Einführung in die Computerlinguistik

WS 2013/14

Manfred Pinkal

Vorläufiges Programm

22.10.13	Einführung
29.10.12	Morphologie und Automaten I
05.11.12	Morphologie und Automaten II
12.11.12	Syntax I
19.11.12	Syntax II
26.11.12	Parsing
03.12.13	Merkmalsstrukturen
10.12.13	Semantik
17.12.13	Statistische Modelle

Im neuen Jahr: Statistische Verfahren, Spracherkennung, Anwendungen

Technisches: Vorlesung und Übung

- Vorlesungsskript (auf der Homepage des Kurses)

  `http://www.coli.uni-saarland.de/courses/I2CL-13/`
- Ausgewählte Kurztexte in englischer und deutscher Sprache
- Übungsaufgaben:
  - Ausgabe: Übungsblatt wird zur Vorlesung am Dienstag auf die Homepage gestellt (tendenziell wöchentlich)
  - Einreichen der Lösungen: bis zum Dienstag der folgenden Woche (Vorlesungsbeginn), als PDF oder auf Papier
  - Besprechung: in der nächsten Übungssitzung am Freitag
- Übungsgruppen

Technisches: Prüfungsvoraussetzungen

Prüfungsvoraussetzung: Schriftliche Bearbeitung der Übungsaufgaben, das heißt genauer:

1. Alle Aufgabenblätter (mit höchstens einer Ausnahme) müssen bearbeitet sein. Aufgabenblatt zählt als bearbeitet, wenn für alle Aufgaben ein ernsthafter Lösungsversuch vorliegt. Dies schließt Teilaufgaben ein, wenn sie unabhängig gelöst werden können.
2. Insgesamt müssen mindestens 50% der Punkte erreicht sein.
3. Aufgaben können in Gruppen mit bis zu drei Studierenden bearbeitet werden (Näheres auf der Homepage des Kurses)

   Abschreiben ist nicht erlaubt: Bei allen Beteiligten wird das Blatt als nicht bearbeitet gewertet. **Das führt im Wiederholungsfall automatisch zum Verlust der Klausurzulassung.**

Technisches: Klausur

- Prüfungsleistung: Klausur über den Stoff der Vorlesung, der
  - im Vorlesungsskript
  - den Übungen und
  - den Lektüretexten

  vorkommt.

- Klausur in der ersten Woche der vorlesungsfreien Zeit

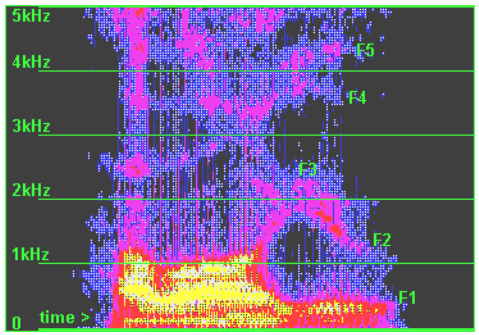
- **Anmeldung zur Prüfung, wird in Kürze bekanntgegeben**

  Wichtig: Ohne fristgerechte Meldung keine Teilnahme möglich!

Einführungsliteratur und andere Informationsquellen

- Eine ausgezeichnetes englisch-sprachiges Einführungswerk: Jurafsky, D./ Martin, J.: Speech and Language Processing, 2009
- Ein aktuelles deutsches Handbuch der Computerlinguistik: Carstensen, Kai-Uwe et al.: Computerlinguistik und Sprachtechnologie - Eine Einführung, 2009
- Ein linguistisches Wörterbuch: H. Bussmann: Lexikon der Sprachwissenschaft, 2008
- Das Online-Wörterbuch: LEO
- Und: Die WikiPedia

Was ist Sprachverarbeitung?



