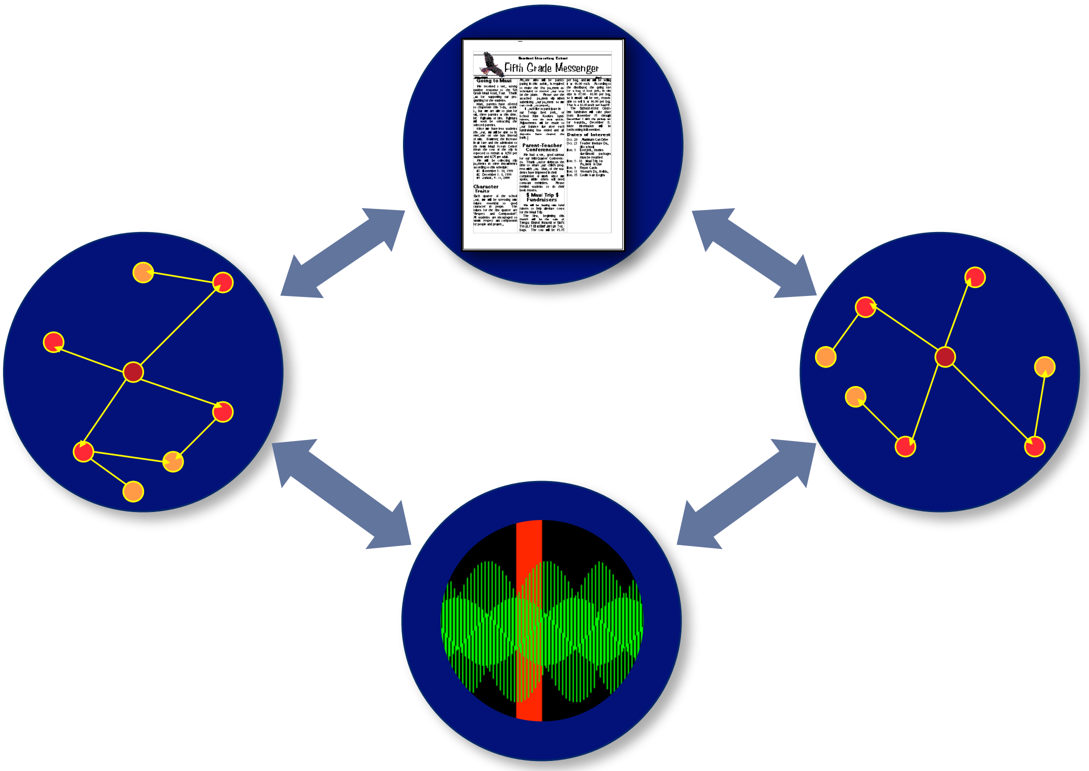
Foundations of Language Science and Technology

