LST Prep Course: General Overview

Manfred Pinkal

Unviersität des Saarlandes

09-10-2006

Course Schedule

	09.10.2006 Monday	10.10.2006 Tuesday	11.10.2006 Wednesday	12.10.2006 Thursday	13.10.2006 Friday
09.15 – 10.45	Introductions M. Pinkal	Syntax M. Pinkal	Semantics M. Pinkal	Pragmatics M. Pinkal	Psycholinguistics P. Knoeferle
11.15 – 12.45	General Overview M. Pinkal	Grammar Formalism B.Crysmann	Semantic Formalisms M. Pinkal	Text and Dialog Structure M. Wolska	Wrap Up and Conclusion M. Pinkal
15.15 – 16.45	Phonetics and Phonology J. Trouvain	Student Papers	Demonstration of LT Systems	Student Papers	
19.00 – 21.00	Party C7 2 - Foyer				

Textbooks

- Jurafsky, Daniel and Martin James H.: *Speech and Natural Language Processing.* Prentice Hall.
- Manning, Christopher D. and Schütze, Hinrich: *Foundations of Statistical Natural Language Processing.* MIT Press.
- Fromkin, Victoria and Rodman, Robert: *An Introduction to Language*. Harcourt Brace.
- Akmajian, Adrian et al.: *An Introduction to Language and Communication.* MIT Press.

Also recommended

- Crystal, David: *The Cambridge Encyclopedia of the English Language.* Cambridge University Press.

Textbooks

UNIVERSITÄT DES SAARLANDES

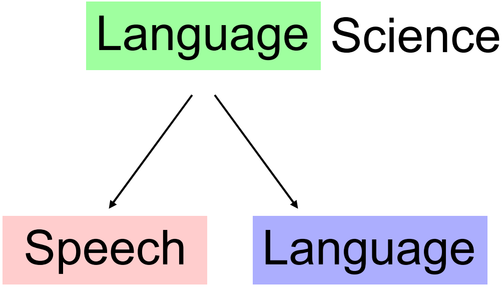
- Objectives: The development of Language Technology software applications:
  - Information Management Applications
  - Multilingual Applications
  - Speech-based Applications
- Interdisciplinary Collaboration with:
  - Computer Science
  - Information Science
  - Electrical Engineering/ Signal Processing