Schallsignal

Spracherkennung

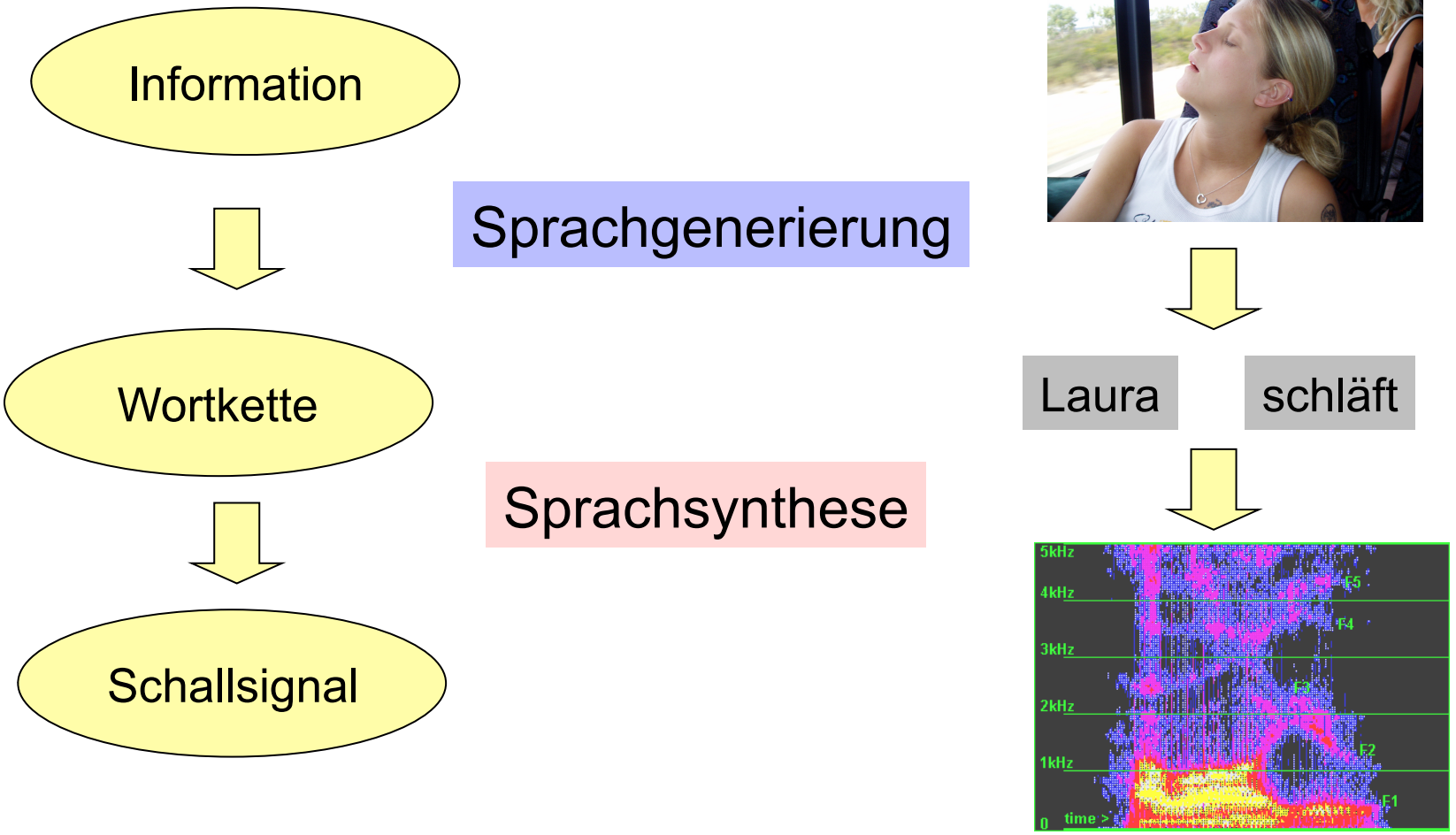
Wortkette

Laura schläft

Sprachverstehen

Information

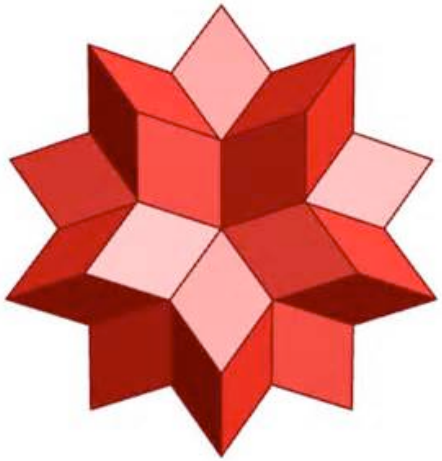
Was ist Sprachverarbeitung?



Aufgaben der Computerlinguistik

- Modellierung und Implementierung komplexer Zusammenhänge und Abläufe bei:
  - Sprachverstehen
  - Sprachproduktion
  - Spracherwerb
- Entwicklung von Formalismen und Werkzeugen für die Repräsentation, Verarbeitung und den Erwerb von sprachlichem Wissen der verschiedenen Ebenen:
  - Phonetik und Phonologie
  - Morphologie und Syntax
  - Semantik
  - Pragmatik, Text, Dialog
- Entwicklung von natürlich-sprachlichen Anwendungssystemen.

Informationszugriff und -management



Google™

Informationszugriff und -management

- Information Retrieval
- Relations-Extraktion
- Frage-Beantwortung (Question Answering)
- Automatische Zusammenfassung (Summarisation)
- Dokumentklassifikation

Gesprochene Sprache

Anwendungen für gesprochene Sprache

- Diktiersysteme, Spracheingabe für medizinische Diagnose, technische Wartungssysteme
- Telefonie-Dialogsysteme: Call-Center, Telefon-Banking, Fahrplanauskunft, ...
- Gerätebedienung: Smartphone, Auto, Haushalt
- Interaktion mit virtuellen Agenten und mit Robotern

