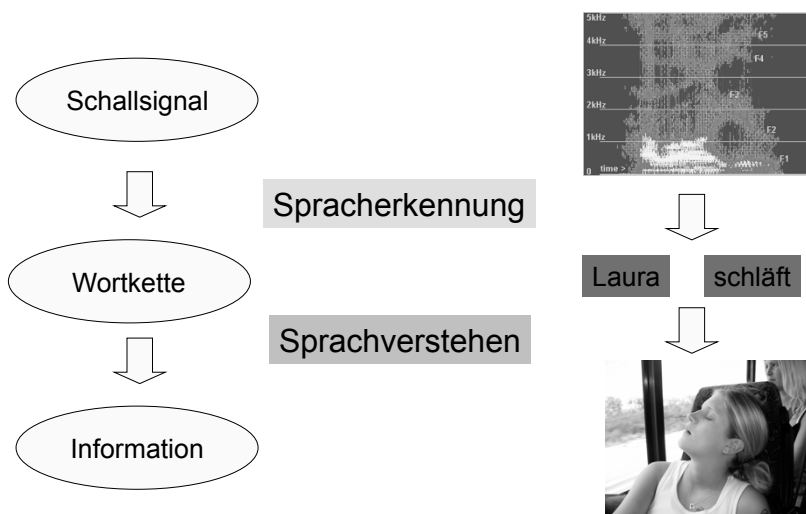


## Literatur

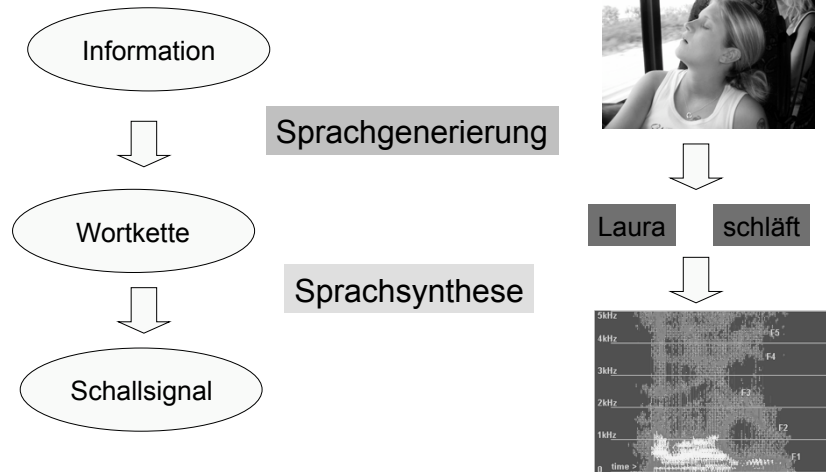
- Eine sehr gutes (wenn auch methodenlastiges) englischsprachiges Einführungswerk: Jurafsky, Daniel/ Martin, James H. 2000. Speech and Language Processing. Prentice-Hall.
- Englischsprachiges Handbuch über die statistischen Aspekte der Computerlinguistik: Christopher Manning and Hinrich Schütze. Foundations of Statistical Natural Language Processing. MIT Press.
- Ein aktuelles deutsches Handbuch zur Computerlinguistik. Carstensen, Kai-Uwe et al. 2001. Computerlinguistik und Sprachtechnologie - Eine Einführung. Heidelberg: Spektrum Akademischer Verlag.



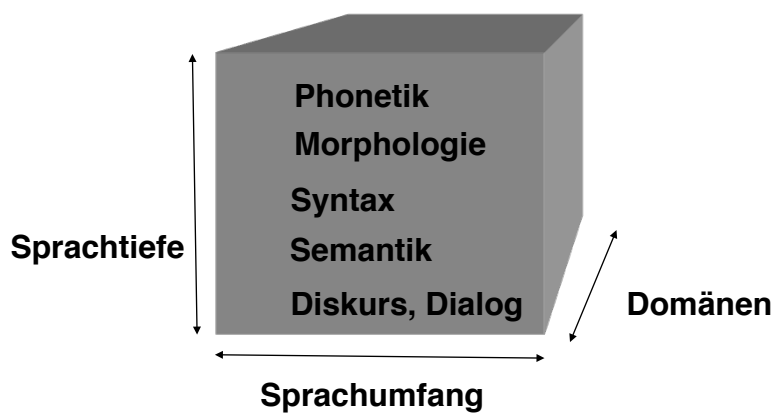
## Was ist Sprachverarbeitung?



