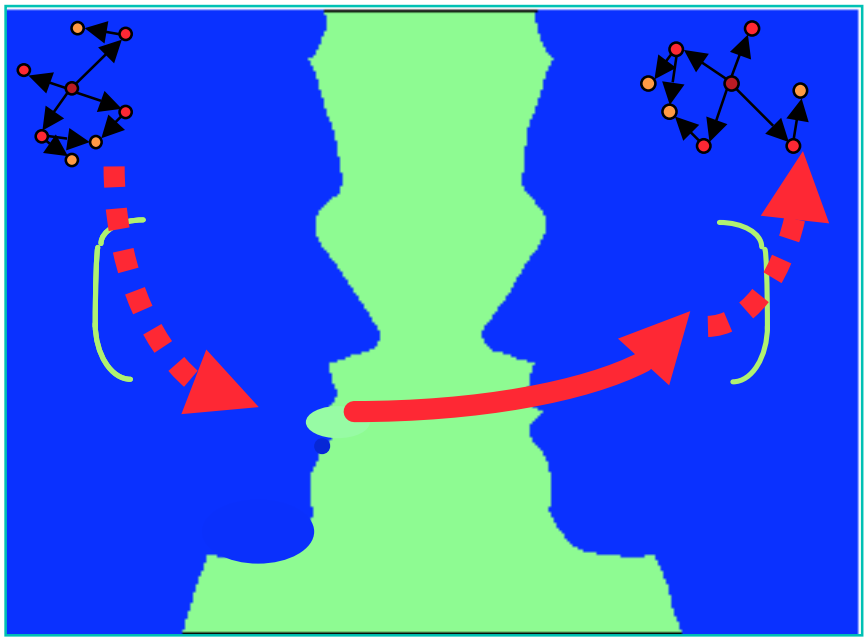
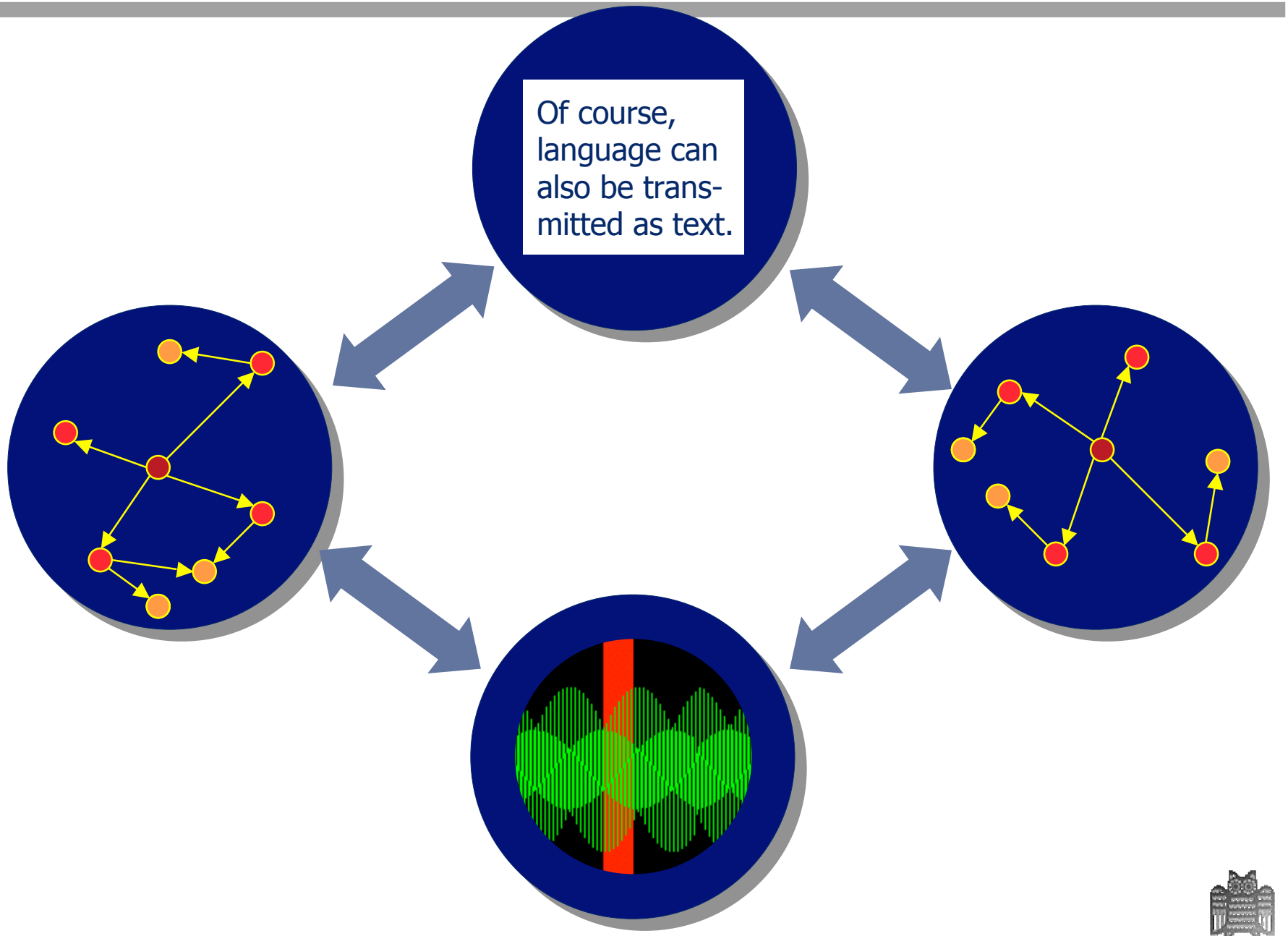
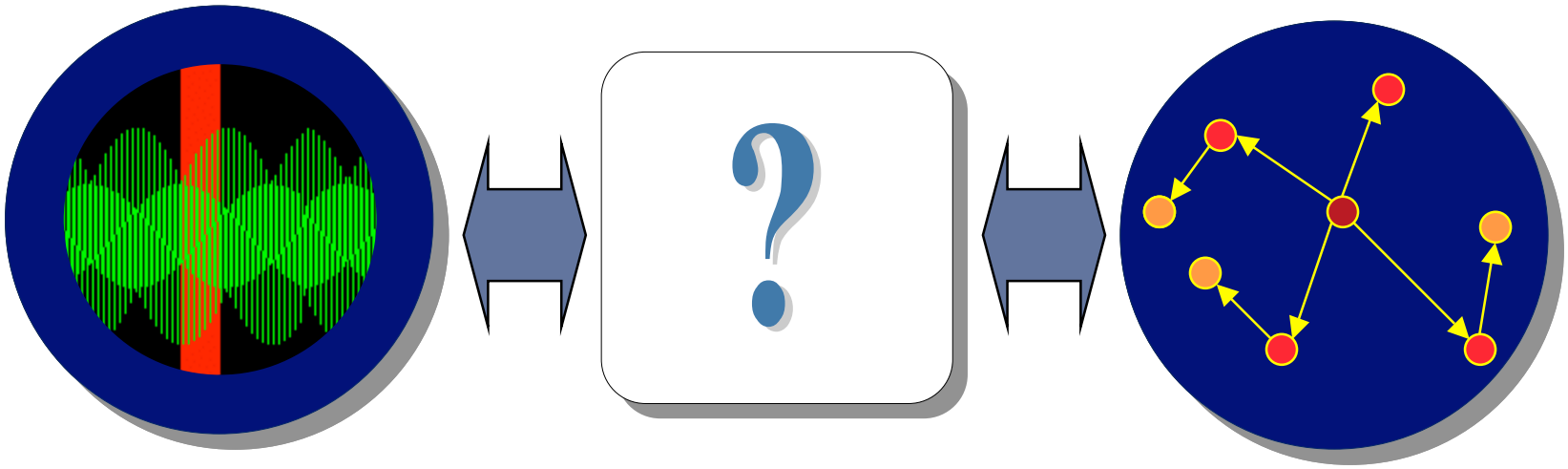
FOUNDATIONS OF LANGUAGE SCIENCE AND TECHNOLOGY



Language is the Medium



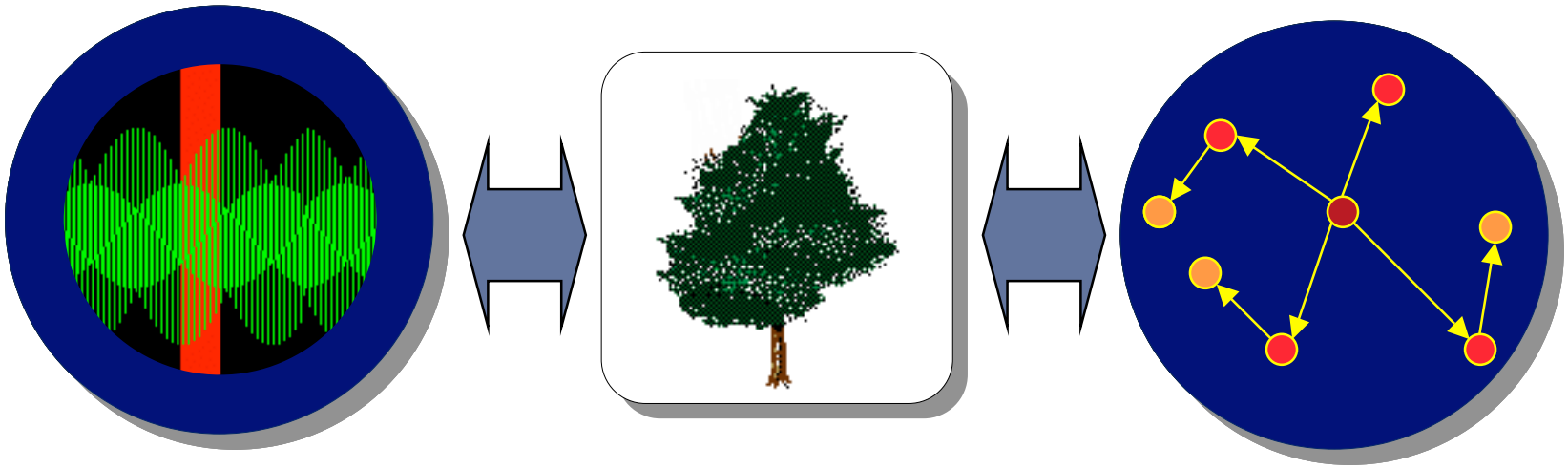


sound waves

Grammar

activation of concepts

WHAT HAPPENS IN BETWEEN?

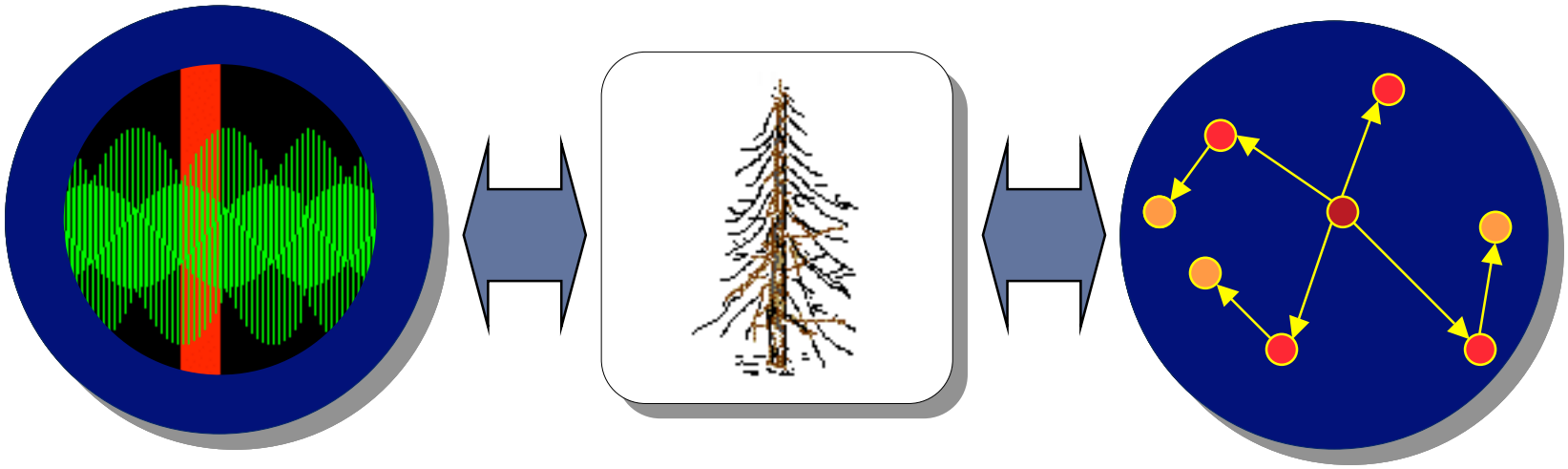


sound waves

Grammar

activation of concepts

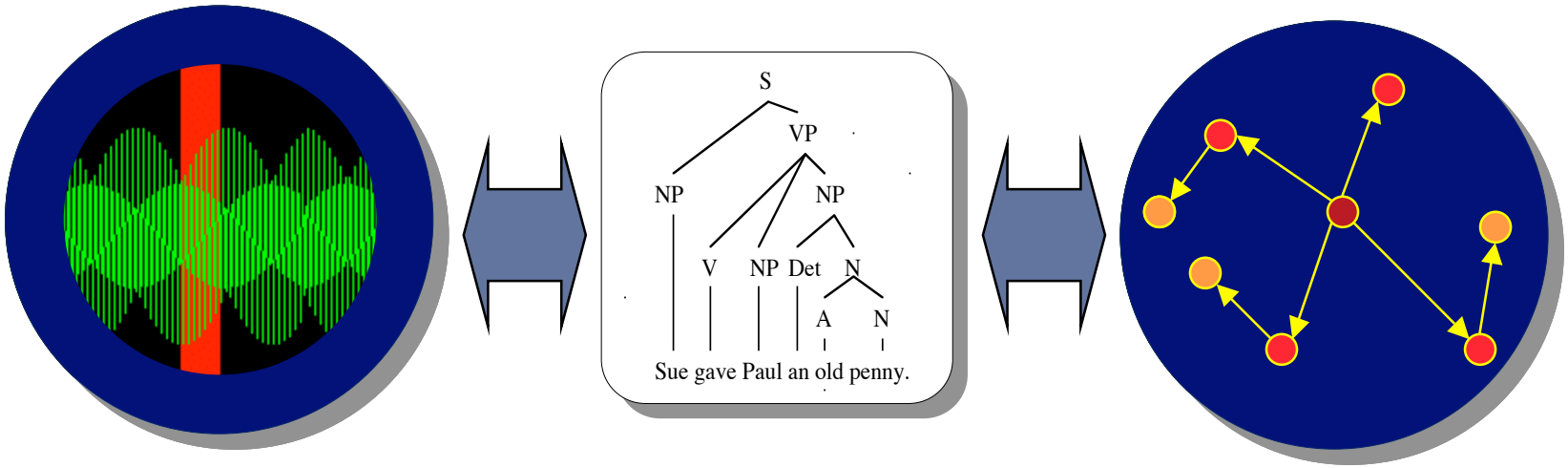
WHAT HAPPENS IN BETWEEN?



