Hybrid NLP

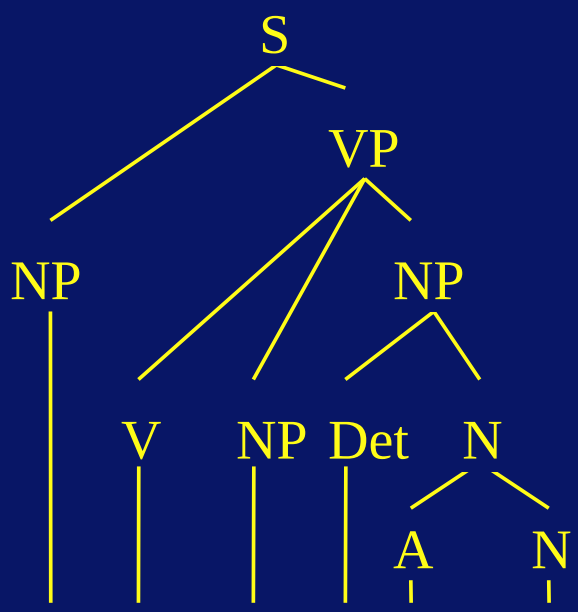
OUTLINE

- **Problems of Deep and Shallow Processing**
- **Hybrid Architectures**
- **An Advanced Platform for Hybrid NLP: Deep Thought**
- **Applications for Hybrid Processing**
- **Conclusion and Outlook**

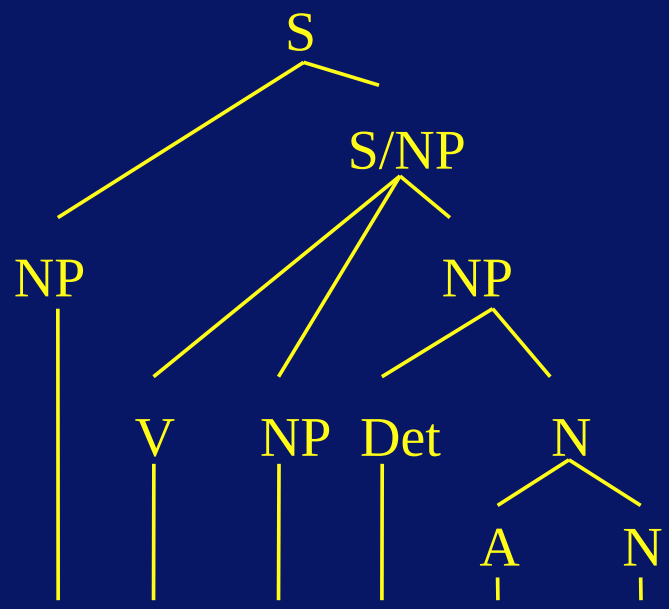
DEEP & SHALLOW PROCESSING

- **deep methods for morphological - syntactic - semantic processing exploit our knowledge about the structure of human language**
- **as opposed to shallow methods such as pattern matching grammars, n-gram language models**
- **deep methods are needed for getting at the meaning of language input**
- **shallow methods perform a partial or heavily under-specified analysis sufficient for certain applications**

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



Sue gave Paul an old penny.



Sue gab Paul einen alten Pfennig.

ONCE UPON A TIME

- **Broad industrial research in deep parsing**
  - **Xerox - LFG**
  - **Siemens - LFG**
  - **IBM Germany - HPSG**
  - **Hewlett Packard - GPSG and HPSG**
  - **IBM USA - PLNLP and Slot Grammar**
- **Very large projects**
  - **EUROTRA**
  - **LILOG**
  - **LS-GRAM**

GRAMMAR FRAMEWORKS

- **Head-Driven Phrase Structure Grammar (HPSG)**
- **Lexical Functional Grammar (LFG)**
- **Tree-Adjunction Grammar (TAG)**
- **Categorial Grammar (CG)**
- **Dependency Grammar (DG)**
- **GB-Minimalist Program**

HPSG

- **Head-Driven Phrase Structure Grammar by Pollard and Sag**
- **Uniform formalism: typed feature structures**
- **High degree of lexicalization: very few PS-rules, rich lexicon structure**
- **Ontological structure: Multiple inheritance type hierarchy**

Problems with Deep Analysis

- **Coverage (Development Time)**
- **Robustness (Coping with Out-of-Grammar Input)**
- **Efficiency (Runtime and Space Efficiency)**
- **Specificity (Selection among Readings)**

Problems with Shallow Analysis

Accuracy

- **Problems with embeddings, grammatical control, anaphora and modal as well as negative contexts.**

  According to SVP Raul Lopez, Slator expected him to be appointed CEO of Crawford Inc. at the upcoming share holders meeting.

  After the retirement of Peter Smith, Mary Hopp was introduced by VP Brown as the new director of the marketing division.

  After every former US based vicepresident except Lisa Ronell served as Chairman of the Board, the shareholders for the first time appointed a non-US Chairperson.

THE CORE MACHINERY

LKB
Development Platform

PET
Runtime Parser

English
Grammar

German
Grammar

Japanese
Grammar

Application

Open Source
Public Domain

HOWEVER

- **Back to the problems of**
  - **robustness**
  - **coverage**
  - **specificity**

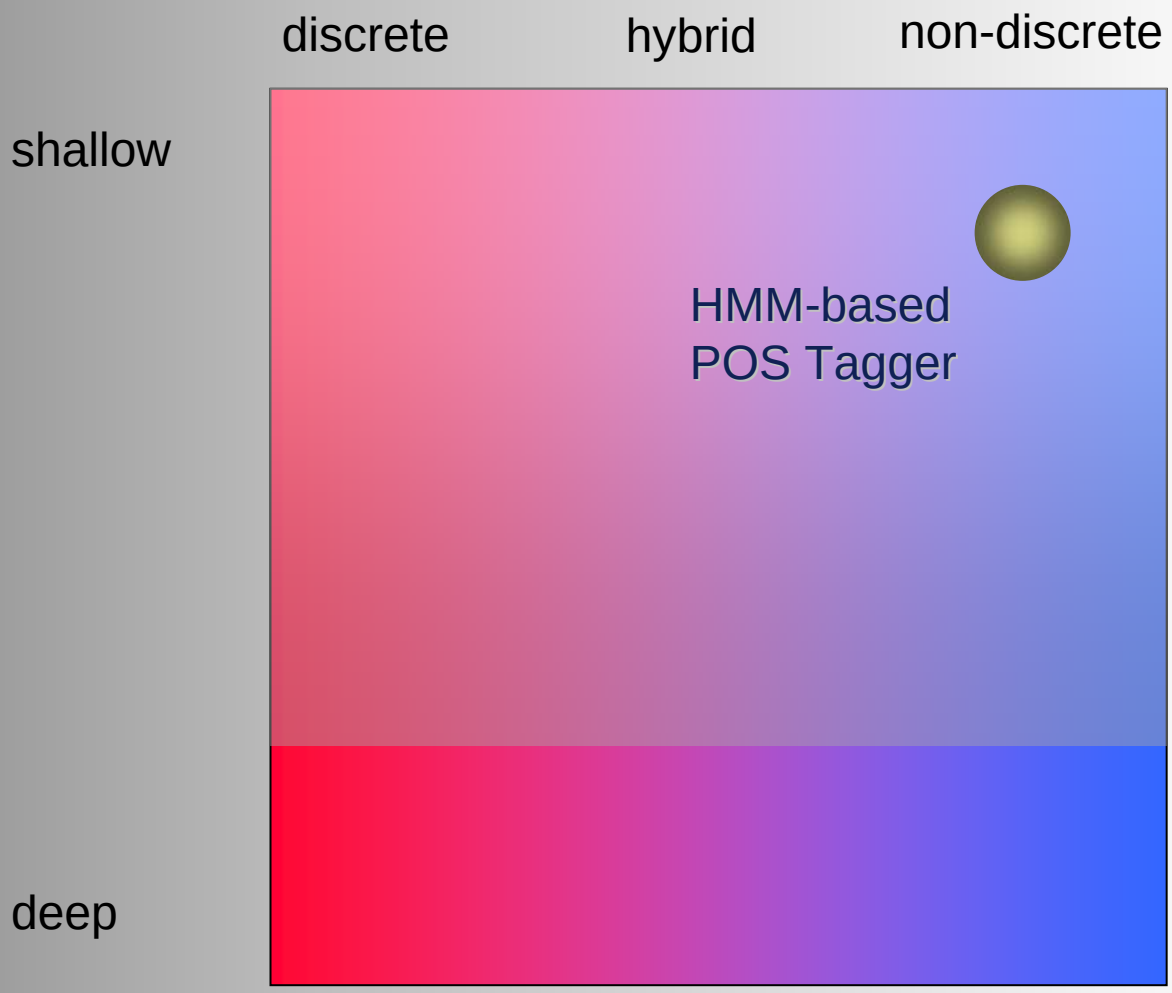
ASSUMPTIONS

- **Information extraction is not an alternative to deep processing but a continuum between classification and "full" semantic analysis**
- **Information Extraction via Text Enrichment**
- **We can detect topics, names, binary relations, complex relations, answers, etc.**
- **Question: At what point is deep processing needed?**