Multilinguale Anwendungen

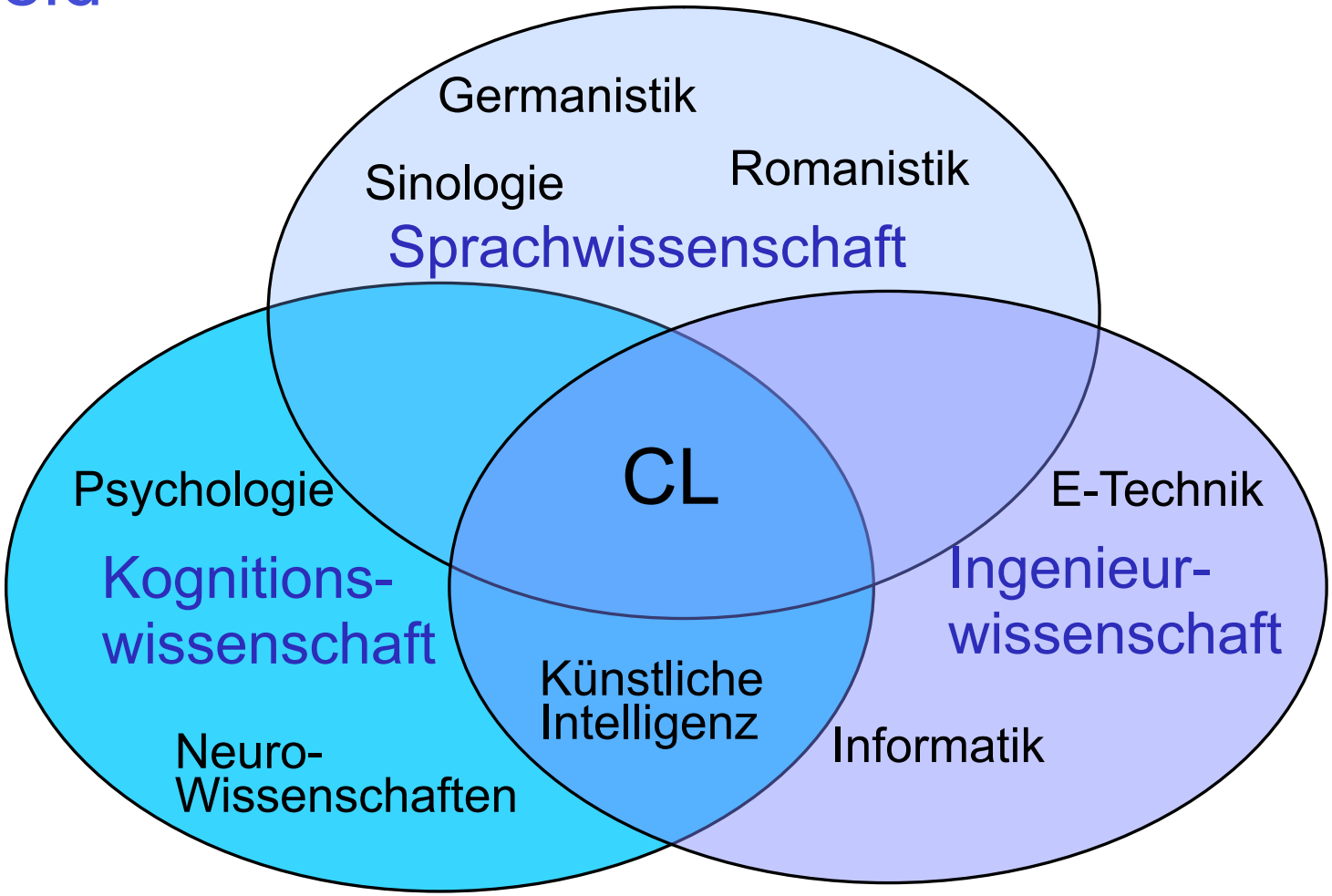
Google translate

translate.google.com

Multilinguale Anwendungen

- Maschinelle Übersetzung
- Multilinguale Informationssuche
- Sprachlernsysteme

Computerlinguistik im interdisziplinären Umfeld



Frühgeschichte der Computerlinguistik

- In den späten 50er Jahren begann in den USA aus politischer Motivation die Förderung der Maschinellen Übersetzung (Russisch-Englisch) als erster computerlinguistischer Anwendungsdisziplin. Die Resultate waren extrem schlecht, die Förderung wurde eingestellt.

- In den späten 60er und den frühen 70er Jahren entstand die zweite Generation von computerlinguistischen Systemen: Dialogsysteme

  Die beiden vermutlich berühmtesten Systeme:

  – ELIZA, 1966, von Joseph Weizenbaum, und
  – SHRDLU, 1970, von T. Winograd

ELIZA: Beispieldialog [1]

(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 seen very talkative today*

ELIZA: Beispieldialog [2]

(7) U: I guess what is really bothering me is that I don`t have any money

S: *Can you elaborate on that*

(8) U: The problem is that I know psychiatrists are expensive and I may not be able to pay you

S: *Please continue, this may be interesting*

(9) U: I thought you might think so

S: *Perhaps in your fantasy we thought each other*

Was steckt hinter ELIZA?

TEMPLATE: `I _X1_`
RESPONSE: `You say you _X1_`

TEMPLATE: `_X1_ my _X2_(category family) _X3_`
RESPONSE: `Who else in your family _X3_ ?`

TEMPLATE: `_X1_ if _X2_`
RESPONSE: `Do you think its likely that _X2_ ?`

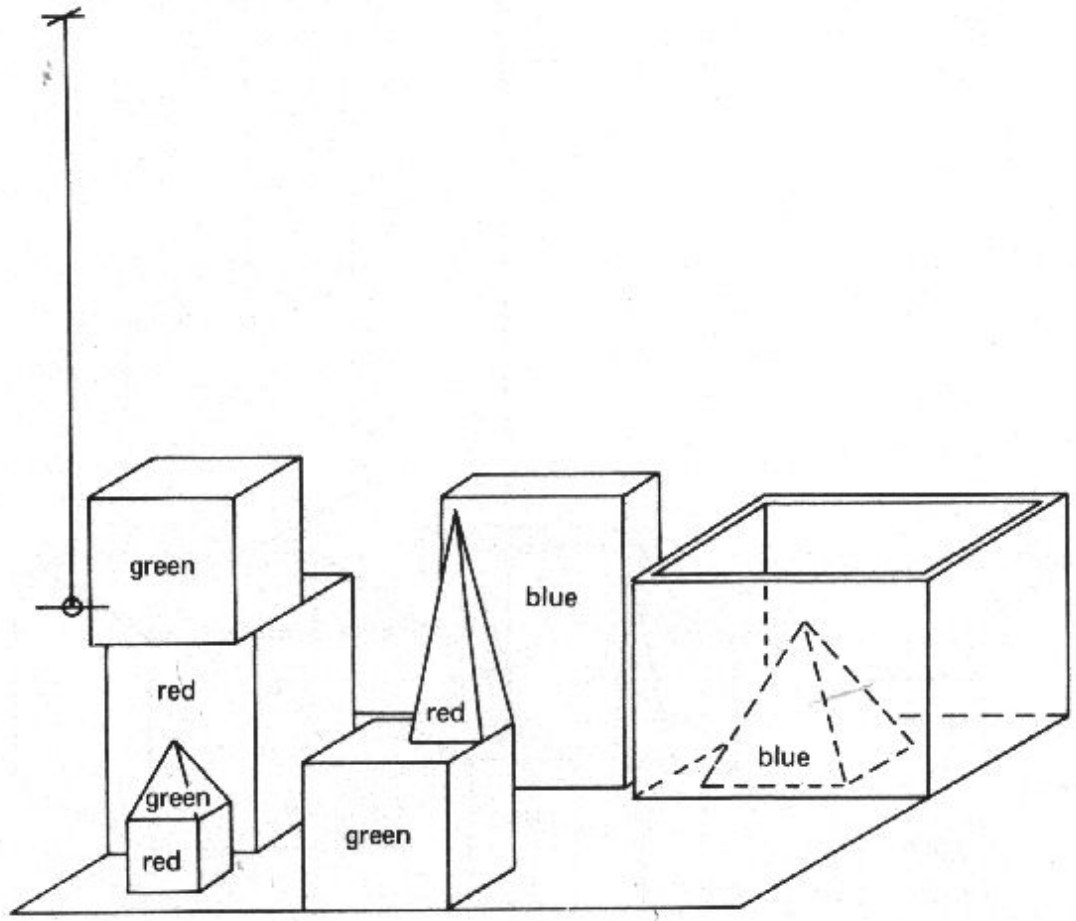
TEMPLATE: `_X1_`
RESPONSE: `You're being somewhat short with me.`

ELIZA: Ein sprachverstehendes System?

