Map of the Field

October 2008

Dietrich Klakow

based on slides by Hans Uszkoreit

WHAT IS COMPUTATIONAL LINGUISTICS?

- ***Computational linguistics (CL)*** is a discipline between linguistics and computer science which is concerned with the computational aspects of the human language.
- It belongs to the cognitive sciences and overlaps with the field of ***artificial intelligence (AI)***, a branch of ***computer science*** aiming at computational models of human cognition.
- Computational linguistics has applied and theoretical components.

Human language is a most exciting and demanding puzzle

- ***Theoretical CL*** takes up issues in ***theoretical linguistics*** and ***cognitive science.*** It deals with formal theories about the linguistic knowledge that a human needs for generating and understanding language. Today these theories have reached a degree of complexity that can only be managed by employing computers. Computational linguists develop formal models simulating aspects of the human language and implement them as computer programmes. These programmes constitute the basis for the evaluation and further development of the theories.

- In addition to linguistic theories, findings from ***cognitive psychology*** play a major role in simulating linguistic competence. Within psychology, it is mainly the area of ***psycholinguistics***that examines the cognitive processes constituting human language use. The relevance of computational modelling for psycholinguistic research is reflected in the emergence of a new subdiscipline: computational psycholinguistics.

We teach computers to communicate with people

- ***Applied CL*** focusses on the practical outcome of modelling human language use. The methods, techniques, tools and applications in this area are often subsumed under the term ***language engineering*** or ***(human) language technology.*** Although existing CL systems are far from achieving human ability, they have numerous possible applications.

- The goal is to create software products that have some knowledge of human language. Such products are going to change our lives. They are urgently needed for improving human-machine interaction since the main obstacle in the interaction beween human and computer is a communication problem. Today's computers do not understand our language but computer languages are difficult to learn and do not correspond to the structure of human thought. Even if the language the machine understands and its domain of discourse are very restricted, the use of human language can increase the acceptance of software and the productivity of its users.

Friendly software should listen and speak

- ☆ Natural language interfaces enable the user to communicate with the computer in French, English, German, or another human language. Some applications of such interfaces are database queries, information retrieval from texts, so-called expert systems, and robot control.

- ☆ Current advances in the recognition of spoken language improve the usability of many types of natural language systems. Communication with computers using spoken language will have a lasting impact upon the work environment, completely new areas of application for information technology will open up.

- ☆ However, spoken language needs to be combined with other modes of communication such as pointing with mouse or finger. If such multimodal communication is finally embedded in an effective general model of cooperation,we have succeeded in turning the machine into a partner.

Machines can also help people communicate with each other

- ☆ Much older than communication problems between human beings and machines are those between people with different mother tongues.
- ☆ One of the original aims of applied computational linguistics has always been fully automatic translation between human languages. From bitter experience scientists have realized that they are still far away from achieving the ambitious goal of translating unrestricted texts.
- ☆ Nevertheless computational linguists have created software systems that simplify the work of human translators and clearly improve their productivity. Less than perfect automatic translations can also be of great help to information seekers who have to search through large amounts of texts in foreign languages.

Language is the fabric of the web

- ☆ The rapid growth of the Internet/WWW and the emergence of the information society poses exciting new challenges to language technology.
- ☆ Although the new media combine text, graphics, sound and movies, the whole world of multimedia information can only be structured, indexed and navigated through language. For browsing, navigating, filtering and processing the information on the web, we need software that can get at the contents of documents.
- ☆ Language technology for content management is a necessary precondition for turning the wealth of digital information into collective knowledge.
- ☆ The increasing multilinguality of the web constitutes an additional challenge for our discipline. The global web can only be mastered with the help of multilingual tools for indexing and navigating. Systems for cross-lingual information and knowledge management will surmount language barriers for e-commerce, education and international cooperation.

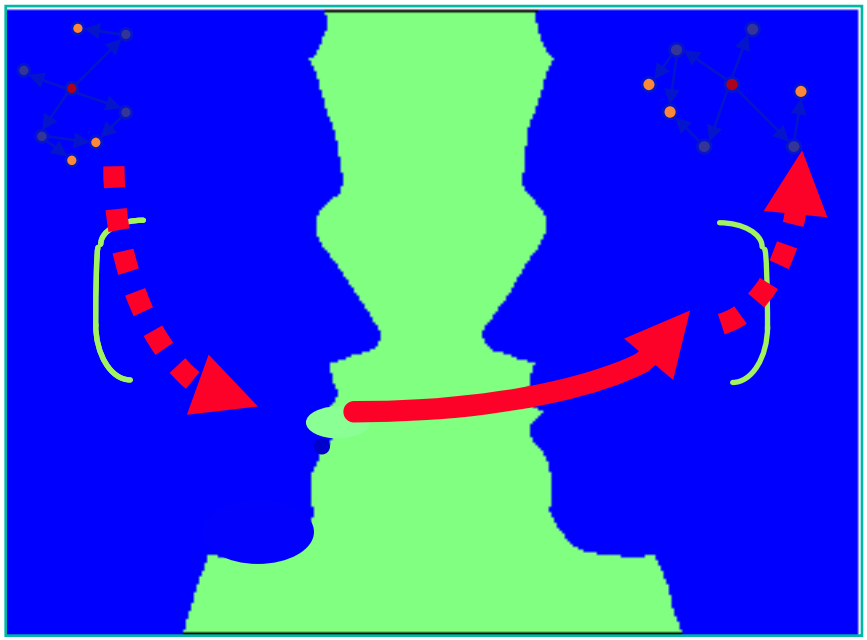
CL combines ambitious visions and realistic applications

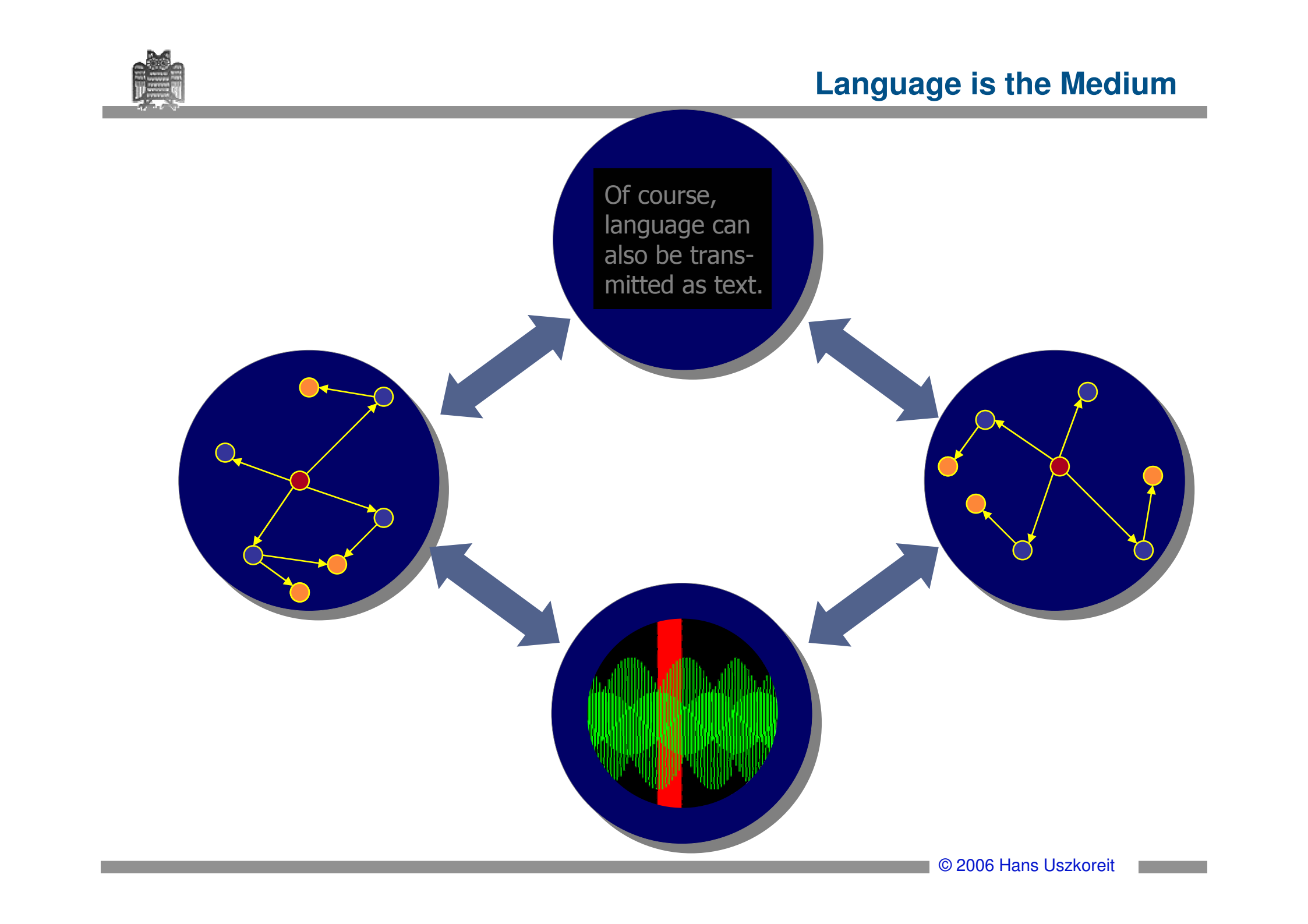
- ☆ We still do not know very well how people produce and comprehend language. Yet our understanding of the intricate mechanisms that underly human language processing keeps growing.
- ☆ Modelling such mechanisms on a computer also helps us to discover and formally describe hidden properties of human language that are relevant for any kind of language processing including many useful software applications.
- ☆ Our long term goal is the deep understanding of human language and powerful intelligent linguistic applications. However, even today's language technologies full of clever short cuts and shallow processing techniques can be turned into badly needed software products.