sound waves

Grammar

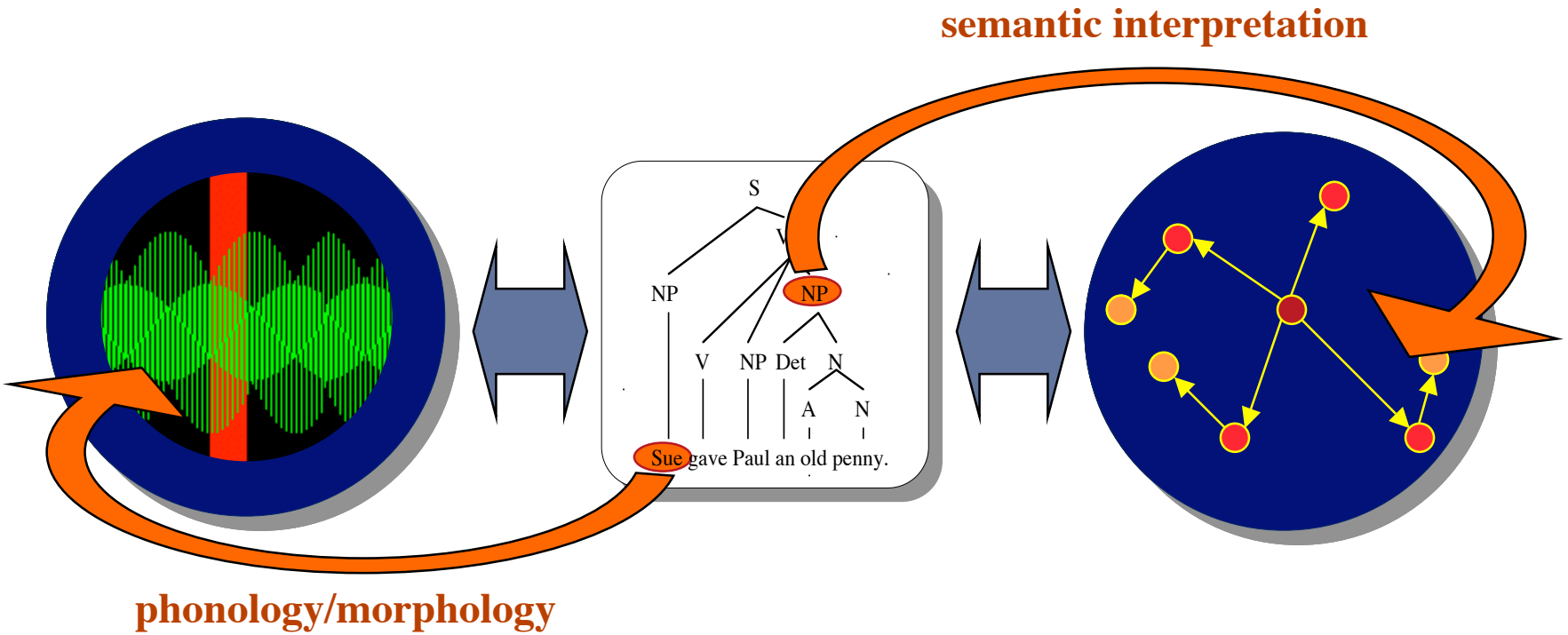
activation of concepts

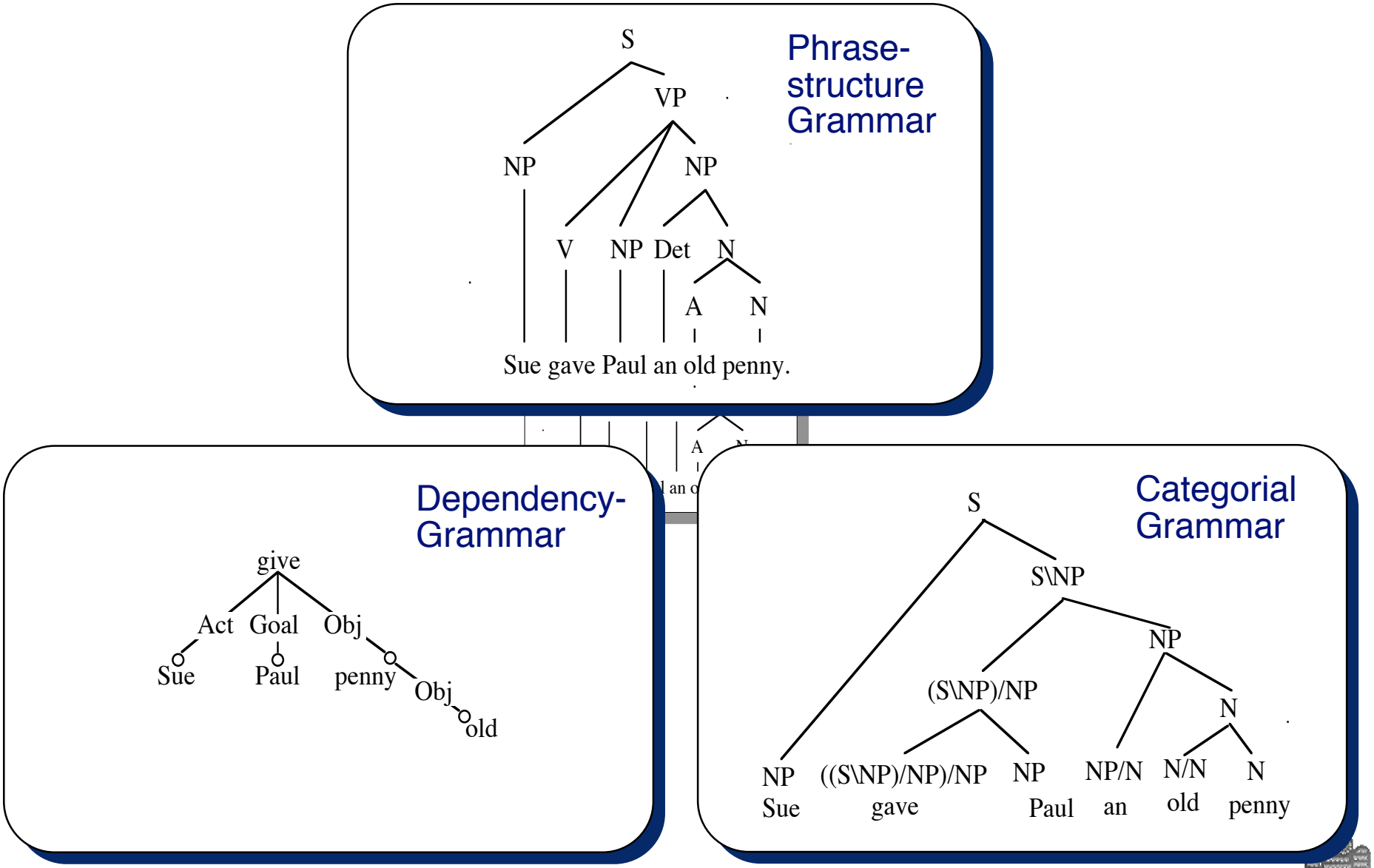


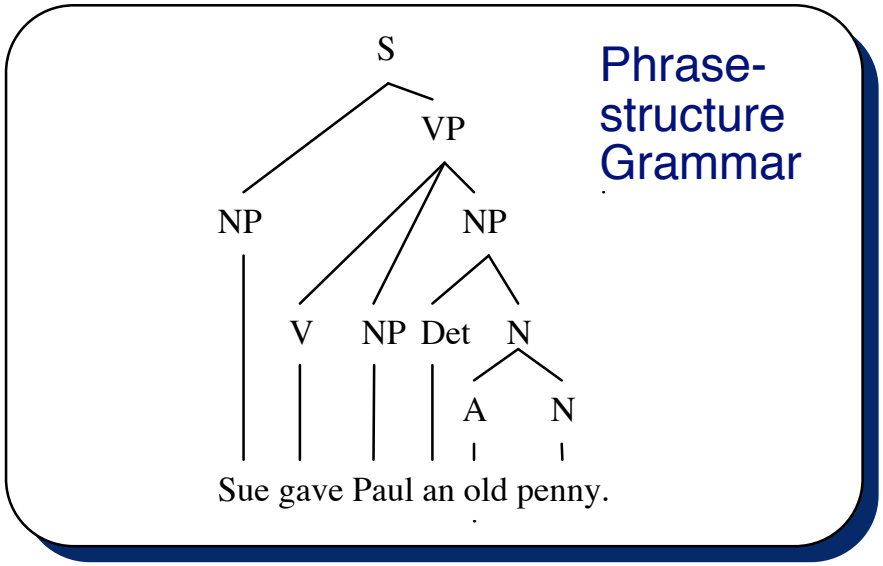
sound waves

Grammar

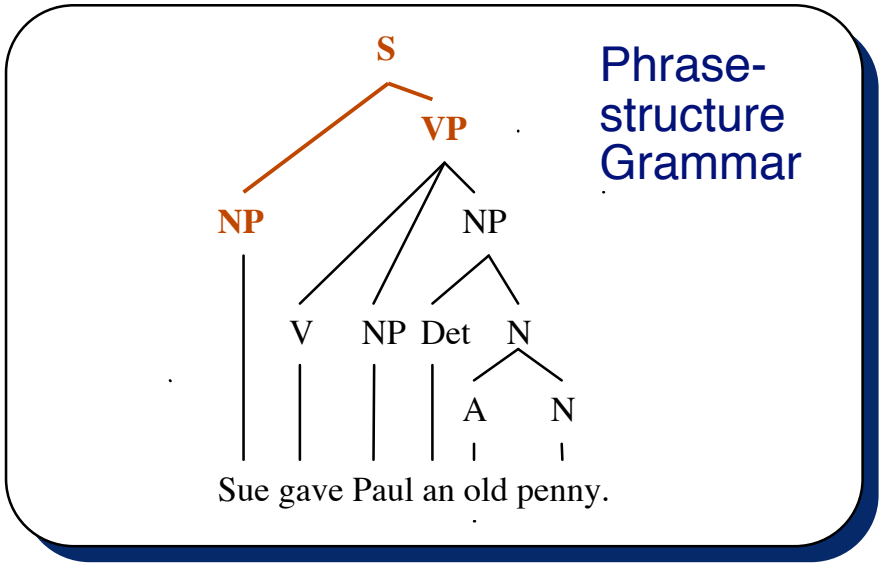
activation of concepts



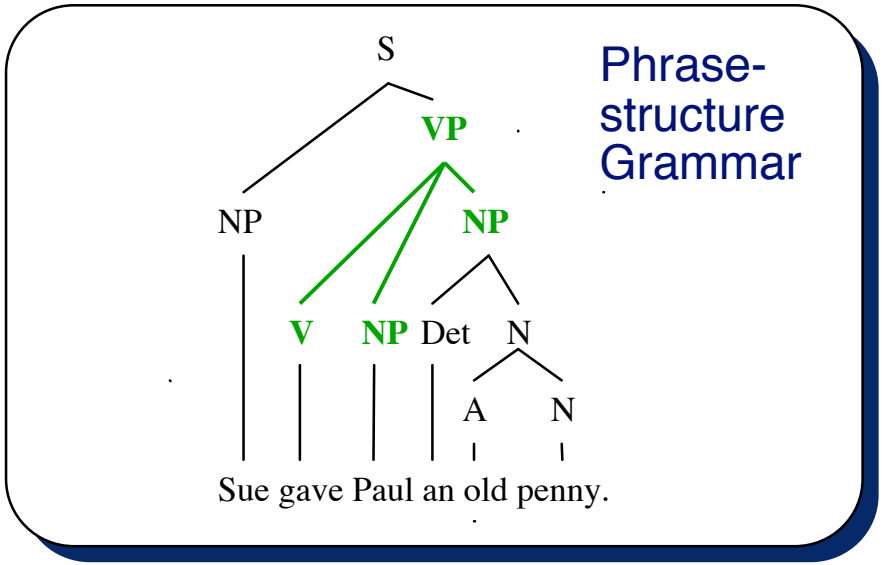




S → NP VP

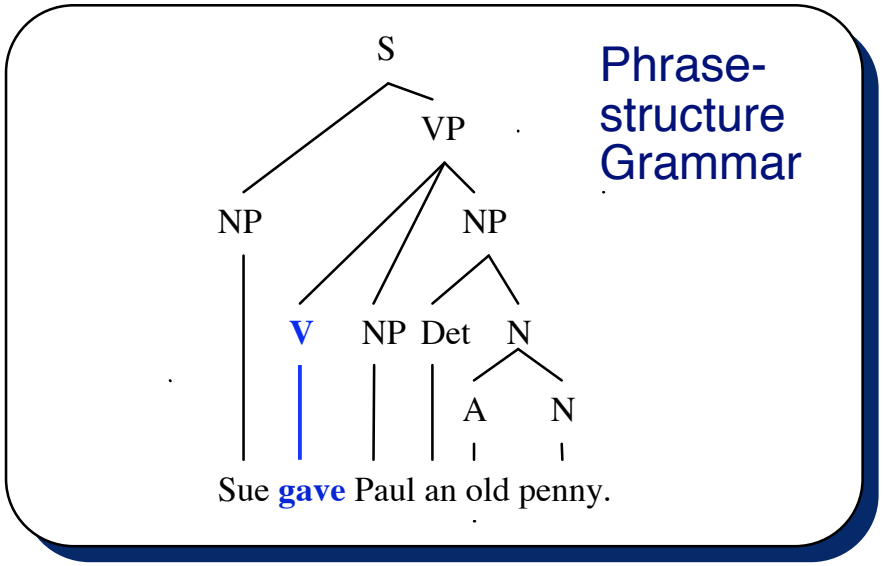


S → NP VP



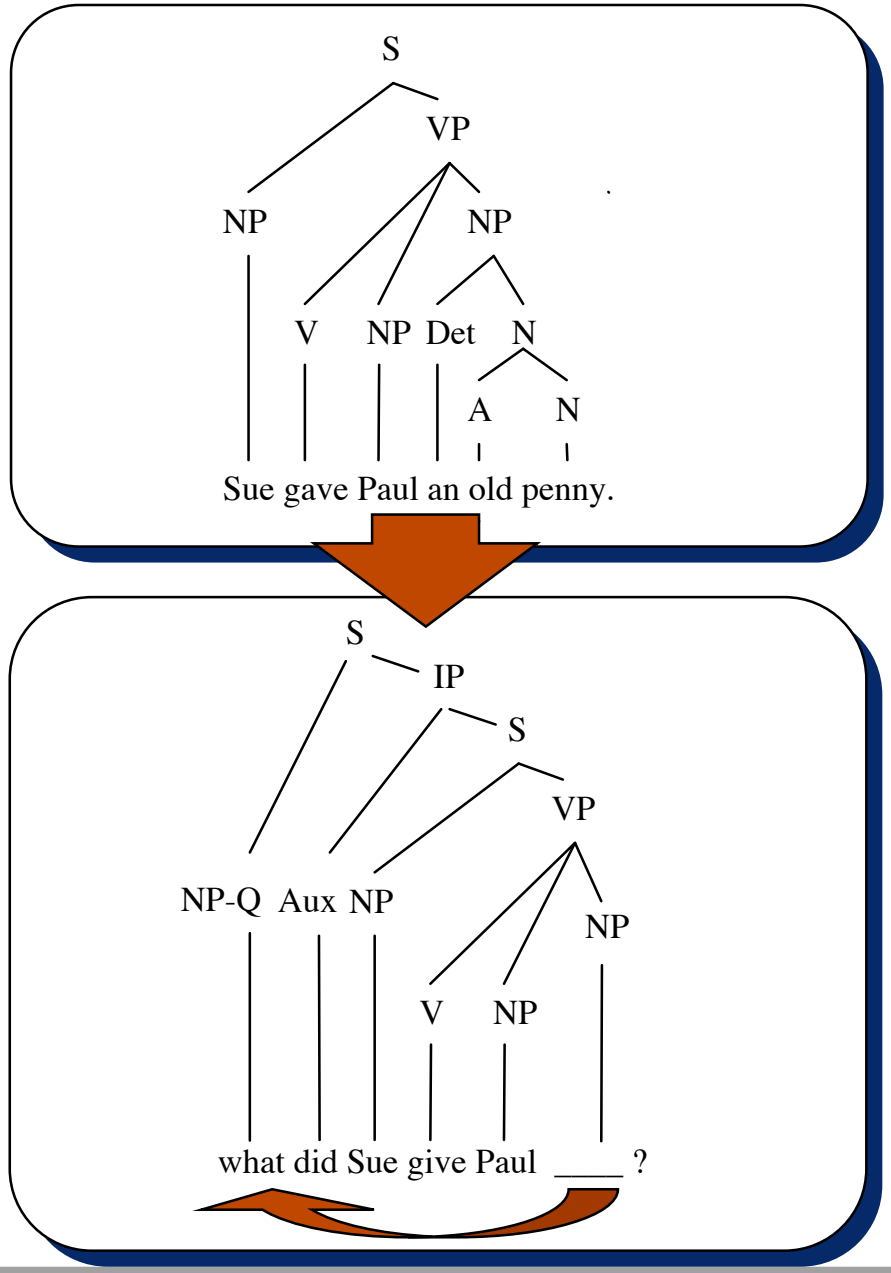
S → NP VP

VP → V NP NP



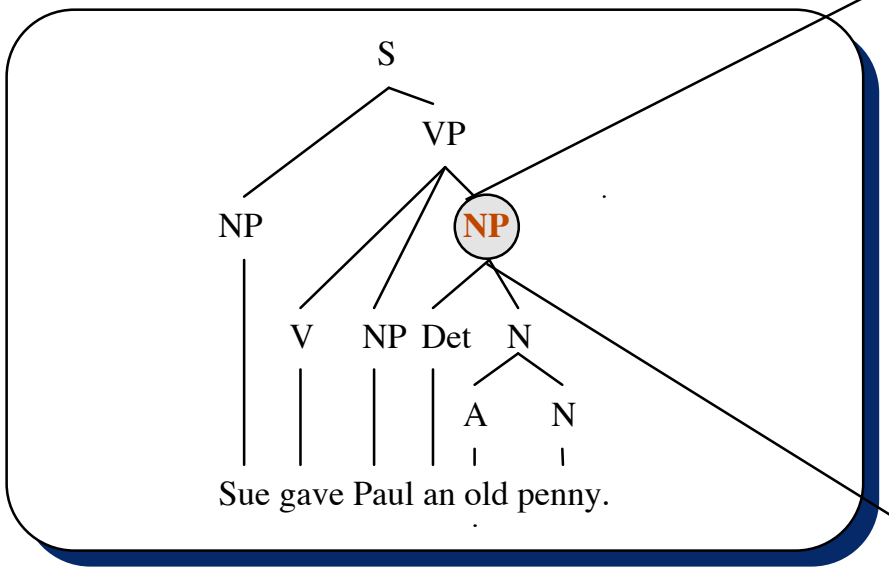
S → NP VP
VP → V NP NP

V → gave



Transformation Grammar

Unification Grammar



$$\begin{bmatrix} \text{PHON} \langle \textit{a old penny} \rangle \\ \text{SYN} \begin{bmatrix} \text{CAT } \textit{NP} \\ \text{HEAD} \begin{bmatrix} \text{CASE } \textit{objective} \\ \text{NUMBER } \textit{sing} \\ \text{PERSON } \textit{third} \end{bmatrix} \\ \text{VALENCE } \textit{vstruc} \end{bmatrix} \\ \text{SEM} \begin{bmatrix} \text{QUANT } \textit{exist} \\ \text{VAR } X_1 \\ \text{RESTR} \begin{bmatrix} \text{REL } \textit{old}' \\ \text{VAR } X_1 \\ \text{ARG } \textit{penny}' \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

- How large is the grammar.
- Let's start with the lexicon.

