

Vorläufiges Programm

Einführung in die Computerlinguistik

WS 2009/2010

Manfred Pinkal

26.10.10	Einführung
2.11.10	Morphologie und Automaten
9.11.10	Morphologie und Automaten
16.11.10	Syntax
23.11.10	Parsing
30.11.10	Merkmale und Unifikation
7.12.10	Grammatische Verarbeitung
14.12.10	Semantik

Im neuen Jahr: Statistische Verfahren, Anwendungen

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

Technisches

Zur Vorlesung gehören:

- Das **Vorlesungsskript** (auf der Homepage des Kurses)
<http://www.coli.uni-saarland.de/courses/I2CL-10/>
- Ausgewählte **Kurztexte** in englischer und deutscher Sprache
- **Übungsaufgaben**: Sie werden (tendenziell wöchentlich) zur Vorlesung am Dienstag auf die Homepage gestellt, sind bis zum Montag der folgenden Woche einzureichen und werden in der darauf folgenden Übungssitzung besprochen.

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

Technisches

Scheine erwirbt man in folgender Weise:

- **Prüfungsvoraussetzung**: Schriftliche Bearbeitung der Übungsaufgaben, d.h.:
 1. Alle Aufgabenblätter (mit höchstens einer Ausnahme) müssen bearbeitet sein. Aufgabenblatt zählt als bearbeitet, wenn für alle Aufgaben ein ernsthafter Lösungsversuch vorliegt.
 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)
- **Anmeldung zur Prüfung bis zum 15.1.2010**
Wichtig: Ohne fristgerechte Meldung keine Teilnahme möglich!
- **Prüfungsleistung**: **Klausur** über den Stoff der Vorlesung, der im Vorlesungsskript, den Übungen und den Lektüretexten vorkommt. Klausurtermin: letzte Semesterwoche oder erste Woche der vorlesungsfreien Zeit (wird unter Berücksichtigung anderer Klausurtermine Anfang Januar festgelegt)

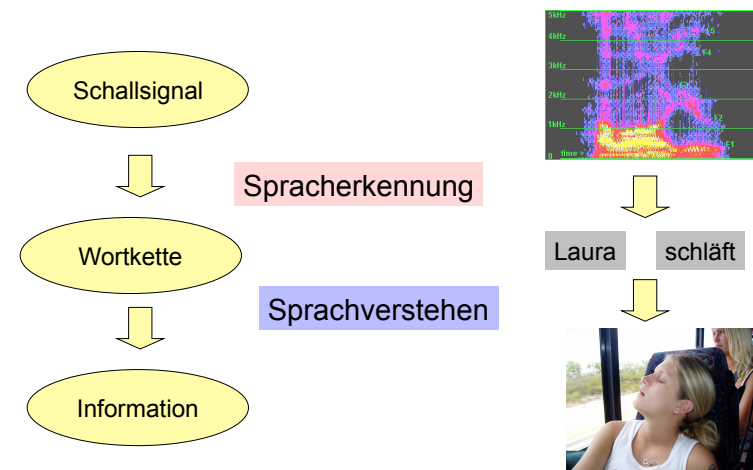
Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

Einführungsliteratur und andere Informationsquellen

- Eine ausgezeichnetes englisch-sprachiges Einführungswerk: Jurafsky, Daniel/ Martin, James H.: Speech and Language Processing, 2009
- Ein aktuelles deutsches Handbuch zur 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 (DE oder EN)

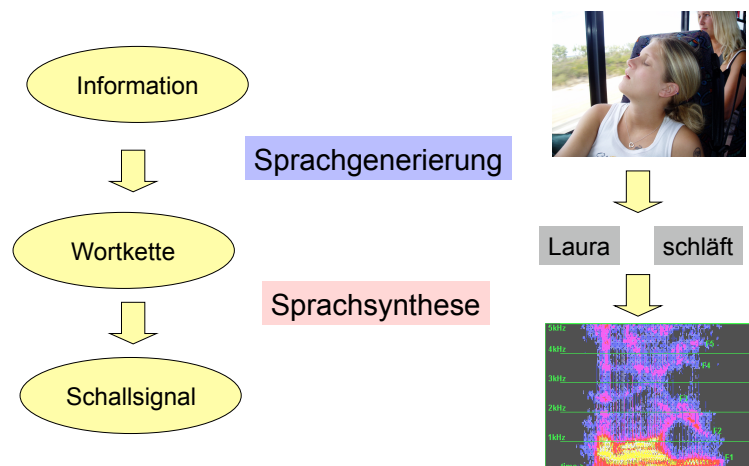
Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

Was ist Sprachverarbeitung?



Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

Was ist Sprachverarbeitung ?



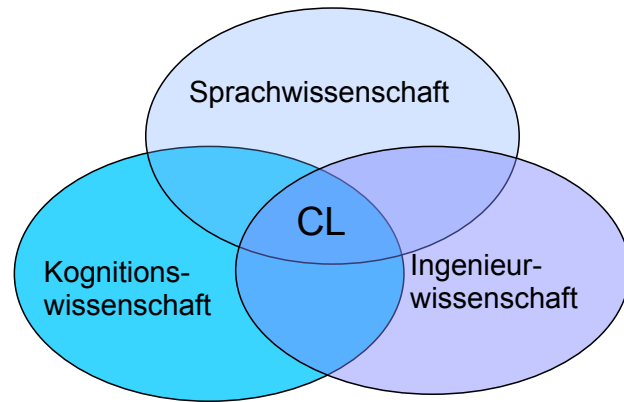
Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

Aufgaben 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 Diskurs
- Die Modellierung und Implementierung der komplexen Zusammenhänge und Abläufe bei:
 - Sprachverstehen
 - Sprachproduktion
 - Spracherwerb
- Die Entwicklung von **natürlich-sprachlichen Anwendungssystemen**.

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

Was ist Computerlinguistik?



Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS 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

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

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:

- kognitive Psychologie
- Neuropsychologie
- Künstliche Intelligenz

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

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

- Informatik/ Informationstechnologie
- Elektrotechnik/ Signalverarbeitung

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

Sprachtechnologie

Wichtige Teilbereiche der Sprachtechnologie:

- Informationsmanagement
- Gesprochene Sprache
- Multilinguale Anwendungen

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

Multilinguale Anwendungen

- Maschinelle Übersetzung
- Multilinguale Informationssuche
- Sprachlernsysteme

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

Informationszugriff und -management

- Information Retrieval
- Informations-Extraktion
- Question Answering (Frage-Antwort-Systeme)
- Summarisation (Automatische Zusammenfassung)
- Dokumentklassifikation

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

Anwendungen für gesprochene Sprache

- Diktiersysteme, Spracheingabe für medizinische Diagnose, technische Wartungssysteme
- Telefonie-Dialogsysteme: Call-Center, Telebanking, Fahrplanauskunft, ...
- Gerätebedienung: Sprachbedienung für Auto, mobile Geräte, Haushalt, Robotersteuerung, ...

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

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

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

ELIZA: Beispieldialog [1]

- | | |
|--|--|
| (1) U: I have been very depressed lately
S: <i>You say you have been very depressed lately.</i> | (4) U: My father calls me nasty names
S: <i>Who else in your family calls you nasty names</i> |
| (2) U: Yes, my mother beats me and my father calls me names
S: <i>You're being somewhat short with me</i> | (5) U: I think the dog would if he could
S: <i>Do you think its likely that he could</i> |
| (3) U: What about my mother beating me?
S: <i>Tell me more about your family</i> | (6) U: No, you do
S: <i>You don't seem very talkative today</i> |

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

ELIZA: Beispieldialog [2]

- | | |
|--|--|
| (7) U: I guess what is really bothering me is that I don't have any money
S: <i>Can you elaborate on that</i> | (9) U: I thought you might think so
S: <i>Perhaps in your fantasy we thought each other</i> |
| (8) U: The problem is that I know psychiatrists are expensive and I may not be able to pay you
S: <i>Please continue, this may be interesting</i> | |

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

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 it's likely that _X2_ ?
- TEMPLATE: _X1_
RESPONSE: You're being somewhat short with me.

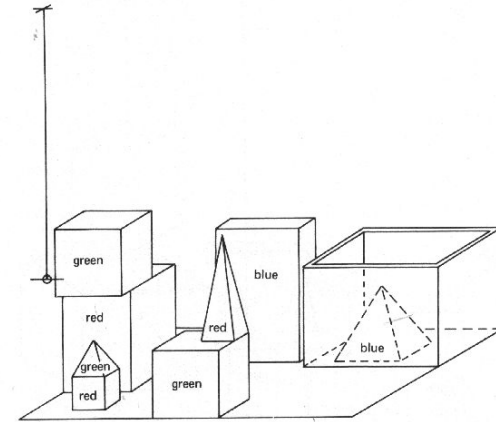
Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

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?