The young field's appeal is its diversity

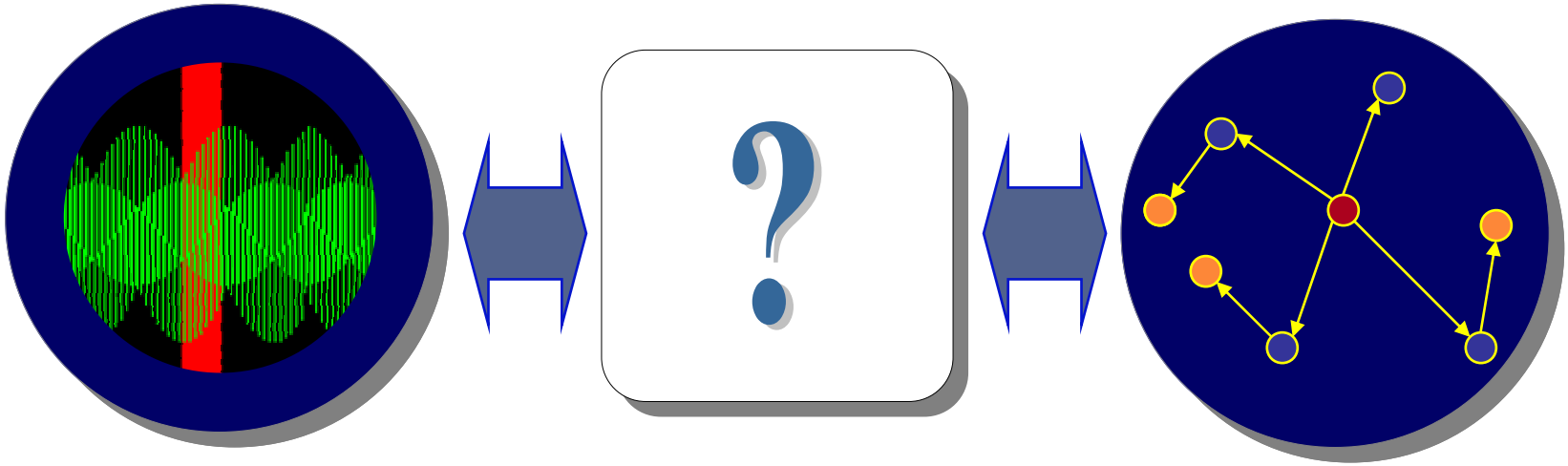
- ☆ For many students and practitioners of computational linguistics the special attraction of the discipline is the combination of expertise from the humanities, natural and behavioural sciences, and engineering.
- ☆ Scientific approaches and practical techniques come from linguistics, computer science, psychology, and mathematics.
- ☆ At some universities the subject is taught in computer science at others it belongs to linguistics or cognitive science. In addition there is a small but growing number of programs and departments dedicated solely to computational linguistics.

THE MIRACLE




Language is the Medium
Of course, language can also be trans-mitted as text.
© 2006 Hans Uszkoreit

WHAT HAPPENS IN BETWEEN?

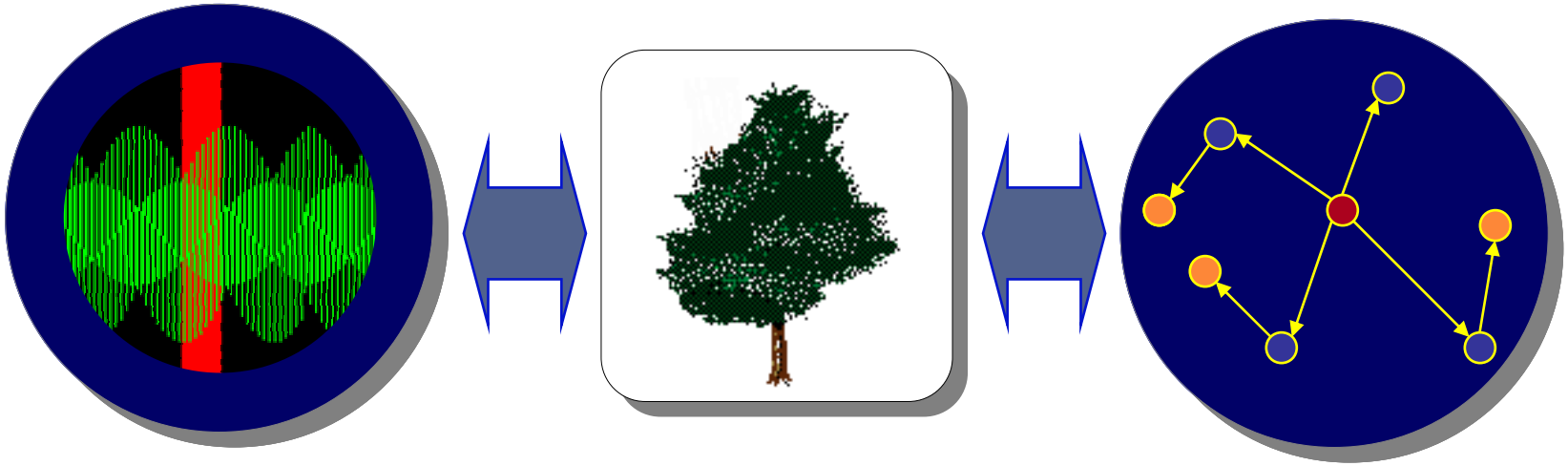


sound waves

Grammar

activation of concepts

WHAT HAPPENS IN BETWEEN?

