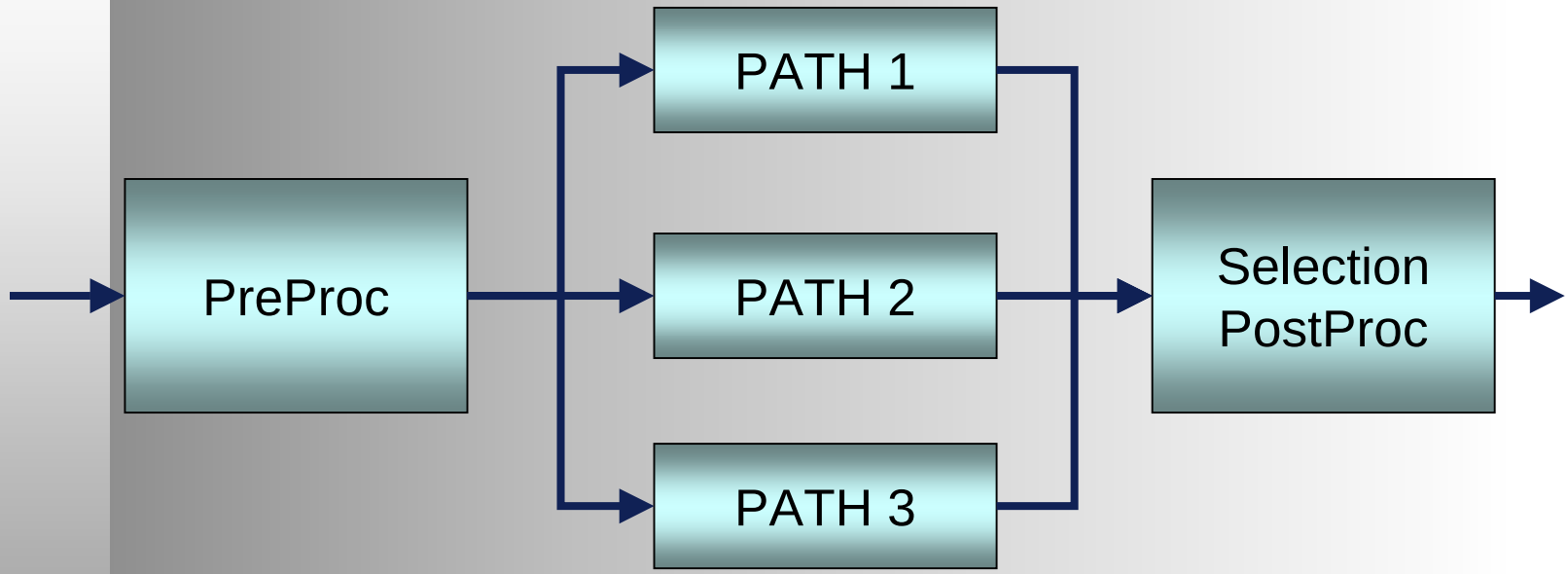
Hybrid NLP

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**

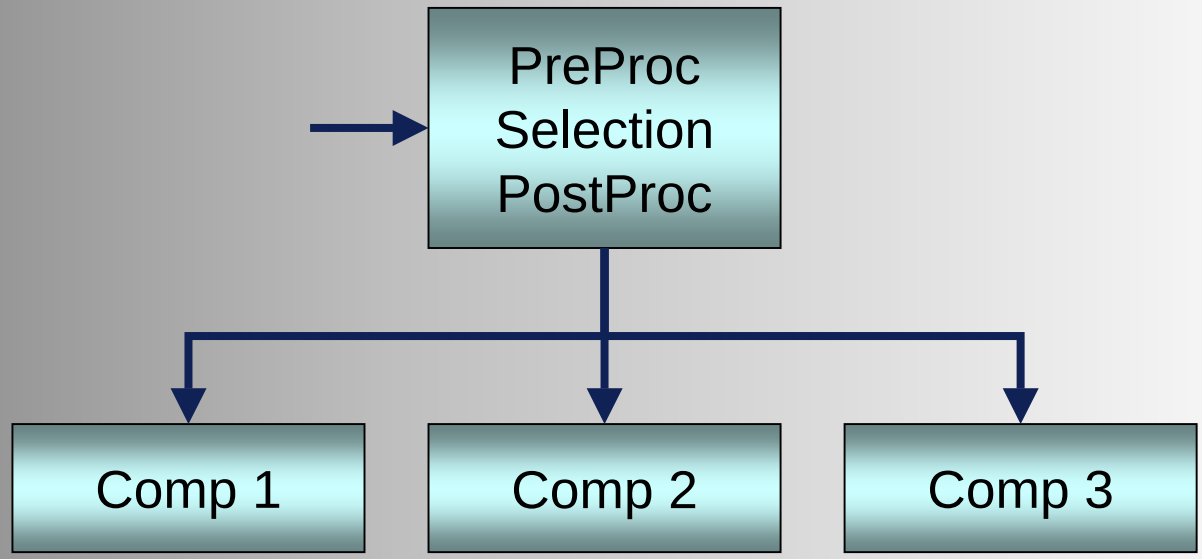


Selection Mechanism

- **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**