Estimates for English

- Shakespeare actively used 29.000 word forms mapping to about 25.000 head words

- common estimates of the vocabluary of a college graduate:
  20.000 words active -- 25.000 words passive

- David Crystal's estimate
  60.000 words active -- 75.000 words passive

- Total Size of English Vocabulary

  1 million words without special scientific and technical terms
  2 million words including all scientific and technical terms

A million-word-corpus of American English exhibits about 38.000 head words.

- LinGO - English Resource Grammar

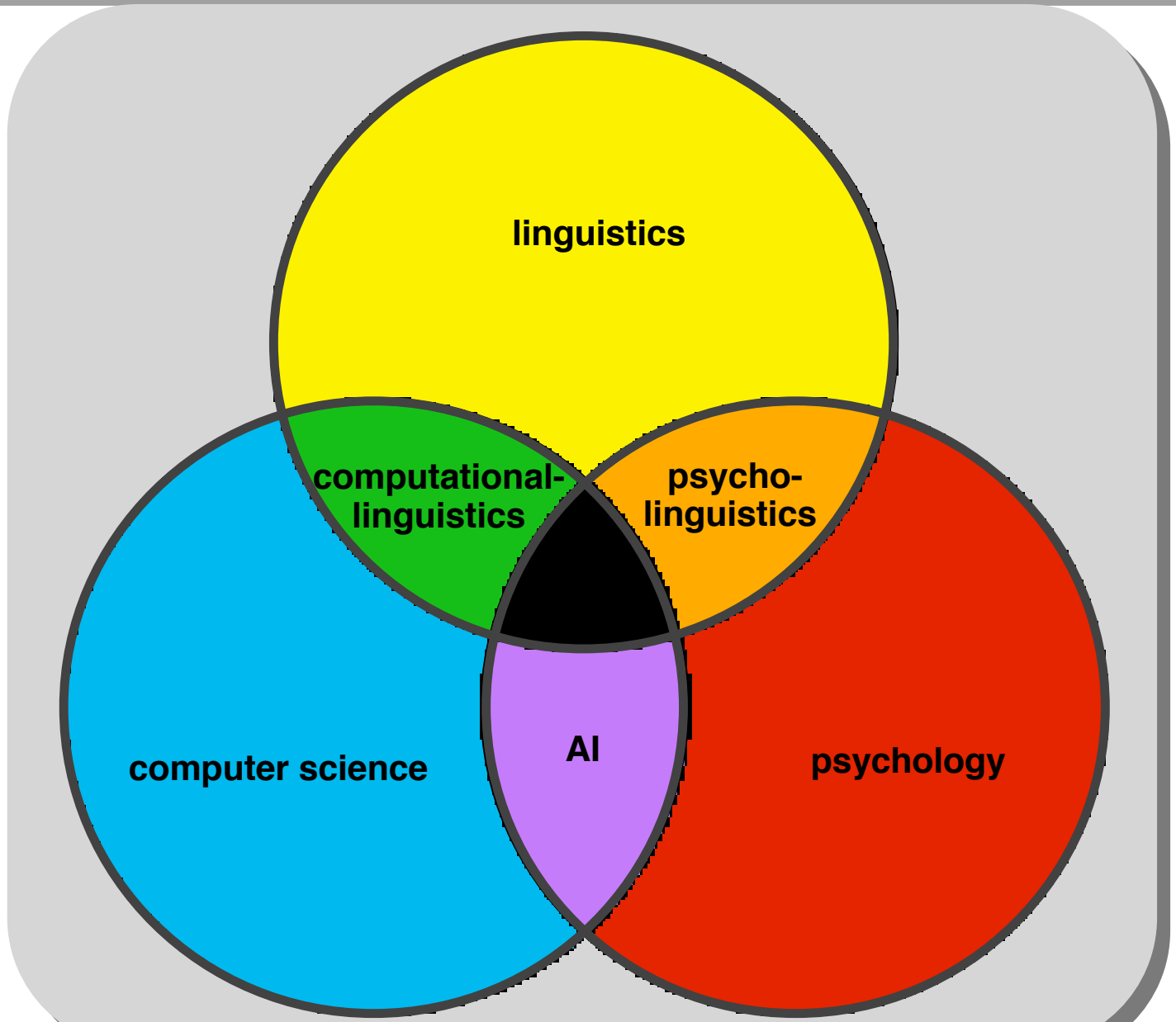
  (60% coverage of newspaper texts)

  - 8.000 types
  - 100.000 lines of code
  - average feature structure > 300 nodes

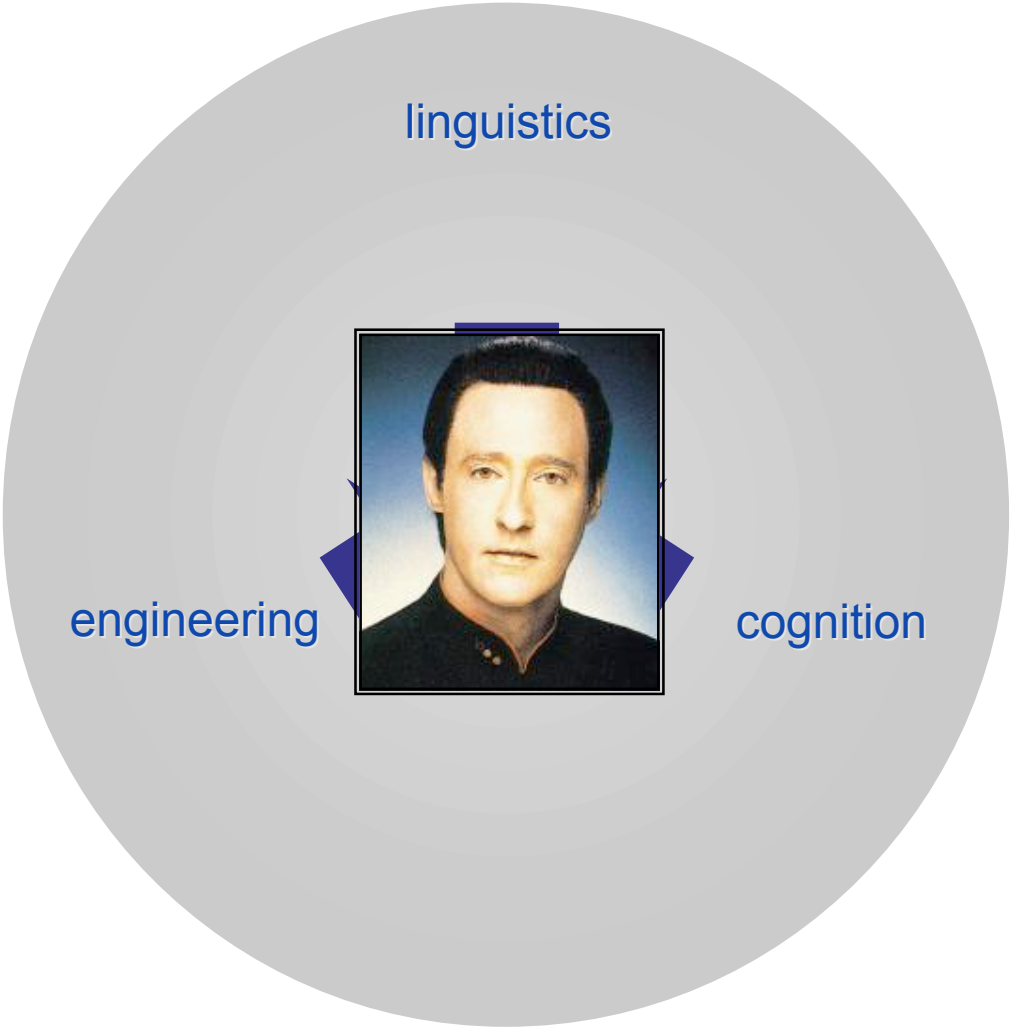
How Many Languages ?

- According to Ethnologue 6,809 languages
- 230 in Europe, 2197 in Asia (832 in Papua-New Guinea)
- Bible translations exist for 2.200 languages
- 250 families of languages (such as Indoeuropean Languages)

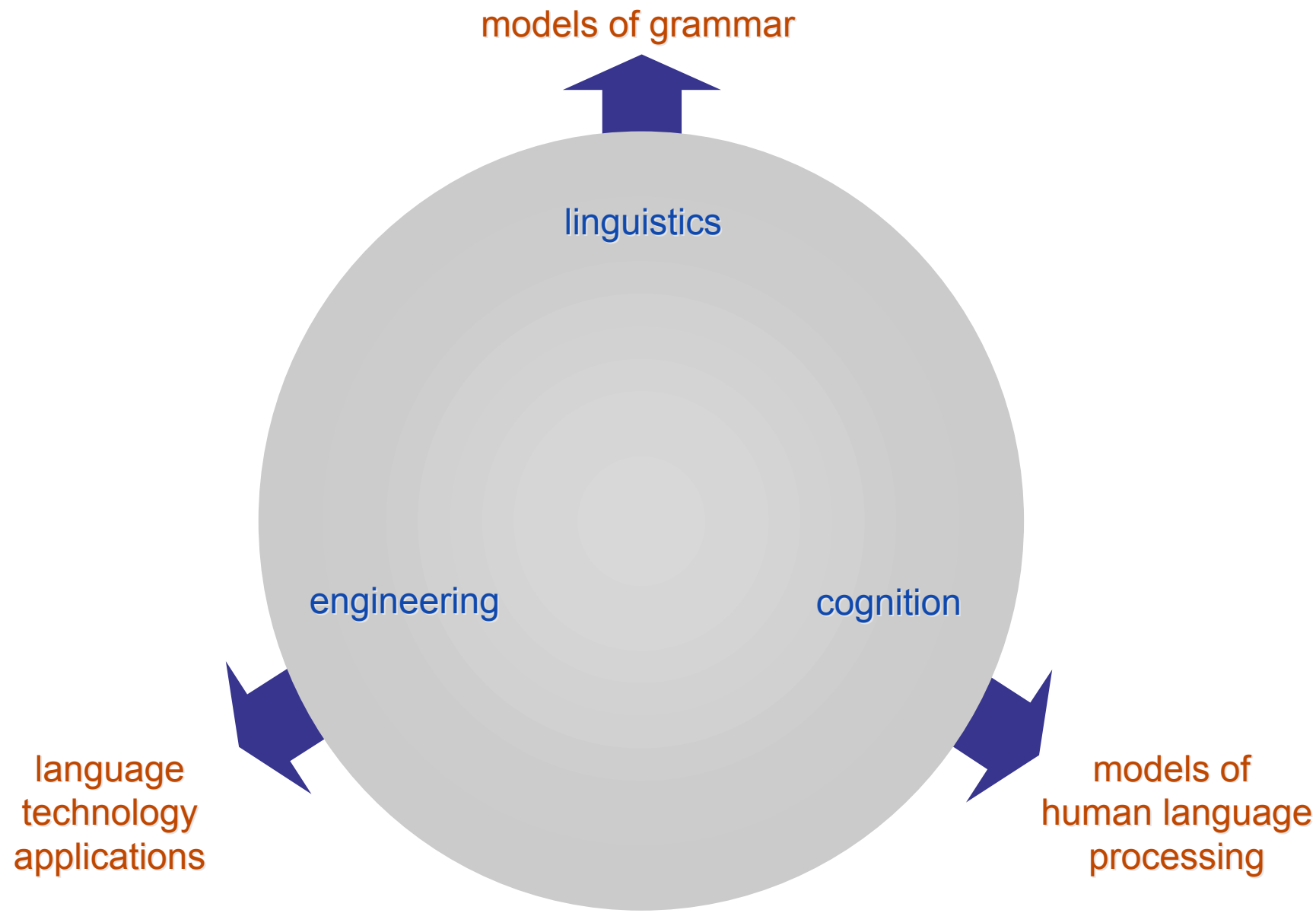
Transdisciplinary Interests



Motivations



Motivationen



Central Questions of Language Research

- LINGUISTIC KNOWLEDGE

  What are the contents and structures of this knowledge

- LANGUAGE PROCESSING

  How do we produce and comprehend linguistic utterances?

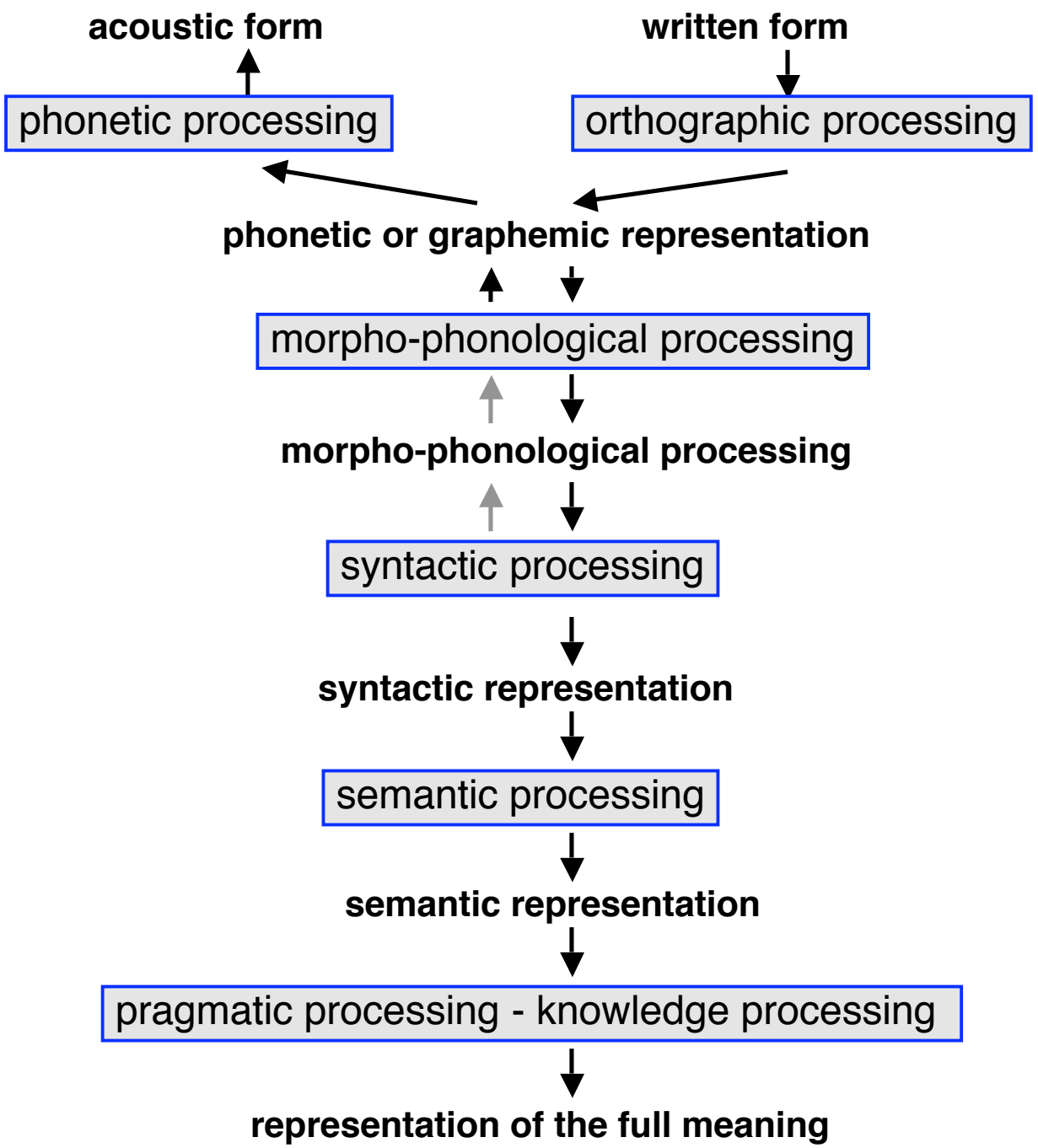
- LANGUAGE ACQUISITION

  How does the child learn his mother tongue?