Language Science and Technology

UNIVERSITÄT DES SAARLANDES

Language Science and Technology

Language Science and Technology

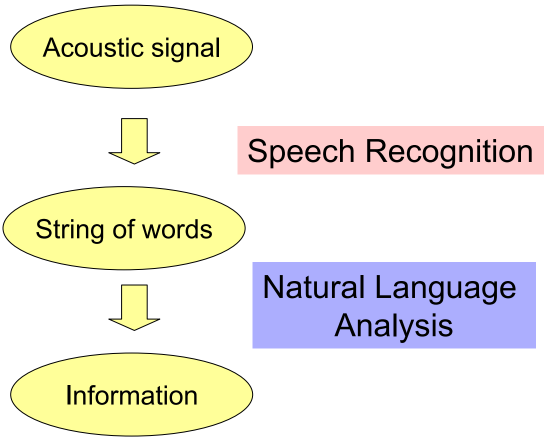


Language Science and Technology

Speech

Language

Speech and Language Processing



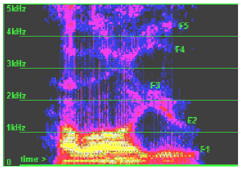
Acoustic signal

Speech Recognition

String of words

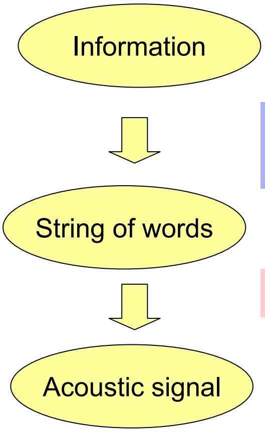
Natural Language Analysis

Information



Laura is sleeping

Speech and Language Processing


Information

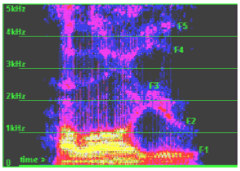
String of words

Acoustic signal

Natural Language Generation

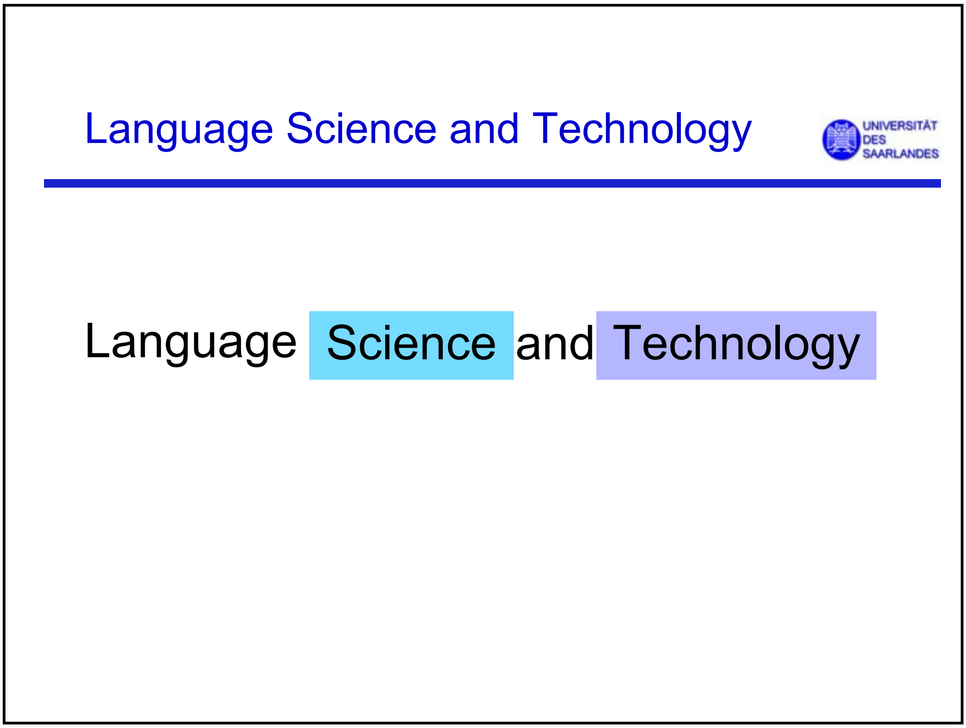
Speech Synthesis

Laura is sleeping



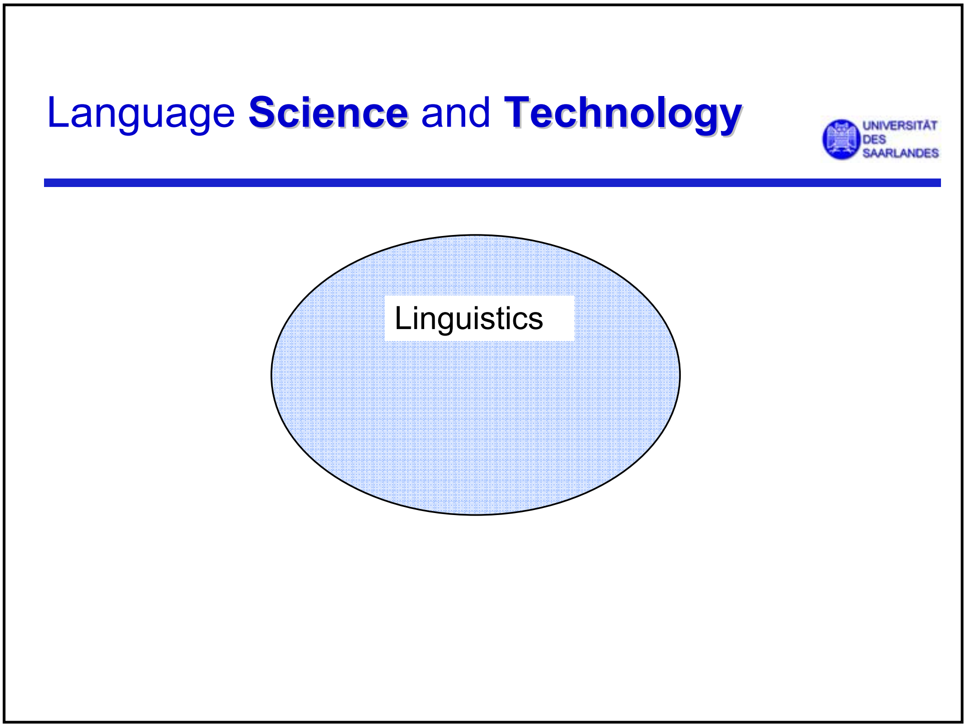
Language Science and Technology

Language Science and Technology



Language Science and Technology

Language Science and Technology

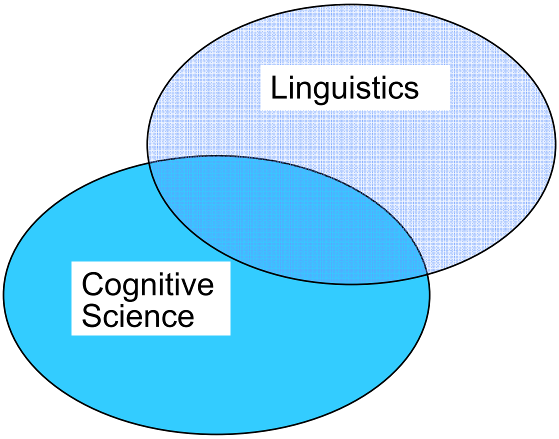


Language **Science** and **Technology**

The Linguistic Aspect of LST

- Objectives: The development of formalisms, theories, and software tools for the representation, processing, and acquisition of linguistic information of the different layers of linguistic structure:
  - Phonetics & Phonology
  - Morphology & Syntax
  - Semantics
  - Pragmatics , Text & Discourse Structure
- Interdisciplinary collaboration with:
  - Theoretical Linguistics
  - Phonetics
  - Chinese, English, German, Japanese, Romance … Linguistics

Language **Science** and **Technology**