## Was ist Sprachverarbeitung?



## Schichtenmodell der Sprache



## Gegenstände der Computerlinguistik

- Die Entwicklung von Formalismen und Werkzeugen für die Repräsentation, Verarbeitung und Akquisition von linguistischem Wissen der verschiedenen Ebenen:
  - Phonetik und Phonologie
  - Morphologie und Syntax
  - Semantik
  - Pragmatik und Diskursstruktur
- Die Modellierung und Implementierung der komplexen Zusammenhänge und Abläufe bei:
  - Sprachverstehen
  - Sprachproduktion
  - Spracherwerb
- Die Entwicklung von natürlichsprachlichen Anwendungssystemen.

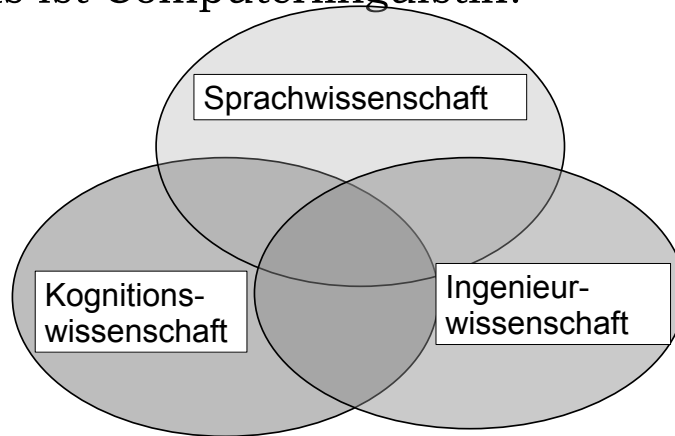


## Vorlesungsplan

23.04.09	Gesprochene Sprache	<b>Phonetik</b>
28.04.09	Morphologie, Automaten 1	<b>Morphologie</b>
05.05.09	Morphologie, Automaten 2	
12.05.09	Syntax 1	<b>Syntax</b>
19.05.09	Syntax 2 (Parsing)	
26.05.09	Korpuslinguistik	
09.06.09	Statistische Modellierung	
16.06.09	Statistisches Parsing	
23.06.09	Semantik 1	<b>Semantik</b>
30.06.09	Semantik 2	
07.07.09	Informationszugriff	<b>Sprachtechnologie</b>
14.07.09	MÜ	



## Was ist Computerlinguistik?



## Computerlinguistik als Sprachwissenschaft

Eine wesentliche Voraussetzung für die Computerlinguistik ist die systematische und einheitliche Beschreibung von sprachlichem Wissen und sprachlichen Strukturen. Umgekehrt stellt die Computerlinguistik für die Erhebung und Erfassung komplexer sprachlicher Struktur Theorien und Werkzeuge zur Verfügung. Insofern gehört Computerlinguistik zu den sprachwissenschaftlichen Disziplinen, zusammen mit

- ▶ Theoretischer Linguistik / allgemeiner Sprachwissenschaft
- ▶ Historischer und vergleichender Sprachwissenschaft
- ▶ Phonetik
- ▶ Germanistischer, romanistischer, japanischer ... Sprachwissenschaft

## Computerlinguistik als Kognitionswissenschaft

---

Das übergeordnete Erkenntnisziel der Computerlinguistik ist die Erforschung der menschlichen Sprachfähigkeit: Wie ist sprachliches Wissen beim Menschen organisiert, und wie wird Sprache produziert, verstanden, und gelernt? Insofern gehört die Computerlinguistik zu den Kognitionswissenschaften, die die "kognitiven" Fähigkeiten des Menschen erforschen, zusammen mit den Fächern und Forschungsbereichen:

- ▶ Psycholinguistik
  - ▶ kognitive Psychologie
  - ▶ Neuropsychologie
  - ▶ Künstliche Intelligenz (in der Informatik)
  - ▶ Sprachphilosophie
- 