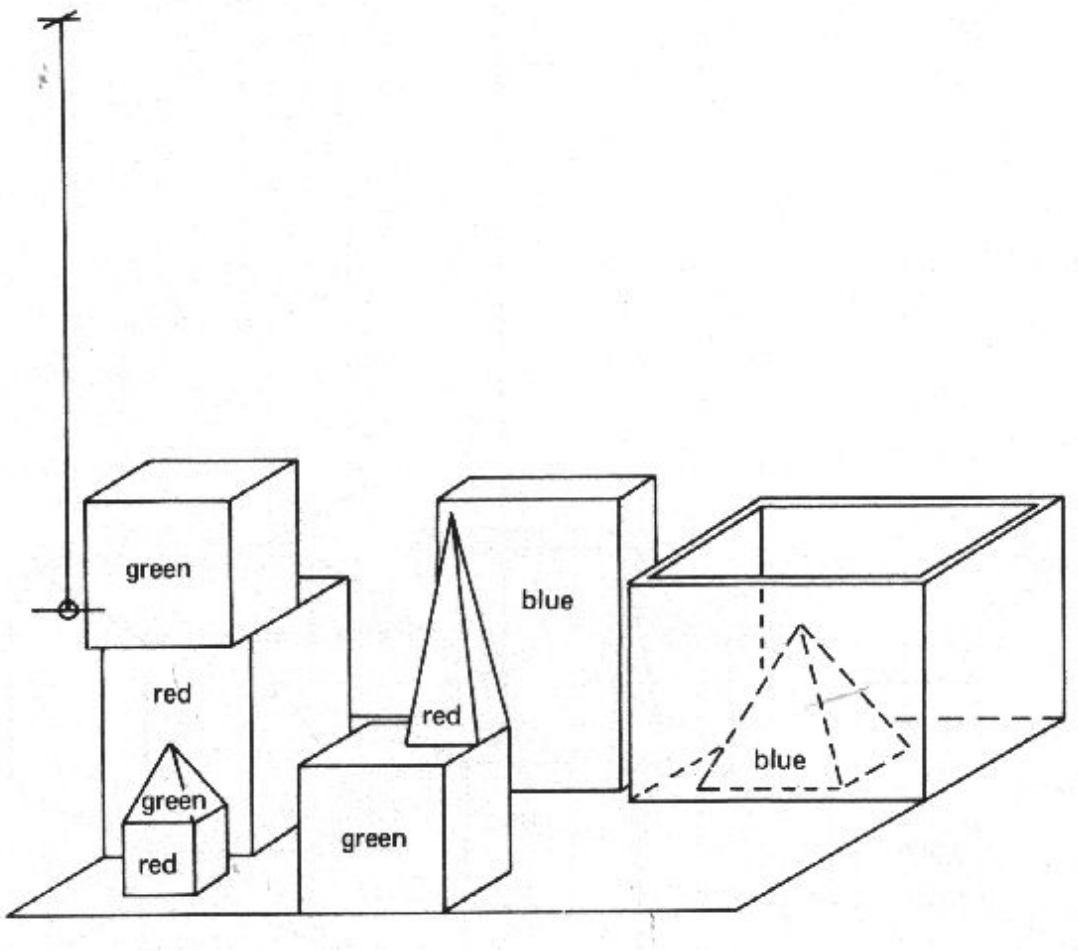
- ELIZA ist ein Dialogsystem, das beliebig komplexe Eingaben mit beliebigem Wortschatz zu beliebigen Themen akzeptiert.
- ELIZA arbeitet mit einfachen Mustervergleichs-Techniken (Pattern Matching), ohne Einsatz von Wissen:
  - Templates: Muster mit variablen Teilen, die mit der Benutzereingabe abgeglichen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

- ELIZA funktioniert besonders gut mit englischem Dialog und dem Psychotherapie-Szenario. Wieso?