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

SHRDLU: Ein wissensbasiertes Dialogsystem



Winograds "Blocks World"

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

SHRDLU

SHRDLU beantwortet Fragen, führt Anweisungen aus und lernt Begriffe.

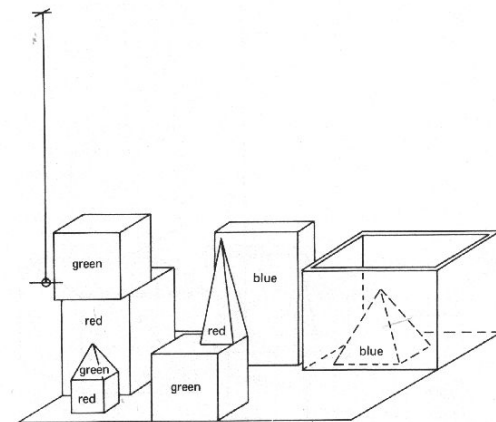
Wichtige Programmkomponenten von SHRDLU sind:

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

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

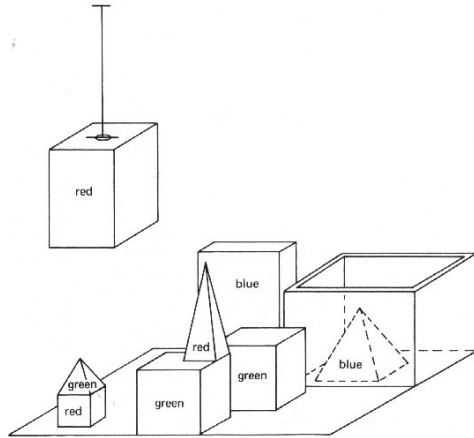
Interessant ist die Interaktion von Analyse und Planung; die Generierungskomponente ist sehr einfach (patternbasiert); die Grafik ist computerlinguistisch nicht sehr interessant, hat aber zum durchschlagenden Erfolg des Systems beigetragen.

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

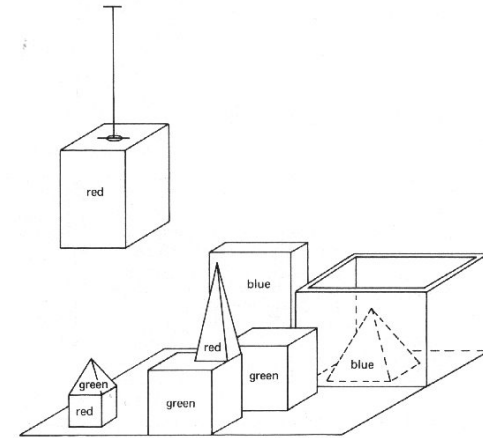


U: Pick up a big red block
S: OK.

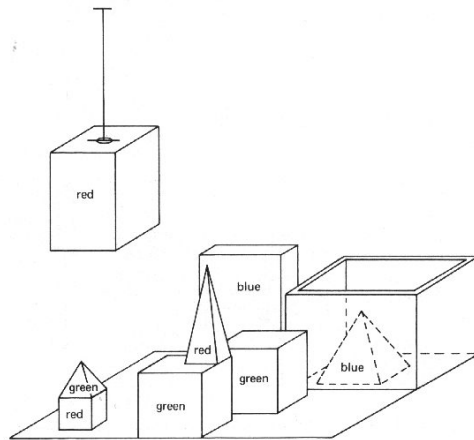
Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik



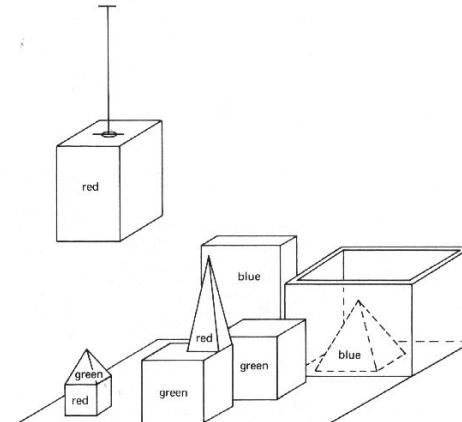
(Pick up a big red block)



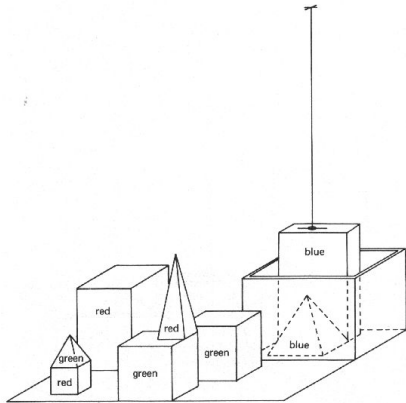
U: Grasp the pyramid.