Introduction

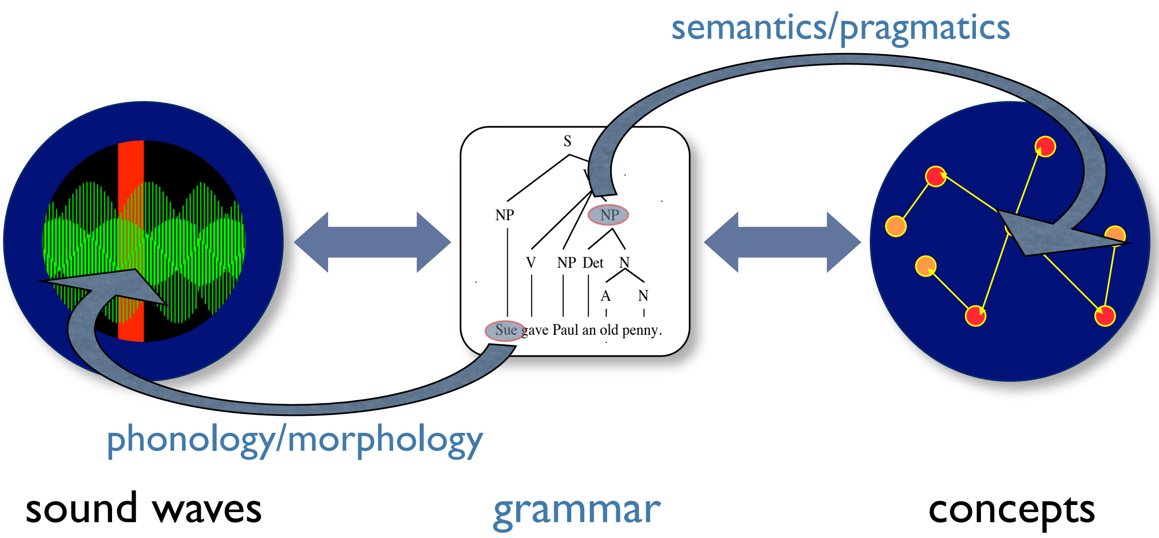
Alexander Koller
October 24, 2008

based in part on slides by Hans Uszkoreit

Language is the Medium



What happens in between?



semantics/pragmatics

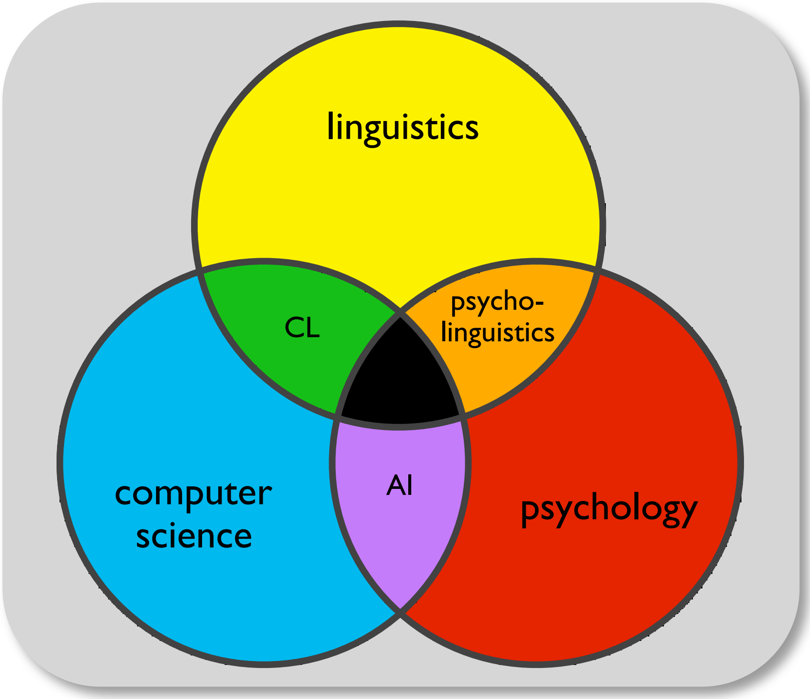
phonology/morphology

sound waves

grammar

concepts

Interdisciplinary Landscape



linguistics

CL

psycho-linguistics

computer science

AI

psychology

Uszkoreit’s Island Ambiguity

„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit
in the past produced the women of the islands on the weekends scarves with

Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem
floral patterns that their husbands on the following Mondays on the

Markt im Zentrum der Hauptinsel verkauften.“
market in the center of the main island sold.