## Computerlinguistik als Ingenieurwissenschaft

---

Die praktische Zielsetzung der Computerlinguistik ist die Realisierung von Computersystemen, die sprachliches Wissen und sprachliche Fertigkeiten einsetzen, um den Menschen in der Kommunikation, beim Verwenden von Sprache und beim Umgang mit sprachlichen Dokumenten zu unterstützen. Computerlinguistik als Sprachtechnologie gehört in den Bereich der Informationstechnologie, zusammen mit den Fächern und Forschungsbereichen

- ▶ Angewandte Informatik
  - ▶ (Maschinelle Lernverfahren)
  - ▶ Informationswissenschaft
  - ▶ Signalverarbeitung (in der Elektrotechnik)
- 



## Sprachtechnologie

---

- Wichtige Teilbereiche der Sprachtechnologie
  - Informationszugriff und -management
  - Multilinguale Anwendungen
  - Gesprochene Sprache und Dialog



## Informationszugriff und -management

---

- Information Retrieval
- Informationsextraktion
- Frage-Antwort-Systeme (Question Answering)
- Zusammenfassung (Summarization)
- Dokumentklassifikation



## Multilinguale Anwendungen

- Maschinelle Übersetzung
- Multilinguale Informationssuche
- Sprachlernsysteme

## Anwendungen für gesprochene Sprache

- Sprachsynthese: Text-to-Speech Systeme (TTS)
- Spracherkennung: Diktiersysteme
- Sprachdialog:
  - Telefonie-Anwendungen: Fahrplan-Information, Call-Center, Telebanking
  - Dialog für Gerätebedienung: Mensch-Computer-Interaktion, Sprache im Fahrzeug (In-car), Intelligentes Haus, Robotersteuerung

---

# Einführung in die Computerlinguistik

## II – Die Rolle von Wissen

SS 2009

Sebastian Pado

---



---

## Der Turing-Test

- ▶ Alan Turing (1950): Wann ist ein Computer intelligent?
  - ▶ Wenn er nicht von anderen Menschen unterscheidbar ist
  - ▶ **Kommunikation**
  
- ▶ Die zweite Generation von computerlinguistischen Systemen (60er/70er Jahre) bestand v.a. aus Dialogsystemen  
Die beiden vermutlich berühmtesten Systeme:
  - ▶ ELIZA, 1966, von Joseph Weizenbaum
    - ▶ Ein „automatischer Psychater“
  - ▶ SHRDLU, 1970, von Terry Winograd
    - ▶ Ein System, das Fragen über einen Tisch mit Bauklötzen beantwortet

Was ist schwieriger?

---



## ELIZA: Beispieldialog

---

(1) U: I have been very depressed lately  
S: *You say you have been very depressed lately.*

(2) U: Yes, my mother beats me and my father calls me names  
S: *You're being somewhat short with me*

(3) U: What about my mother beating me?  
S: *Tell me more about your family*

(4) U: My father calls me nasty names  
S: *Who else in your family calls you nasty names*

(5) U: I think the dog would if he could  
S: *Do you think its likely that he could*

(6) U: No, you do  
S: *You don't seem very talkative today*

---

## ELIZA – was dahintersteckt

---

TEMPLATE: I *\_X1\_*

RESPONSE: You say you *\_X1\_*

TEMPLATE: My *\_X1\_* (*category family*) *\_X2\_*

RESPONSE: Who else in your family *\_X2\_* ?

TEMPLATE: *\_X1\_* if *\_X2\_*

RESPONSE: Do you think it's likely that *\_X2\_* ?

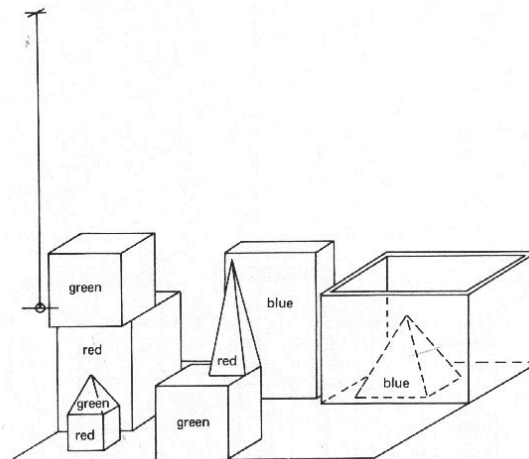
TEMPLATE: *\_X1\_*

RESPONSE: You're being somewhat short with me.