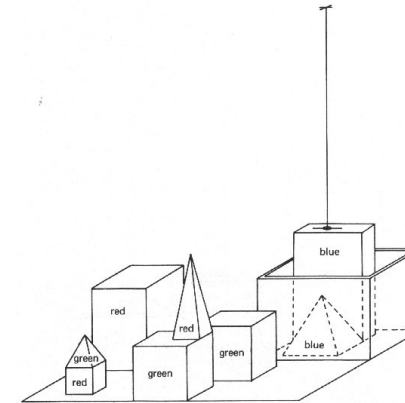
U: Grasp the pyramid.
 S: I don't understand which pyramid you mean.



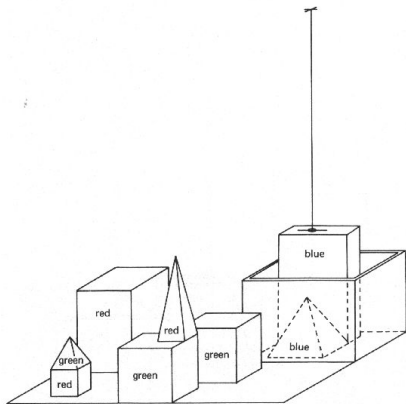
U: Find a block which is taller than the one you are holding and put it into the box.
 S: By „it“, I assume you mean the block which is taller than the one I am holding. OK



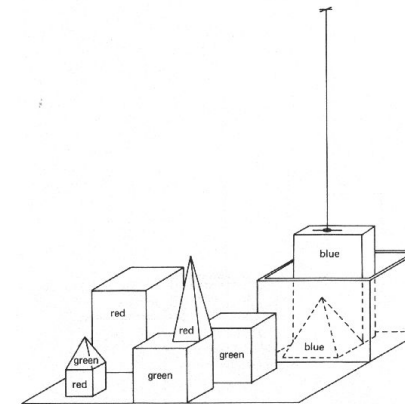
(U: Find a block which is taller than the one you are holding and put it into the box.)



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 [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 unter A und unter N fallen *grasp* bezeichnet eine Handlung, ...

Sprachliches Wissen in SHRDLU: Beispiele

Grammatik	Lexikon
	Morphologisches Wissen:
regelmäßige Verben bilden Präteritum auf -ed	<i>grasp</i> ist regelmäßiges Verb <i>put</i> ist unregelm. Verb mit Prät. <i>put</i>
	Syntaktisches Wissen:
In Imperativen steht das Verb an erster Stelle	<i>grasp</i> ist transitives Verb <i>stop</i> ist intransitives Verb
	Semantisches Wissen:
A+N in attributiven Konstruktionen bezeichnet Dinge, die unter A und unter N fallen	<i>red</i> bezeichnet rote Objekte <i>pyramid</i> ist Unterbegriff von <i>block</i> <i>grasp</i> 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.
- Achtung:
 - Es gibt keine scharfe Grenze zwischen systematischer grammatischer Information und wortspezifischer 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"*

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

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

Das zentrale Problem ist die **Mehrdeutigkeit (Ambiguität)** auf allen Ebenen: Wie kommen wir zu einer **eindeutigen Abbildung (Disambiguierung)**?

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik

Explizites und implizites Wissen

Zwei Optionen:

- **Manuelle Grammatik- und Lexikon-Entwicklung**, Erstellung von extralinguistischen Wissensbasen (Ontologien)
 - Verlässliche Information
 - Erlaubt die Modellierung komplexer struktureller Zusammenhänge
 - Sehr aufwändig, deshalb Abdeckungsprobleme
 - Wenig flexibel (z.B. in Bezug auf fehlerhafte Eingaben)
 - Große Probleme mit der Disambiguierung
- **Statistische Modellierung:**
 - Automatische Erkennung von wiederkehrenden Mustern in Sprachdaten
 - Vergleichsweise preiswert und effizient
 - Robuste Verfahren mit hoher Abdeckung
 - Nur approximativ korrekt, die Verlässlichkeit nimmt mit zunehmender Komplexität der linguistischen Strukturen ab

Vorlesung "Einführung in die CL" 2009/2010 © M. Pinkal UdS Computerlinguistik