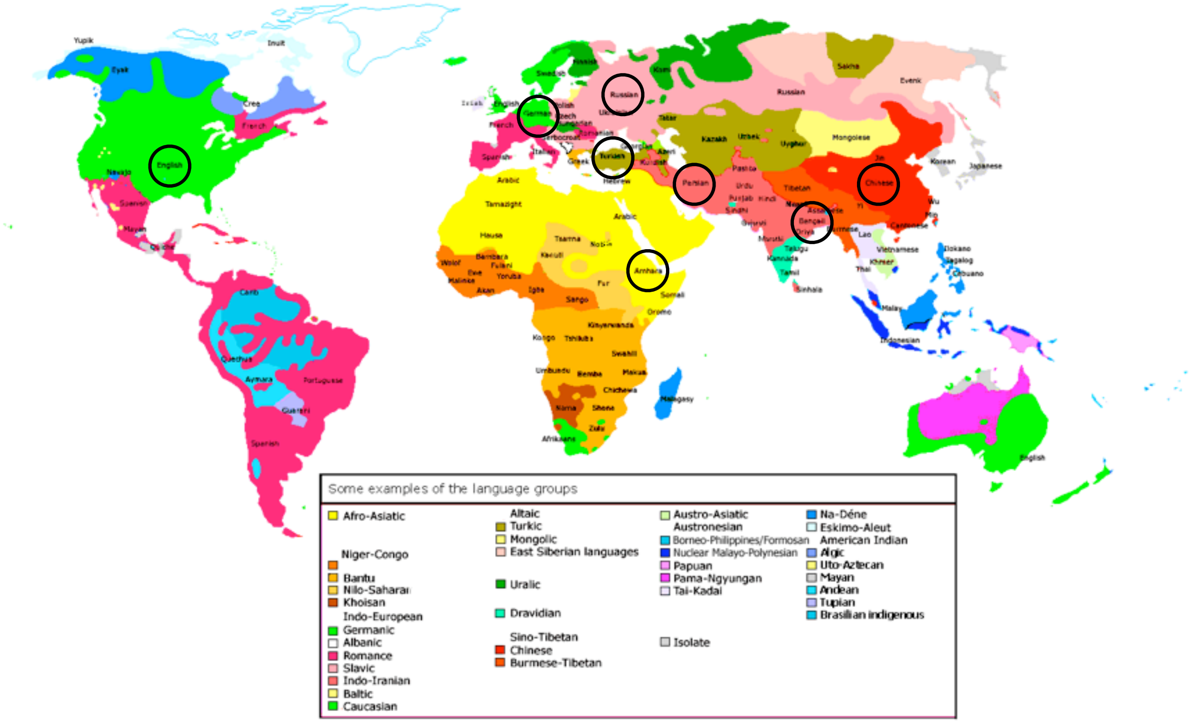
(Hans Uszkoreit)

The sentence exhibits a total of 13 lexical, syntactic, and referential ambiguities.

2 x 2 x 2 x 3 x 3 x 2 x 4 x 2 x 4 x 2 x 2 x 7 x 2
= **258,048 readings**

Your Turn!

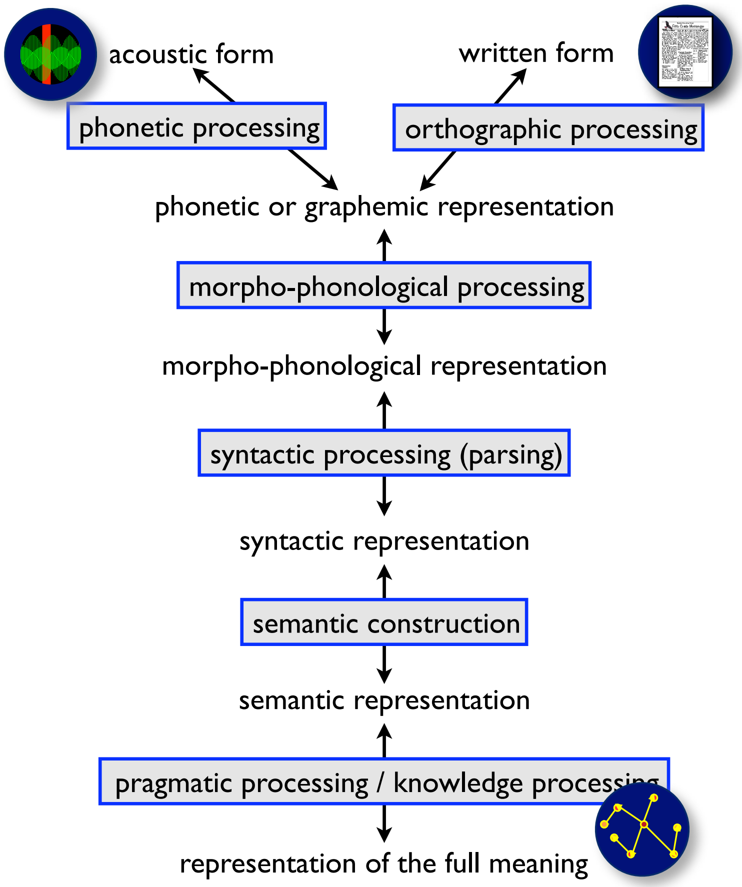
Your languages



Language Technology

- Machine translation
- Question answering
- Information extraction & retrieval
- Dialogue systems
- Generation systems