---

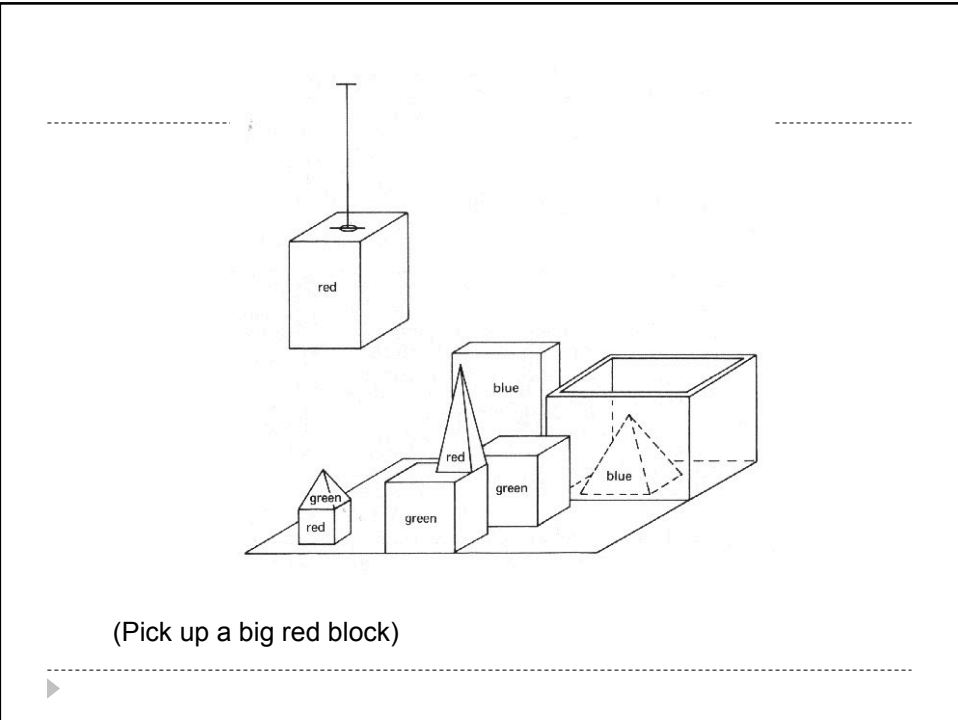
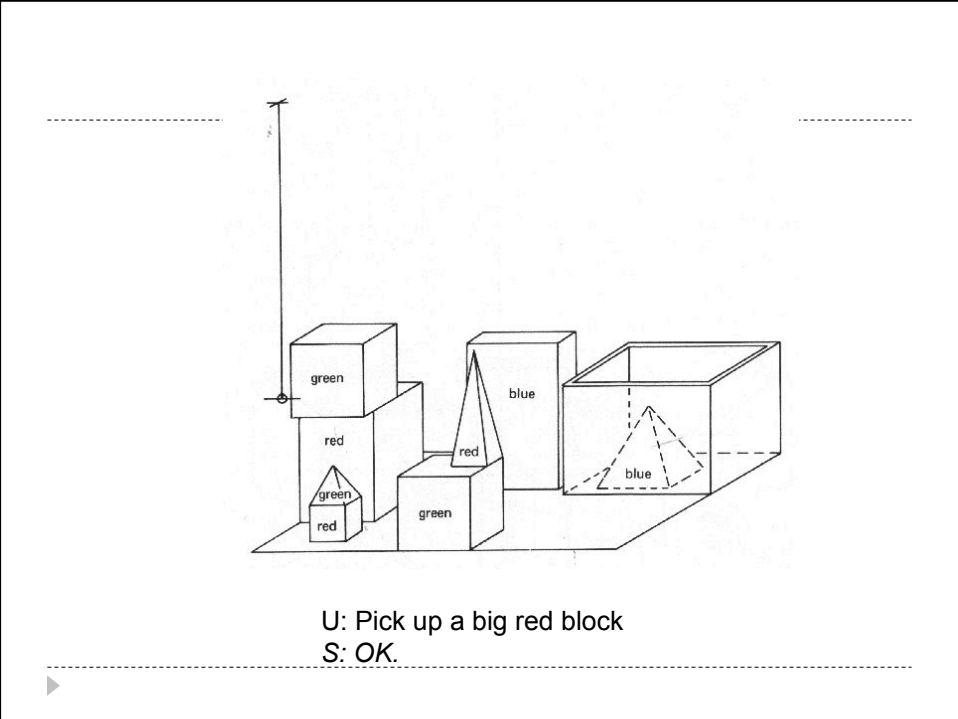
## ELIZA