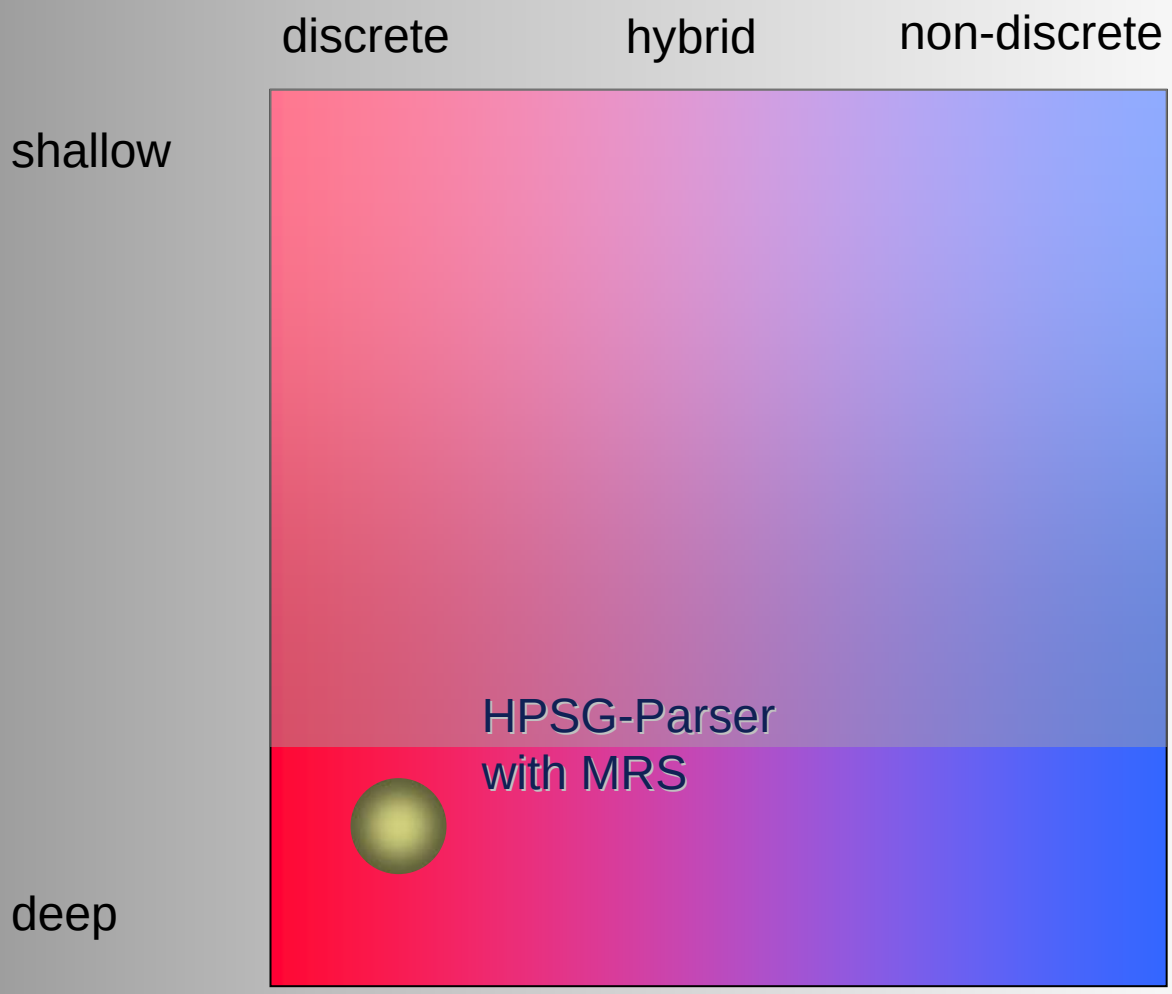
APPROACH

- **Lack of robustness and coverage remains a serious problem for deep processing.**
- **So we need to find applications, where deep processing can improve detection without spoiling the performance.**
- **Example: Relation extraction.**
- **Let deep processing assist shallow methods.**