Levels of Processing



acoustic form

written form

phonetic processing

orthographic processing

phonetic or graphemic representation

morpho-phonological processing

morpho-phonological representation

syntactic processing (parsing)

syntactic representation

semantic construction

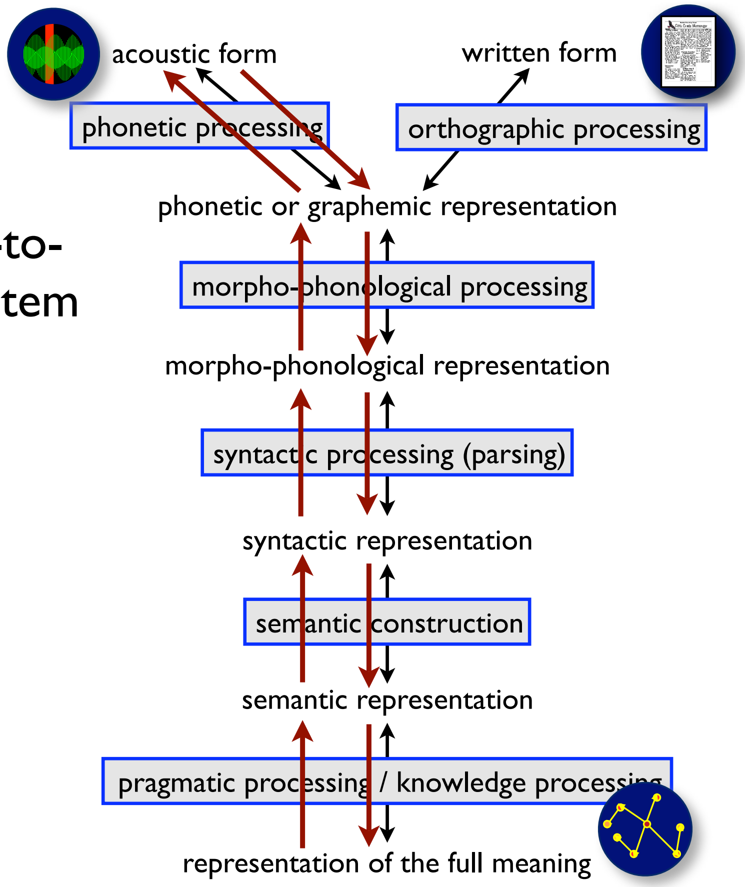
semantic representation

pragmatic processing / knowledge processing

representation of the full meaning

Levels of Processing

... in a speech-to-speech MT system



acoustic form

written form

phonetic processing

orthographic processing

phonetic or graphemic representation

morpho-phonological processing

morpho-phonological representation

syntactic processing (parsing)