- LANGUAGE CHANGE

  How do languages (dialects, sociolects) emerge, change, evolve?

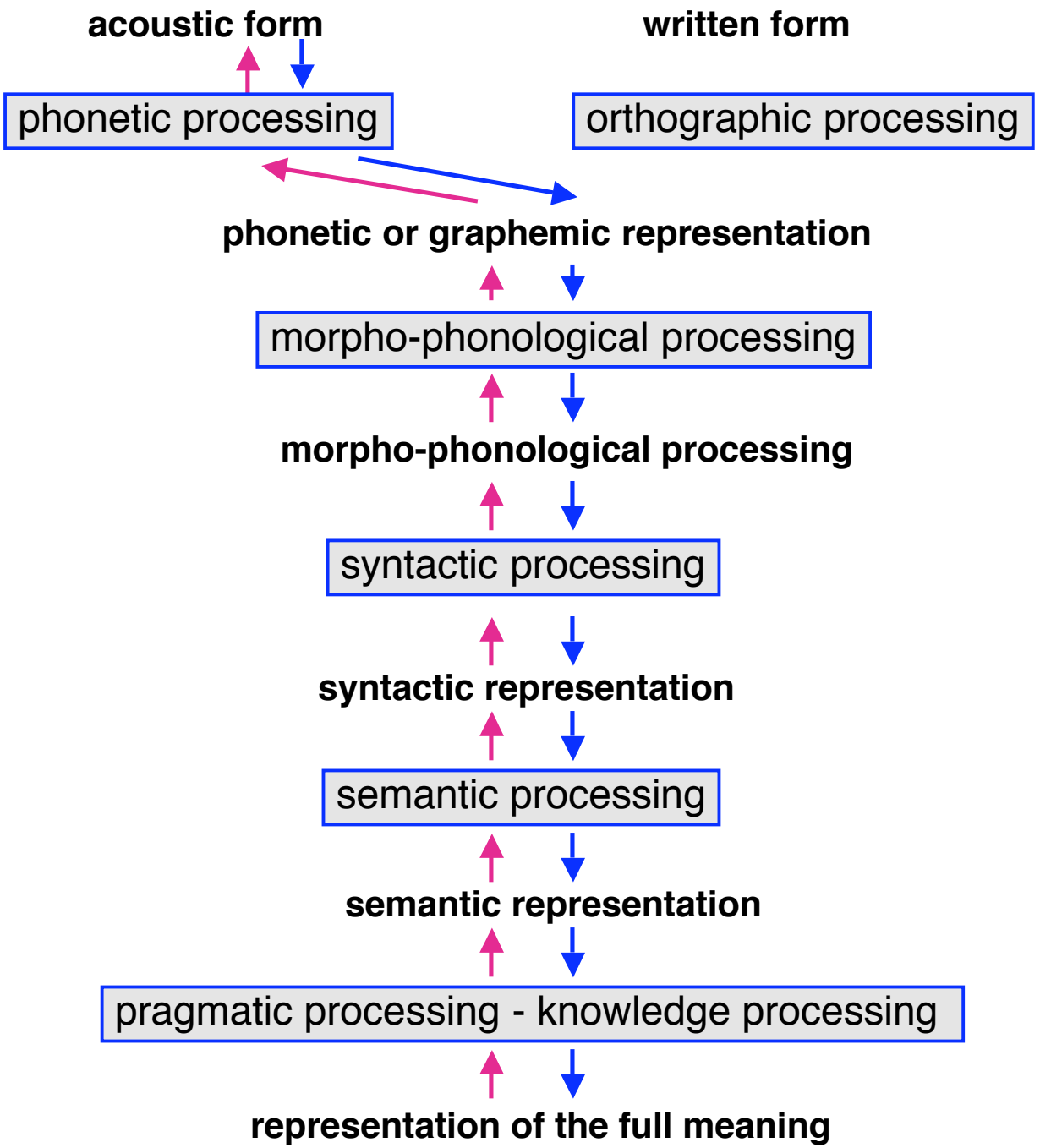
Text-to-Speech System



Why do we need deep processing for simple text-to-speech conversion

(I) The girls will read the paper. (*reed*)
(2) The girls have read the paper. (*red*)
(3) Will the girls read the paper? (*reed*)
(4) Have any men of good will read the paper? (*red*)
(5) Have the executors of the will read the paper? (*red*)
(6) Have the girls who will arrive next week read the paper yet? (*red*)
(7) Please have the girls read the paper. (*reed*)
(8) Have the girls read the paper? (*red*)

Speech Translation



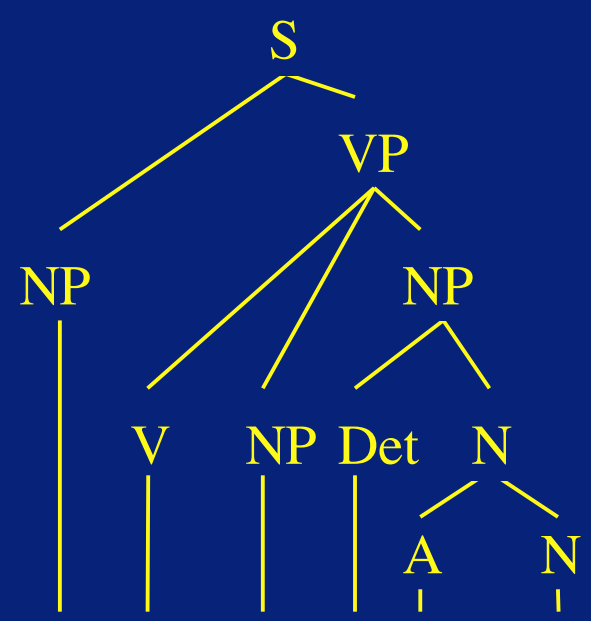
If you can walk
you can dance.

If you can talk
you can sing.

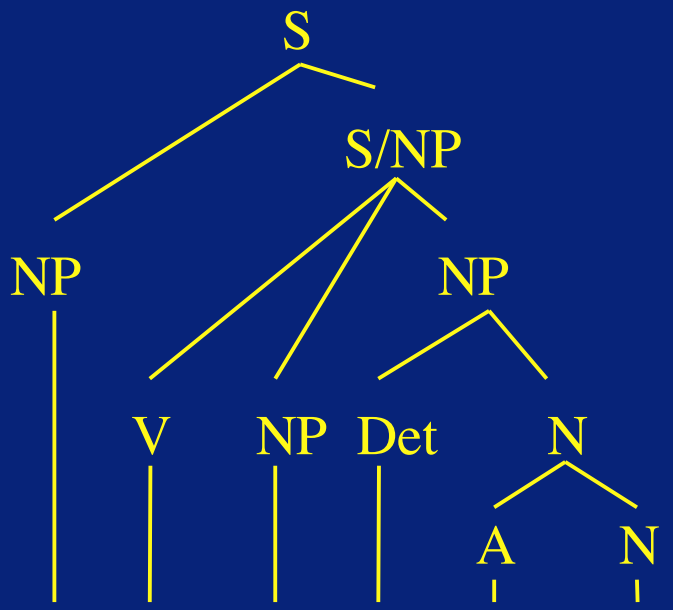
If you can parse,
you can understand.

(Proverb from Zimbabwe, China,
and maybe other places)

$\exists x[(old'(penny'))\ (x) \wedge (Past(give'(sue', paul', x)))]$

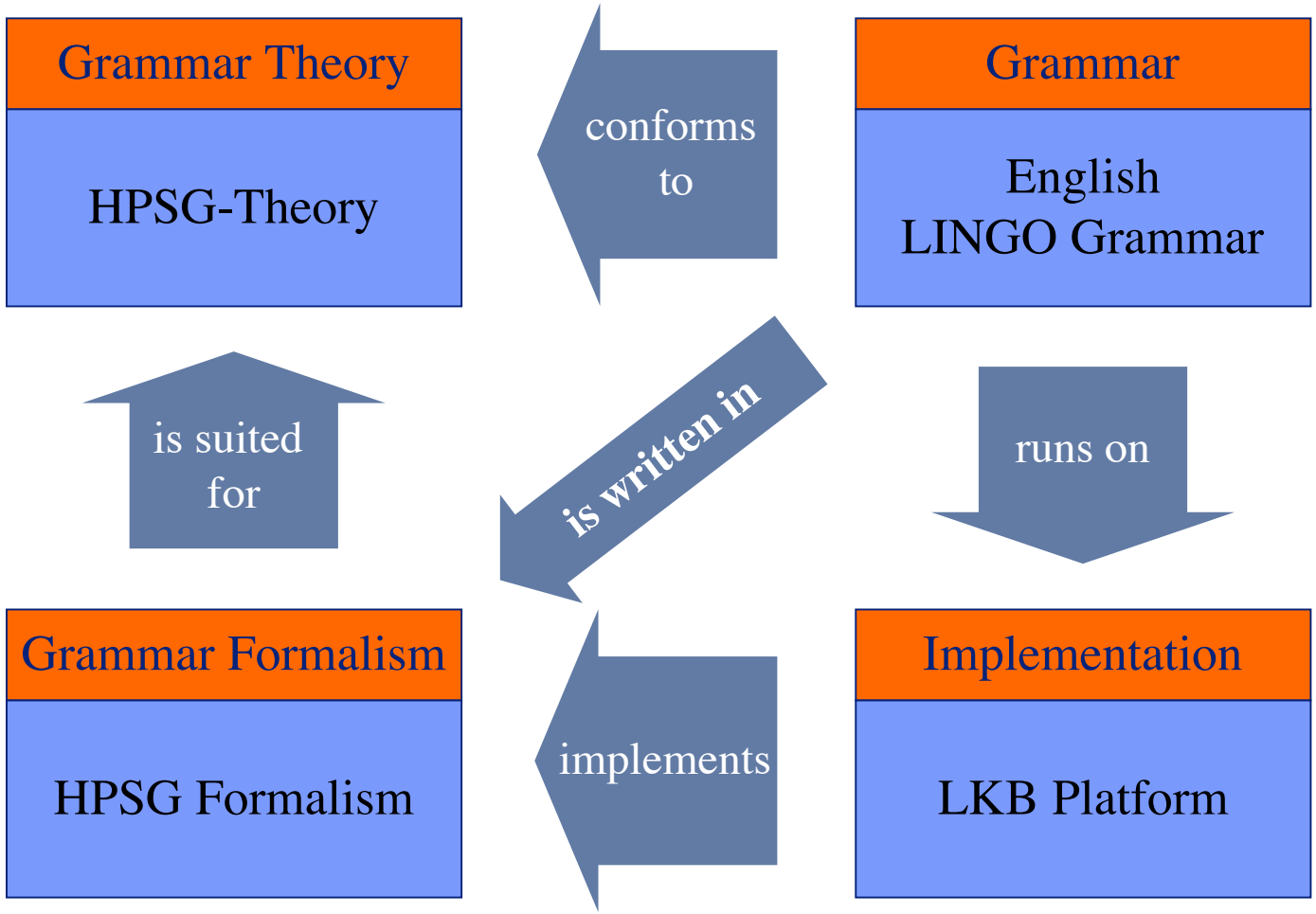


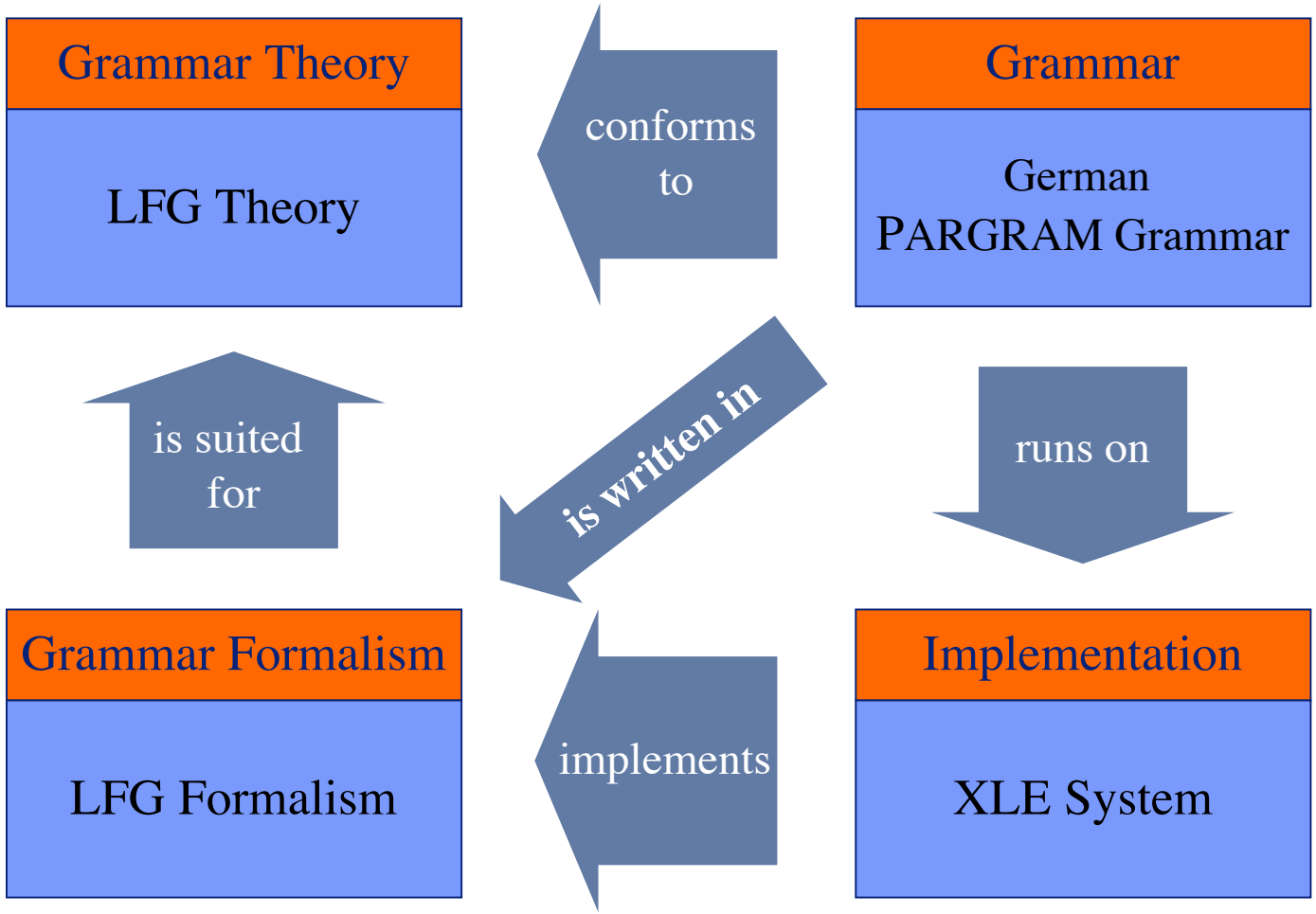
Sue gave Paul an old penny.