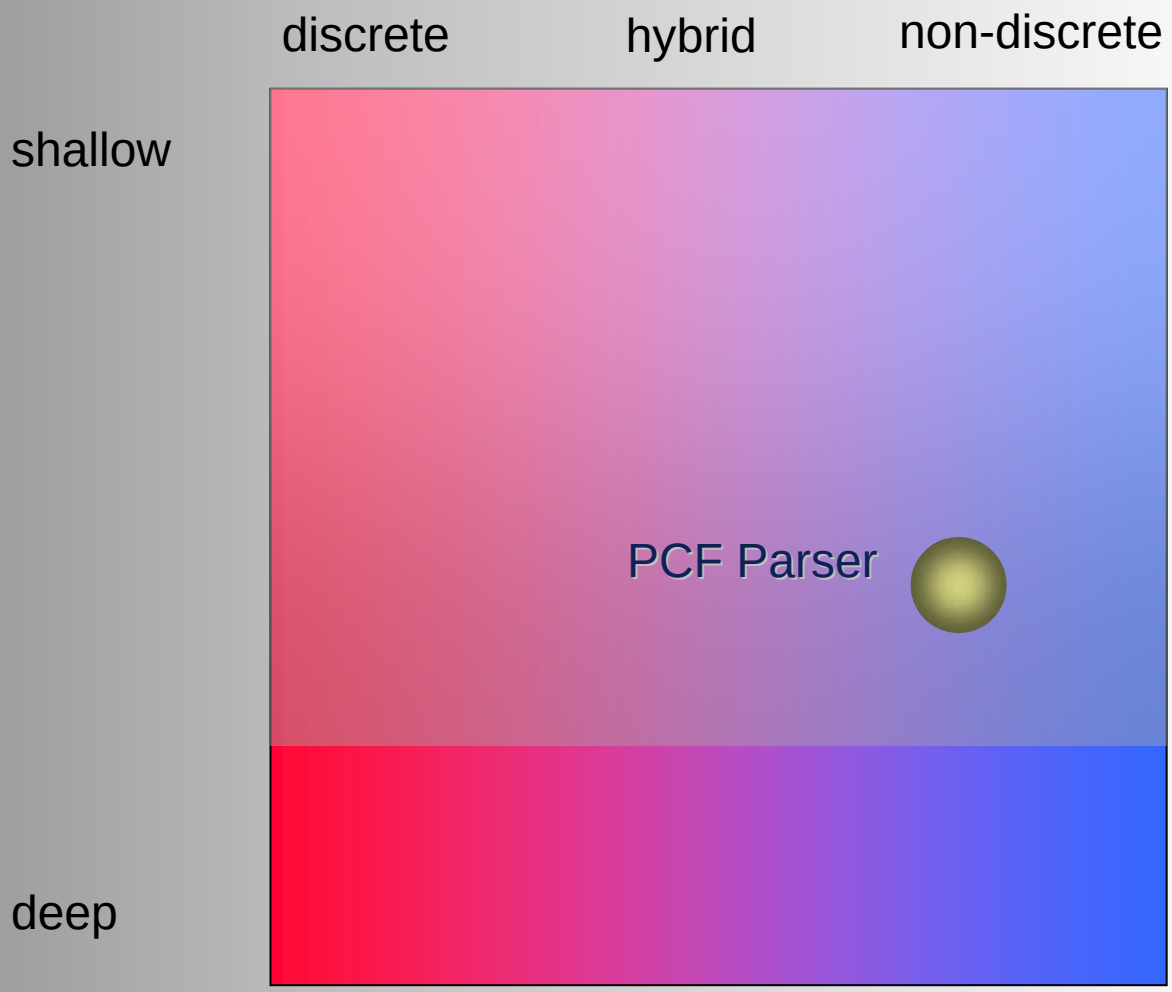
LT METHODS


discrete
hybrid
non-discrete
shallow
deep
HMM-based
POS Tagger

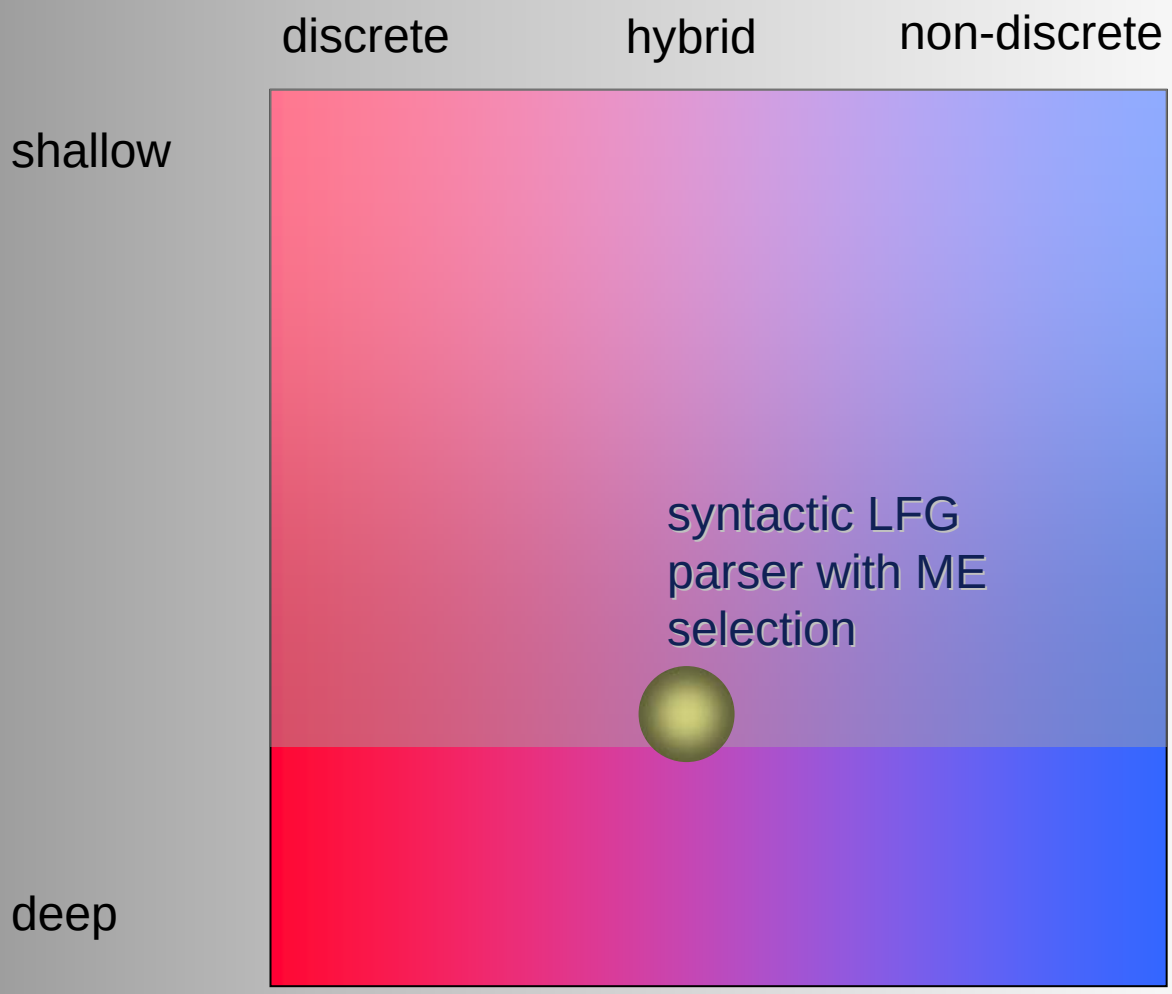
LT METHODS


discrete
hybrid
non-discrete
shallow
deep
HPSG-Parser
with MRS

LT METHODS



LT METHODS

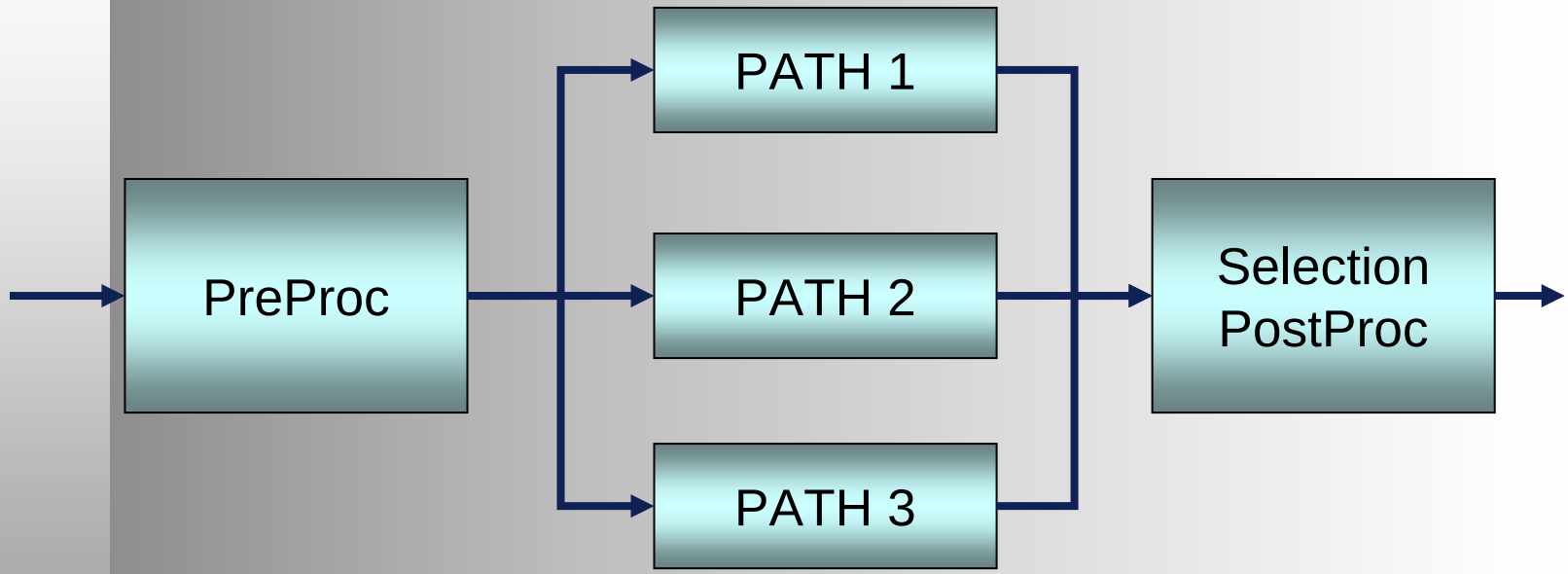

discrete
hybrid
non-discrete
shallow
deep
syntactic LFG
parser with ME
selection

METHODS OF COMBINATION

- **Separate systems running in parallel with a selection function based on preference, confidence or assessment metrics**
- **All in one integrated system (such as a blackboard or manager architecture)**
- **Separate systems annotating the same input with different control schemes (whiteboard or pool architecture)**

Systems Running in Parallel

- **Separate systems running in parallel with a selection function based on preference, confidence or assessment metrics**

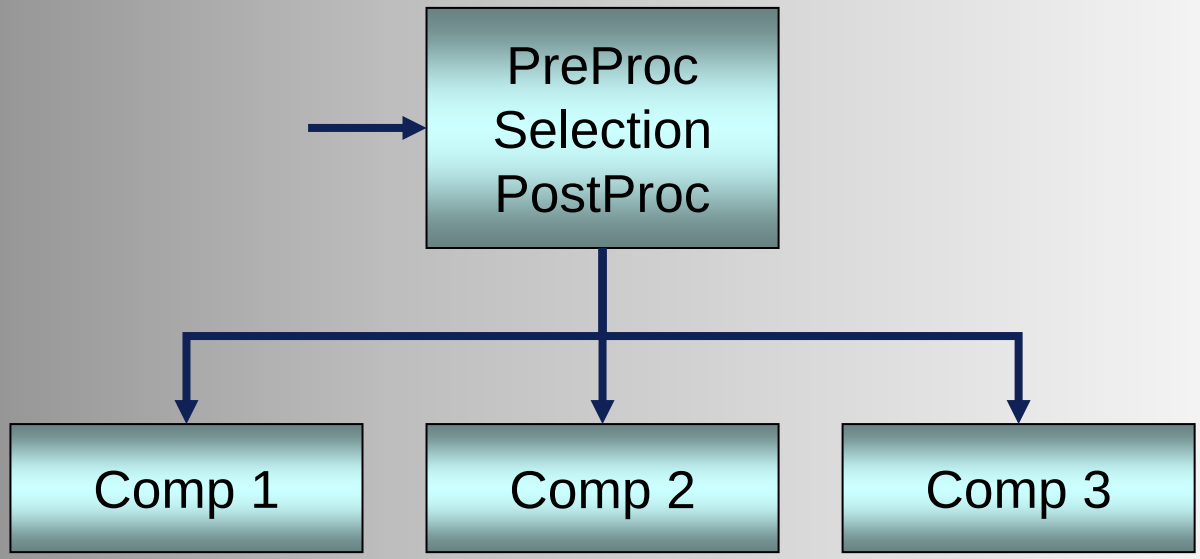


Selection Mechanism

- **Preference:**
  most probable component wins (deep < shallow)
- **Confidence:**
  most confident wins (each analysis gets a confidence measure)
- **Probability:**
  most probable analysis wins (calibrated probabilities)
- **Mixed Approaches**