SHRDLU: Ein wissensbasiertes Dialogsystem

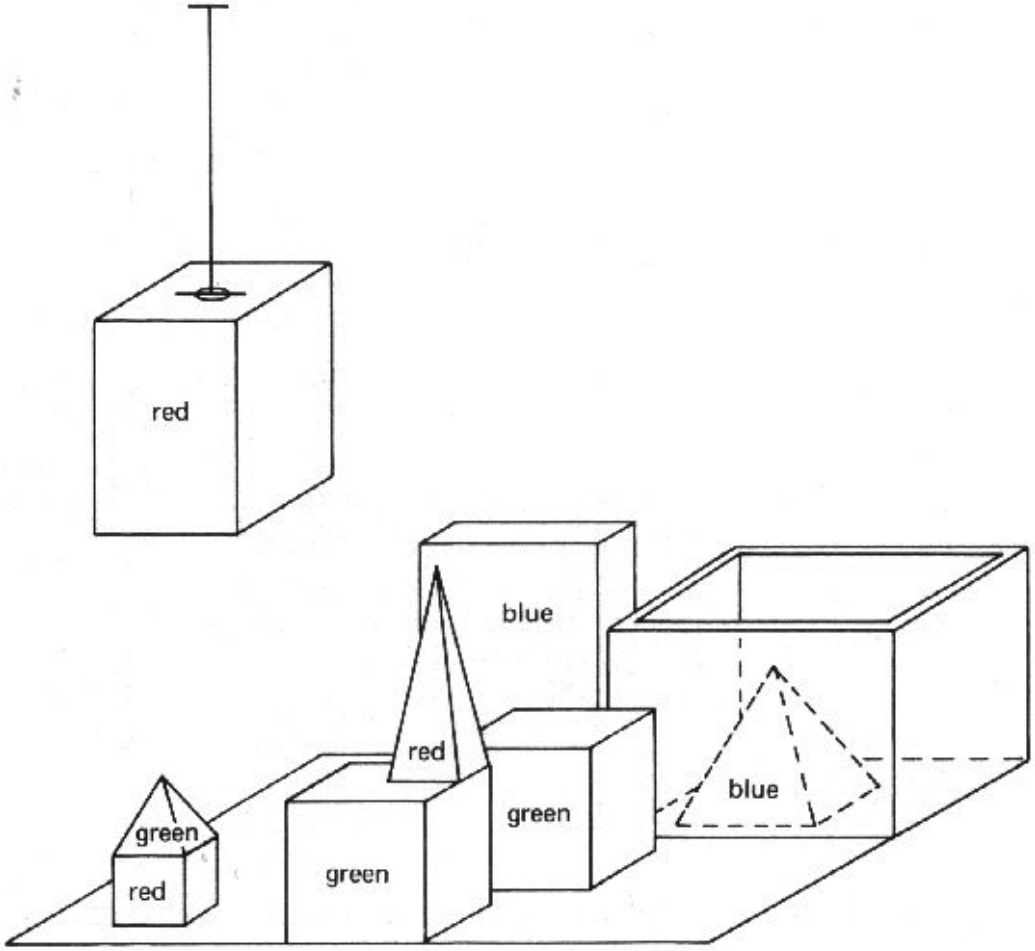


Winograds “Blocks World”