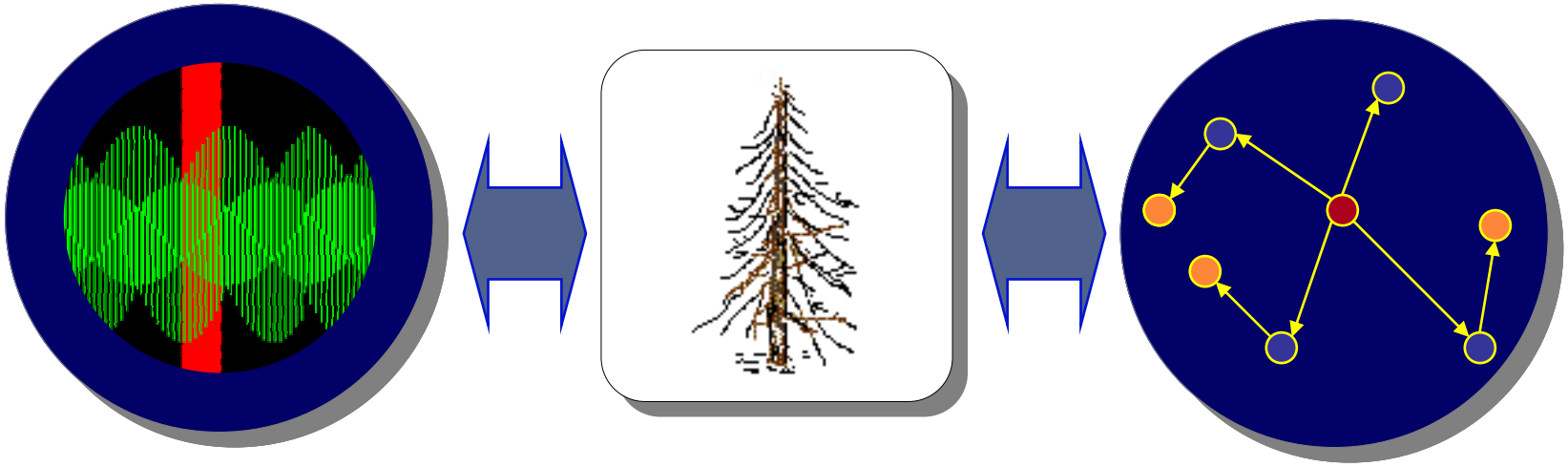


sound waves

Grammar

activation of concepts

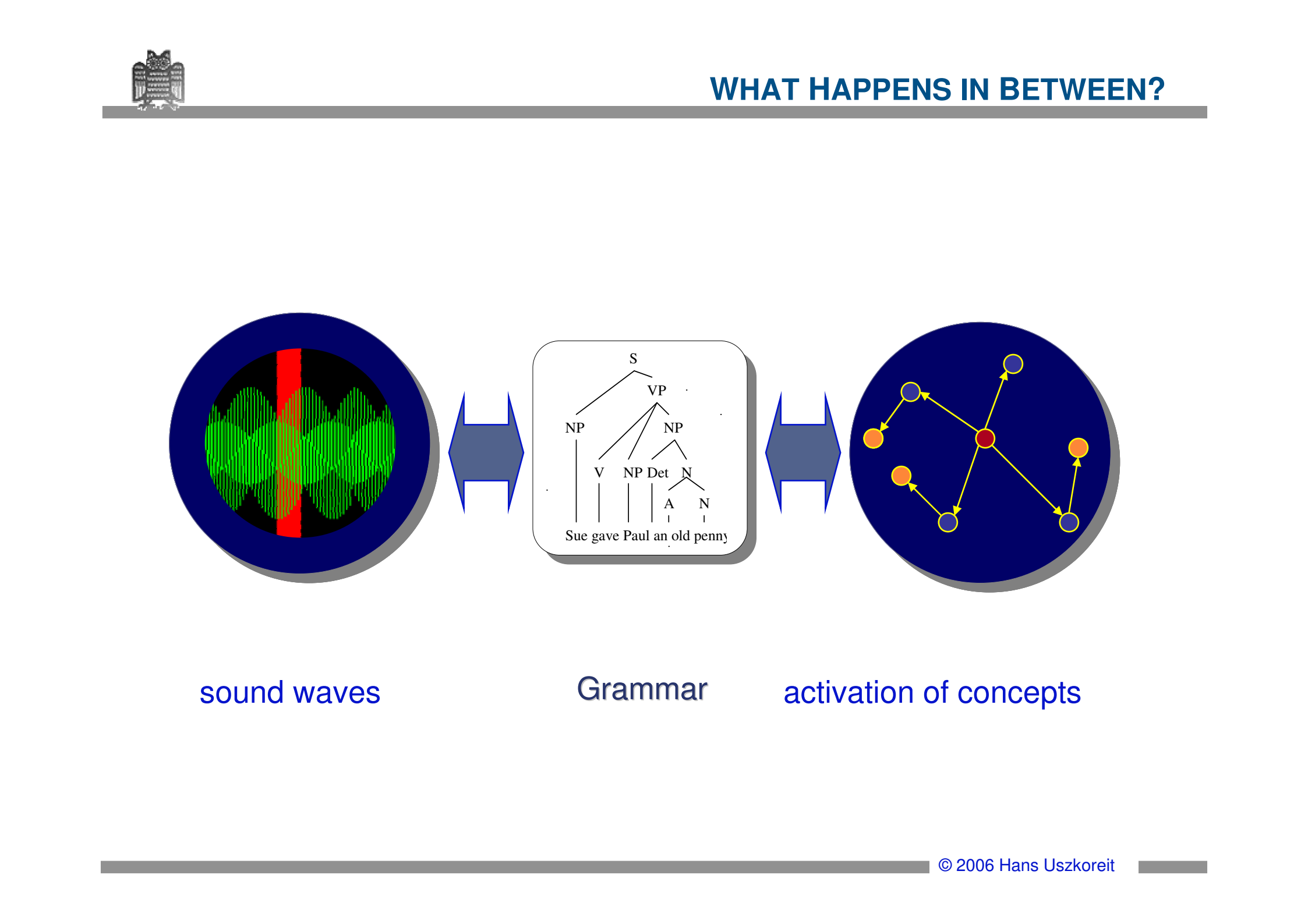
WHAT HAPPENS IN BETWEEN?



sound waves

Grammar

activation of concepts



WHAT HAPPENS IN BETWEEN?

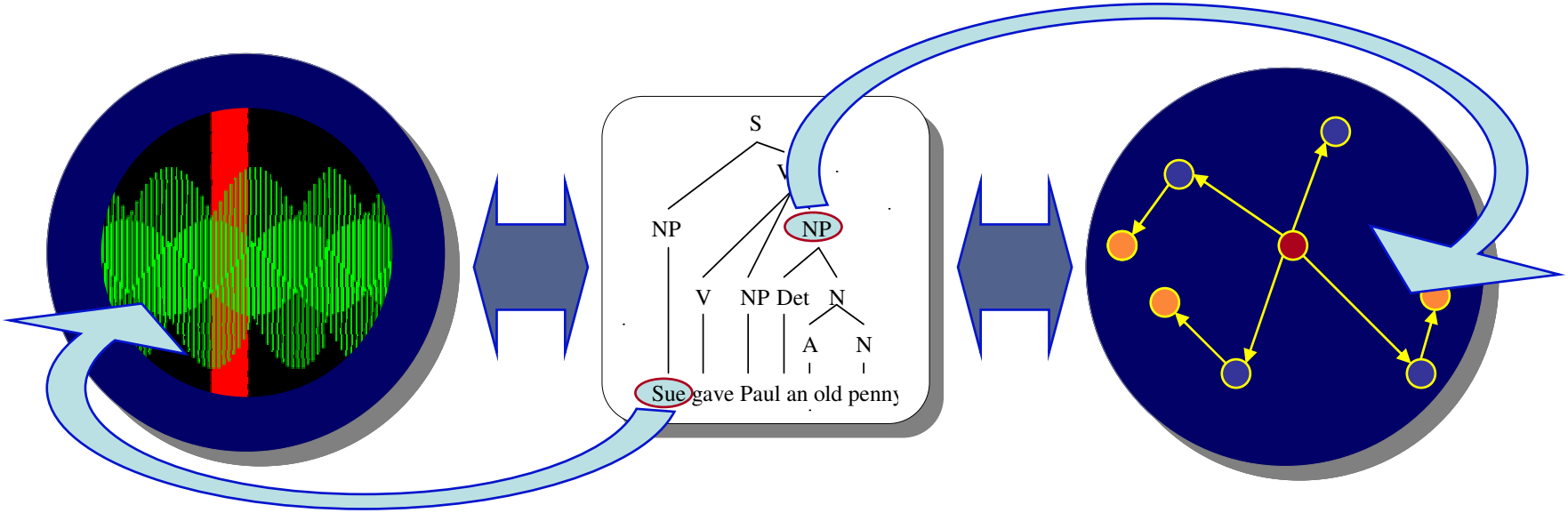
sound waves

Grammar

activation of concepts

WHAT HAPPENS IN BETWEEN?