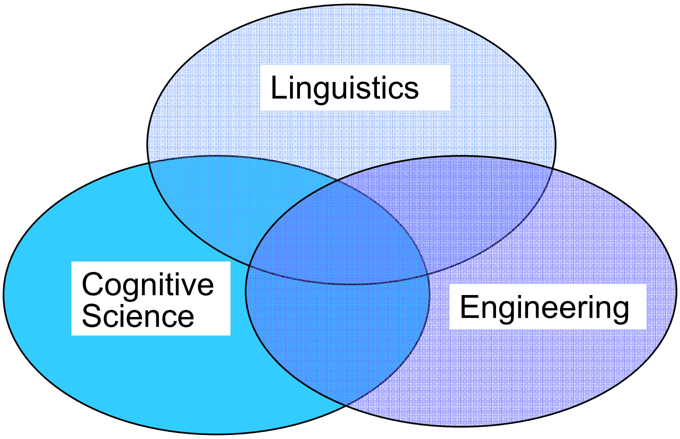
The Cognitive Aspect of LST

UNIVERSITÄT DES SAARLANDES

- Objectives: Modelling, implementation and empirical validation of
  - Language Comprehension
  - Language Production
  - Language Acquisition
- Interdisciplinary Collaboration with:
  - Psycholinguistics
  - Cognitive and Neuropsychology
  - Artificial Intelligence
  - Philosophy of Mind

Language **Science** and **Technology**

UNIVERSITÄT DES SAARLANDES



The Engineering Aspect of LST

- Objectives: The development of Language Technology software applications:
  - Information Management Applications
  - Multilingual Applications
  - Speech-based Applications
- Interdisciplinary Collaboration with:
  - Computer Science
  - Information Science
  - Electrical Engineering/ Signal Processing

Information Search and Management

- Information Retrieval
- Information Extraction
- Question Answering
- Summarisation
- Document Classification

Multilingual Applications

- Machine Translation
- Multi-lingual Information Search
- Tutoring systems for language learning

Speech and Dialogue Applications

- Text-to-Speech Systems (TTS)
- Dictation Systems
- Dialogue Systems
  - Phone-based Applications: Information, Call-Center, Telebanking Systems
  - Dialogue interfaces for device control: Human-computer, In-car,Intelligent Home, Robots

The Early History of CL

- Around 1960: Russian-English machine translation, politically motivated.