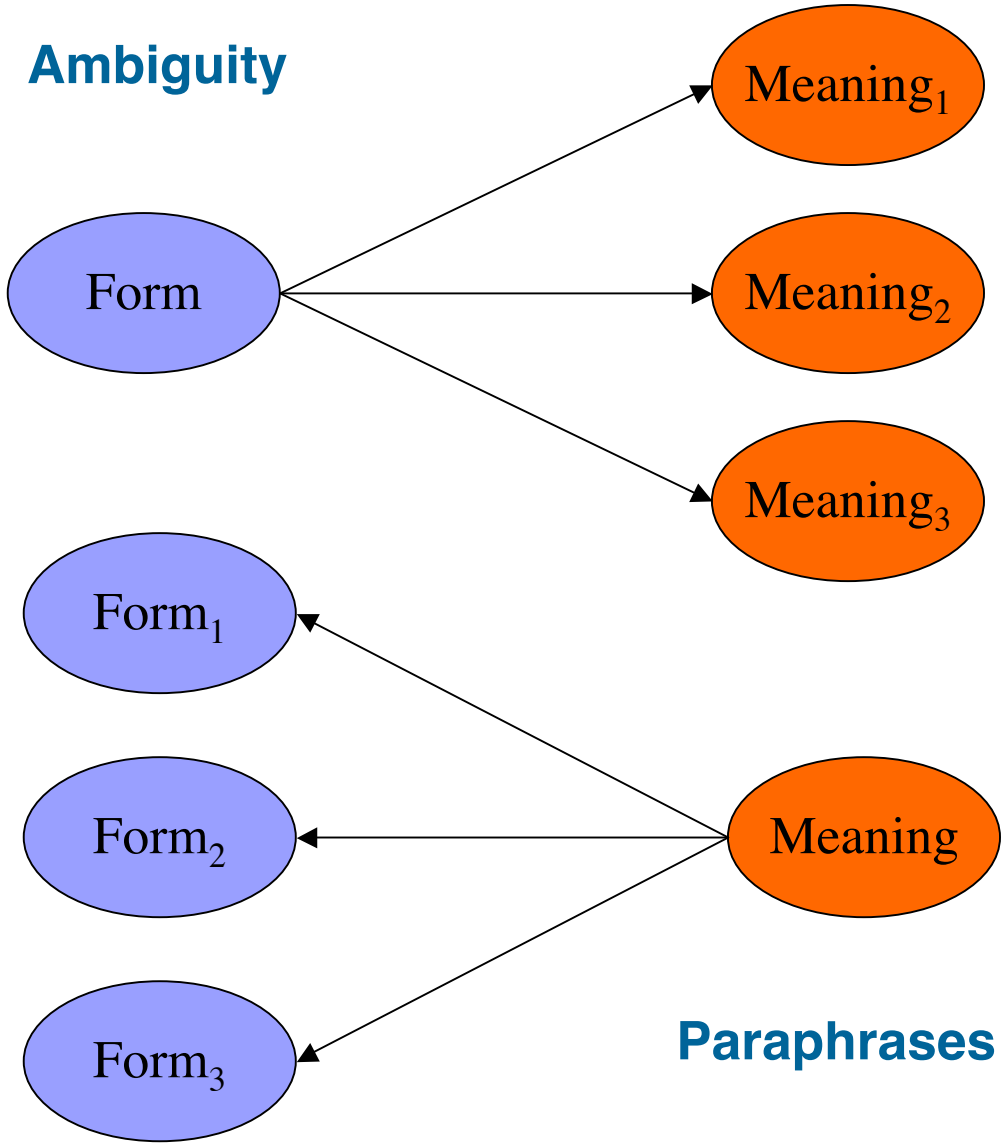


Sue gab Paul einen alten Pfennig.

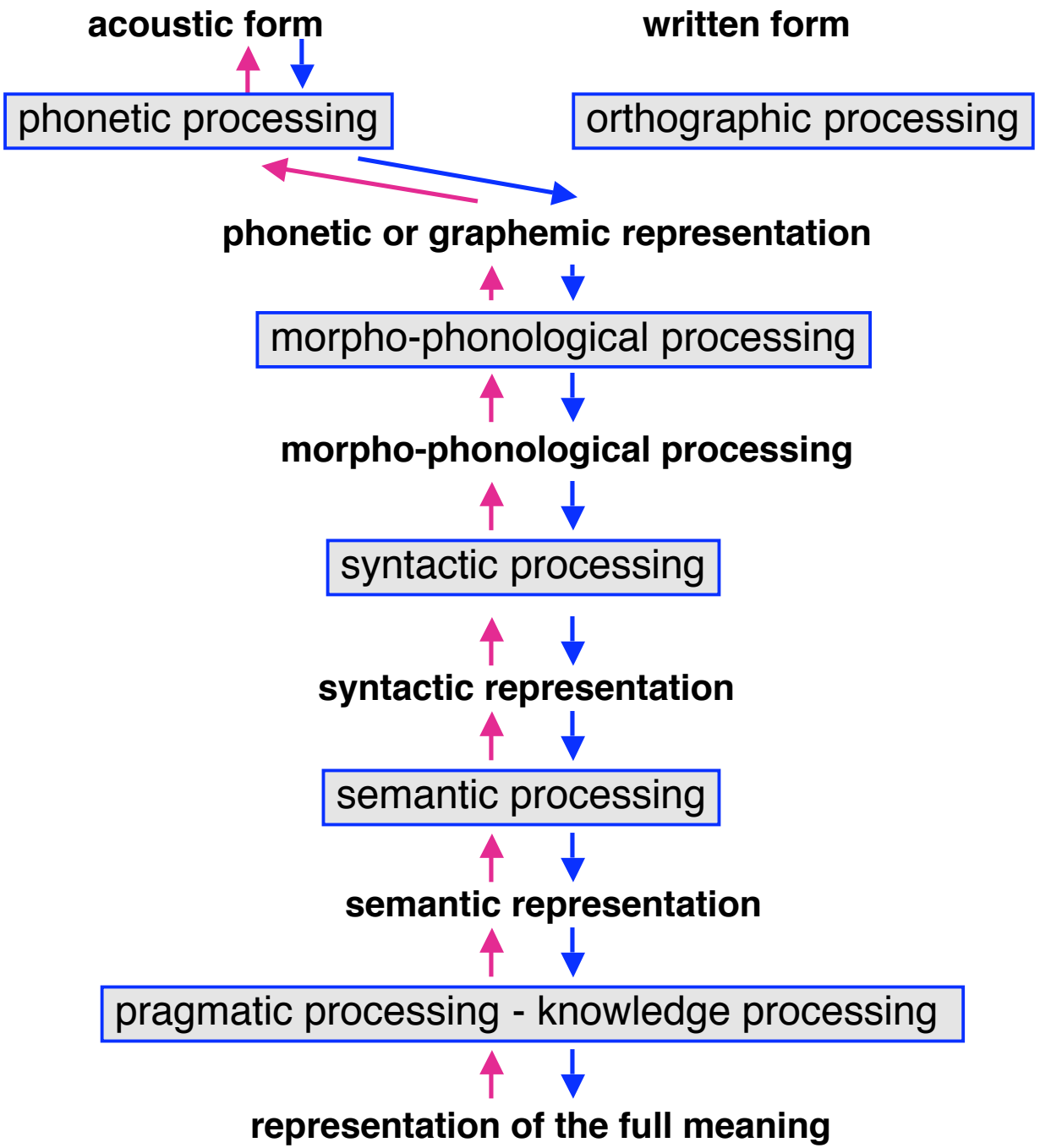
FOUR SENSES OF GRAMMAR







Speech Translation



phonetic (homophony):

their	there
toe	tow

orthographic (homography):

read	read
undoable	undoable

lexical (homonymy):

bank	bank
ball	ball

syntactic

With the naked eye she
couldn´t see much.
So she watched the man
with a telescope.

She couldn't watch
all suspects
So she watched the man
with a telescope.

semantic

The three selected special agents
speak two foreign languages
nearly without an accent.
Namely French and Russian.

The three selected special agents
speak two foreign languages
nearly without an accent.
But only two of them master
Russian.

pragmatic

Could you translate this text?
I need it tomorrow.

Could you translate this text?
I even wonder if anybody could do it.

Certain readings are less preferred than others:

Where is a bank?

Do you like plants?

The preference can be influenced by context.

The goal keeper opened the ball. **vs.** ***The Mayor opened the ball.***

The astronomer married a star. **vs.** ***The movie director married a star.***

Ambiguity (a pathological example)

„Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit

in the past produced the women of the islands on the weekends scarfs with

Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem

flower 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.

In the past the women of the islands produced scarfs with flower patterns on the weekends that were sold by their husbands on the following Mondays on the market in the center of the main island.

The sentence exhibits a total of 13 lexical, syntactic and anaphoric 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

phonetic:	
réport	repórt
ínnovative	innóvative
orthographic:	
summarization	summarisation
co-ordination	coordination
lexical:	
access road	feeder road
Fall	Autumn

syntactic:

I will very slowly pull out.	I will pull out very slowly.
He was recognized by all of us.	All of us recognized him.

semantic:

Not all students could attend.	Some students could not attend.

pragmatic:

Could you translate this text.	Please translate this text.

- Meaning is expressed by a formula in some logic
- candidates:
  - first order predicate logic (FOPL)
  - subsets of FOPL (Horn-clause logics, Description Logics)
  - some higher order logics such as modal logics
- some logical reasoning is performed by humans with great ease
  - If you skip classes, I have to flunk you. $p \rightarrow q$
  - Either you don't skip classes or I have to flunk you $\neg p \vee q$
- If Berlusconi becomes the next president of the European Commission, I´ll become the emperor of China

- Go to the table with many screws lying on top.
- Are there many screws on the table?
- Hand me the big hammer?
- Is the hammer big?

- Linking of data and theories
- shared data, joint tasks and comparative evaluation
- replicability
- layers of annotations
- comments

- How is processing being realized?
- How is linguistic knowledge encoded?
- What are the representations?
- What are the algorithms for processing?

Class of Approaches	heydays
Direct Implementation	1960-1970
Specialized Algorithms and Methods	1970-1980
Declarative Grammar Formalisms	1980-1990
Statistical Approaches	1990-2000
Hybrid Processing Systems	since 2000

Drawbacks of logical approaches

- lack of learning from data (by frequencies)
- lack of mechanisms for soft constraints
- lack of mechanisms for vagueness
- missing robustness w.r.t. ill-formed or unexpected input
- lack of efficiency

Drawbacks of statistical approaches

- lack of general competence models
- lack of accuracy
- no reasoning
- lack of understanding the learned knowledge

Linguistics -- by tradition -- is not an exact or empirical science. Modern linguists have attempted to transform linguistics step by step into a science.

An exact science needs formalized models and provable methods for verifying (or more often falsifying) theories.

Empirical science needs a methodology of how to obtain, process, evaluate data and how to exploit data for the verification (falsification) of theories.

An exact empirical science needs to establish the correspondence between data and formal models. Therefore data need to be interpreted. Quantitative data require methods and tools for measurement.

In linguistics, the quantitative branch of the discipline has been disconnected from the theoretical core of the field for many decades, since quantitative linguists could not measure phenomena that were in the focus of discussion. It was language technology that finally brought them together.

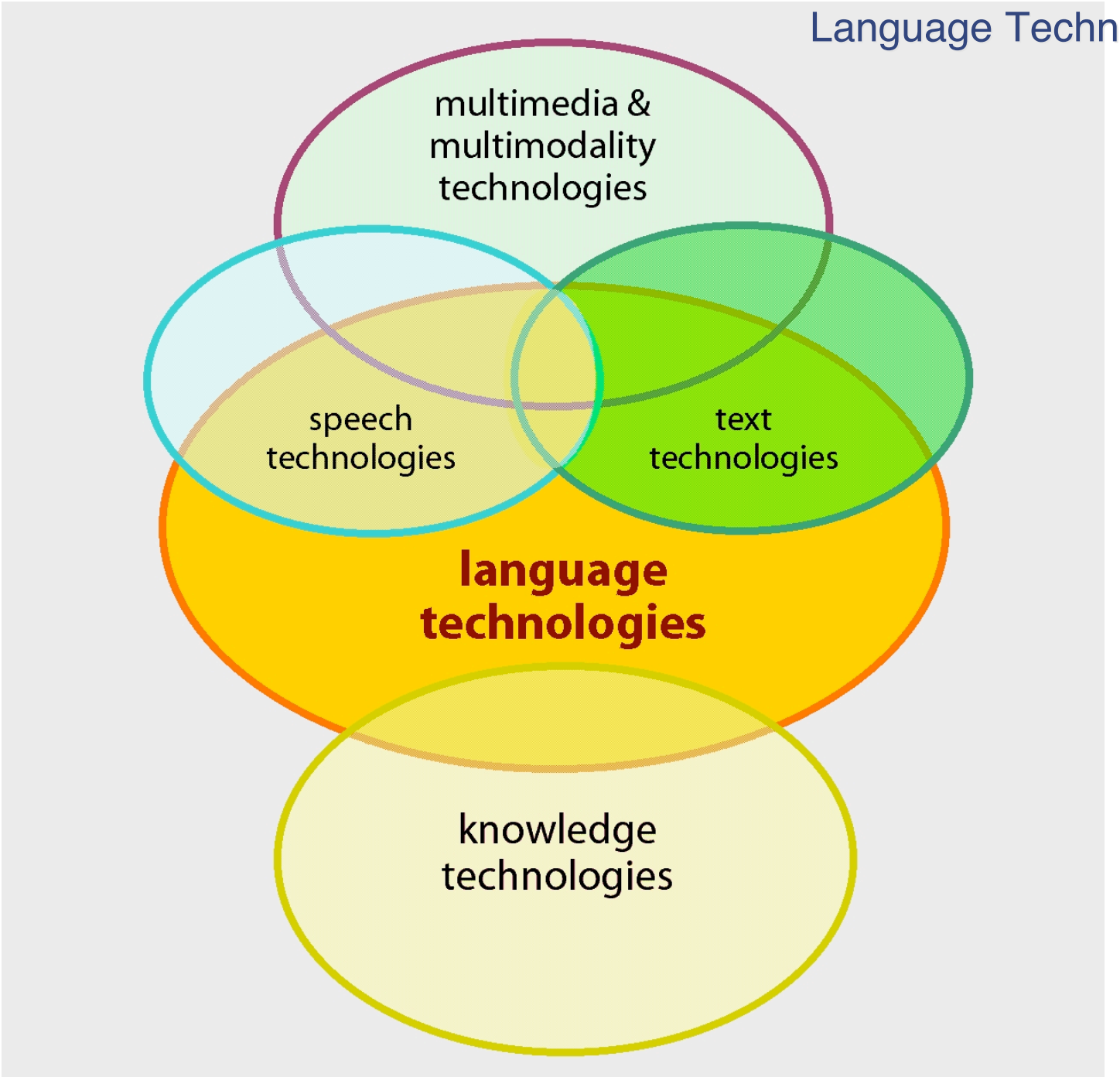
Example: Astronomy

Photographs and spectral analyses of distant heavenly bodies are scientific data. However without their interpretation in relationship with the formal models, they are rather useless.

☆ Theme: How can we build useful applications without any working methods for automatic language understanding?

☆ Which applications do not require full understanding?