syntactic representation

semantic construction

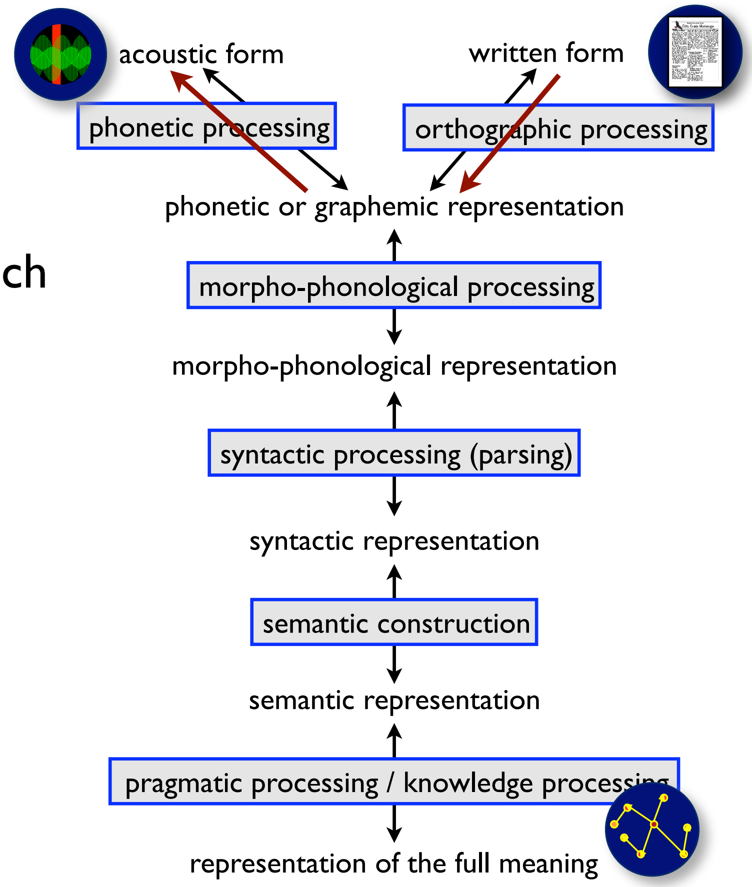
semantic representation

pragmatic processing / knowledge processing

representation of the full meaning

Levels of Processing

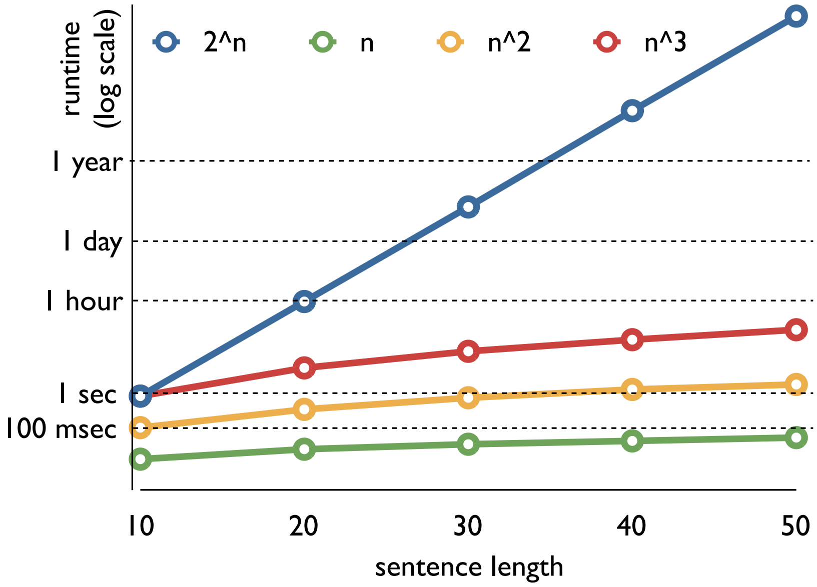
... in a text-to-speech system


acoustic form
written form
phonetic processing
orthographic processing
phonetic or graphemic representation
morpho-phonological processing
morpho-phonological representation
syntactic processing (parsing)
syntactic representation
semantic construction
semantic representation
pragmatic processing / knowledge processing
representation of the full meaning

Combinatorial Explosions

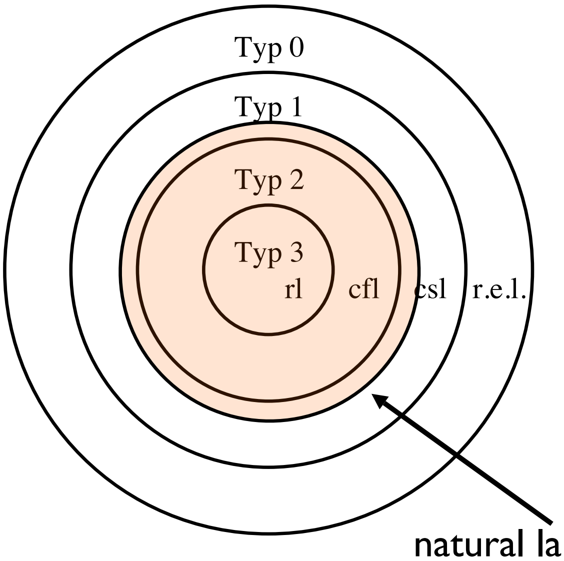
- Let’s say a sentence has n ambiguities with two readings each that can be combined freely.
- Total number of readings: 2^n
- Combinatorial explosion = extremely fast growth of number of readings with number of ambiguities.

A thought experiment



(Assumption: One parse per millisecond.)

Complexity of natural language



Chomsky Hierarchy:

type 0: recursively enumerable
type 1: context-sensitive
type 2: context-free
type 3: regular languages

natural languages: just beyond context-free
- Shieber 1987: Swiss German
- Mildly context-sensitive grammar formalisms
- Can be parsed in O(n^6)

Example: The RTE Challenge

- RTE ("Recognizing Textual Entailment"): Given a pair of sentence, decide whether second "follows from" first.

T: About two weeks before the trial started, I was in Shapiro's office in Century City.

H: Shapiro works in Century City.