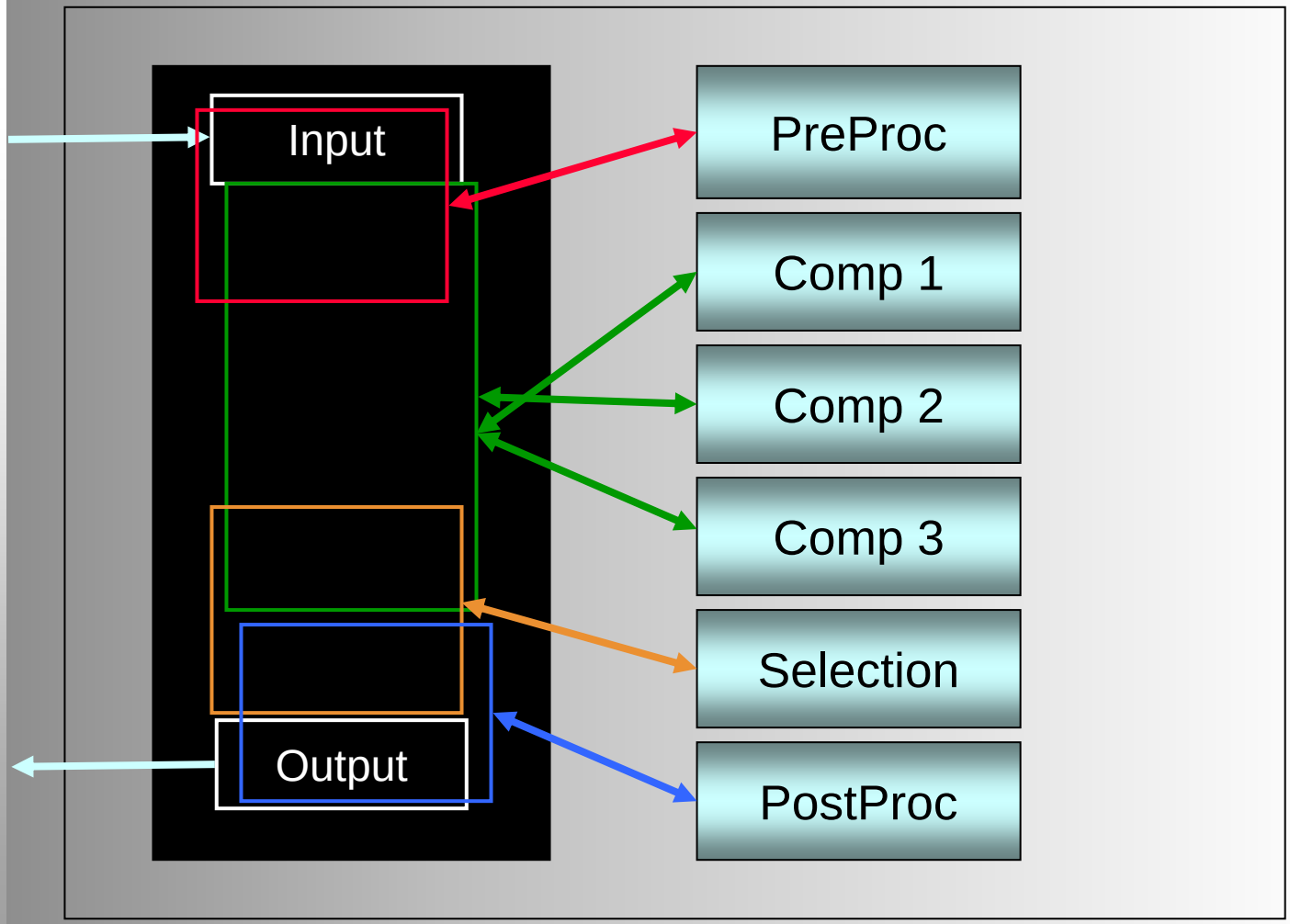
One Integrated System

- **Manager Architecture**



One Integrated System

- **Blackboard Architecture**

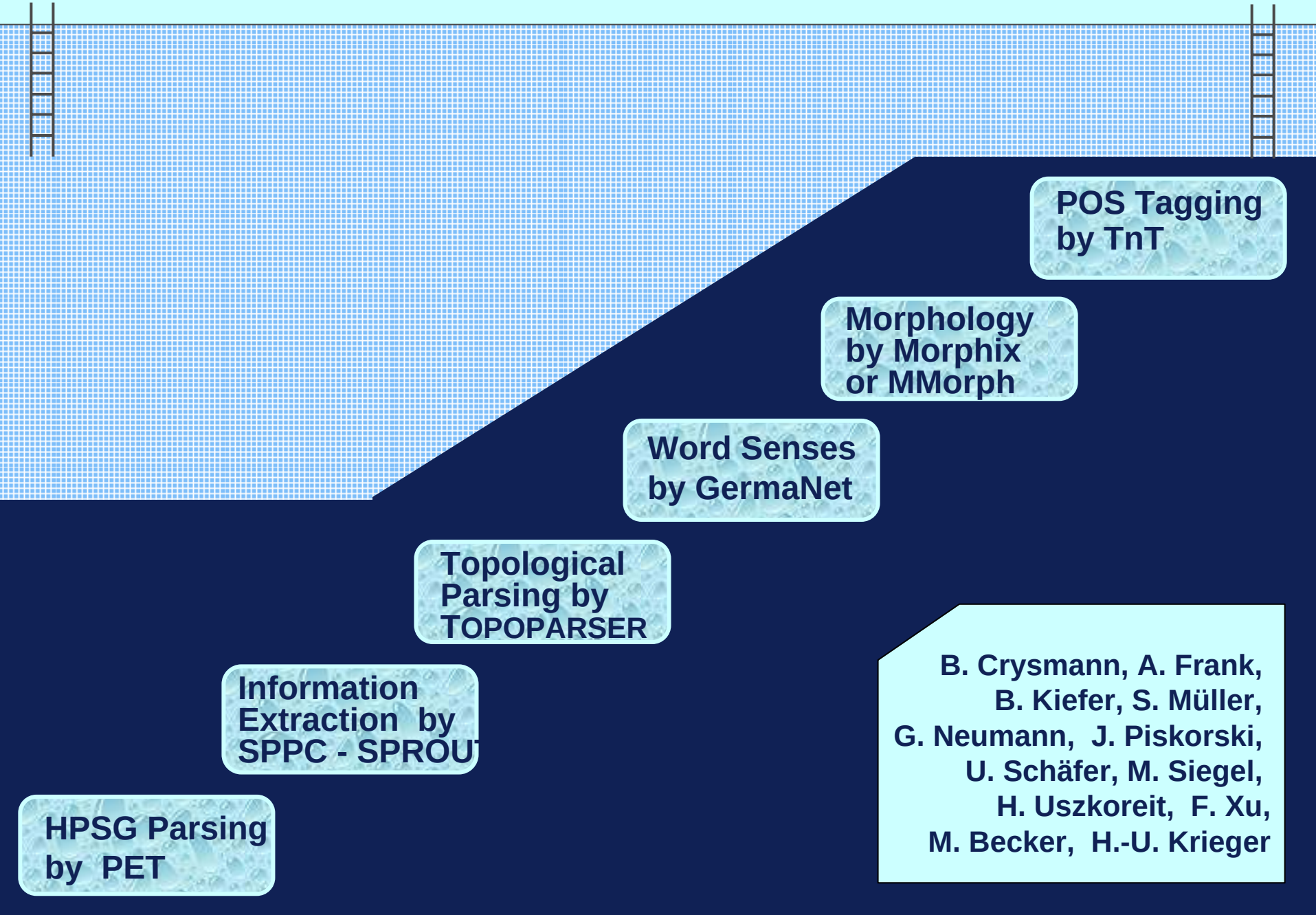


THE APPROACH

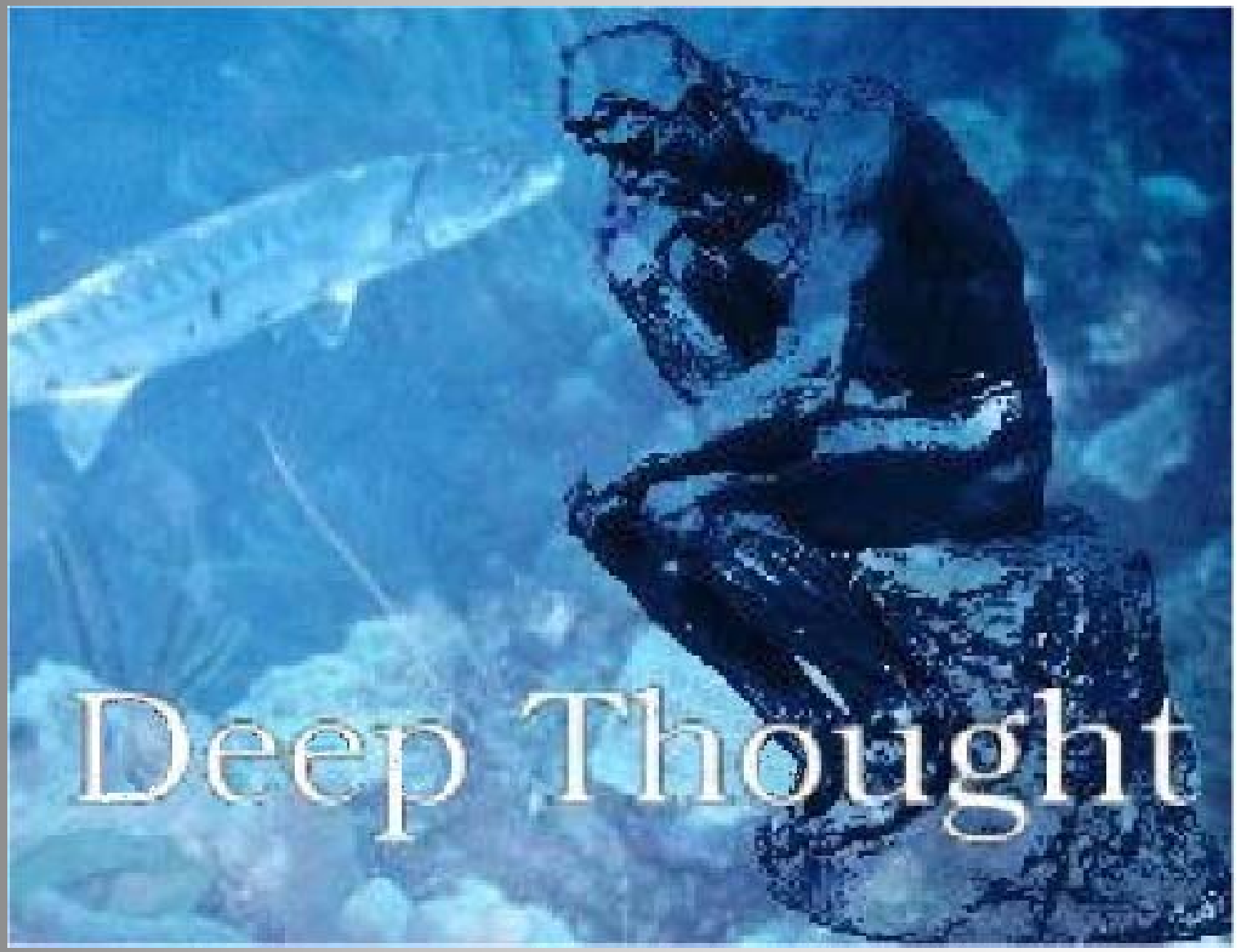
WHITEBOARD

- **Combination of shallow and deep processing by incremental text enrichment**
- **Shallow Methods**
  - **POS Tagging -- TnT**
  - **Morphology -- Morphix and MMorph**
  - **Word Nets -- GermaNet**
  - **Information Extraction -- SPPC and SPROUT**
- **Deep Methods**
  - **HPSG Parsing -- PET**

POOL ARCHITECTURE



An EU IST Project



The Consortium

DeepThought

Saarland University (Coordinator)

U. of Cambridge

U. of Sussex

U. of Trondheim

Stanford University

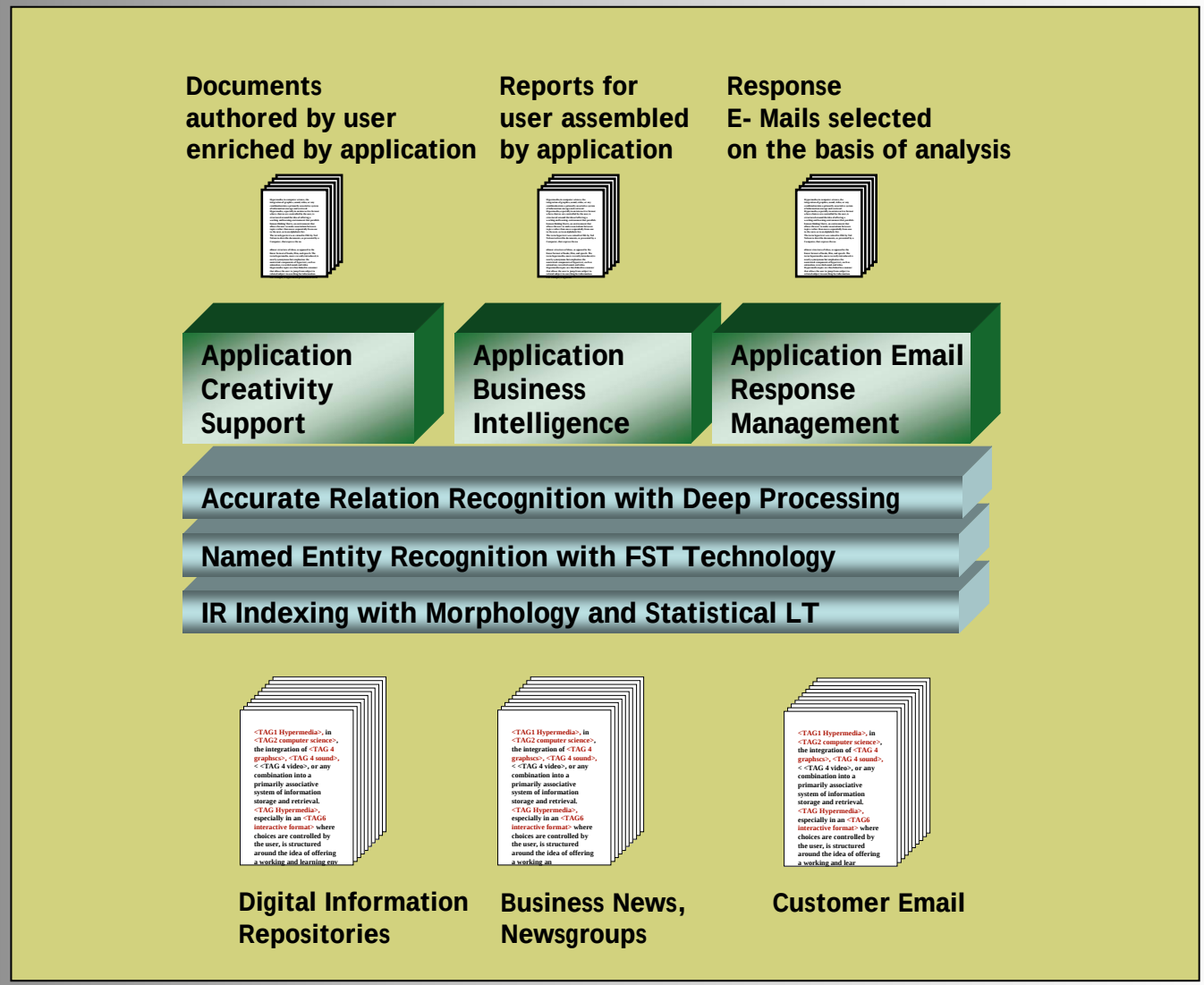
CELI Company of Torino

XtraMind GmbH of Saarbrücken

acrolinx gmbh, Berlin

DFKI Saarbrücken

Three Applications



Project Objectives

- **The lack of robustness and coverage is a serious problem for Information Extraction based on deep processing alone.**
- **Information Extraction with shallow processing methods is efficient and robust, but:**
  - **The lack of subcategorization information precludes extraction of conventional semantic representations.**
  - **There is insufficient information to recover scope information.**

Hybrid approach to multilingual information extraction

NL Input

Every experienced vice president served on a committee.