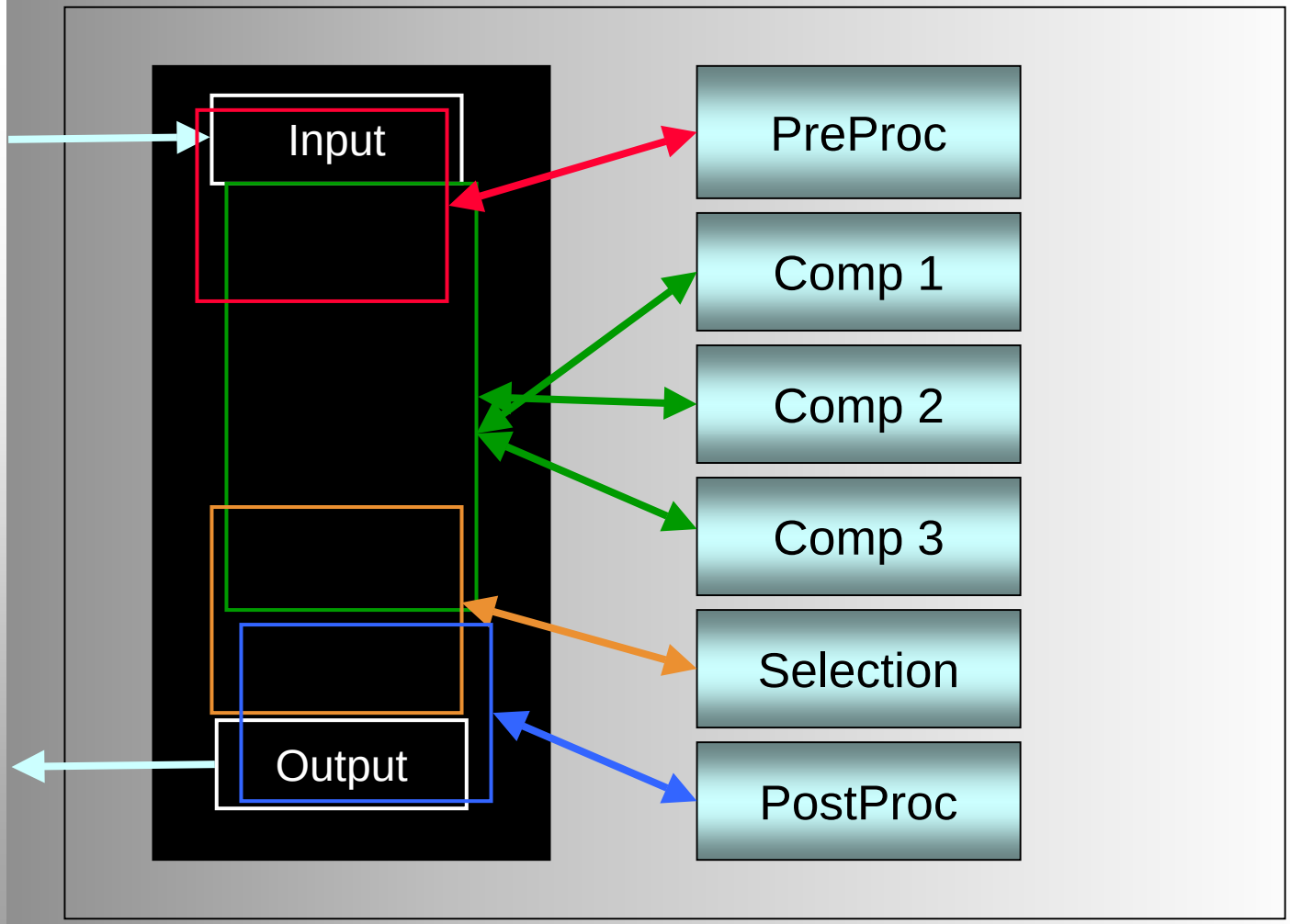
One Integrated System

- **Manager Architecture**



One Integrated System

- **Blackboard Architecture**

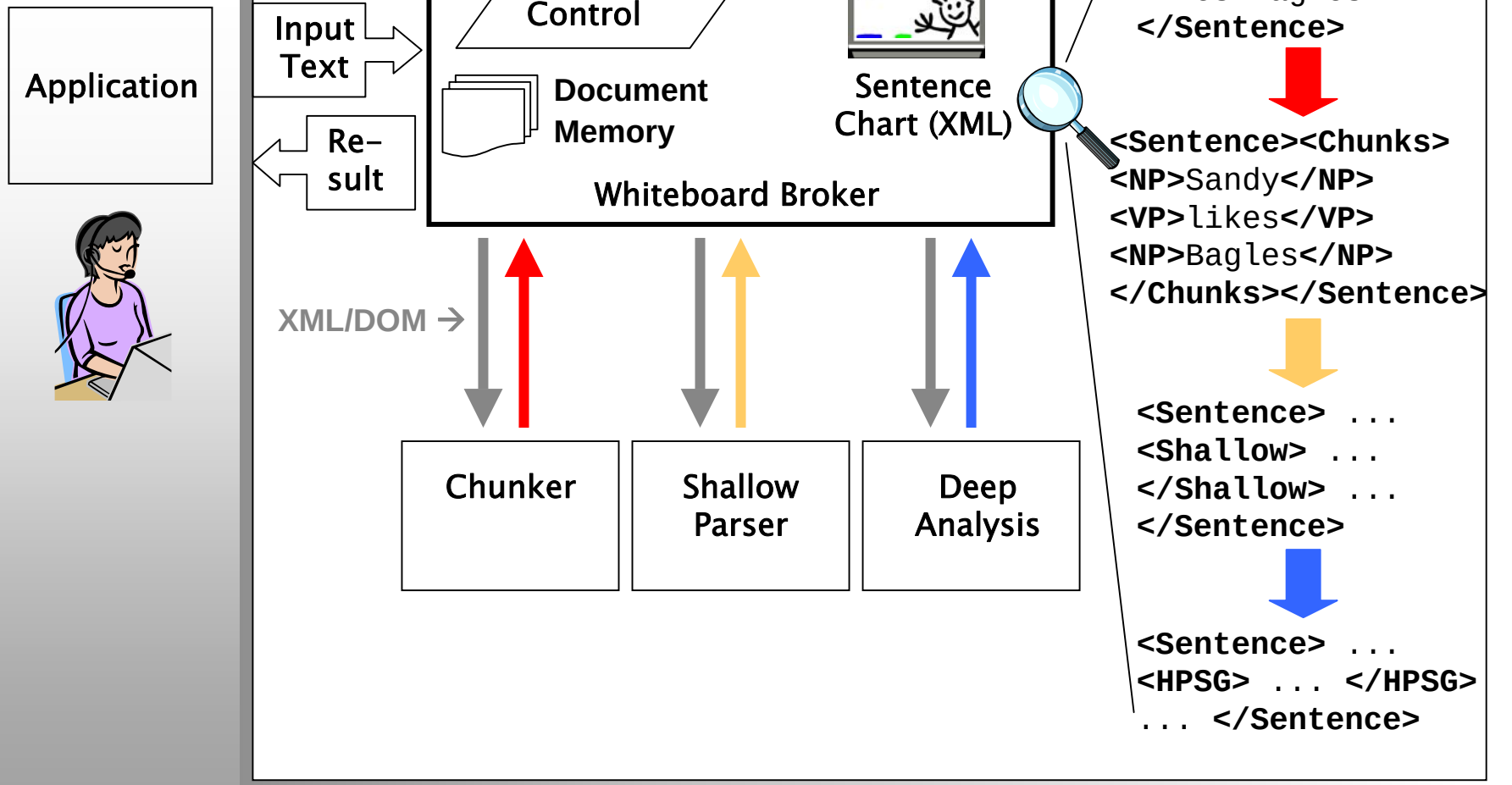


THE APPROACH

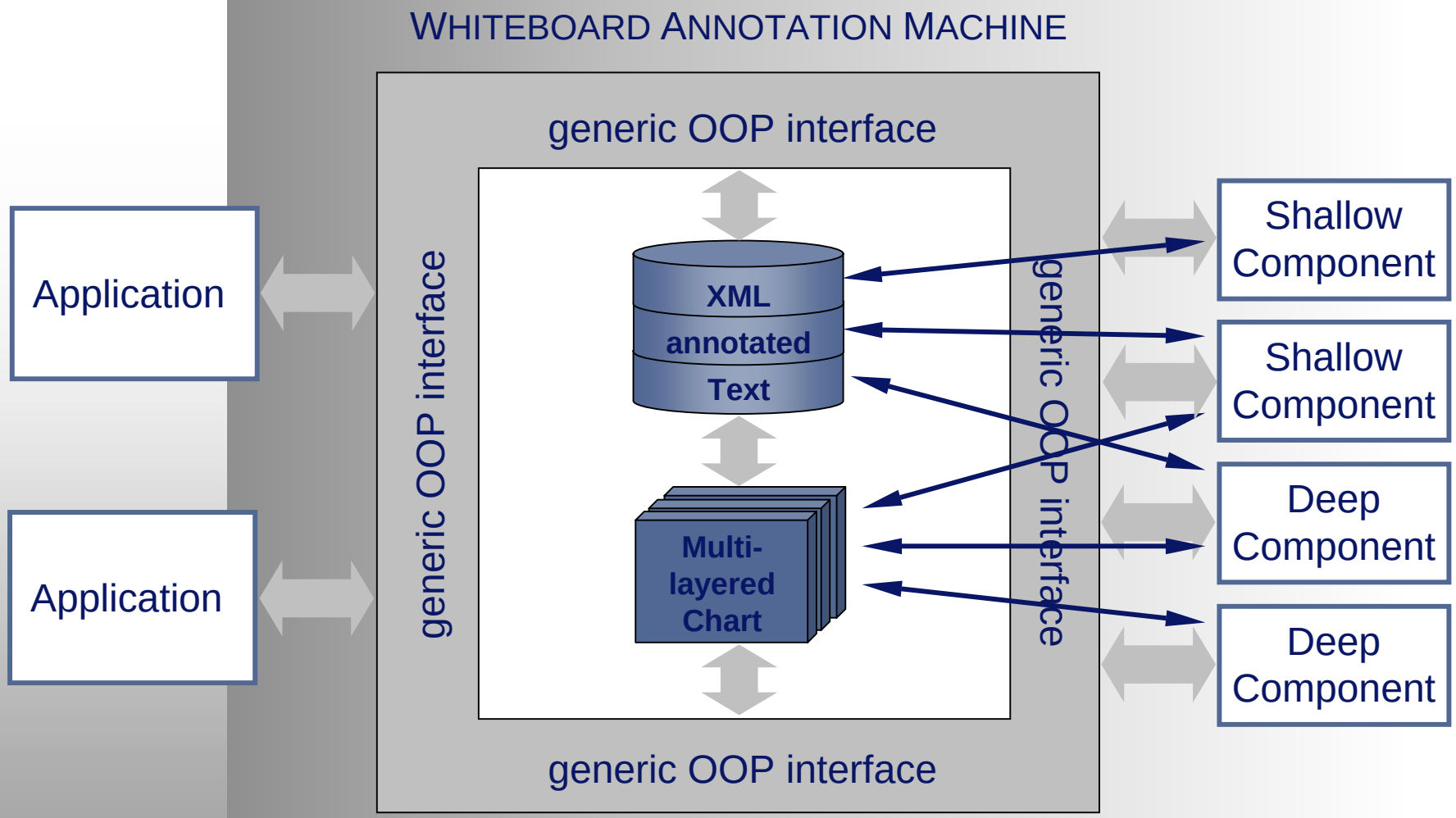