  The pen is in the box.

  The box is in the pen.

The Early History of CL

- Around 1960: Russian-English machine translation, politically motivated.
- Around 1970: First dialogue systems. The most famous ones are:
  - ELIZA (1966, Joseph Weizenbaum)
  - SHRDLU (1970, Terry Winograd)

ELIZA: Example Dialogue [1]

UNIVERSITÄT DES SAARLANDES

(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: Example Dialogue [2]

UNIVERSITÄT DES SAARLANDES

(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*

Inside ELIZA

TEMPLATE: You _*X1*_
RESPONSE: You say you _*X1*_

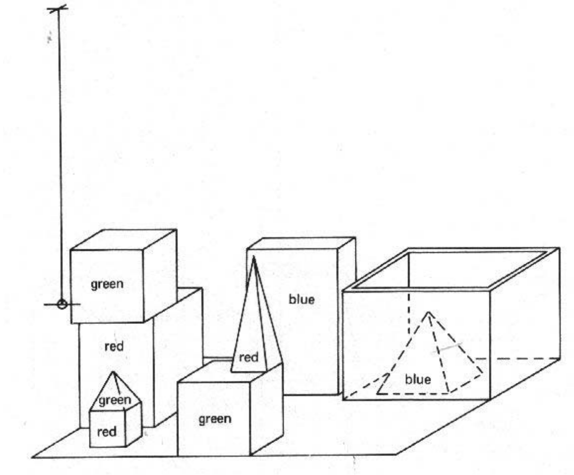
TEMPLATE: _*X1*_ your _*X2*_ *(category family)* _*X3*_
RESPONSE: Who else in your family _*X3*_ ?

TEMPLATE: _*X1*_ if _*X2*_
RESPONSE: Do you think ist likely that _*X2*_ ?

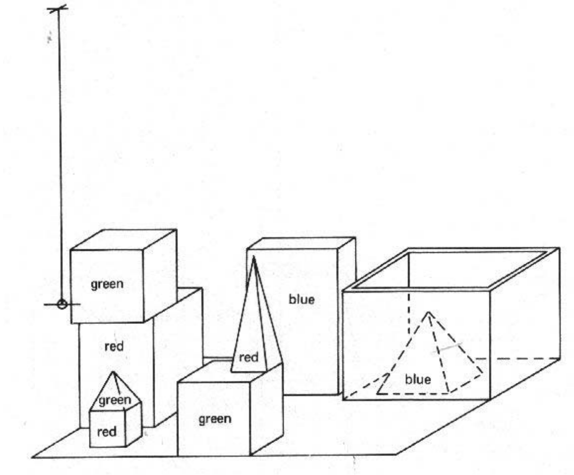
TEMPLATE: _*X1*_
RESPONSE: You‘re being somewhat short with me.

The Early History of CL

- Around 1960: Russian-English machine translation, politically motivated.
- Around 1970: First dialogue systems. The most famous ones are:
  - ELIZA (1966, Joseph Weizenbaum)
  - SHRDLU (1970, Terry Winograd)

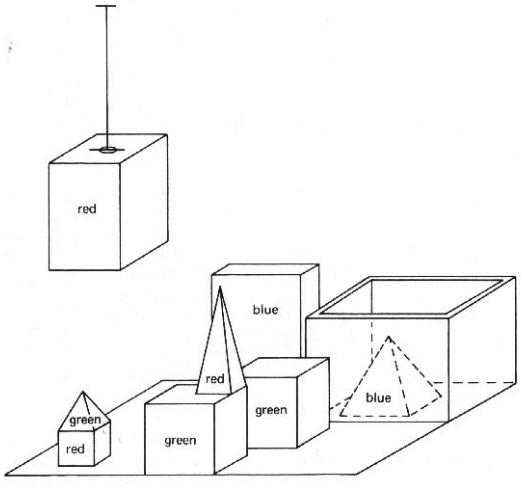


SHRDLU: The “Blocks World”