Event-Based LF

every(*x*), experienced(*x*),VP(*x*),serve1(*e,x*), spast(*e*), on2(*e',e,y*), some(*y*), committee(*y*).

Robust Minimal Recursion Semantics (RMRS)

every(*x1*), experienced(*x2*),VP(*x3*),serve1(*e4*), spast(*e4*), on2(*e5*), some(*x6*), committee(*x7*)

ARG1(e4,x1), ARG1(e5,e4),ARG2(e5,x6),

x1=x2, x2=x3, x6=x7.

DeepThought

NL Input

Every experienced vice president served on a committee.

POS Tagger

every_DT(x1), experienced_AJ(x2),vice_president_N(x3),served_V(e4), past(e4), on_PRP(e5), a_DT(x6), committee_N(x7)

Deep Processing

every(*x1*), experienced(*x2*),VP(*x3*),serve1(*e4*), spast(*e4*), on2(*e5*), some(*x6*), committee(*x7*)

ARG1(e4,x1), ARG1(e5,e4),ARG2(e5,x6),

x1=x2, x2=x3, x6=x7.

The NP chunk parser outputs an (R)MRS representation in which the parts of NPs are connected by equations.

NL Input

Every experienced vice president served on a committee.

NP Chunker

every_DT(*x1*), experienced_AJ(*x2*),vice_president_N(*x3*),served_V(*e4*), past(*e4*), on_PRP(*e5*), a_DT(*x6*), committee_N(*x7*),

x1=x3, x2=x3, x6=x7.

Certain grammatical relations or attachment may not be detected by the shallow parser. (Such as topicalization in the example below).

NL Input

That proposal every experienced committee member supported.

NP Chunker

that_DT(*x1*), proposal_N(*x2*), every_DT(*x3*), experienced_AJ(*x4*), committee_member _N(*x5*), support_V(*e6*), past(*e6*),

ARG1(e6,x5), *x1=x2, x3=x4, x4=x5*

Main Idea

- **RMRS provides a common representation for linguistic processing on all levels and for all languages. Shallow and deep processing produce compatible semantic representations.**

Main results

- **New architecture for hybrid integrated multilingual natural language processing.**
  - **Combining efficient processing modules on different levels of preciseness and robustness and for multiple languages in a variable way.**

Main results: applications

- **Information extraction for business intelligence.**
  - **Implementation where HPSG refines and verifies the output of shallow parsing.**
- **Information lookup for document authoring and creative meetings.**
  - **Server-based implementation with connection to external services (Google/Answerbus) and HoG.**
- **Information extraction in an email autoresponse task.**
  - **Implementation with connection to HoG using the deepest available analysis triggered by clues.**
- **Information extraction for a web portal from hike descriptions.**
  - **Implementation with semantically enriched grammar and connection via HoG .**

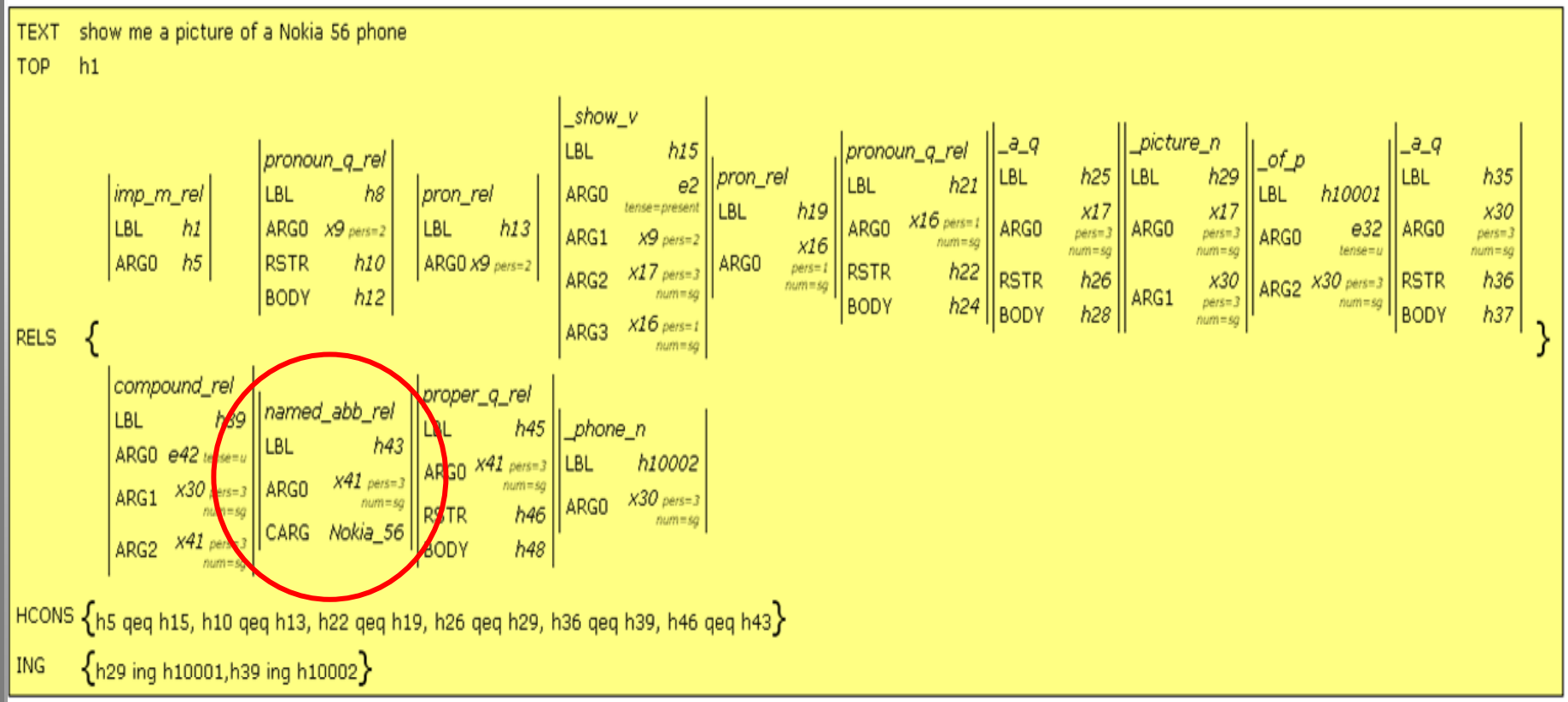
DeepThought

Main results

- **Linguistically sound analyses for a variety of phenomena and languages.**
  - **We have an integrated demonstration of large and wide-coverage grammars for five languages.**
- **Lexical databases.**
- **Methods for unknown words and named entities.**
- **Efficient generation.**

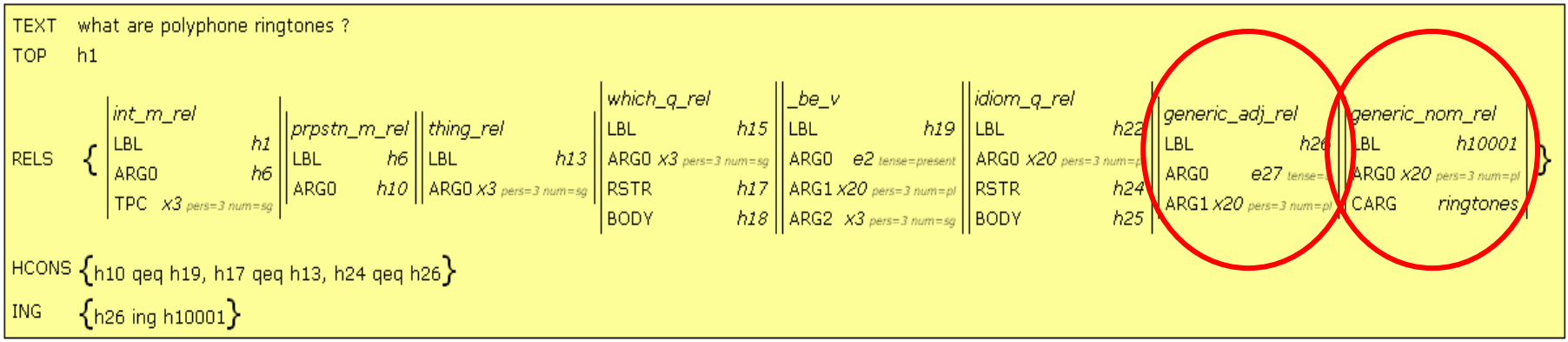
Hybrid Natural Language Processing

- **The analysis results of NLP tools at lower processing levels can be used by components at higher levels.**
  - **For example, the deep linguistic analysis module PET uses default lexicon entries for Named-Entities that Sprout delivers.**



Hybrid Natural Language Processing

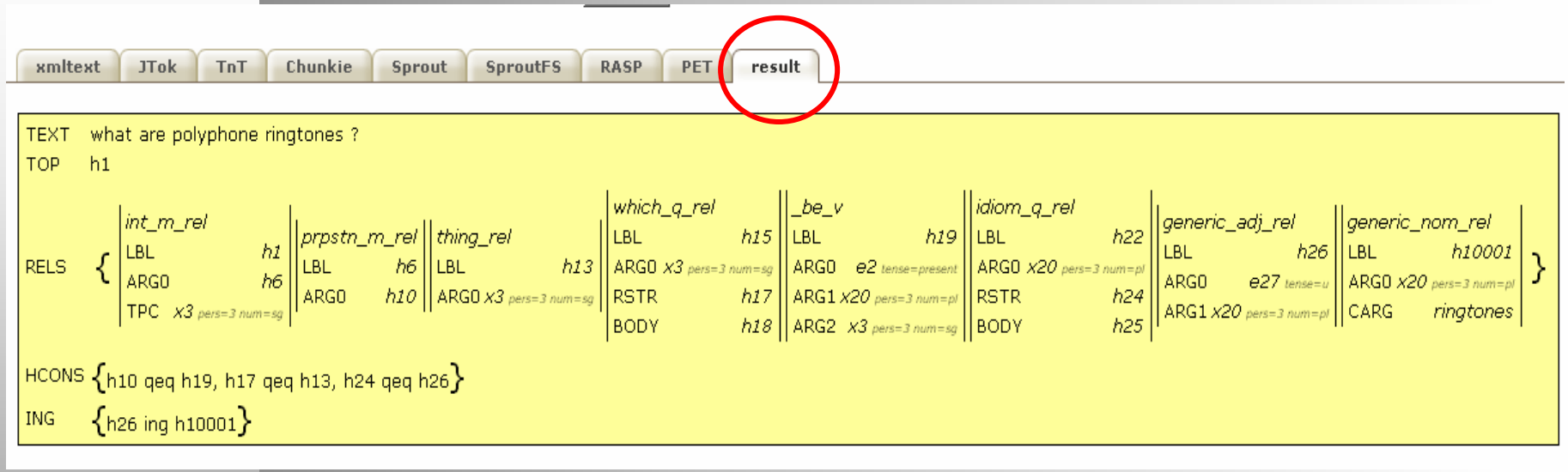
- **The analysis results of NLP tools at lower processing levels can be used by components at higher levels.**
  - **For example, the deep linguistic analysis module PET uses default lexicon entries for Part-of-Speech tags that the POS tagger TnT delivers.**



Hybrid Natural Language Processing

Deliver the deepest result.