semantic interpretation

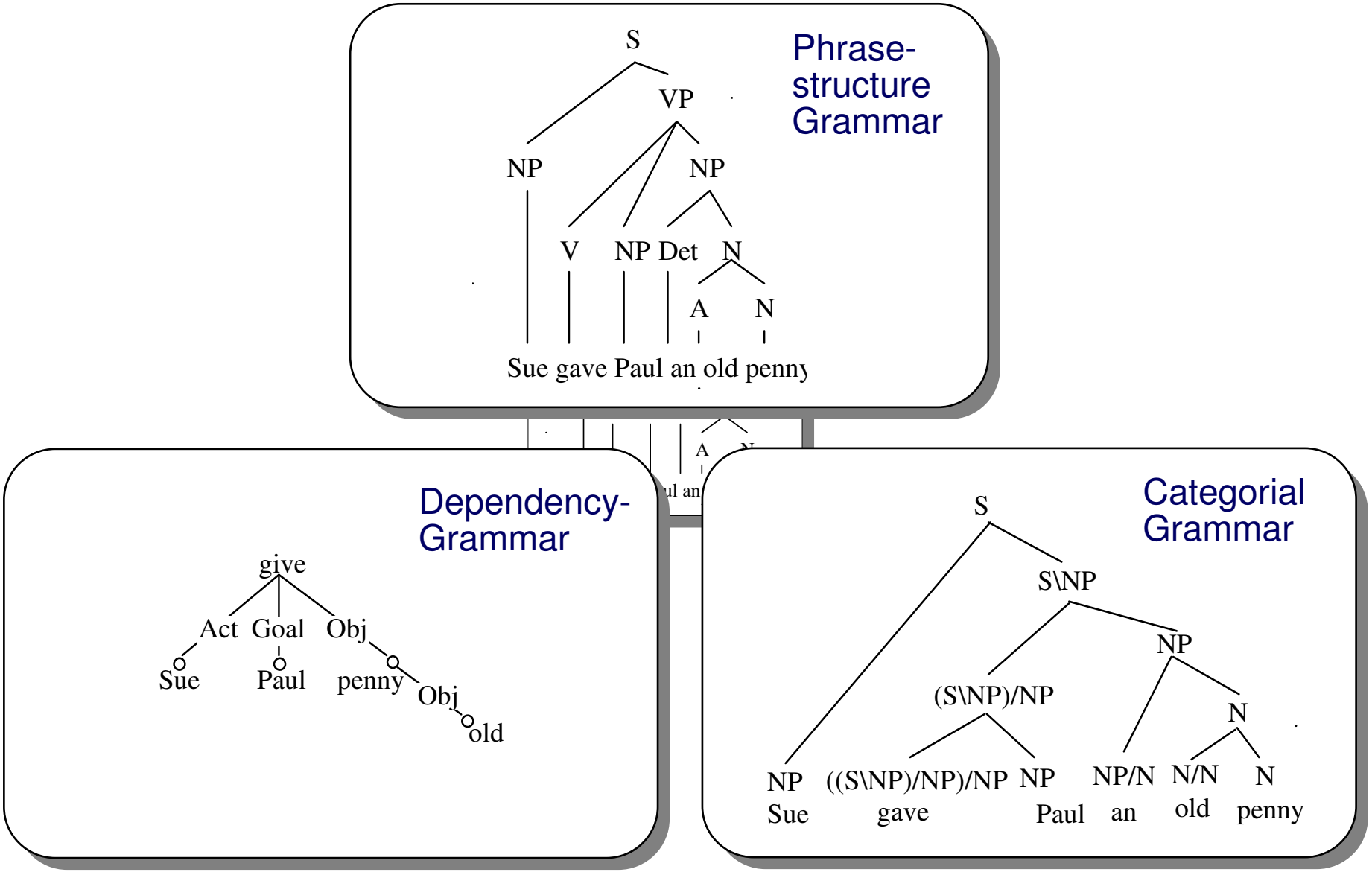


Signal/phonology/morphology

sound waves

Grammar

activation of concepts



Size

- 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

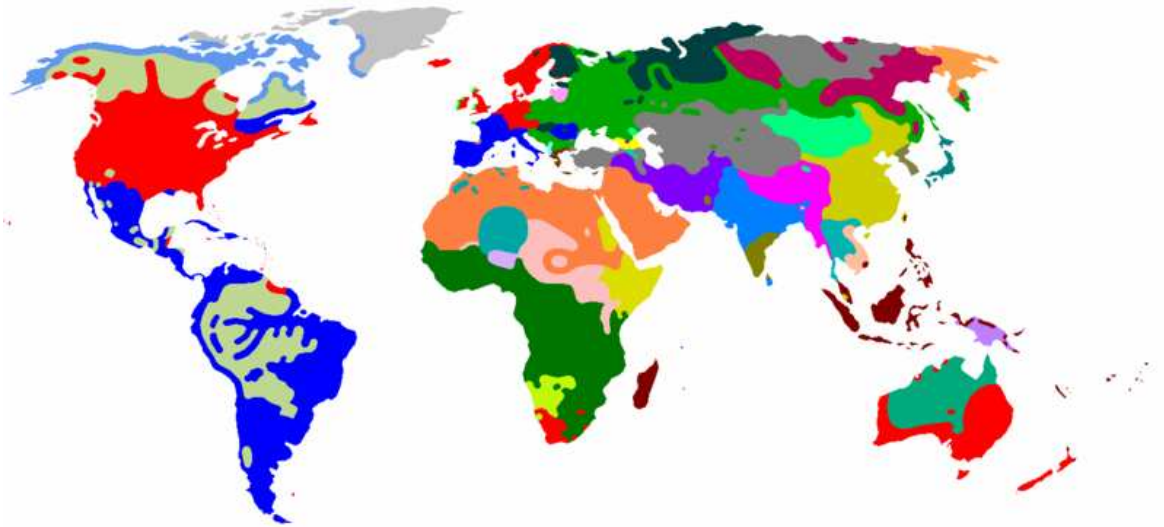
  (60% coverage of newspaper texts)