- ▶ ELIZA ist ein Dialogsystem, das beliebig komplexe Eingaben mit beliebigem Wortschatz zu beliebigen Themen akzeptiert.
- ▶ ELIZA arbeitet mit einfachen Mustervergleichs-Techniken (Pattern Matching), ohne Einsatz von Wissen:
  - ▶ Templates: Muster mit variablen Teilen, die mit der Benutzereingabe abgeglichen werden, und
  - ▶ Template-basierten System-Äußerungen (Prompts)
- ▶ ELIZA hat in gewisser Hinsicht den Turing-Test absolviert, dies aber unter besonderen Rahmenbedingungen.
- ▶ **Kein echtes Verständnis der Eingabe!**
- ▶ (Tip: Emacs, Meta-X doctor)



SHRDLU: Die "Blocks World"





U: Grasp the pyramid.

A diagram showing a robotic arm with a gripper holding a red cube. The gripper is positioned above a table with several objects: a small pyramid (green top, red base), a green cube, a red pyramid, a green cube, a blue cube, and a large transparent cube containing a blue pyramid. A red cube is being held by the gripper. The scene is framed by a dashed line at the top and a dashed line at the bottom with a right-pointing arrow.

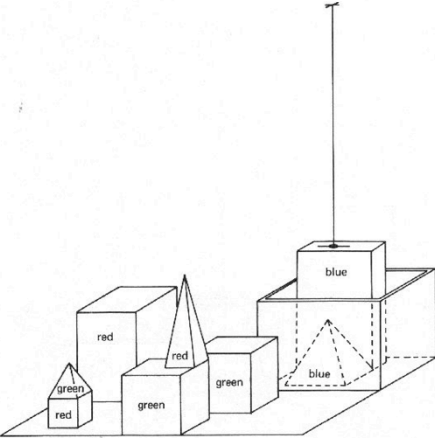
U: Grasp the pyramid.  
S: *I don't understand which pyramid you mean.*

A diagram showing a robotic arm with a gripper holding a red cube. The gripper is positioned above a table with several objects: a small pyramid (green top, red base), a green cube, a red pyramid, a green cube, a blue cube, and a large transparent cube containing a blue pyramid. A red cube is being held by the gripper. The scene is framed by a dashed line at the top and a dashed line at the bottom with a right-pointing arrow.

U: Find a block which is taller than the one you are holding and put it into the box.  
 S: By „it“, I assume you mean the block which is taller than the one I am holding. OK

(U: Find a block which is taller than the one you are holding and put it into the box.)

-----

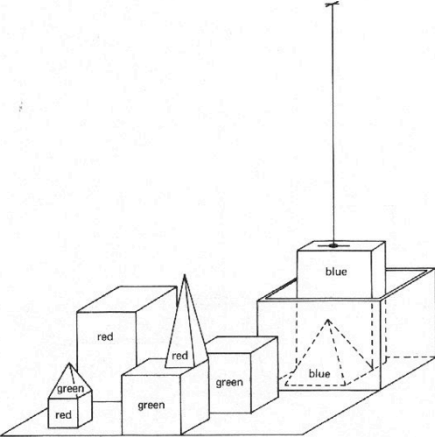


U: What does the box contain?  
S: *The blue pyramid and the blue block*

-----

▶

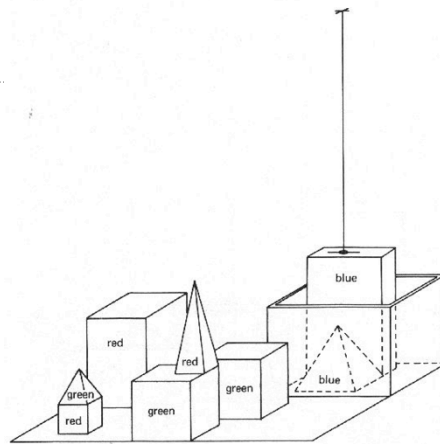
-----



U: What is the pyramid supported by?  
S: *The box*

-----

▶



U: How many blocks are not in the box?  
S: *Four of them*

## SHRDLU [2]

SHRDLU ist ein wissensbasiertes Dialogsystem.  
Im Unterschied z.B. zu ELIZA verwendet es bei der Dialogführung unterschiedliche Arten von Wissen, und zwar