The Email Autoresponse application follows this approach



Hybrid Natural Language Processing

Deliver partial results, whenever a complete analysis is not available. Partial results are taken from the deepest module that delivers results.

xmltext | JTok | TnT | Chunkie | Sprout | SproutFS | RASP | PET | result

TEXT I want a new phone do you ?
TOP h1

RELS {
prpstn_m_rel	
LBL	h1
ARG0	h5

pron_rel	
LBL	h8
ARG0	x9 pers=1 num=sg

pronoun_q_rel	
LBL	h11
ARG0	x9 pers=1 num=sg
RSTR	h12
BODY	h14

_want_v	
LBL	h15
ARG0	e2 tense=present
ARG1	x9 pers=1 num=sg
ARG2	x16 pers=3 num=sg

_a_q	
LBL	h18
ARG0	x16 pers=3 num=sg
RSTR	h19
BODY	h20

_new_a	
LBL	h22
ARG0	e24 tense=u
ARG1	x16 pers=3 num=sg

_phone_n	
LBL	h10001
ARG0	x16 pers=3 num=sg

}

HCONS {h5 qeq h15, h12 qeq h8, h19 qeq h22}
ING {h22 ing h10001}

TEXT I want a new phone do you ?
TOP h1

RELS {
_do_v	
LBL	h1
ARG0	e2 tense=u
ARG1	u4
ARG2	u5
ARG3	x3 pers=2

pron_rel	
LBL	h7
ARG0	x3 pers=2

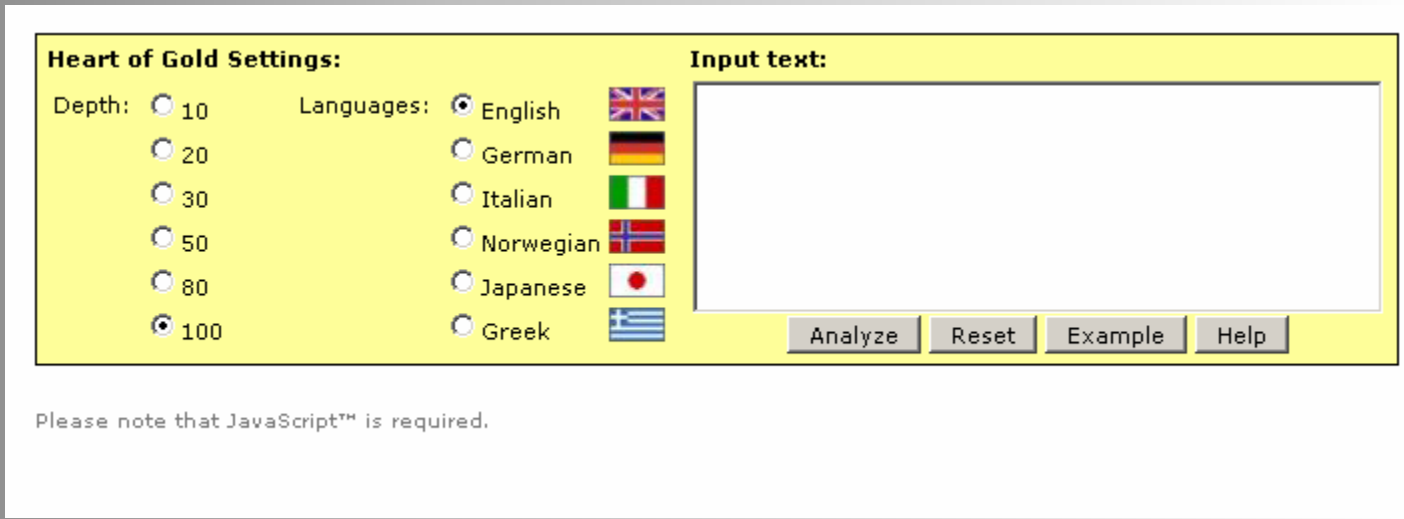
pronoun_q_rel	
LBL	h9
ARG0	x3 pers=2
RSTR	h10
BODY	h12

}

HCONS {h10 qeq h7}
ING {}

Hybrid Natural Language Processing

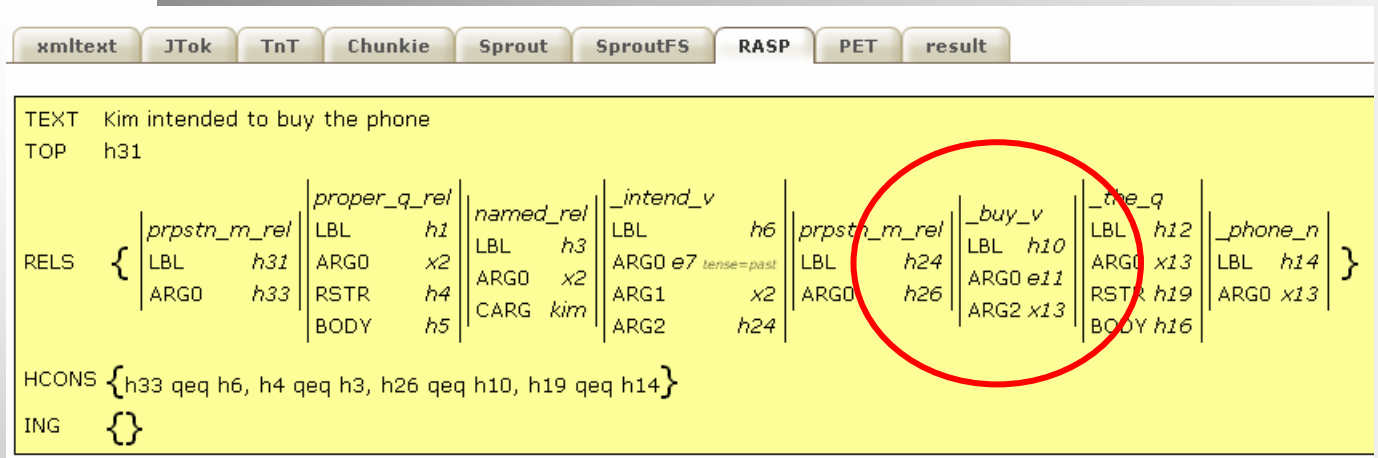
Combine modules and grammars for different languages. Each language has its own configuration of valid modules and grammars.

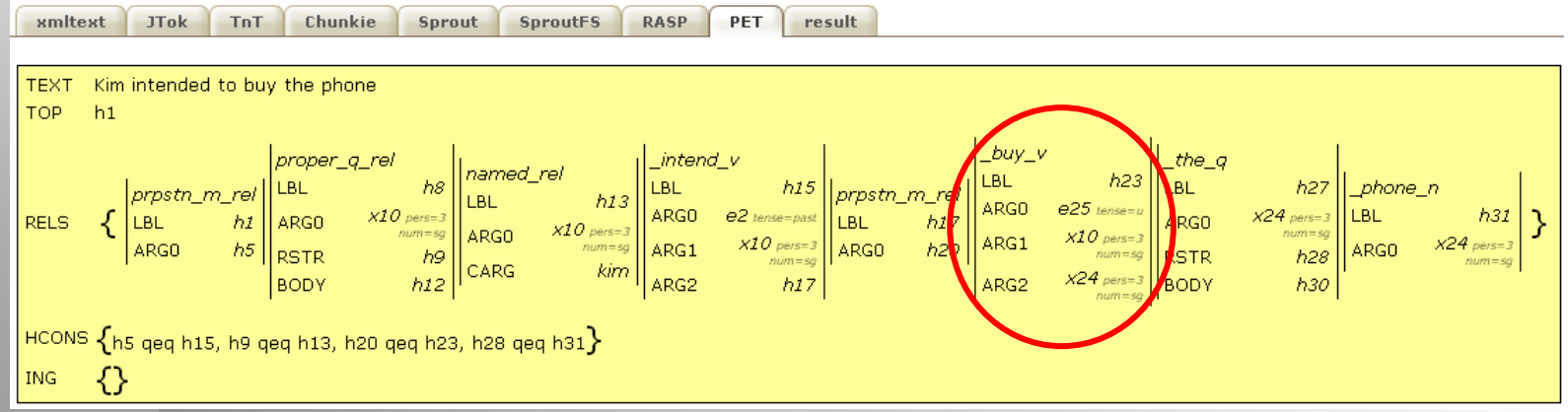


Hybrid Natural Language Processing

The different modules use a compatible output formalism, RMRS.

In case of shallower modules, this robust semantic structure allows for underspecification of, e.g., argument structure.





Multilingual HPSG Grammar Engineering

- **Available HPSG grammars :**
  - **German (50.000 lexical entries)**
  - **English (12.300 lexical entries)**
  - **Japanese (35.000 lexical entries)**
  - **Norwegian (84.240 lexical entries)**
  - **Italian (4.850 lexical entries)**
- **We have a Grammar Matrix that allows an efficient implementation of new grammars with compatible and correct output.**

MULTILINGUAL GRAMMAR DEVELOPMENT

- Existing Grammars in English, German, Japanese
- Sizeable Grammar of Norwegian built in the project Deep Thought by Lars Hellan and others at Trondheim U.
- Italian Grammar by company CELI built in Deep Thought
- Greek grammar being set up by Valia Kordoni and Julia Neu at Saarland University
- Korean grammar being build by Jong-Bok Kim
- New Portuguese Grammar project at University of Lisbon headed by Antonio Branco
- Spanish Grammar converted from ALEP format at U. Barcelona
- New: Beginning of a Chinese Grammar at Saarland U.