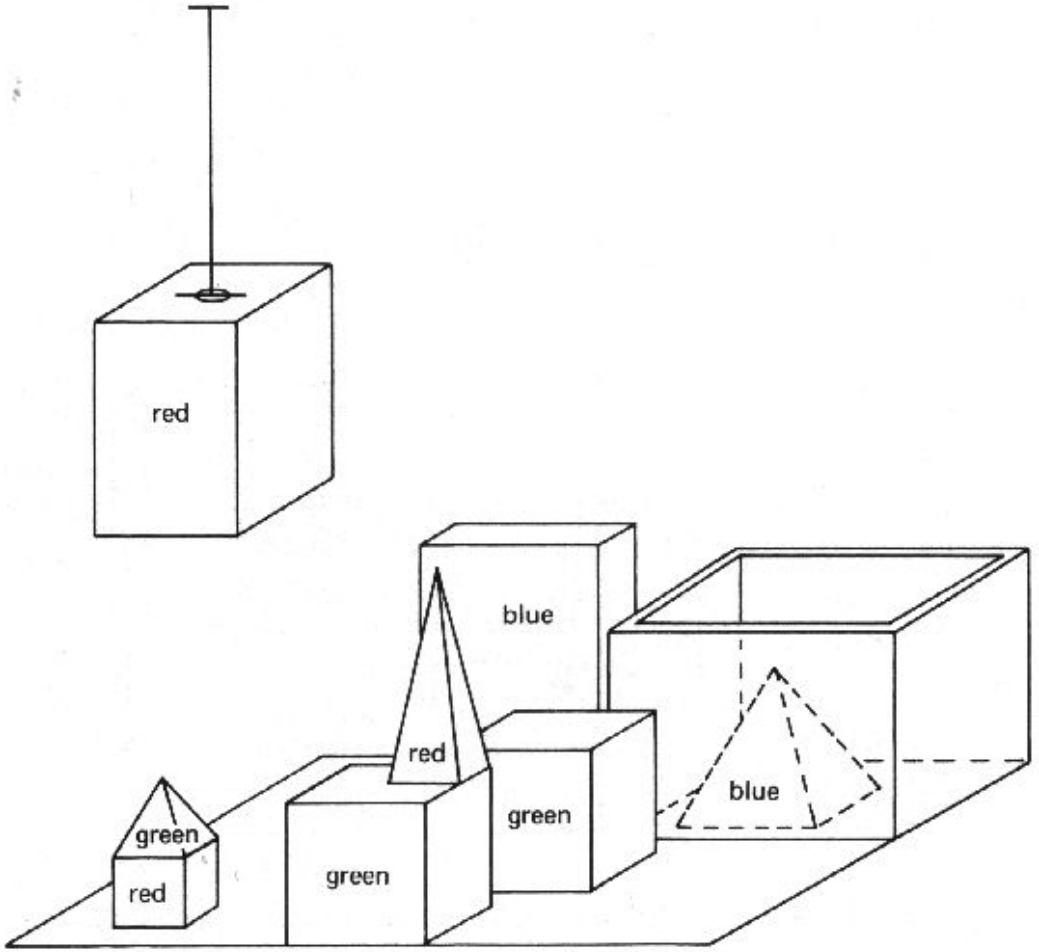


U: Pick up a big red block

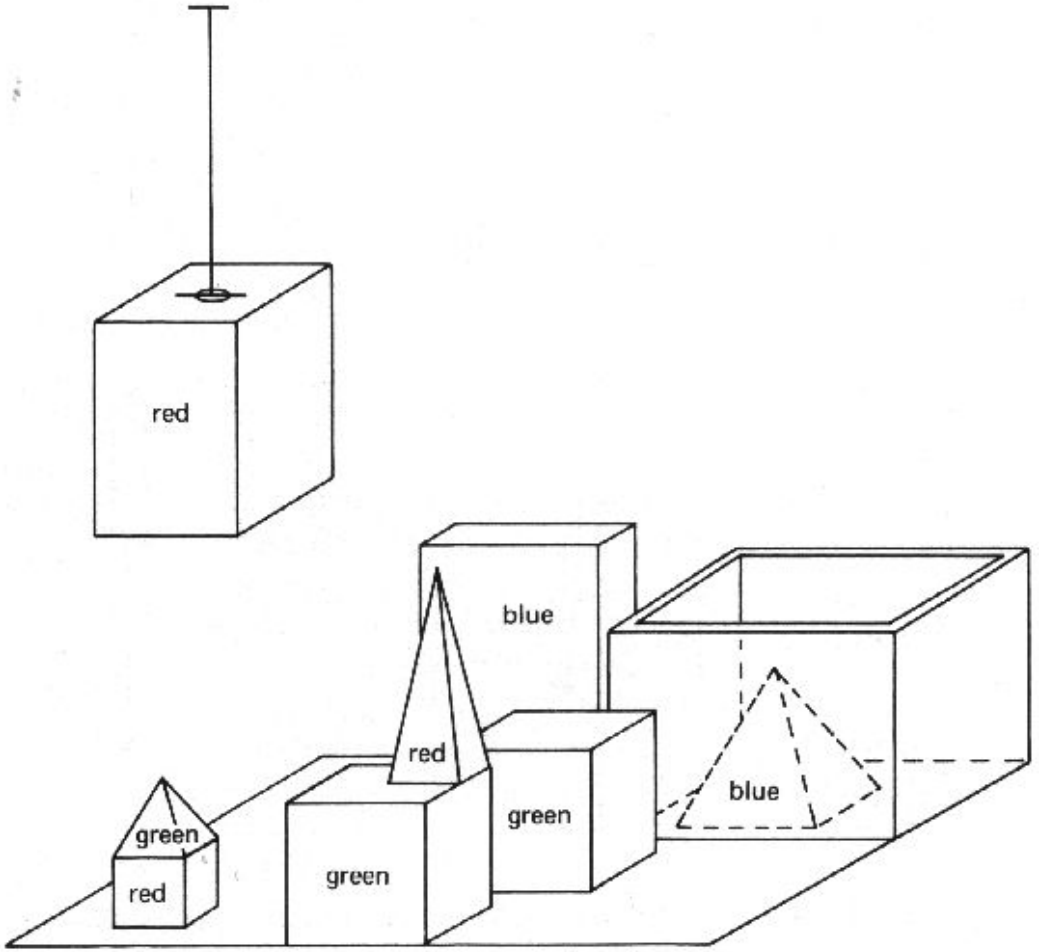
S: OK.



(Pick up a big red block)

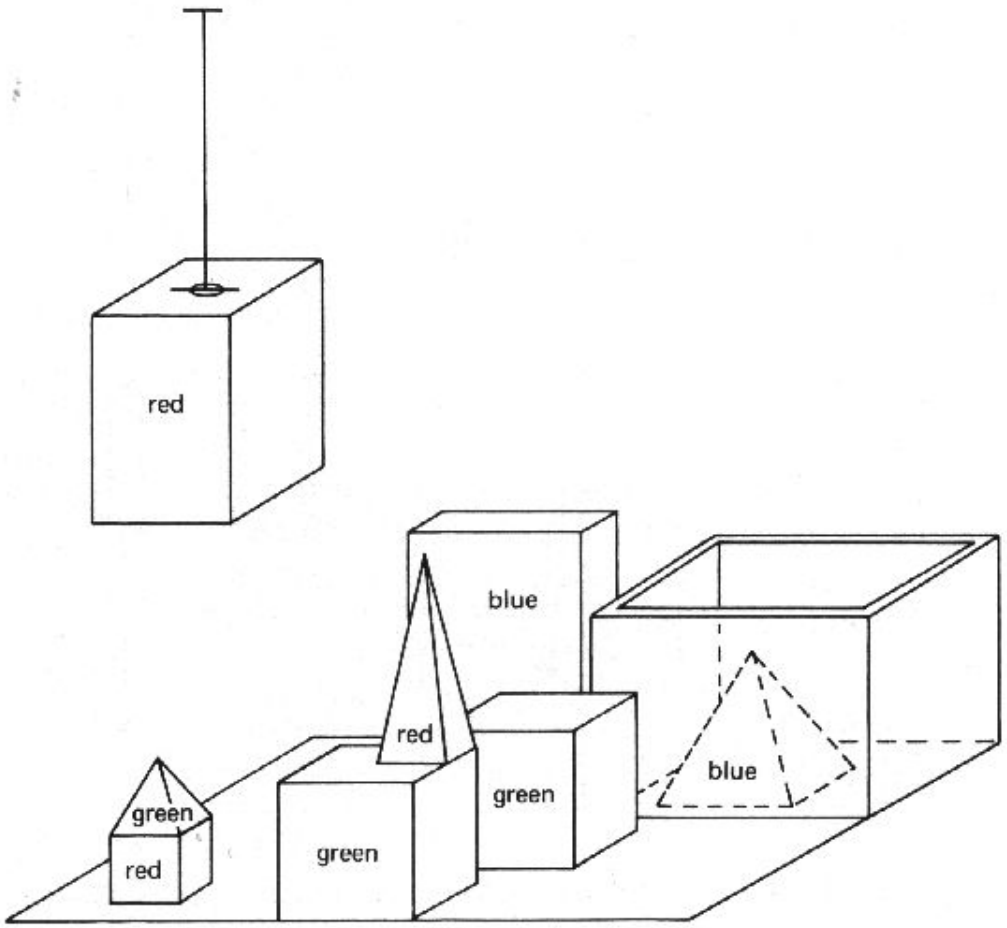


U: Grasp the pyramid.