- ▶ Sprachliches/linguistisches Wissen
- ▶ Kontextwissen
- ▶ Weltwissen

## SHRDLU

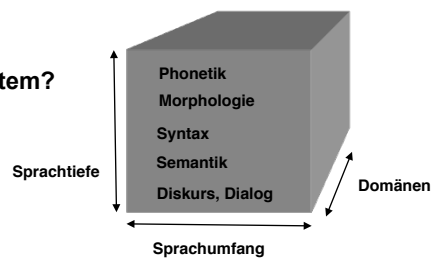
SHRDLU beantwortet Fragen, führt Anweisungen aus und lernt Begriffe.

Wichtige Programmkomponenten von SHRDLU sind:

- ▶ (Linguistische) Analyse
- ▶ Generierung
- ▶ (Handlungs-)Planung
- ▶ (grafische) Visualisierung

**Warum funktioniert das System?**

**Es arbeitet in einer kleinen, eingeschränkten Mini-Welt oder –Domäne („blocks world“).**



<http://hci.stanford.edu/~winograd/shrdlu/>

## Sprachliches Wissen in SHRDLU: Beispiele

### Morphologisches Wissen:

regelmäßige Verben bilden      *grasp* ist regelmäßiges Verb  
Präteritum auf -ed              *put* ist unregelm. Verb mit Prät. *put*

### Syntaktisches Wissen:

In Imperativen steht das      *grasp* ist transitives Verb  
Verb an erster Stelle            *stop* ist intransitives Verb

### Semantisches Wissen:

A+N in attributiven            *red* bezeichnet rote Dinge (?)  
Konstruktionen bezeichnet    *pyramid* ...  
Dinge, die unter A und unter    *grasp* ...  
N fallen

## Sprachliches Wissen in SHRDLU: Beispiele

Grammatik	Lexikon
<b>Morphologisches Wissen:</b> regelmäßige Verben bilden Präteritum auf -ed	<i>grasp</i> ist regelmäßiges Verb <i>put</i> ist unregelm. Verb mit Prät. <i>put</i>
<b>Syntaktisches Wissen:</b> In Imperativen steht das Verb an erster Stelle	<i>grasp</i> ist transitives Verb <i>stop</i> ist intransitives Verb
<b>Semantisches Wissen:</b> A+N in attributiven Konstruktionen bezeichnet Dinge, die unter A und unter N fallen	<i>red</i> bezeichnet rote Dinge (?) <i>pyramid</i> ... <i>grasp</i> ...



## Grammatisches und lexikalisches Wissen

- ▶ Morphologische, syntaktische, semantische Regularitäten sind tendenziell in der Grammatik kodiert
- ▶ Spezielle morphologische, syntaktische, semantische Information über Einzelwörter sind im Lexikon kodiert.
- ▶ Achtung:
  - ▶ Es gibt keine scharfe Grenze zwischen systematischer grammatischer Information und ideosynkratischer lexikalischer Information.
  - ▶ Unterschiedliche linguistische Theorien schlagen eine unterschiedliche Arbeitsteilung zwischen Grammatik und Lexikon vor.



## Außersprachliches Wissen

---

### ▶ Kontextwissen:

- ▶ Sprachlicher Kontext / Dialoggeschichte: Welches Objekt wurden zuletzt erwähnt? (*Put it into the box.*)
- ▶ Situationskontext: Welche Objekte kommen in der Äußerungssituation vor? (*What is the pyramid supported by?*)

### ▶ Weltwissen:

- ▶ Episodisches Wissen: Wissen über Einzelfakten  
*"Es gibt zwei rote Klötze."  
"Die Kiste enthält eine Pyramide"*
- ▶ Regelwissen: Wissen über mathematische, naturwissenschaftliche, gesellschaftliche Regularitäten  
*"Zwei Objekte können nicht den gleichen Platz einnehmen."  
"Ein Objekt muss eine ebene Auflagefläche besitzen, damit ein zweites stabil darauf stehen kann"*



## Lektüre

---

- ▶ Carstensen et al. Kapitel I: "Computerlinguistik, was ist das?"
- ▶ Jurafsky/Martin Kapitel I: "Introduction"