The Grammar Matrix

- **The Matrix for grammars of multiple languages:**
  - **A system of types that is *directly included* into new and existing grammars.**
  - **Reduced start-up costs.**
  - **Common feature descriptions.**
  - ***Shared insights* on analyses of phenomena.**
  - **Support for *multilingual* applications.**
  - **Robust treatment of *real corpora*.**

The Grammar Matrix

- **It contains 19 files and documentation:**
  - **Basic types and features for multilingual HPSG development.**
  - **Basic types and features for multilingual semantic construction.**
  - **Settings for working with LKB, [incr tsdb()] and PET.**
  - **Basic lexical types**
  - **Basic rule types**

Matrix-based multilingual grammar engineering

xmltext | JTok | TnT | Chunkie | Sprout | SproutFS | RASP | PET | result

TEXT the man smiles
TOP h1

RELS {
- *prpstn_m_rel*: LBL *h1*; ARG0 *h5*
- *_the_q*: LBL *h8*; ARG0 *x9* pers=3 num=sg gender=n; RSTR *h10*; BODY *h12*
- *_man_n*: LBL *h13*; ARG0 *x9* pers=3 num=sg gender=n
- *_smile_v*: LBL *h15*; ARG0 *e2* tense=present; ARG1 *x9* pers=3 num=sg gender=n
}

HCONS {h5 qeq h15, h10 qeq h13}
ING {}

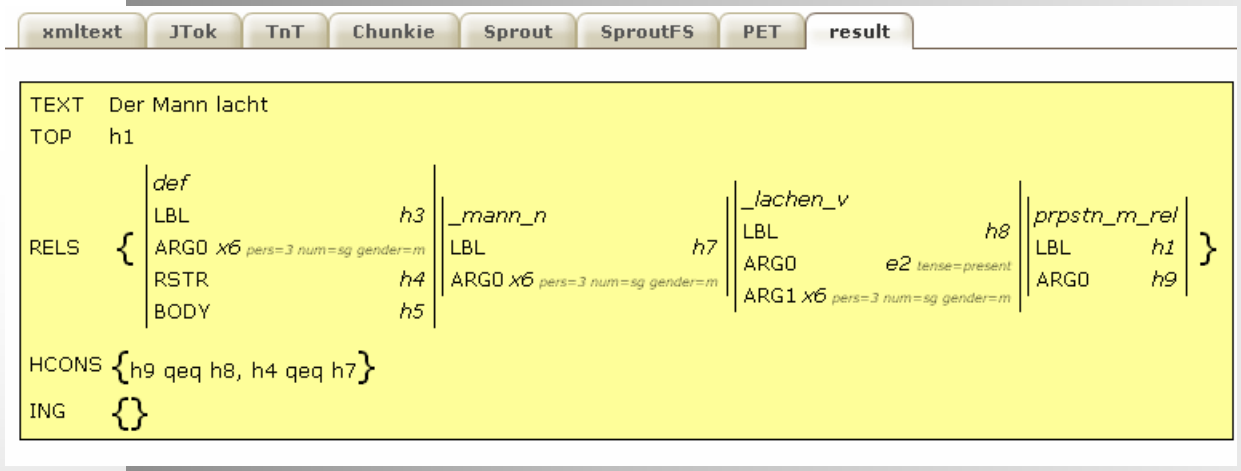
xmltext | PET | result

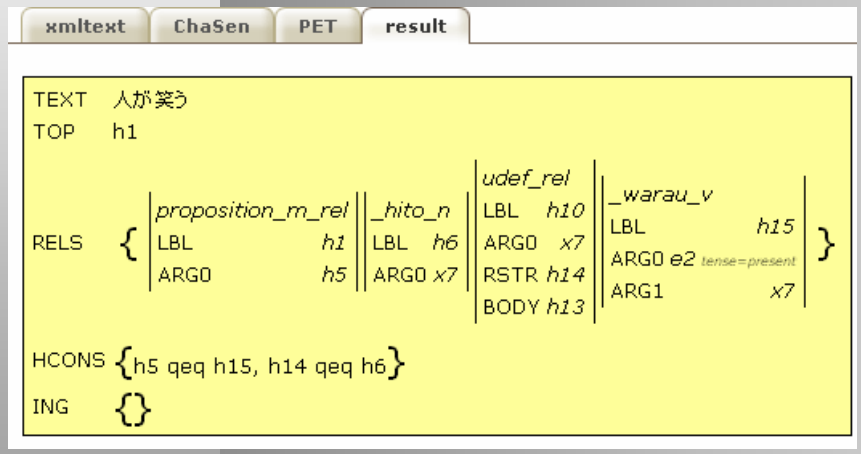
TEXT mannen smiler
TOP h1

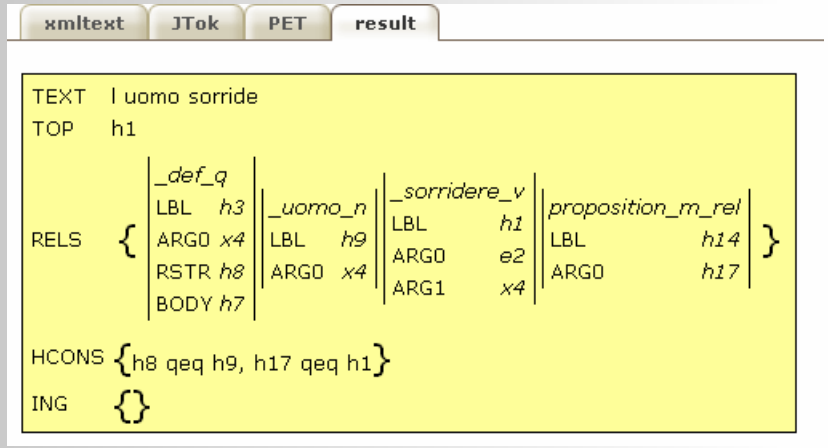
RELS {
- *mann-rel*: LBL *h3*; ARG0 *x4* bounded=+
- *def-q-rel*: LBL *h5*; ARG0 *x4* bounded=+; RSTR *h7*; BODY *h6*
- *card_rel*: LBL *h8*; ARG0 *x4* bounded=+; CARG *1-rel*
- *smile-rel*: LBL *h9*; ARG0 *e2* tense=present; ARG1 *x4* bounded=+
- *prpstn_rel*: LBL *h1*; ARG0 *h10*
}

HCONS {h7 qeq h3, h10 qeq h9}
ING {}

Matrix-based multilingual grammar engineering







HOME 0 html_page
INTRODUCTION 1 html_page
RESEARCH PROGRAM 2 pdf_doc
SYSTEMS & GRAMMARS 3 html_page
PUBLICATIONS 4 html_page
NEWS & EVENTS 5 html_page
LINKS 6 html_page
CONTACT US 7 mail_to
delphinsite

DEEP LINGUISTIC PROCESSING WITH HPSG

Computational linguists of several research sites have joined forces in a collaborative effort aimed at deep linguistic processing of human language. The goal is the combination of linguistic and statistical processing methods for getting at the meaning of texts and utterances. The partners have adopted Head-Driven Phrase Structure Grammar (HPSG) and Minimal Recursion Semantics (MRS), two highly advanced and influential models of formal linguistic analysis. They have also committed themselves to a shared format for grammatical representation and to a rigid scheme of evaluation.

The collaboration involves researchers from the following institutions:

- Cambridge University (UK), Computer Laboratory
- DFKI Saarbrücken GmbH (Germany), Language Technology Lab (co-founder)
- Norwegian University of Science and Technology (Norway), Lingvistisk Institutt
- Saarland University (Germany), Department for Computational Linguistics
- Stanford University (US), LinGO Laboratory at CSLI (co-founder)
- Tokyo University (Japan), Tsujii Laboratory
- University of Oslo (Norway), MT Research Group
- University of Sussex (UK), School of Cognitive and Computing Sciences
- University of Washington (US), Computational Linguistics Laboratory

The DELPH-IN collaboration is open to additional partners who share our ambitious goals and commitments and who can dedicate the necessary resources to the common task.

Scientific Impact: DELPH-IN

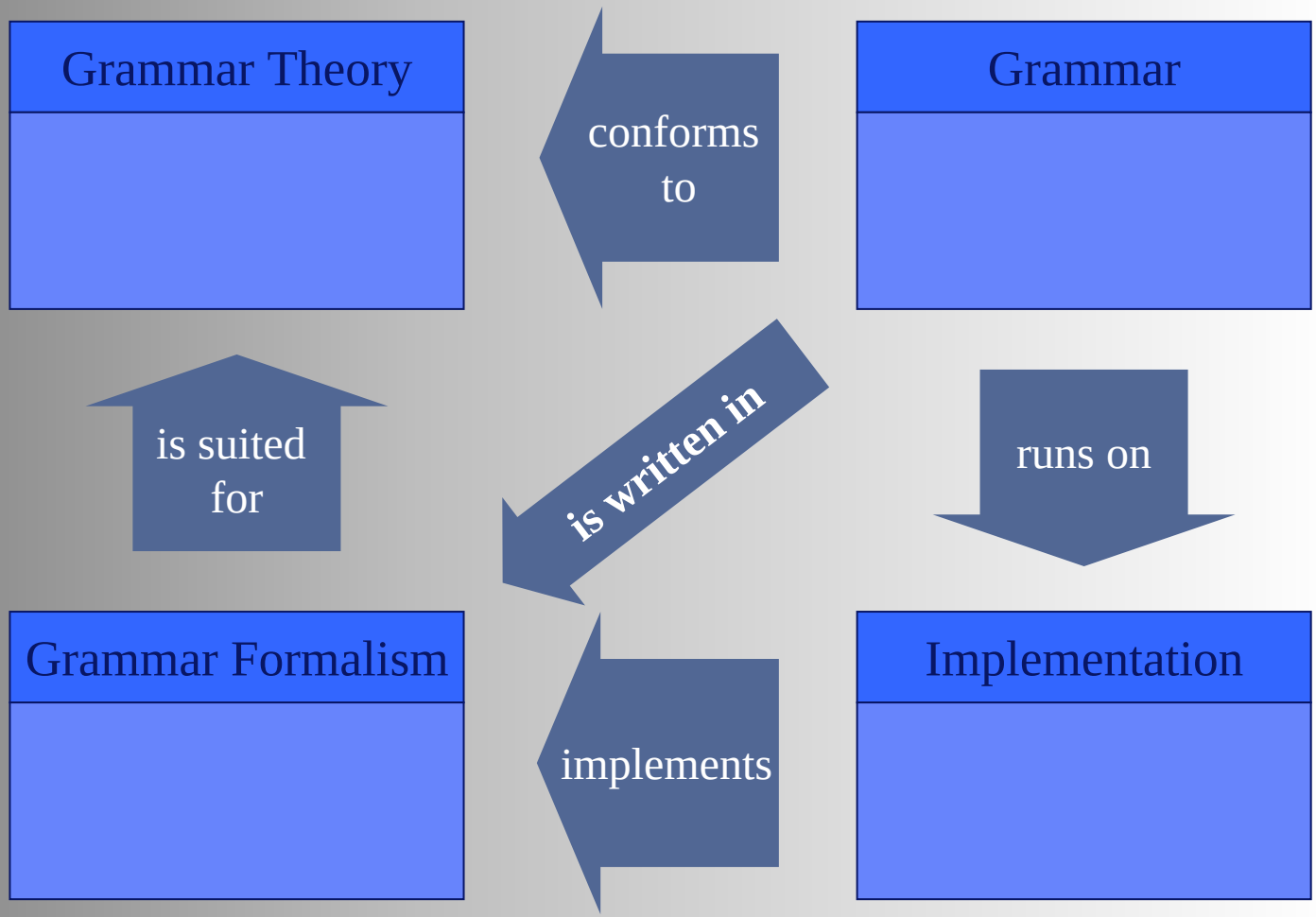
- **Including open-source resources:**
  - **LKB grammar development system (incl. generation)**
  - **PET grammar processing system**
  - **[incr tsdb()] grammar profiling system**
  - **ERG English HPSG**
  - **JACY Japanese HPSG**
  - **NorSource Norwegian HPSG**
  - **Modern Greek Resource Grammar**
  - **Lingo Grammar Matrix**
  - **Redwoods treebank**