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

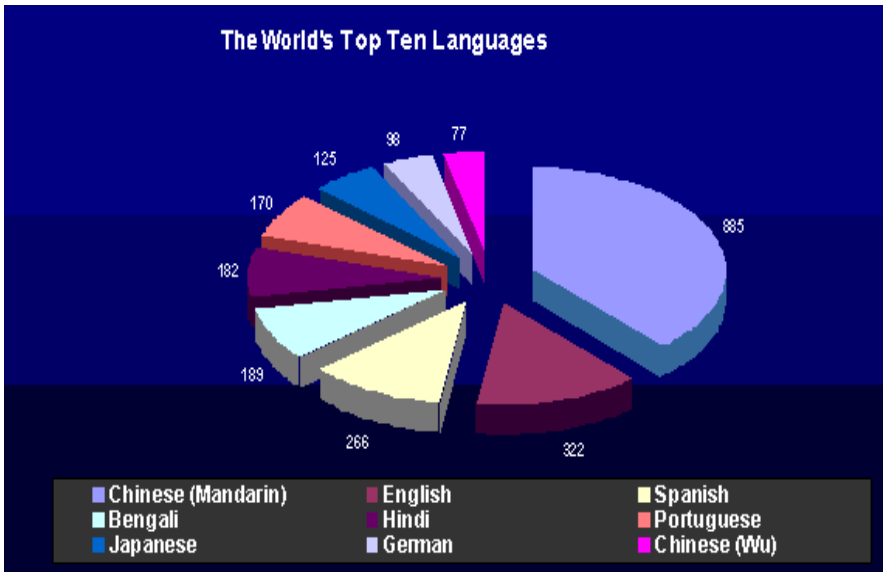
The Tower of Babel

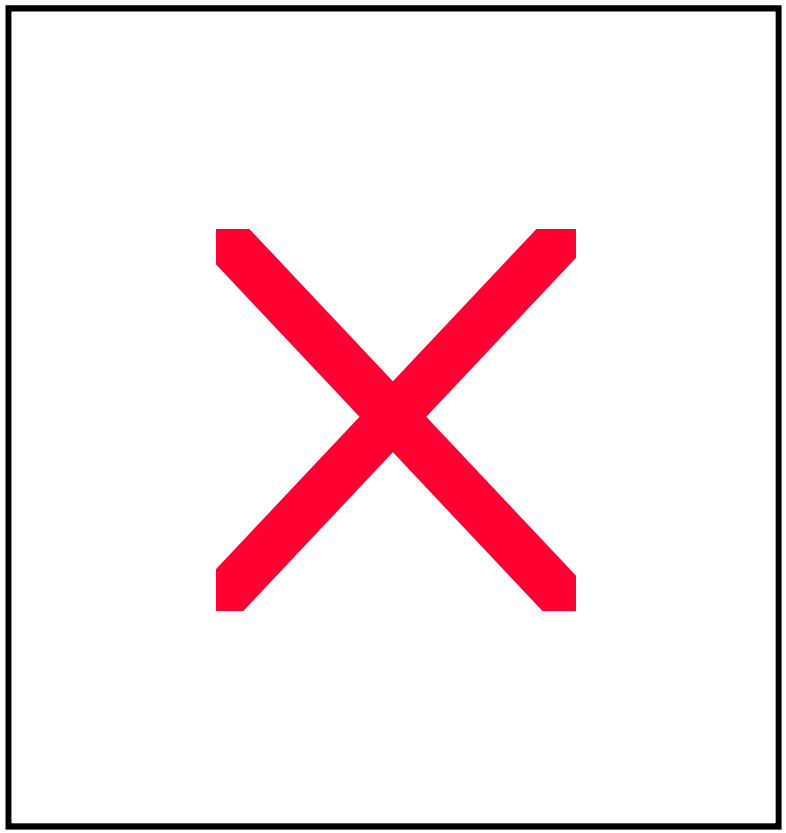
How Many Languages ?

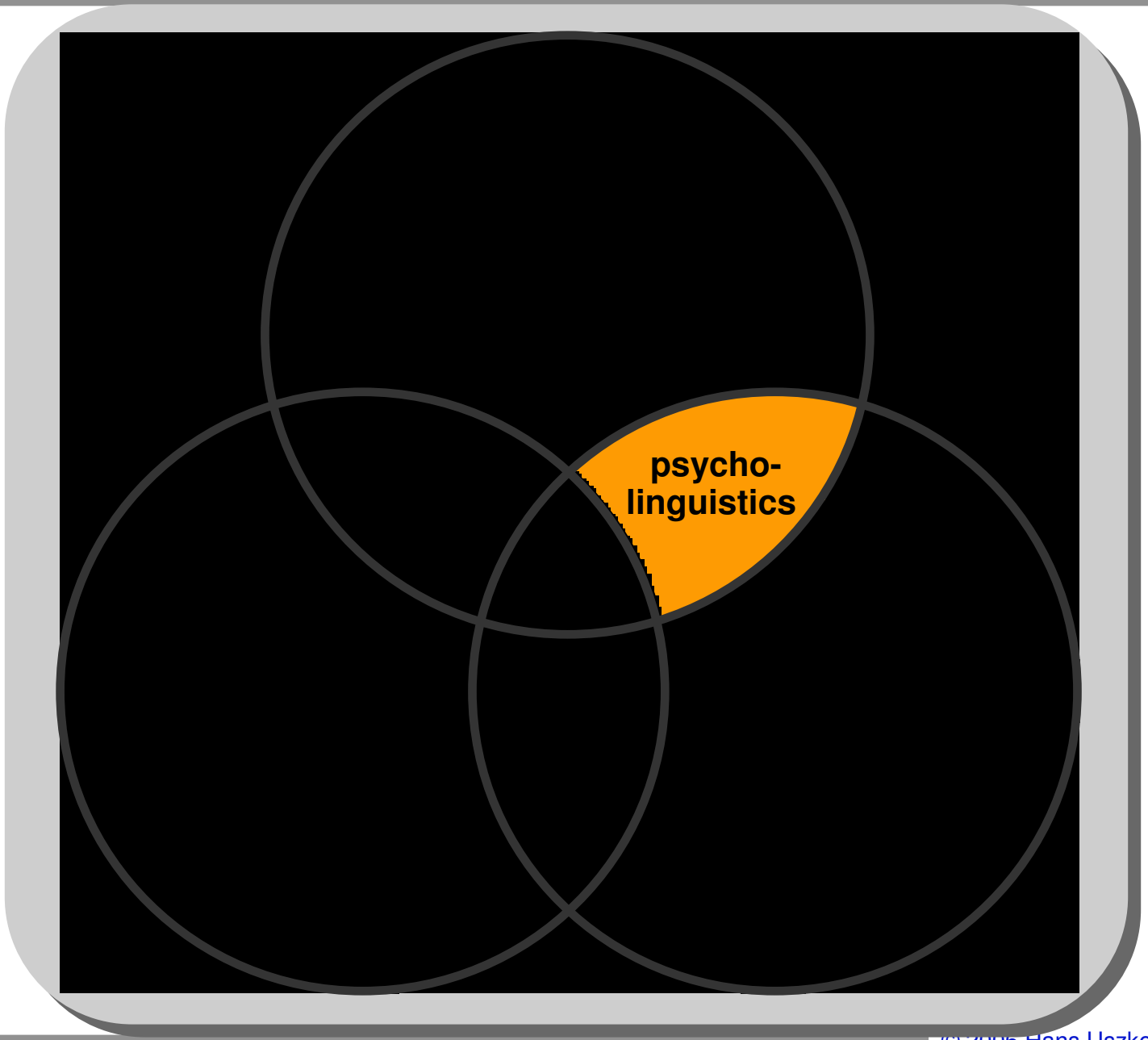
- According to Ethnologue 6,809 languages
- 230 in Europe, 2197 in Asia (832 in Papua-New Guinea)
- Bible translations exist for 2.200 languages
- 250 families of languages (such as Indoeuropean Languages)