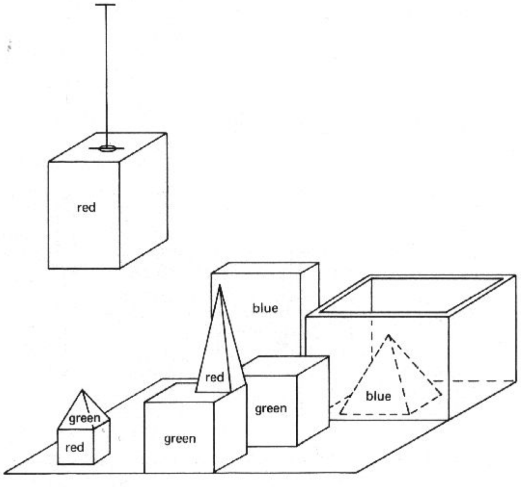


U: Pick up a big red block

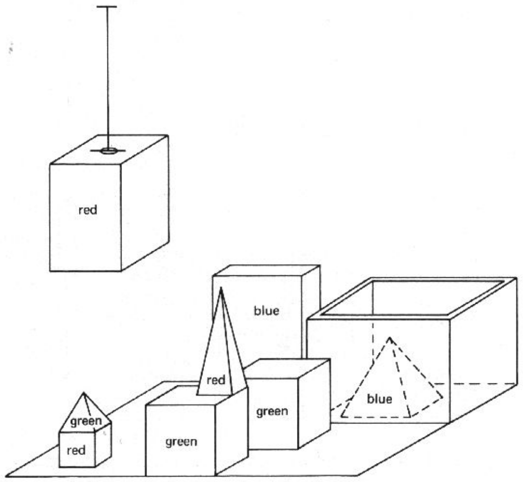
S: OK.



(Pick up a big red block)

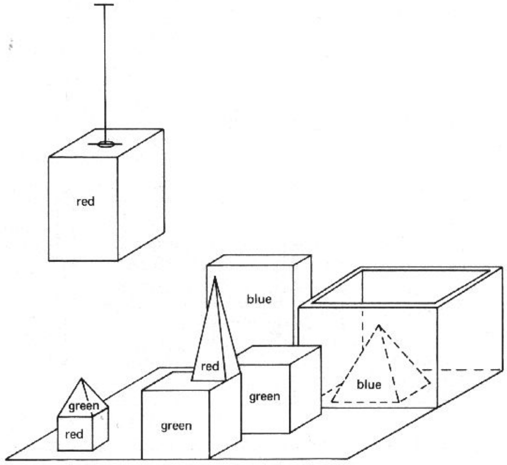


U: Grasp the pyramid.



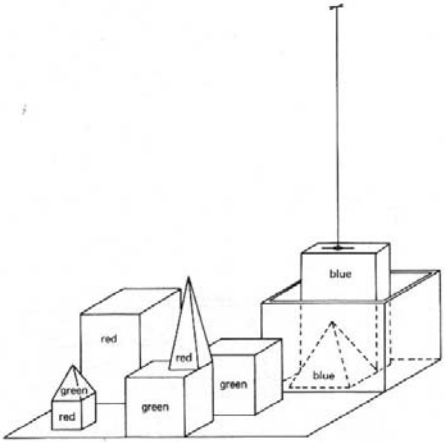
U: Grasp the pyramid.

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

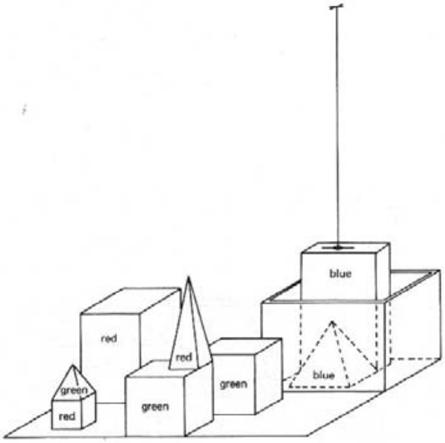


U: Find a block which is taller than the one you are holding and put it into the box.