Language Technologies



Types of Technologies

Communication partners: humans and machines (technology),
humans and humans
humans and infostructure

Modes and media for input and output: text, speech, pictures, gestures

Synchronicity: synchronous vs. asynchronous

Situatedness: sensitivity to context, location, time, plans

Type of linguality: monolingual, multilingual, translingual

Type of processing: Categorization, summarization, extraction, understanding, translating, responding

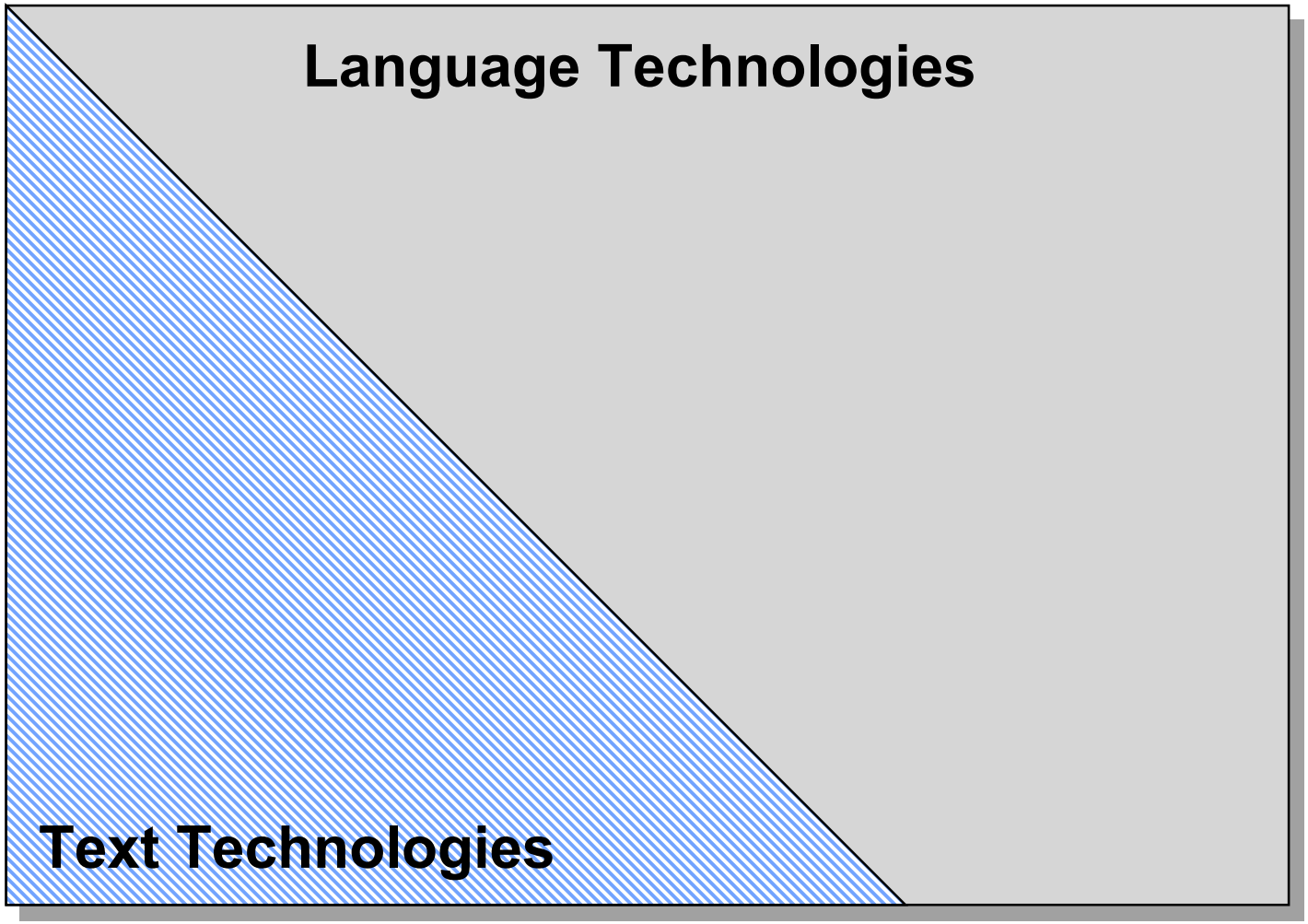
Level of linguistic description: phonology, morphology, syntax, semantics,pragmatics

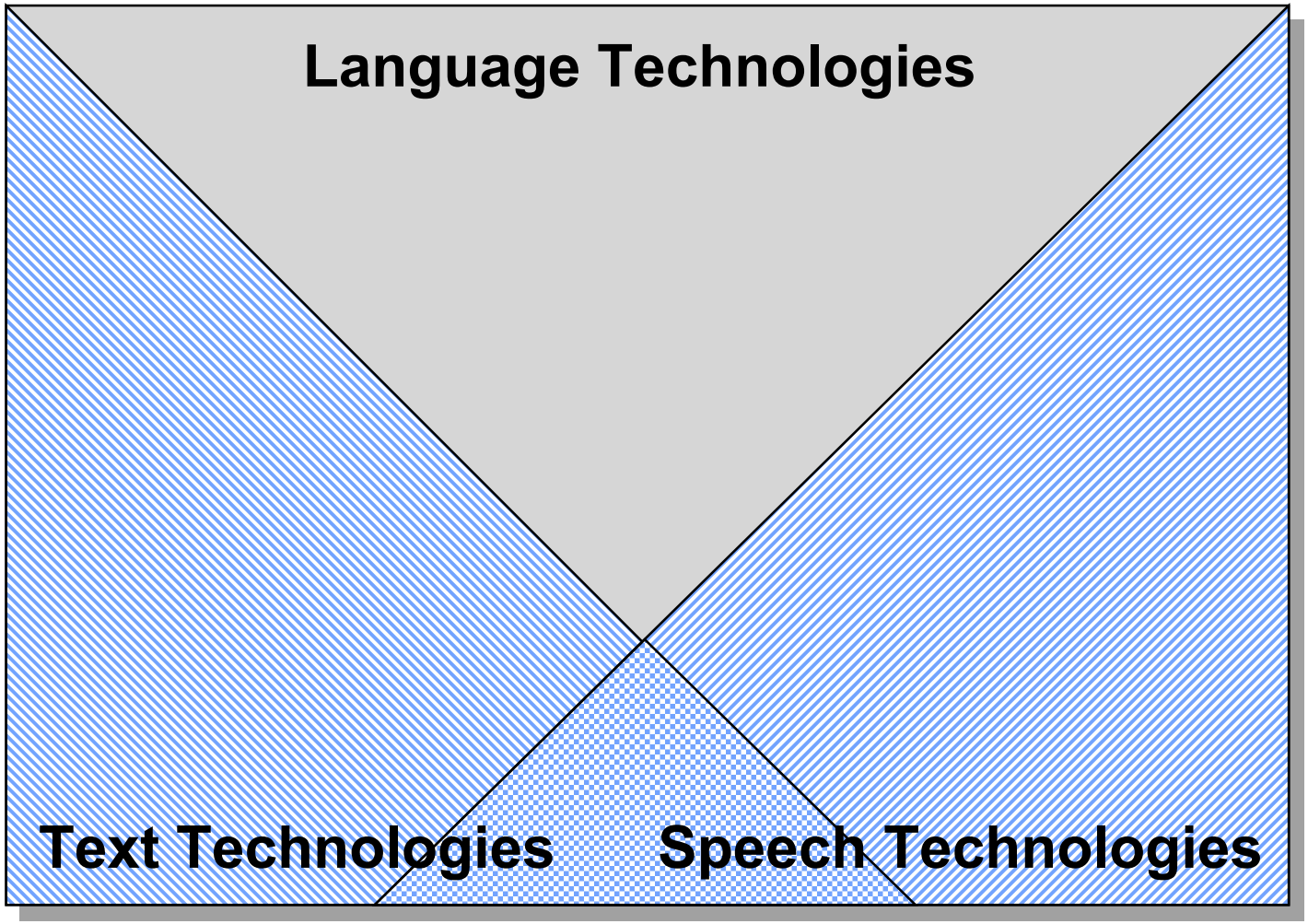
Technology: methods and techniques that together enable some application.

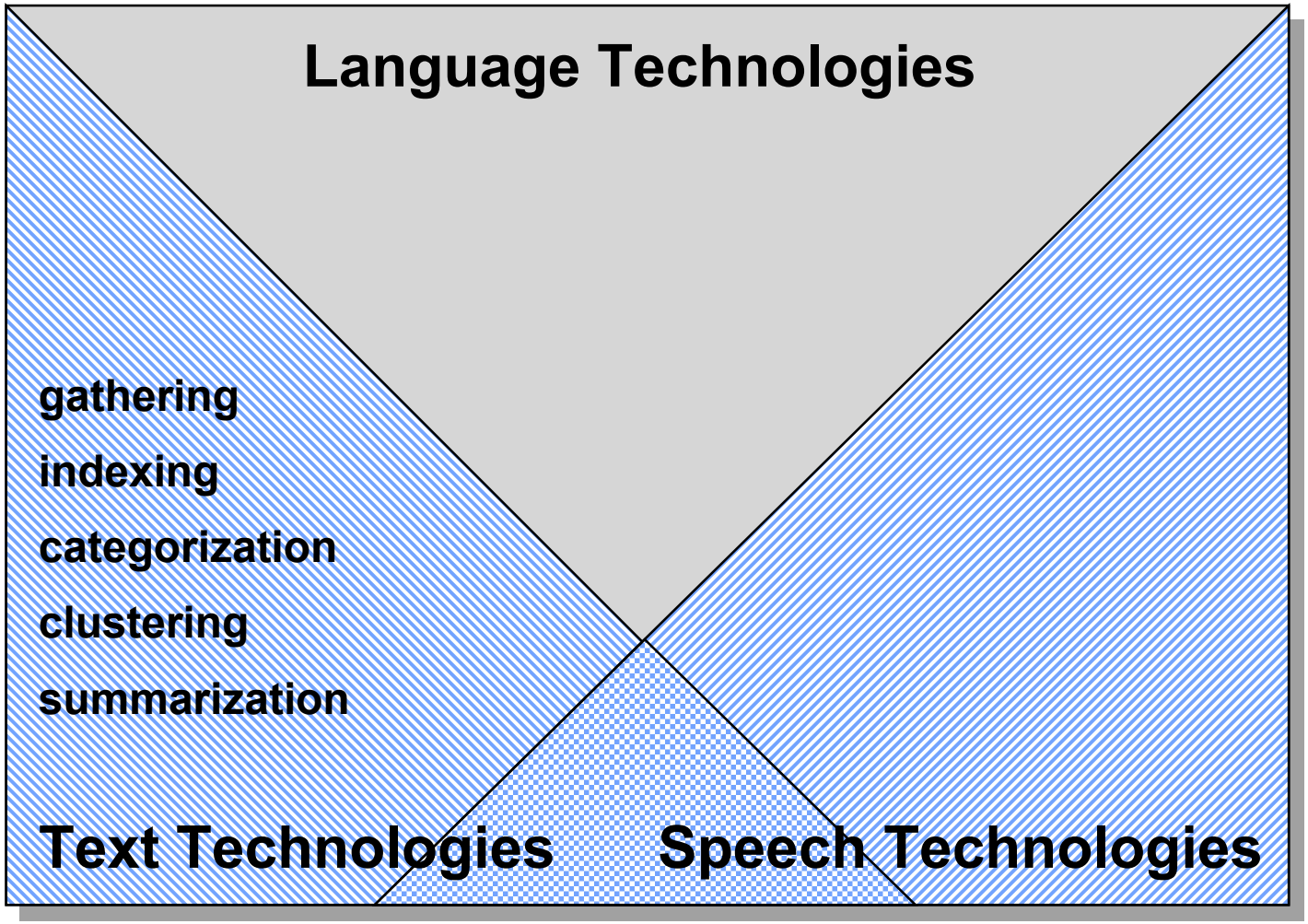
In real life usage of the word there is a continuum between methods and applications.

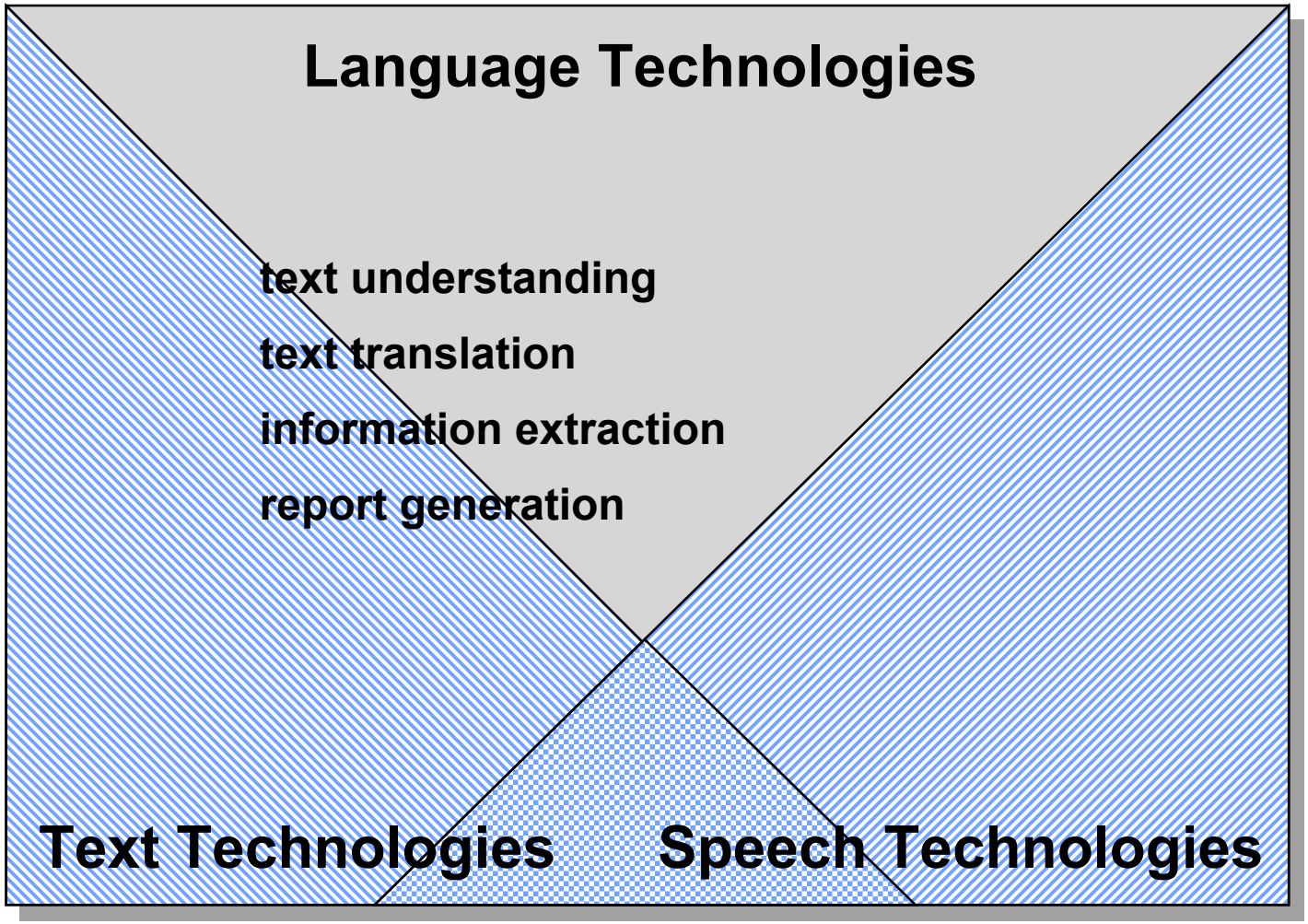
method/technique	finite state transduction
component technology	tokenizer
technology	named entity recognition
	high precision text indexing
application	concept based search engine