WHITEBOARD

- **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**

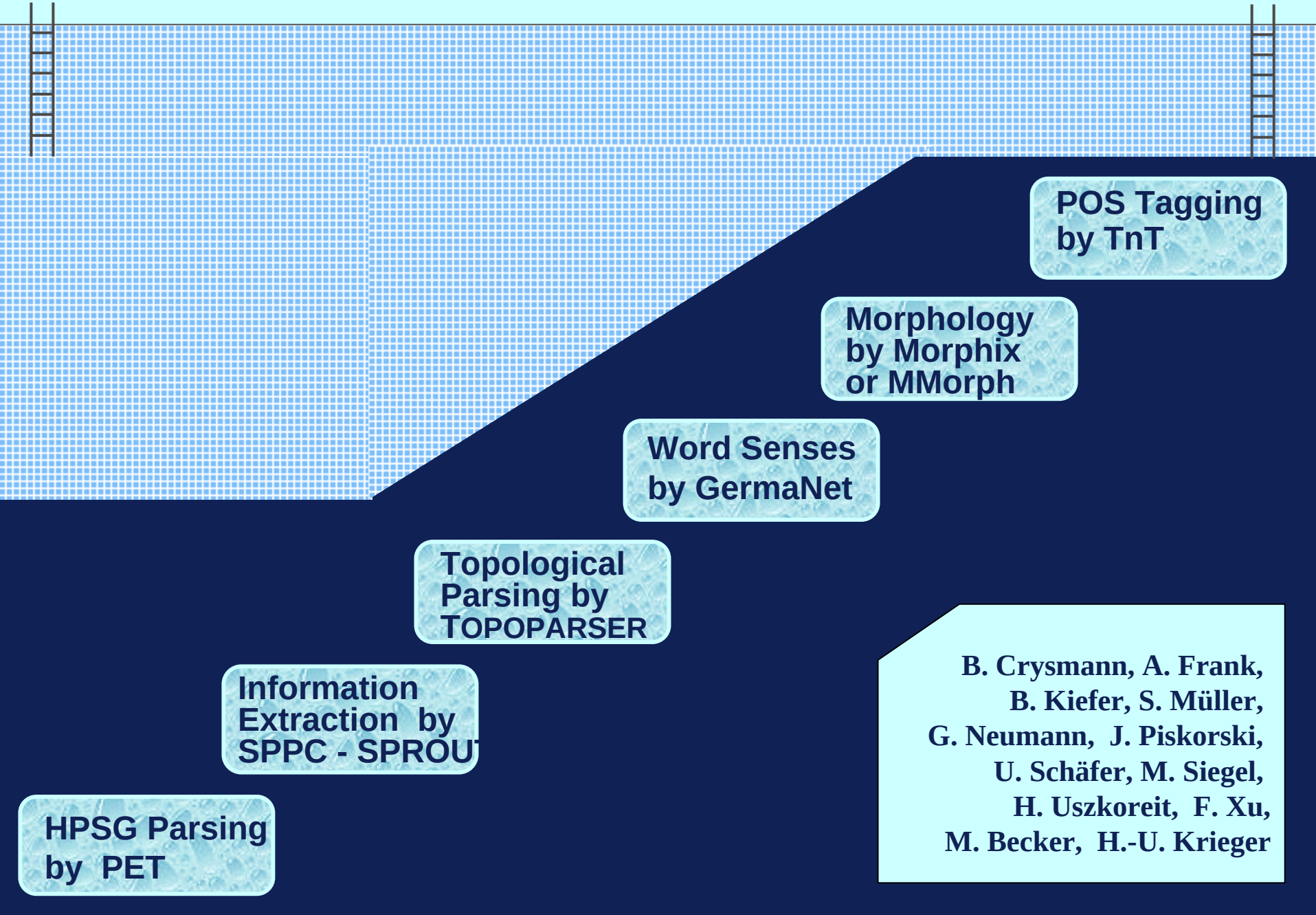
WHITEBOARD


Application
Input Text
Re-sult
Negotiating Control
Document Memory
Sentence Chart (XML)
Whiteboard Broker
XML/DOM →
Chunker
Shallow Parser
Deep Analysis
<Sentence>Sandy likes Bagles </Sentence>
<Sentence><Chunks> <NP>Sandy</NP> <VP>likes</VP> <NP>Bagles</NP> </Chunks></Sentence>
<Sentence> ... <Shallow> ... </Shallow> ... </Sentence>
<Sentence> ... <HPSG> ... </HPSG> ... </Sentence>

POOL ARCHITECTURE



POOL ARCHITECTURE



B. Crysmann, A. Frank,
B. Kiefer, S. Müller,
G. Neumann, J. Piskorski,
U. Schäfer, M. Siegel,
H. Uszkoreit, F. Xu,
M. Becker, H.-U. Krieger