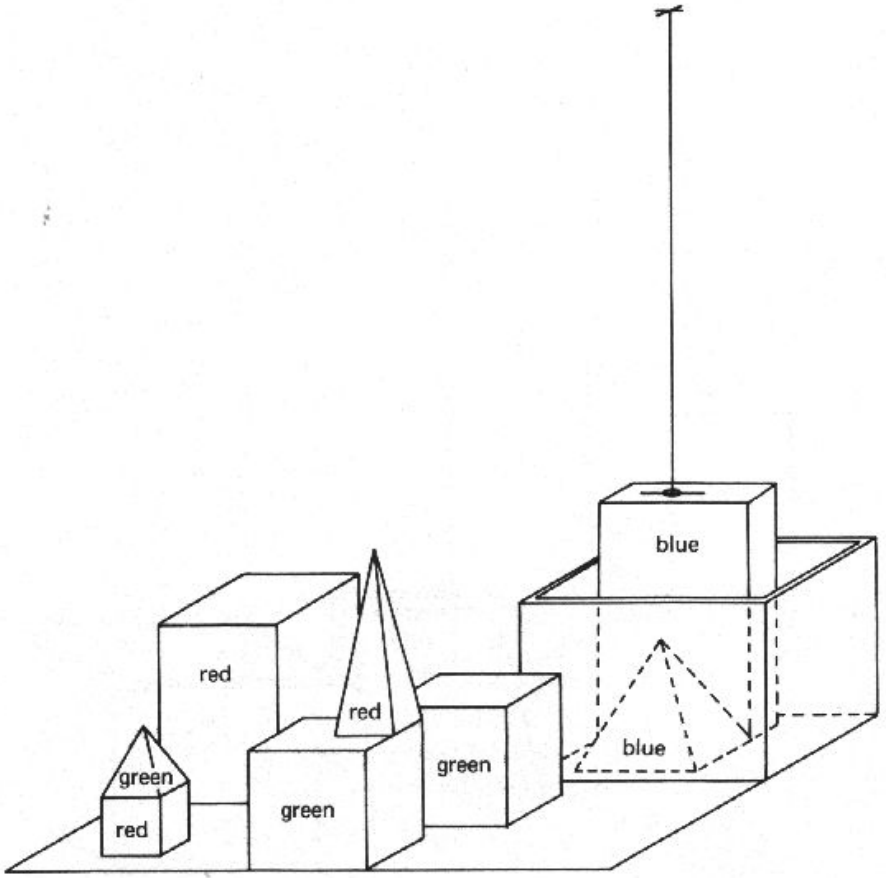
U: Grasp the pyramid.

S: I don't understand which pyramid you mean.

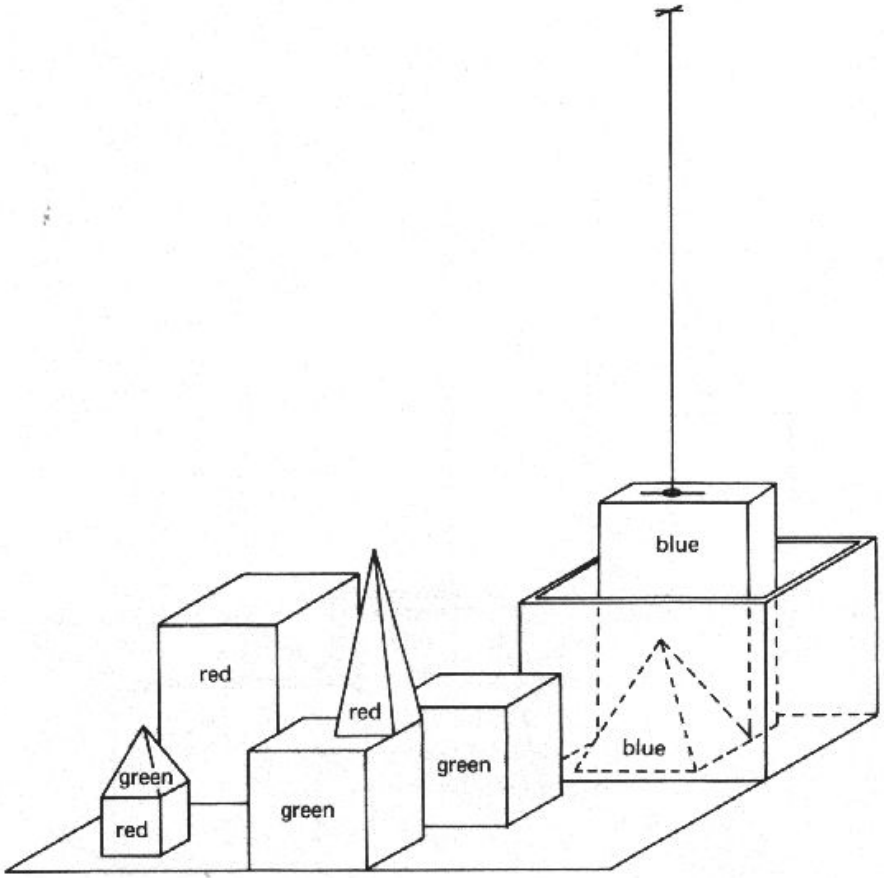


U: Find a block which is taller than the one you are holding and put it into the box.

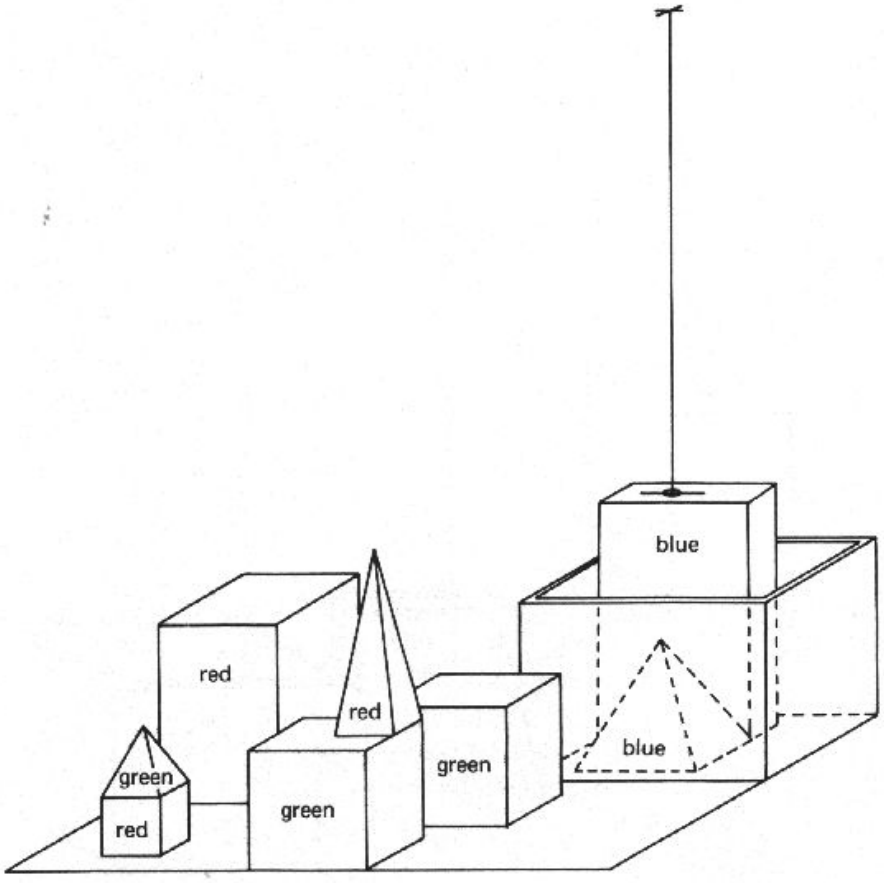
S: By „it“, I assume you mean the block which is taller than the one I am holding. OK