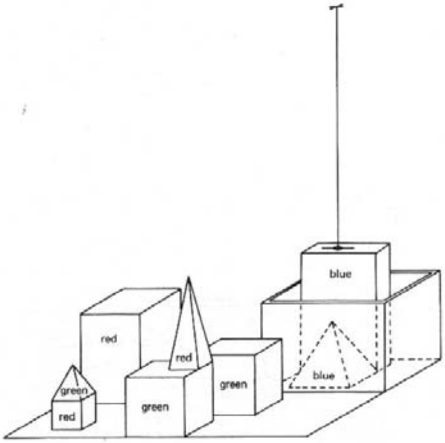
S: By „it“, I assume you mean the block which is taller than the one I am holding. OK



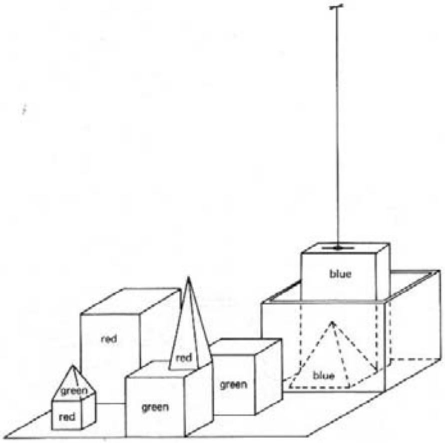
(U: Find a block which is taller than the one you are holding and put it into the box.)



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 [2]

- SHRDLU is a knowledge-based NLP system (in contrast to ELIZA).
- It uses
  - Linguistic Knowledge
  - Context Knowledge
  - World Knowledge

Linguistic Knowledge in SHRDLU

Morphological Knowledge:

regular verbs form past tense with -ed	grasp is a regular verb *put* is irregular verb with past *put*

Syntactic knowledge:

In imperative sentences, the verb is in first position	*grasp* is transitive verb *stop* is intransitive verb

Semantic knowledge:

A+N in attributive constructions denotes objects that are A and B at the same time	*red* denotes red objects *(???)* *pyramid ...* grasp ...

Linguistic Knowledge in SHRDLU

Morphological Knowledge:

regular verbs form past tense with -ed	grasp is a regular verb *put* is irregular verb with past *put*

Syntactic knowledge:

In imperative sentences, the verb is in first position	*grasp* is transitive verb *stop* is intransitive verb

Semantic knowledge:

A+N in attributive constructions denotes objects that are A and B, at the same time	*red* denotes red objects *(???)* *pyramid ...* grasp ...

Grammatical and lexical knowledge

- Grammatical knowledge is about phonological, morphological, syntactic, and semantic regularities of the language.
- Lexical knowledge comprises special morphological, syntactic, and semantic information about single words.
- Note:
  - There is no clear boundary between systematic grammatical and ideosyncratic lexical knowledge.
  - Different grammar theories draw the boundary between grammar and lexicon in different ways.

Extra-linguistic Knowledge

- Context knowledge
  - Linguistic context: Which is the most recently mentioned object? (*Put it into the box.*)
  - Utterance situation: Which objects occur in the visual scene? (*What does the block in the box support?*)
- World knowledge
  - Episodic knowledge

    There are two red blocks

    The box contains one pyramid
  - Rule knowledge

    Two objects cannot occupy the same space

    You can position things only onto objects with a planar top

How do we get at the knowledge?

- Development of grammars, lexica, extra-linguistic databases (ontologies) by hand
  - Reliable
  - Appropriate to model complex structure, but
  - lack of coverage and flexibility
- Automatic extraction of information from corpora with statistical / machine learning techniques
  - supports high coverage, robust processing
  - only approximatively correct, decreasing reliability with increasing complexity of linguisti structure

Knowledge in NLP

- Linguistic knowledge is only implicitly contained in statistical models, that relate, e.g.,
  - text words to parts of speech (POS-Taggers)
  - sentences of a source language to sentences of a target language (statistical Machine Translation)
- There is a trend towards hybrid NLP systems: Systems combining knowledge-based and statistical, data-intensive methods.

Deep and shallow techniques in Language Technology

- The central question in traditional NLP: What kind of knowledge do we need to achieve general, full, and reliable understanding of language?
- A practically more helpful question: What can we achieve with certain kinds and amounts of knowledge?