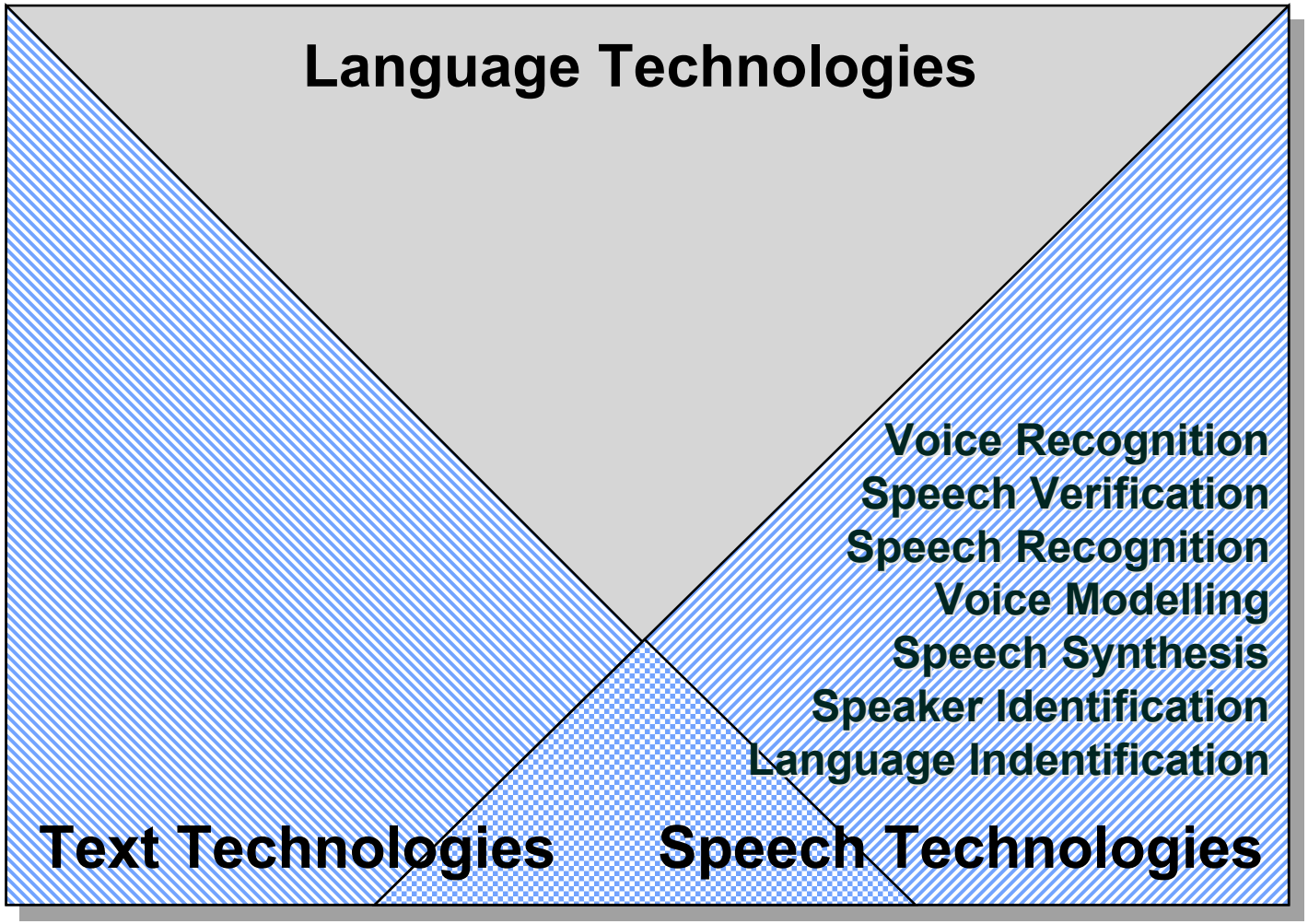
Language Technologies

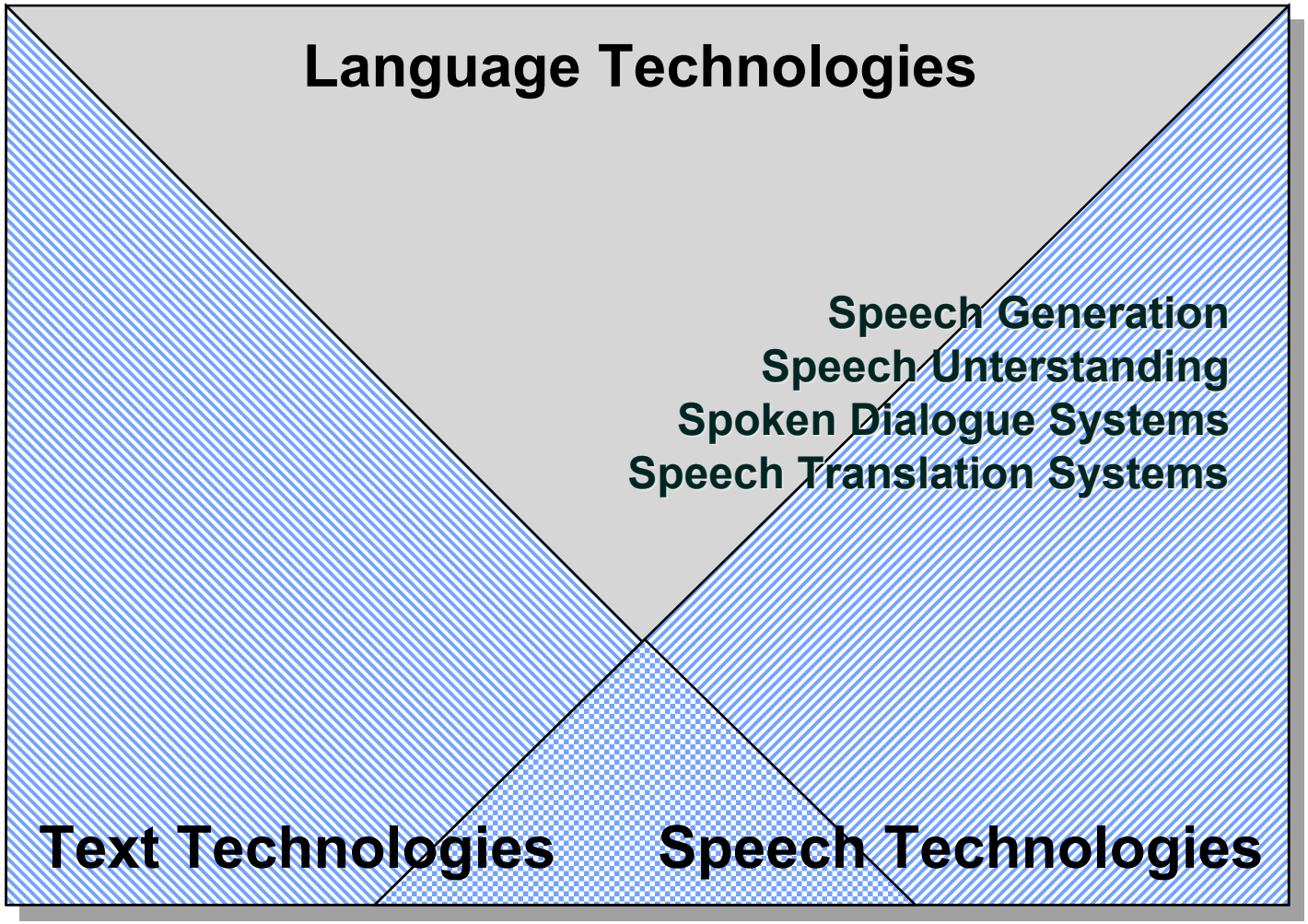


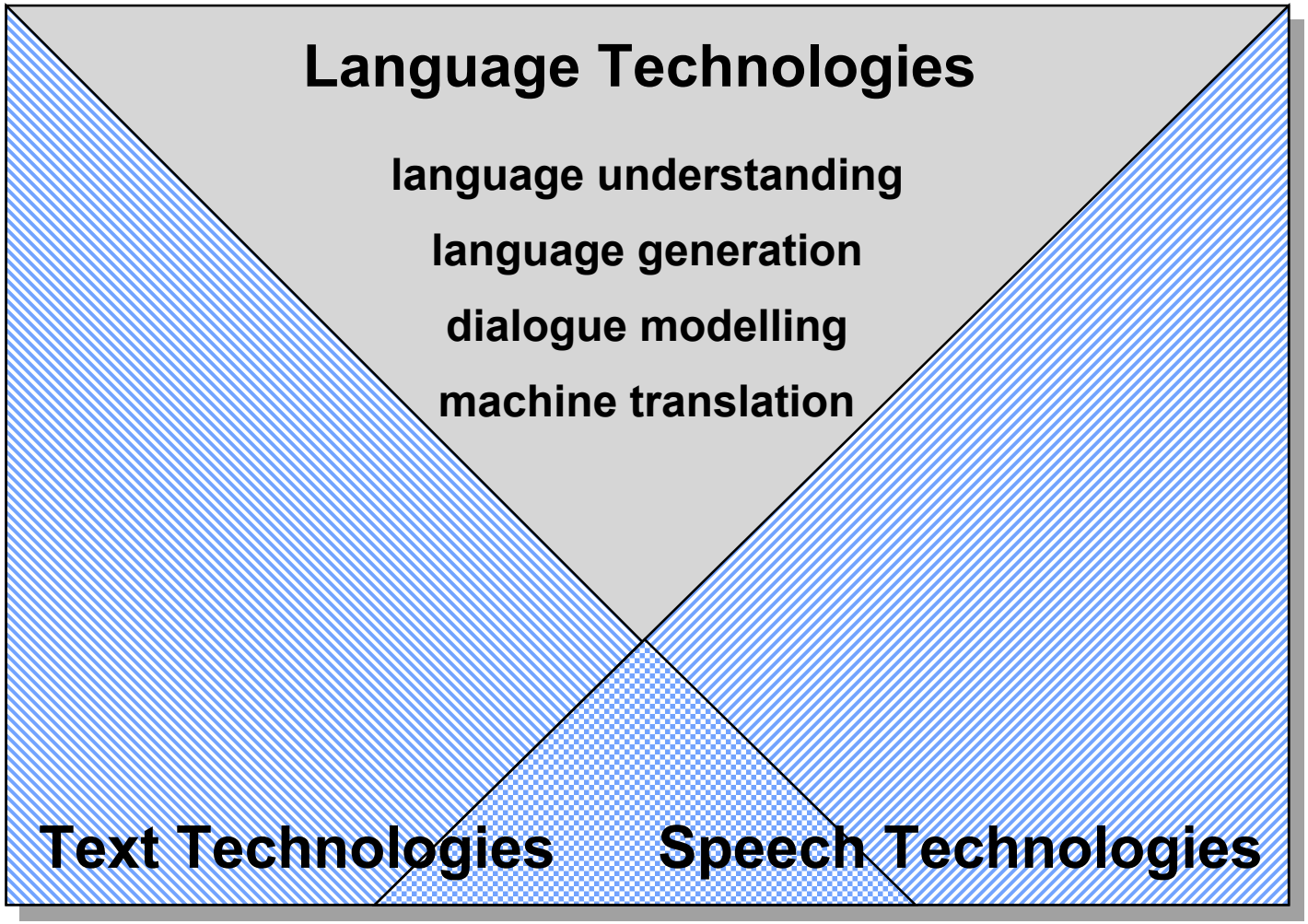



Language Technologies
gathering
indexing
categorization
clustering
summarization
Text Technologies
Speech Technologies

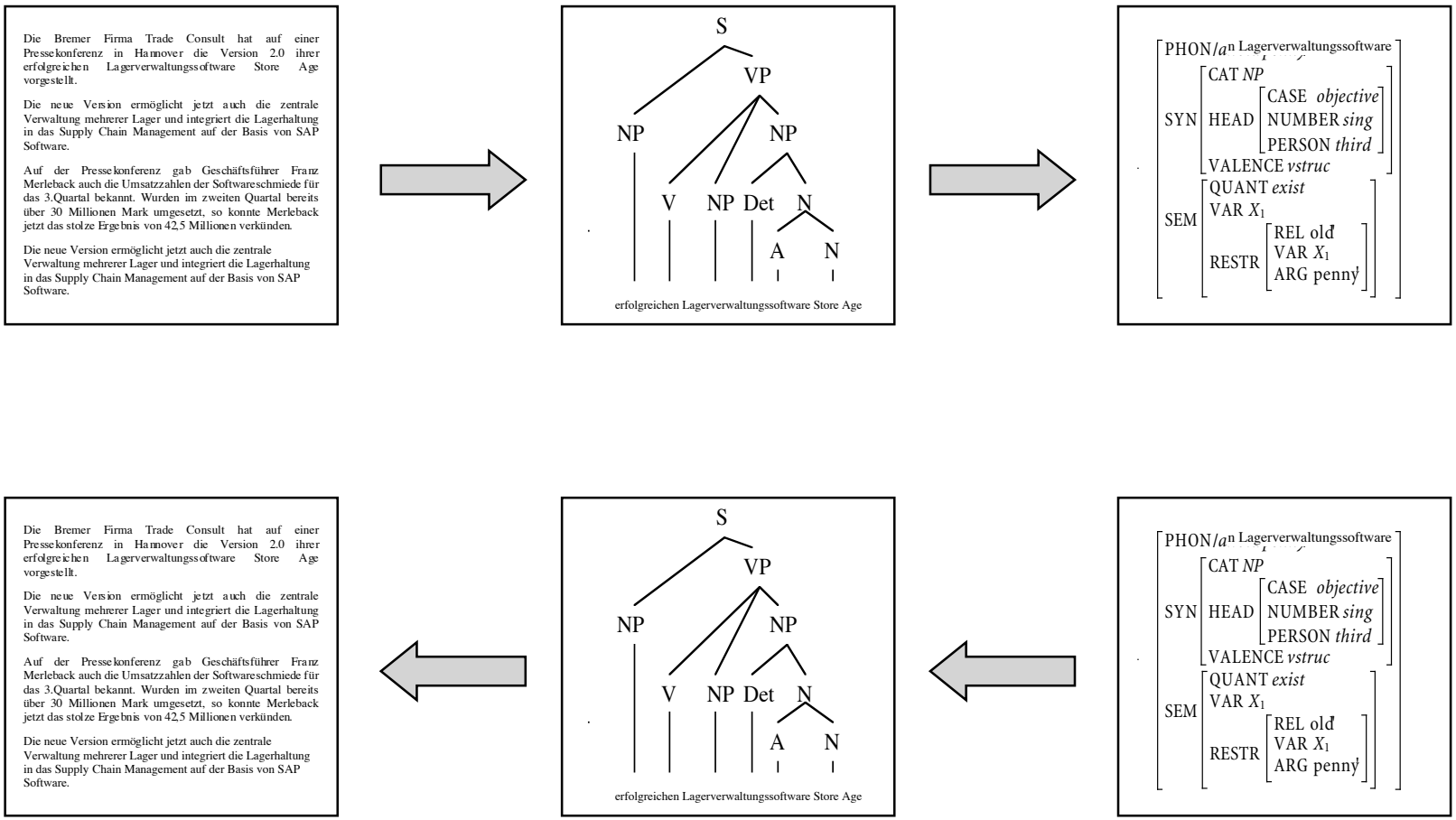

Language Technologies
text understanding
text translation
information extraction
report generation
Text Technologies
Speech Technologies


Language Technologies
Voice Recognition
Speech Verification
Speech Recognition
Voice Modelling
Speech Synthesis
Speaker Identification
Language Indentification
Text Technologies
Speech Technologies


Language Technologies
Speech Generation
Speech Unterstanding
Spoken Dialogue Systems
Speech Translation Systems
Text Technologies
Speech Technologies



Analysis and Generation



Language Technologies

UNIVERSITÄT DES SAARLANDES

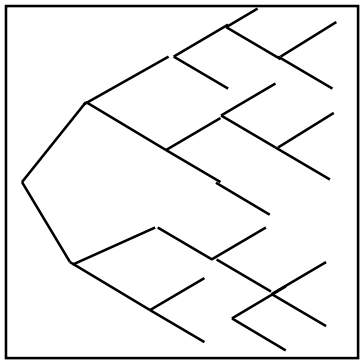
Die Bremer Firma Trade Consult hat auf einer Pressekonferenz in Hannover die Version 2.0 ihrer erfolgreichen Lagerverwaltungssoftware Store Age vorgestellt.

Die neue Version ermöglicht jetzt auch die zentrale Verwaltung mehrerer Lager und integriert die Lagerhaltung in das Supply Chain Management auf der Basis von SAP Software.

Auf der Pressekonferenz gab Geschäftsführer Franz Merleback auch die Umsatzzahlen der Softwareschmiede für das 3.Quartal bekannt. Wurden im zweiten Quartal bereits über 30 Millionen Mark umgesetzt, so konnte Merleback jetzt das stolze Ergebnis von 42,5 Millionen verkünden.

Die neue Version ermöglicht jetzt auch die zentrale Verwaltung mehrerer Lager und integriert die Lagerhaltung in das Supply Chain Management auf der Basis von SAP Software.

building an index



Die Bremer Firma Trade Consult hat auf einer Pressekonferenz in Hannover die Version 2.0 ihrer erfolgreichen Lagerverwaltungssoftware Store Age vorgestellt.