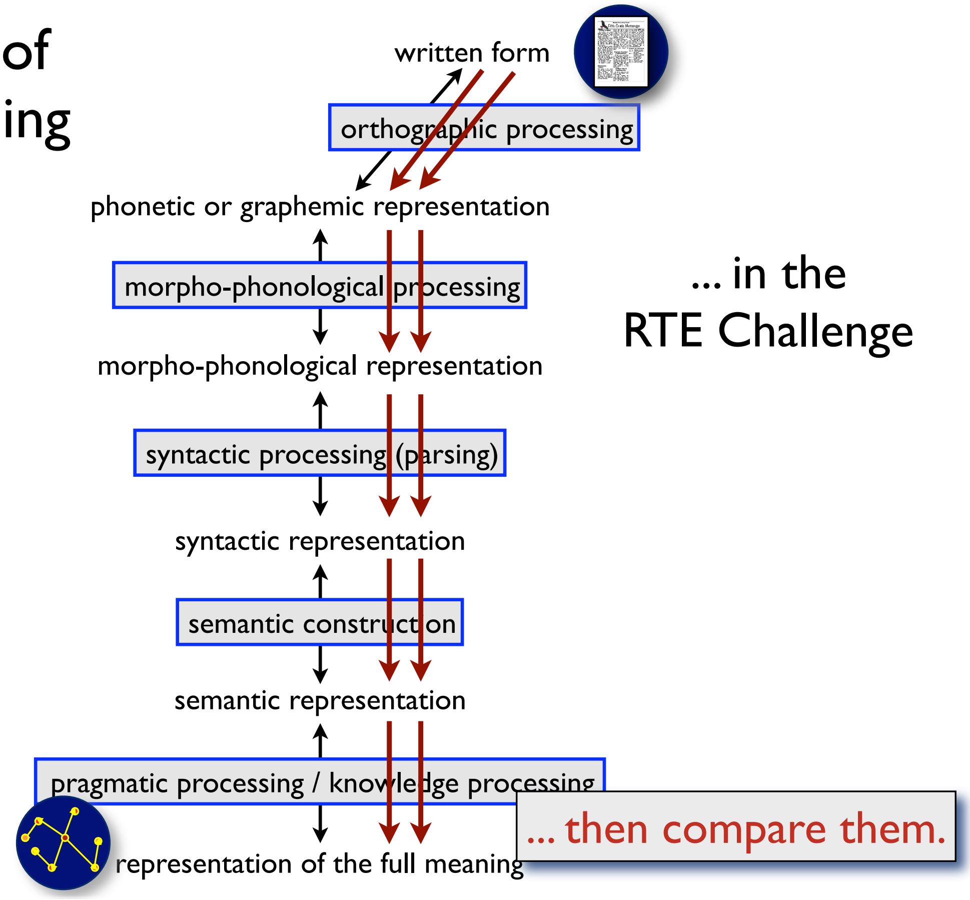
YES

T: Drew Walker, NHS Tayside's public health director, said: "It is important to stress that this is not a confirmed case of rabies."

H: A case of rabies was confirmed.

NO

Levels of Processing


written form
orthographic processing
phonetic or graphemic representation
morpho-phonological processing
morpho-phonological representation
syntactic processing (parsing)
syntactic representation
semantic construction
semantic representation
pragmatic processing / knowledge processing
representation of the full meaning

... in the RTE Challenge

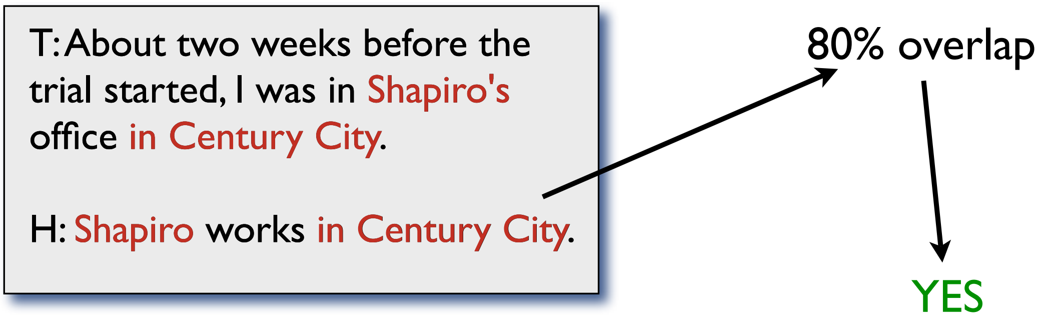
... then compare them.