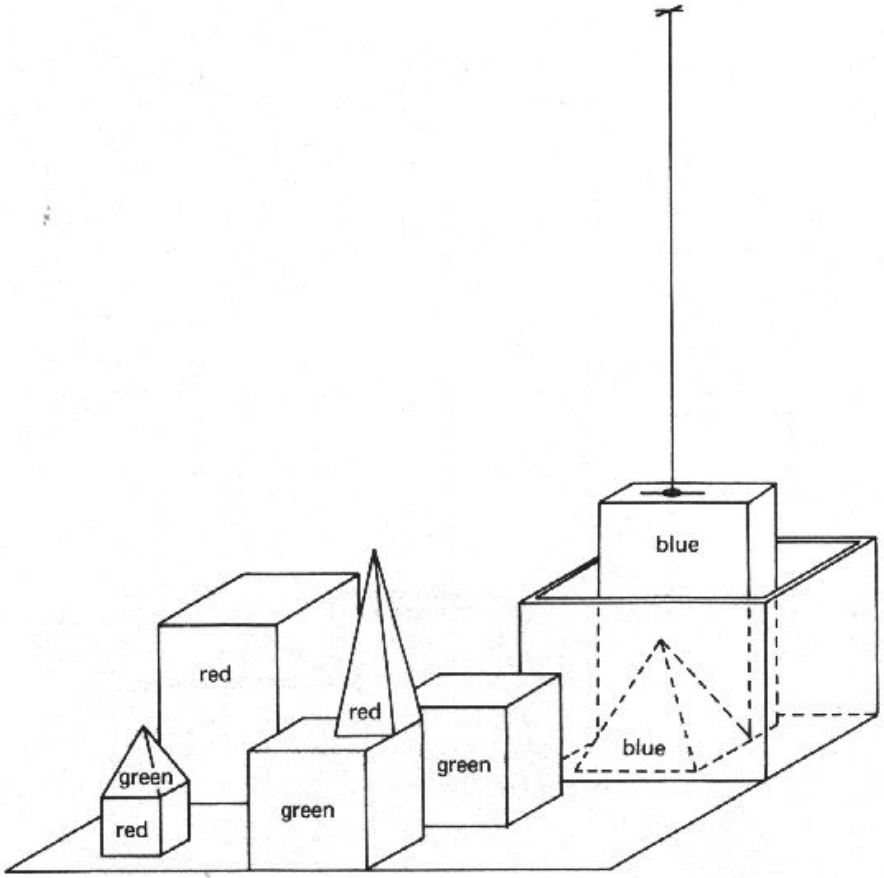
(U: Find a block which is taller than the one you are holding and put it into the box.)



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

SHRDLU beantwortet Fragen, führt Anweisungen aus und lernt Begriffe.

Wichtige Programmkomponenten von SHRDLU sind:

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

Winograds SHRDLU-System arbeitet in einer kleinen, eingeschränkten Mini-Welt oder -Domäne ("Blocks World").

Interessant ist die Interaktion von Analyse und Planung; die Generierungskomponente ist sehr einfach (patternbasiert); die Grafik ist computerlinguistisch nicht sehr interessant, hat aber zum durchschlagenden Erfolg des Systems beigetragen.

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

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 Objekte
Konstruktionen bezeichnet	*pyramid* ist Unterbegriff von *block*
Dinge, die gleichzeitig unter	*grasp* bezeichnet eine Handlung, ...
A und N fallen	

Sprachliches Wissen in SHRDLU: Beispiele

Grammatik	Lexikon
	Morphologisches Wissen:
regelmäßige Verben bilden Präteritum auf -ed	*grasp* ist regelmäßiges Verb *put* ist unregelm. Verb mit Prät. *put*
	Syntaktisches Wissen:
In Imperativen steht das Verb an erster Stelle	*grasp* ist transitives Verb *stop* ist intransitives Verb
	Semantisches Wissen:
A+N in attributiven Konstruktionen bezeichnet Dinge, die unter A und unter N fallen	*red* bezeichnet rote Objekte *pyramid* ist Unterbegriff von *block* *grasp* bezeichnet eine Handlung, ...

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.
- Es gibt keine scharfe Grenze zwischen grammatischer Information und 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 Regulariten

    "Zwei Objekte können nicht den gleichen Platz einnehmen."

    "Ein Objekt muss eine ebene Auflagefläche besitzen, damit ein zweites stabil darauf stehen kann"

Wozu wird Wissen eingesetzt?

Wissen wird in der – menschlichen und maschinellen – Sprachverarbeitung eingesetzt, um – linguistische und extralinguistische – Strukturen unterschiedlicher Arten und Ebenen aufeinander abzubilden:

- Speech →Text
- Text → Speech
- Wortkette → Bedeutungsinformation
- Bedeutungsinformation → Handlungsplan
- Bedeutungsinformation → Wortkette
- deutscher Satz → englischer Satz

Zentrale Probleme:

- Wo kommt das Wissen her?
- Wie beseitigen wir Mehrdeutigkeit?

Wissensakquisition

Zwei Optionen:

- Manuelle Entwicklung von grammatischen, lexikalischen und extralinguistischen Wissensbeständen

- Statistische Modellierung von Wissen durch maschinelle Lernverfahren