Die neue Version ermöglicht jetzt auch die zentrale Verwaltung mehrerer Lager und integriert die Lagerhaltung in das Supply Chain Management auf der Basis von SAP Software.

Auf der Pressekonferenz gab Geschäftsführer Franz Merleback auch die Umsatzzahlen der Softwareschmiede für das 3.Quartal bekannt. Wurden im zweiten Quartal bereits über 30 Millionen Mark umgesetzt, so konnte Merleback jetzt das stolze Ergebnis von 42,5 Millionen verkünden.

Die neue Version ermöglicht jetzt auch die zentrale Verwaltung mehrerer Lager und integriert die Lagerhaltung in das Supply Chain Management auf der Basis von SAP Software.

extracting the topic

Trade Consult Umsatzzahlen

Die Bremer Firma Trade Consult hat auf einer Pressekonferenz in Hannover die Version 2.0 ihrer erfolgreichen Lagerverwaltungssoftware Store Age vorgestellt.

Die neue Version ermöglicht jetzt auch die zentrale Verwaltung mehrerer Lager und integriert die Lagerhaltung in das Supply Chain Management auf der Basis von SAP Software.

Auf der Pressekonferenz gab Geschäftsführer Franz Merleback auch die Umsatzzahlen der Softwareschmiede für das 3.Quartal bekannt. Wurden im zweiten Quartal bereits über 30 Millionen Mark umgesetzt, so konnte Merleback jetzt das stolze Ergebnis von 42,5 Millionen verkünden.

Die neue Version ermöglicht jetzt auch die zentrale Verwaltung mehrerer Lager und integriert die Lagerhaltung in das Supply Chain Management auf der Basis von SAP Software.

extracting relations

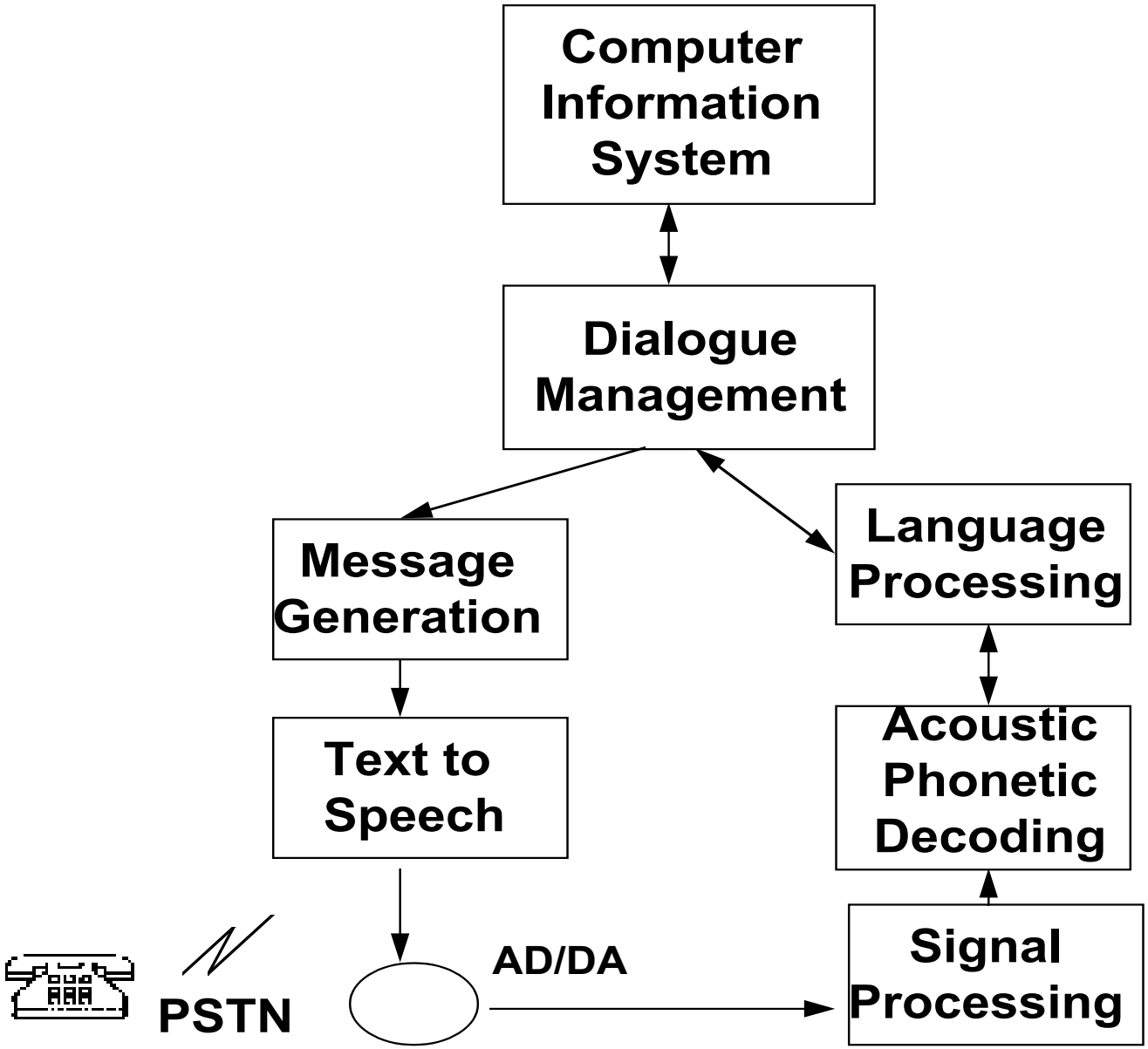
Firma	96Q4	1996	97Q1	97Q2	97Q3	97Q4	1997	Diff
Hahnemann		105 Mio					110Mio	
Trade Consult				30 Mio	42,5Mio			
Z&M					12,0Mio	14 Mio		

IE Result

UNIVERSITÄT DES SAARLANDES

Firma	96Q4	1996	97Q1	97Q2	97Q3	97Q4	1997	Diff
ComSoft		120Mio					110Mio	
Trade Consult				30 Mio	42,5Mio			
Z&M					71,0Mio			

95%-98%	Correct recognition of word categories (part-of-speech-tagging)
85%-98%	recognition of names of people, companies, places, products (named-entity-recognition)
95%	statistical recognition of major phrases (HMM chunk parsing)
91%	parsing of newspaper texts by statistically trained parsers (probibilistic context free parsing)
40%-60%	deep parsing of newspaper texts (HPSG or LFG parsing with large lexicon)



Problematic Areas

- Plan Recognition
- Ellipsis
- Anaphora
- Cooperativity
- Clarification Dialogues

Cognitive Approaches

- Mirror human performance.
- Improve machine performance.
- Understand human processing.
- Understand why language is as it is.

- Linguistic Competence:

  The knowledge a speaker has to possess in order to master a language.

  The system of rules, principles and contraints that constitute the grammar of a language

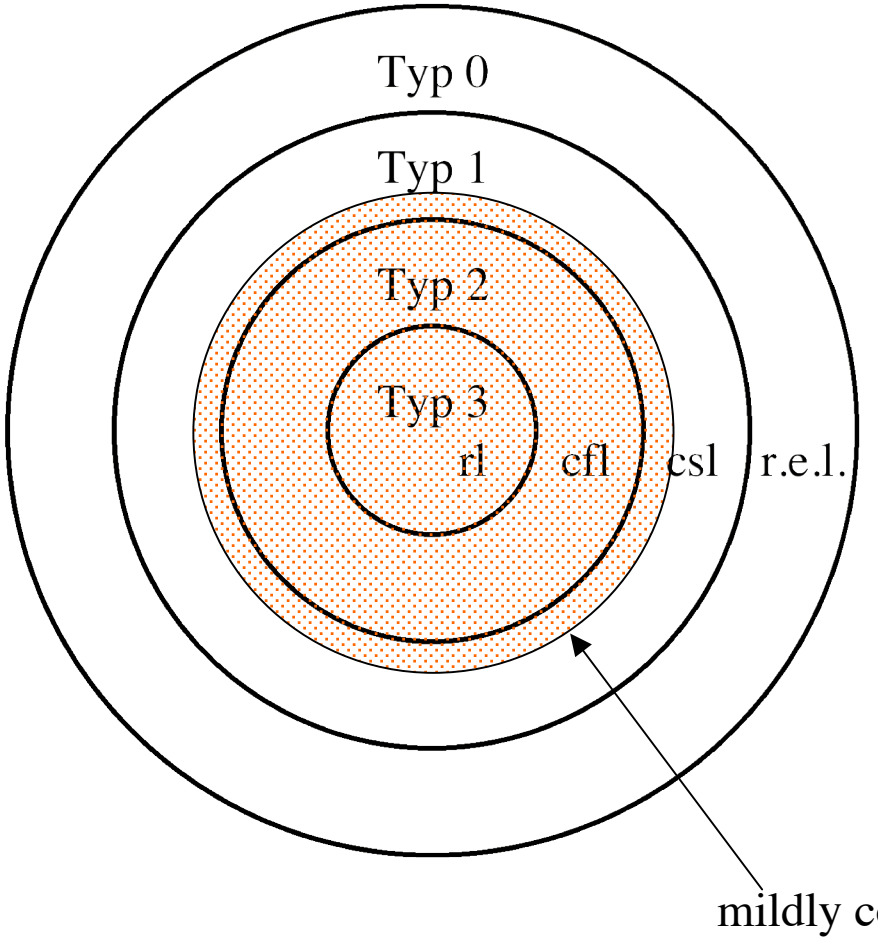
  The finite definition of an infinite natural language.

- Linguistic Performance

  The mechanisms and processes underlying actual human language use, i.e., sentence production and comprehension.

  This includes the influence (assisting or limitating) of other cognitive processes such as reasoning, perception and action as well as other tasks.

Computational Complexity of Human Language



Typ 0: recursively enumerable sets

Typ 1: contextsensitive languages

Typ 2: context-free languages

Typ 3: regular languages

A Paradoxon

The predominant linguistic grammar formalisms have a polynomial or exponential worst-case parsing complexity.

(for CF languages On^3, where O is a constant and n is the length of the sentence)

Certain phenomena increase the parsing complexity.

Humans seem to analyze sentences in real time.

Why does syntax possess phenomena that make life harder?

Competence vs. Performance

The competence-performance distinction was necessary for the development of modern formal linguistics.

The majority of sentences generated (or accepted) by formal grammars cannot be generated or analyzed by human speakers.

Why does the grammar contain syntactic phenomena that make processing harder?

examples:

- long-distance dependencies
- “free” word order
- right-extraposition
- parenthetical constructions

Hypothesis: When we understand the functional reasons for the evolution of these phenomena, our view of grammar and processing will change.

A Performance Model Should Explain...

- why many ungrammatical sentences get produced
  - speech errors, grammar errors
- why many ungrammtical sentences are understood
  - communication with non-native speakers and children
- why many grammtical sentences are never produced
  - preferences in generation
- why many grammatical sentences cannot be understood
  - garden path sentences
- how processing is structured
  - efficiency and flow of control
- which effort do the steps or components require
  - dependence on other cognitive efforts (load)

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.

The horse raced past the barn fell.

but:

The clothes put on the rack smelled.

- Humans produce and comprehend sentences incrementally.
- We understand parts of the sentence while we still listen to the rest of the sentence.
- We already articulate parts of the sentence while we are still thinking about the rest of our statement.
- Both for understanding and generation we would prefer incremental algorithms.
- However, most approaches to language processing are sequential (pipeline) models without feedback to earlier components.

Determinism vs. Underspecification

- The astronomer married a star.
- The canoe floated down the river sank.
- The three selected special agents speak two foreign languages nearly without an accent.

The competence-performance distinction was necessary for the development of modern formal linguistics.

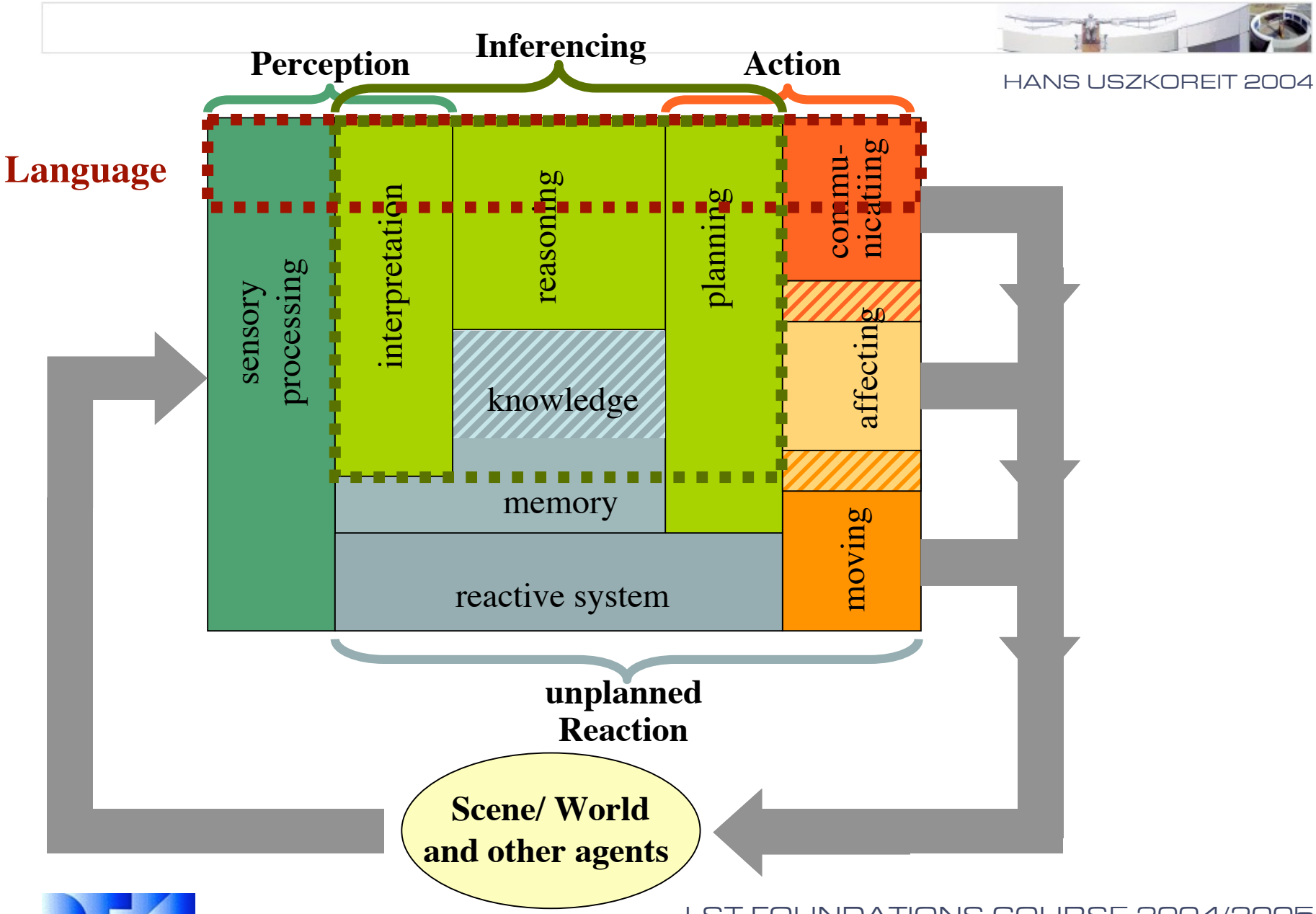
The majority of sentences generated (or accepted) by formal grammars cannot be generated or analysed by human speakers.

Why does the grammar contain syntactic phenomena that make processing harder?

examples:

- long-distance dependencies
- “free” word order
- right-extraposition
- parenthetical constructions

Hypothesis: When we understand the functional reasons for the evolution of these phenomena, our view of grammar and processing will change.


Perception
Inferencing
Action
Language
sensory processing
interpretation
reasoning
planning
commu-nicating
knowledge
affecting
memory
moving
reactive system
unplanned Reaction
Scene/ World and other agents

- ☆ Derive convincing solutions to the ambiguity problem:
  - by fusing constraints from knowledge (reasoning about context and general knowledge), perception, language (context and grammar), intention and attention
- ☆ Study realistic and cognitively plausible models of language learning
  - by combining learning about spatial concepts, objects and language in communicative learning situations
- ☆ Investigate realistic communication situations and get at real language understanding
- ☆ Gain insights on the functional properties of language