Need for resources

- Robustness problem: Grammar may not contain entries for unseen words.
- World knowledge problem: We don't have all the formalized knowledge we need for semantic inferences.
- Hand-written language resources expensive and almost necessarily incomplete.

A shallow alternative

Let's just count word overlap!



T: About two weeks before the trial started, I was in Shapiro's office in Century City.

H: Shapiro works in Century City.

80% overlap

YES

On RTE-3 data, this test gives the correct answer in 60% of cases.

Limits

Shallow processing doesn't always get it right.

T: Drew Walker, NHS Tayside's public health director, said: "It is important to stress that this is not a confirmed case of rabies."

H: A case of rabies was confirmed.

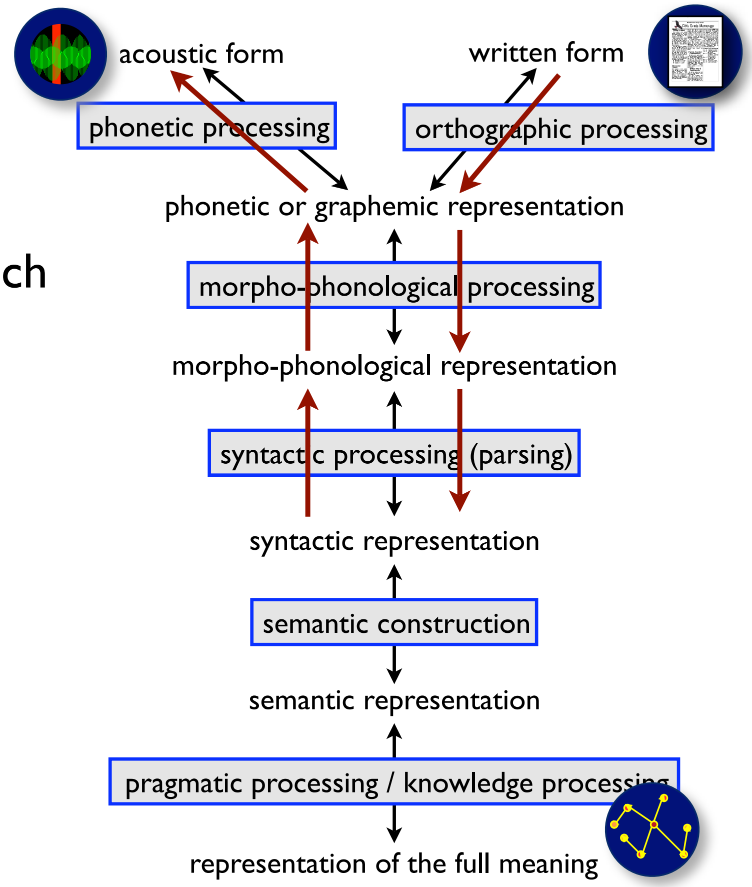
83% overlap

YES

(but should be NO)

Levels of Processing

... in a text-to-speech system


acoustic form
written form
phonetic processing
orthographic processing
phonetic or graphemic representation
morpho-phonological processing
morpho-phonological representation
syntactic processing (parsing)
syntactic representation
semantic construction
semantic representation
pragmatic processing / knowledge processing
representation of the full meaning

Deep processing in TTS

(1) The student will read the paper. (/riːd/)

(2) The students have read the paper. (/rɛd/)

(3) Will the students read the paper? (/riːd/)

(4) Have the students read the paper? (/rɛd/)

(5) Have the students who will arrive next week read the paper yet? (/rɛd/)

(6) Have any citizens of good will read the paper? (/rɛd/)

(7) Please have the students read the paper. (/riːd/)

State of the art

- Deep language processing is too slow for many applications, and we lack resources.
- Shallow language processing can be much faster and doesn't care about ambiguity, but suffers from uninformative analyses.
- Future: Make deep processing faster; make shallow processing more informed; combine them.

Some paradoxes

- Language processing complex, but still you can understand it in real time.
- Language is often ambiguous, but you almost never notice it.
- How is this possible?

Hard-to-understand sentences

- English: “In mud eels are, in clay are none.”
- German: “Mähen Äbte Heu?”
- Garden-path sentences:
  “The canoe floated down the river sank.”
  (vs. “The clothes put on the rack smelled.”)