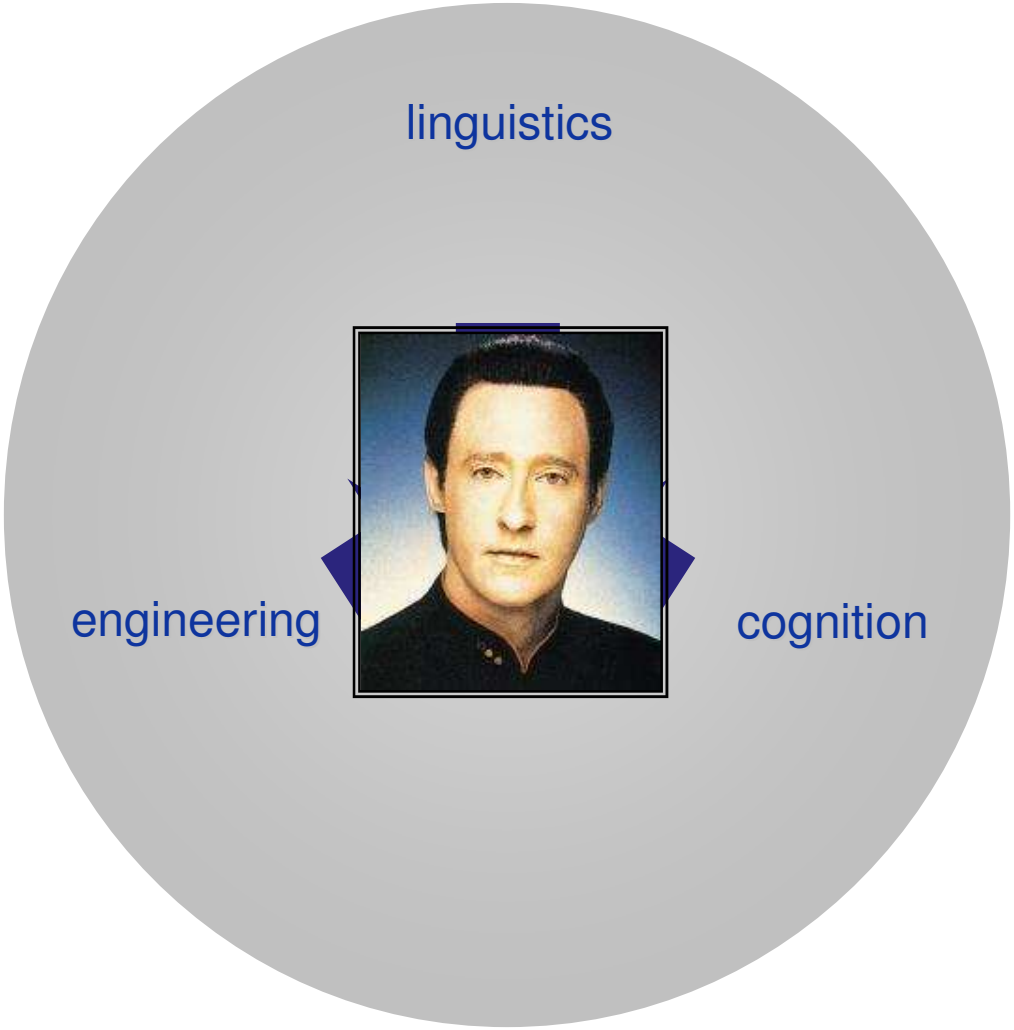
How Many Languages ?

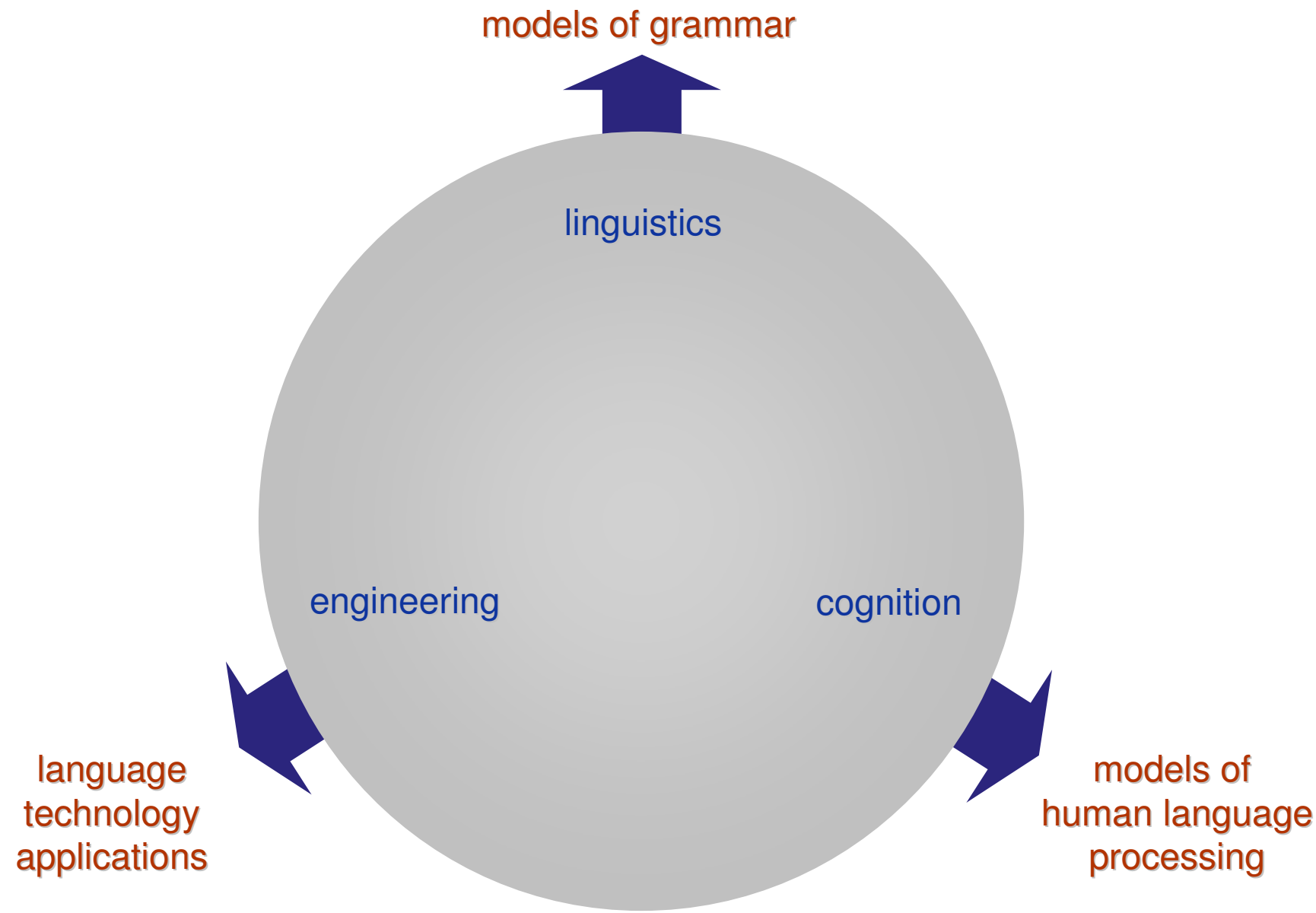






Transdisciplinary Interests





Areas of Linguistics

- According to levels of linguistic description
  - Phonetics
  - Phonology
  - Morphology
  - Syntax
  - Semantics
  - Pragmatics/Discourse
- According to aspects of human language
  - Psycholinguistics
  - Neurolinguistics
  - Historical Linguistics
  - Sociolinguistics
  - Ethnolinguistics
  - Dialectology
  - Applied Linguistics
  - Mathematical Linguistics
  - Computational Linguistics

- According to levels of linguistic description
  - Computational Phonetics
  - Computational Phonology
  - Computational Computational Morphology
  - Computational Syntax
  - Computational Semantics
  - Computational Pragmatics
- According to aspects of human language
  - Computational Psycholinguistics
  - Computational Neurolinguistics
  - Computational Historical Linguistics
  - Computational Sociolinguistics
  - Computational Ethnolinguistics ???
  - Computational Dialectology
  - Computational Applied Linguistics / Applied Computational Linguistics
  - Computational Mathematical Linguistics (funny)