  (DeepThought Heart of Gold will be part of DELPH-IN)

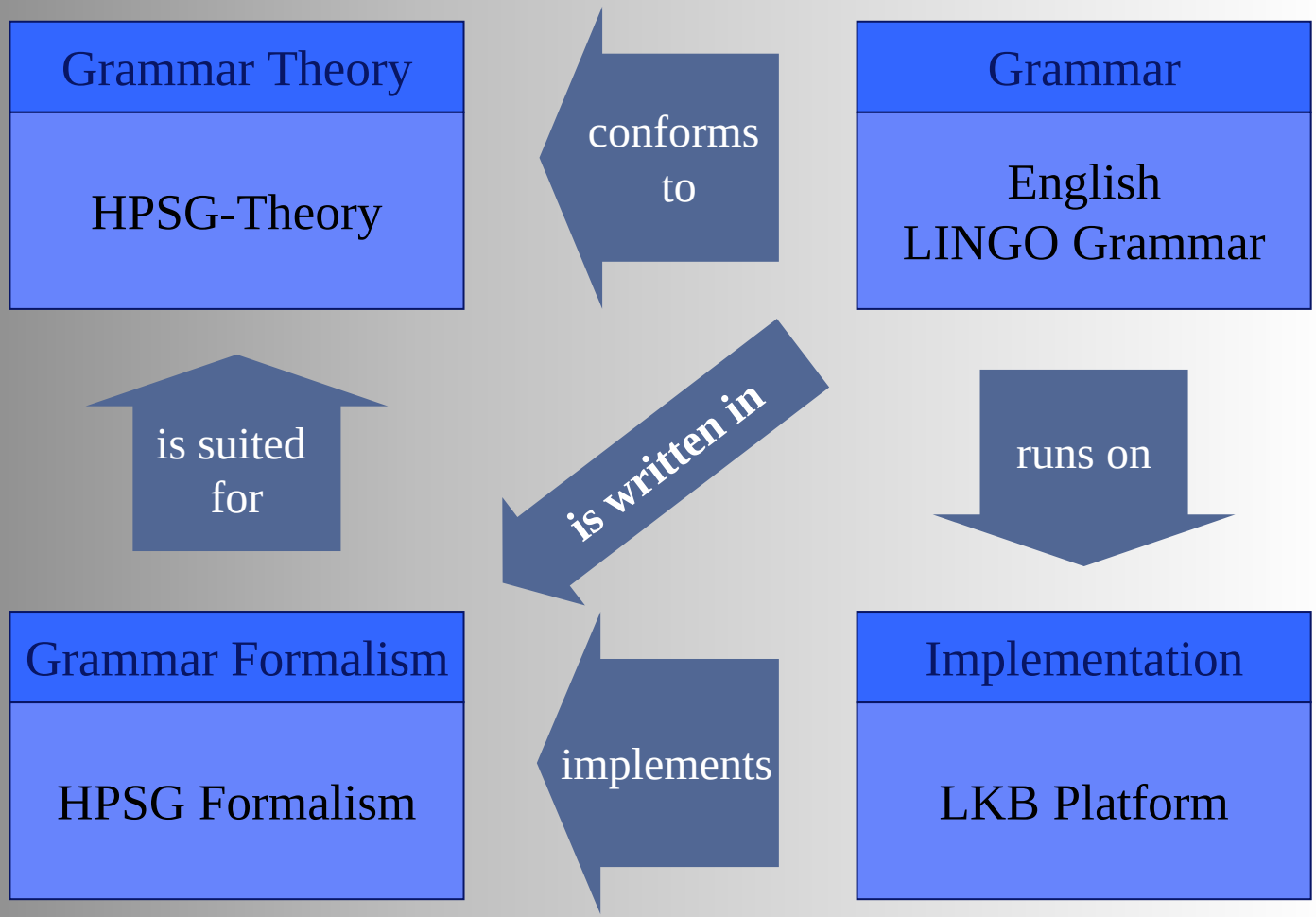
Conclusion and Outlook

- **There has been considerable progress in the area of deep linguistic processing.**
- **However, deep processing methods have to be combined with discrete and non-discrete shallow methods for sufficient performance.**
- **Flexible and scalable platform for the composition of hybrid systems.**
- **Test of the platform in real world applications.**
- **A better integration of statistical and deep linguistic methods is still badly needed.**

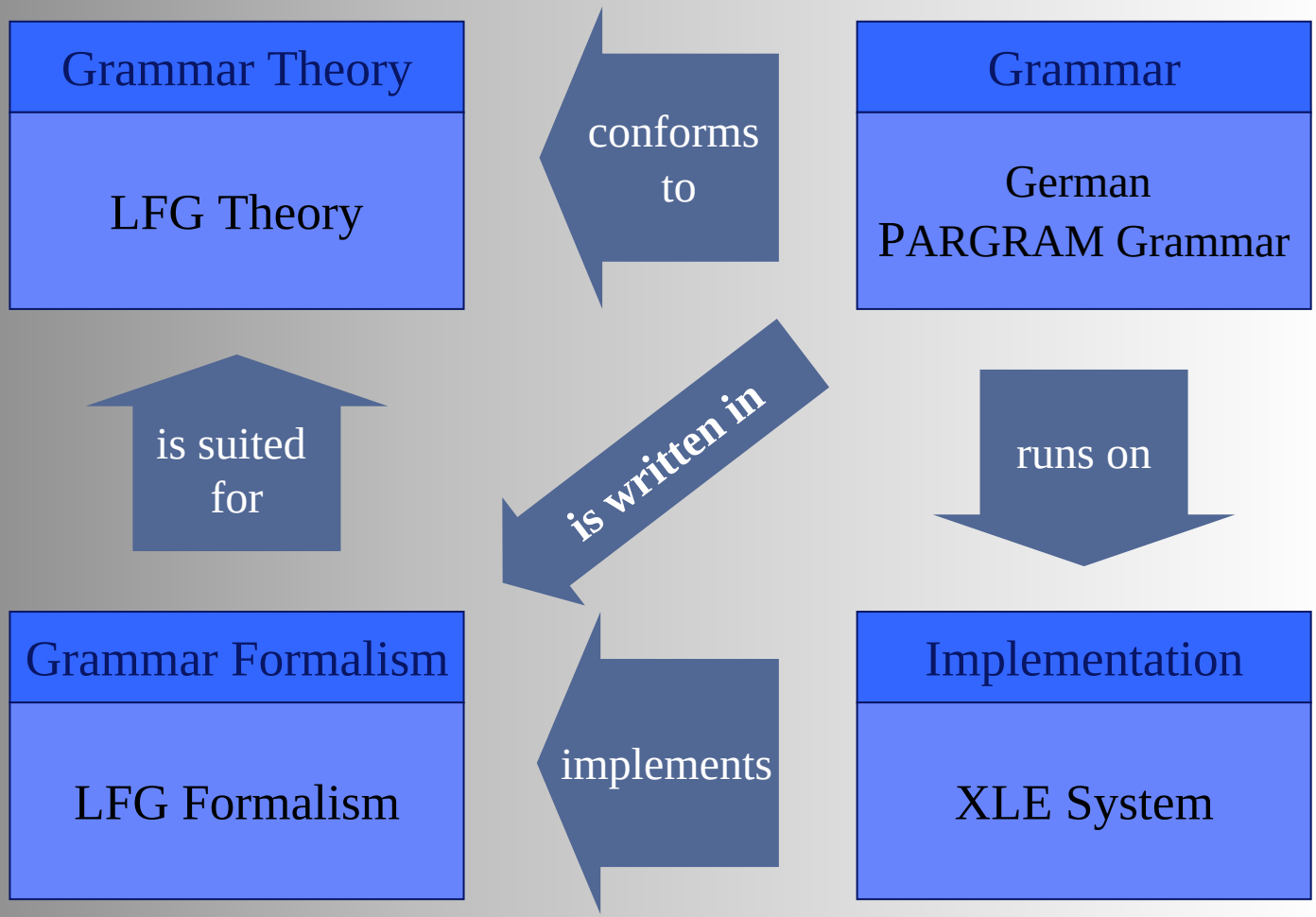
GRAMMAR[4]


Grammar Theory
conforms to
Grammar
is suited for
is written in
runs on
Grammar Formalism
implements
Implementation

GRAMMAR⁴


Grammar Theory
HPSG-Theory
conforms to
Grammar
English
LINGO Grammar
is suited for
is written in
runs on
Grammar Formalism
HPSG Formalism
implements
Implementation
LKB Platform

GRAMMAR⁴


Grammar Theory
LFG Theory
conforms to
Grammar
German PARGRAM Grammar
is suited for
is written in
runs on
Grammar Formalism
LFG Formalism
implements
Implementation
XLE System