Competence vs. Performance

- Linguistic Competence:
  - The knowledge a speaker has to possess in order to master a language.
  - The system of rules, principles and constraints that constitute the grammar of a language
  - The finite definition of an infinite natural language.
- Linguistic Performance:
  - The mechanisms and processes underlying actual human language use (production and comprehension).
  - Language use under the constraints of using a real brain in a real communicative situation.

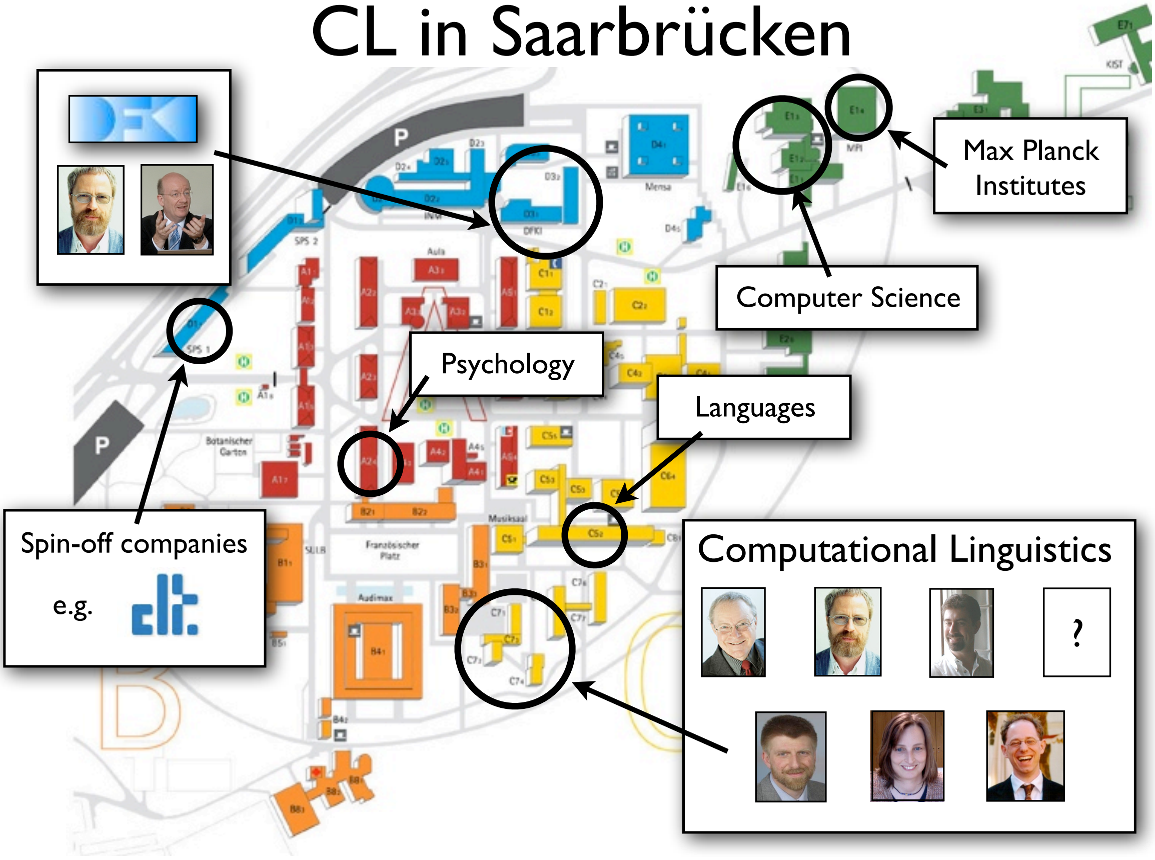
Performance Models

- ... should explain:
  - why many ungrammatical sentences are produced
    (speech errors, grammar errors)
  - why many ungrammatical sentences are understood
    (communication with non-native speakers, children)
  - why many grammatical sentences are never produced
    (preferences in generation)
  - why many grammatical sentences are not understood
    (garden-path sentences)
  - how processing is structured
    (efficiency and control flow)
  - effort required by the components
    (dependence on other cognitive efforts)

Summary

- On Wednesday: Linguistics and ambiguity.
- Combinatorial explosion, efficiency, robustness, world knowledge.
- Deep vs. shallow processing.
- Competence vs. performance.

CL in Saarbrücken



Max Planck Institutes

Computer Science

Psychology

Languages

Spin-off companies

e.g.

Computational Linguistics

?