Levels of Description

acoustic form

written form

phonetic or graphemic representation

morpho-phonological processing

syntactic representation

semantic representation

representation of the full meaning

Levels of Processing

acoustic form **written form**

phonetic processing | orthographic processing

phonetic or graphemic representation

morpho-phonological processing

morpho-phonological processing

syntactic processing

syntactic representation

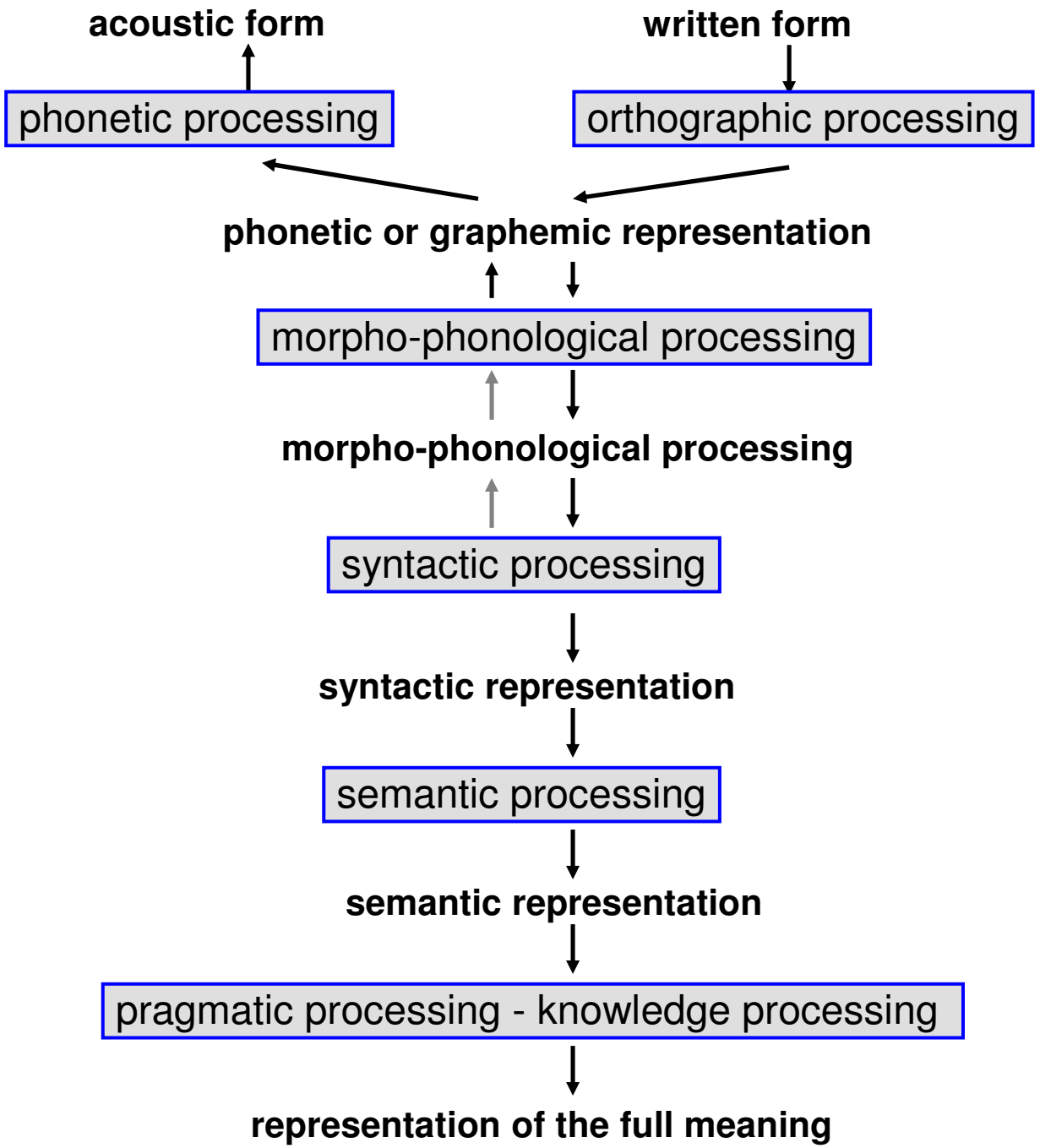
semantic processing

semantic representation

pragmatic processing - knowledge processing

representation of the full meaning

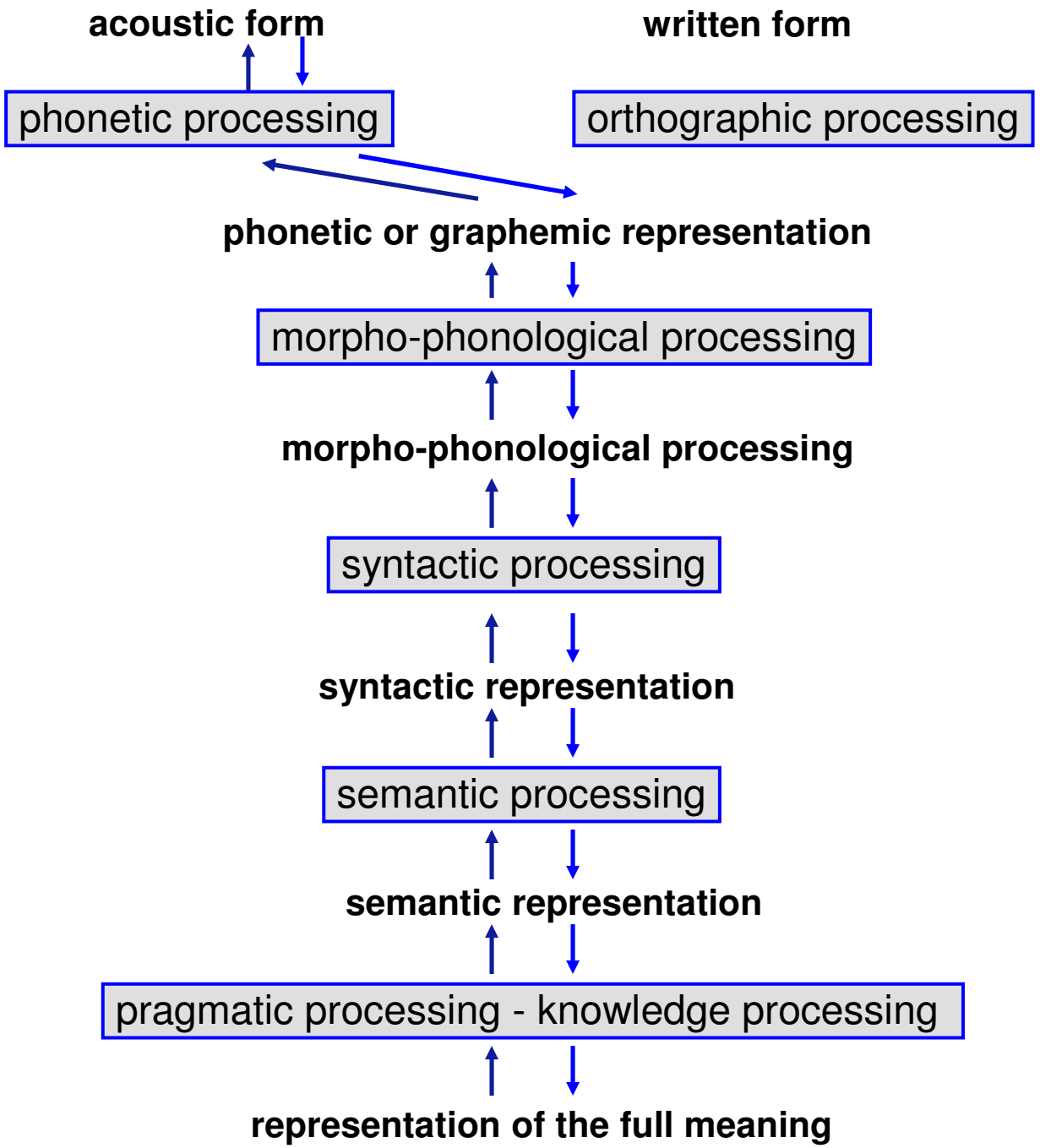
Text-to-Speech System



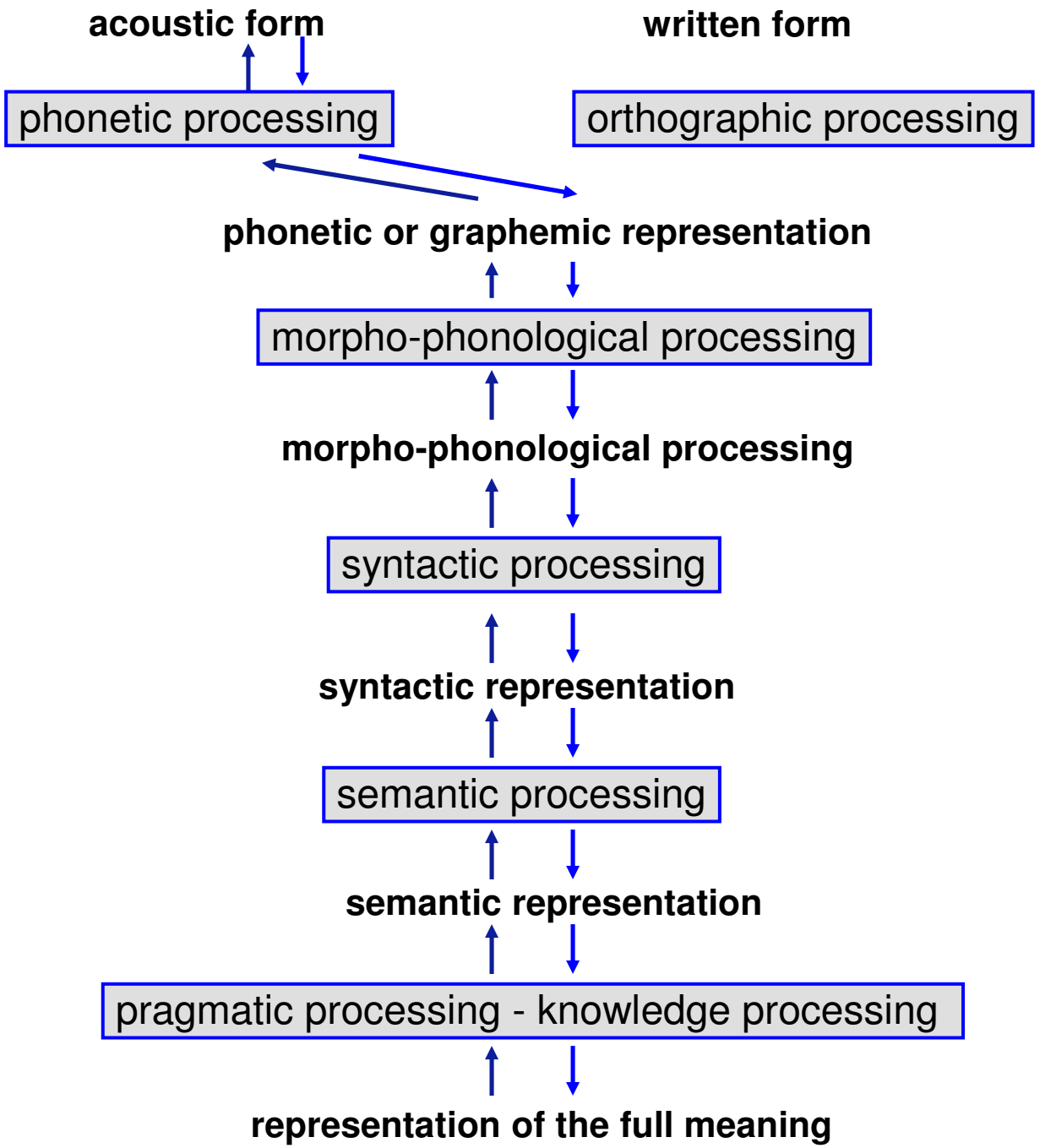
Why do we need deep processing for simple text-to-speech conversion

(I) The student will read the paper. (/ri ː d/)
(2) The students have read the paper. (/rɛd/)
(3) Will the students read the paper? (/ri ː d/)
(4) Have any citizens of good will read the paper? (/rɛd/)
(5) Have the executors of the will read the paper? (/rɛd/)
(6) Have the students who will arrive next week read the paper yet? (/rɛd/)
(7) Please have the students read the paper. (/ri ː d/)
(8) Have the students read the paper? (/rɛd/)

Speech Translation



Speech Translation



phonetic (homophony):

their	there
toe	tow

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

Competence vs. Performance

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

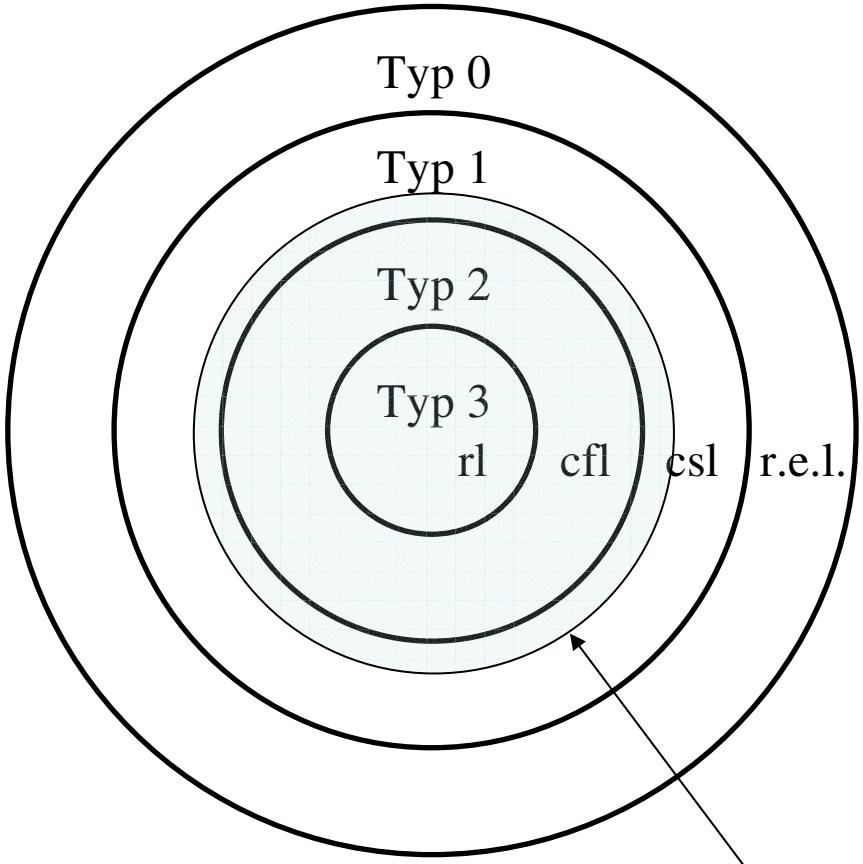
The finite definition of an infinite natural language.

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

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)

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.

Incrementality

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.

Some Important Online Resources

- ☆ LT-World Portal of Language Technology
  http://www.linguistlist.org/
- ☆ The Linguist List
  http://www.linguistlist.org/
- ☆ Ethnologue Catalogue of the World‘s Languages
  http://www.ethnologue.com/
- ☆ Web References for Linguists (U. Aberdeen)
  http://www.abdn.ac.uk/langling/resources/#Reference