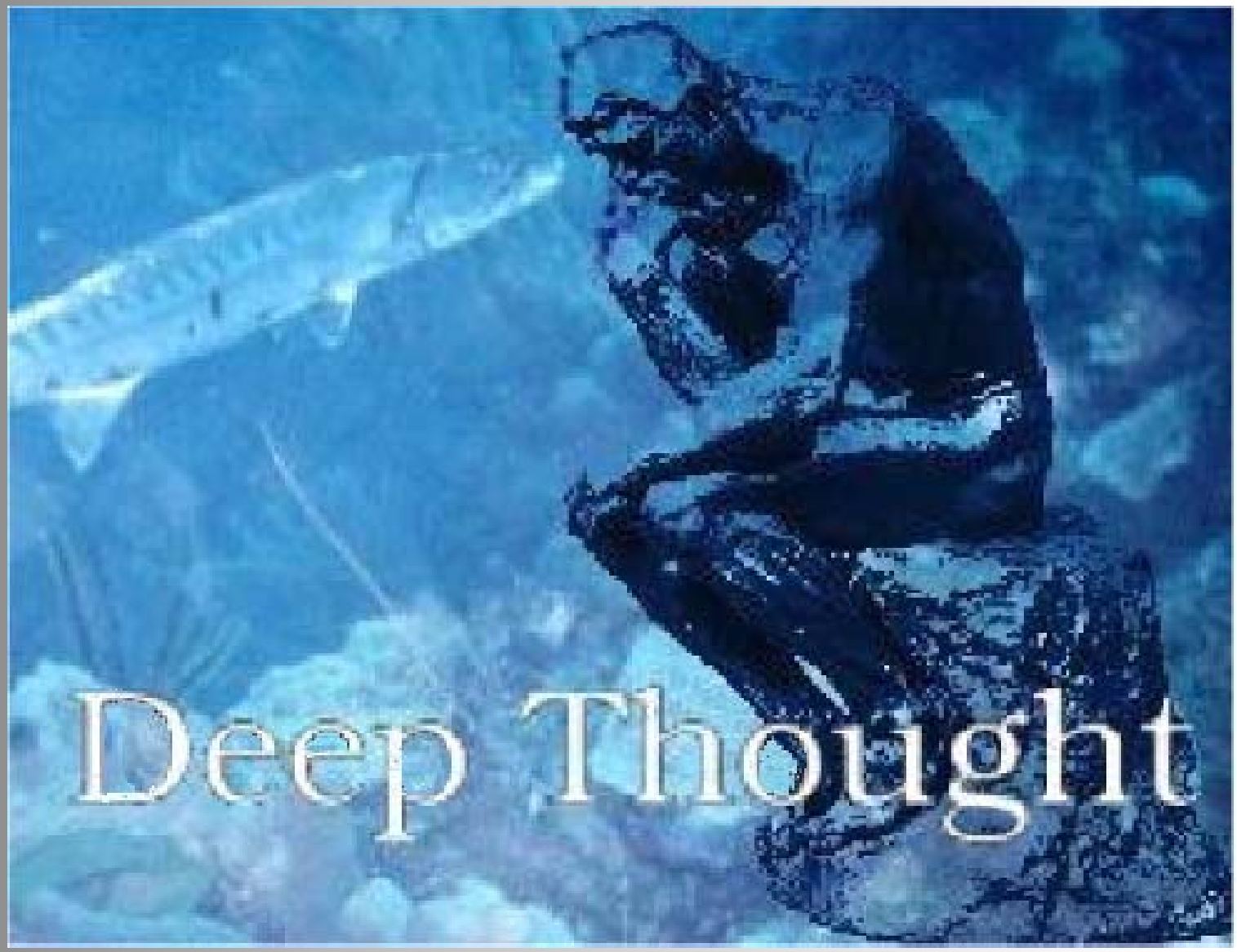
- **Shallow pre-processing and integration with deep analysis.**
- **Exploit HPSG processing for information extraction.**
- **Domain- and task specific extensions to HPSG grammars for IE.**
- **Unified semantic target structures for shallow and deep analysis (RMRS).**

Provide a maximally specific and unified output of an integrated deep and shallow NLP architecture.

- **Dynamic stochastic techniques for**
  - **optimal selection from ambiguous system output**
  - **improving robustness and efficiency**
- **Comparative evaluation of parse selection methods**
  - **Unsupervised training methods based on CFG approximations of HPSGs.**
  - **Corpus-based methods**

Develop a dynamic stochastic HPSG modelling which is robust, efficient and domain-adaptive, thus suited for real-world applications.

An EU IST Project



The Consortium

DeepThought

Saarland University (Coordinator)

U. of Cambridge

U. of Sussex

U. of Trondheim

Stanford University

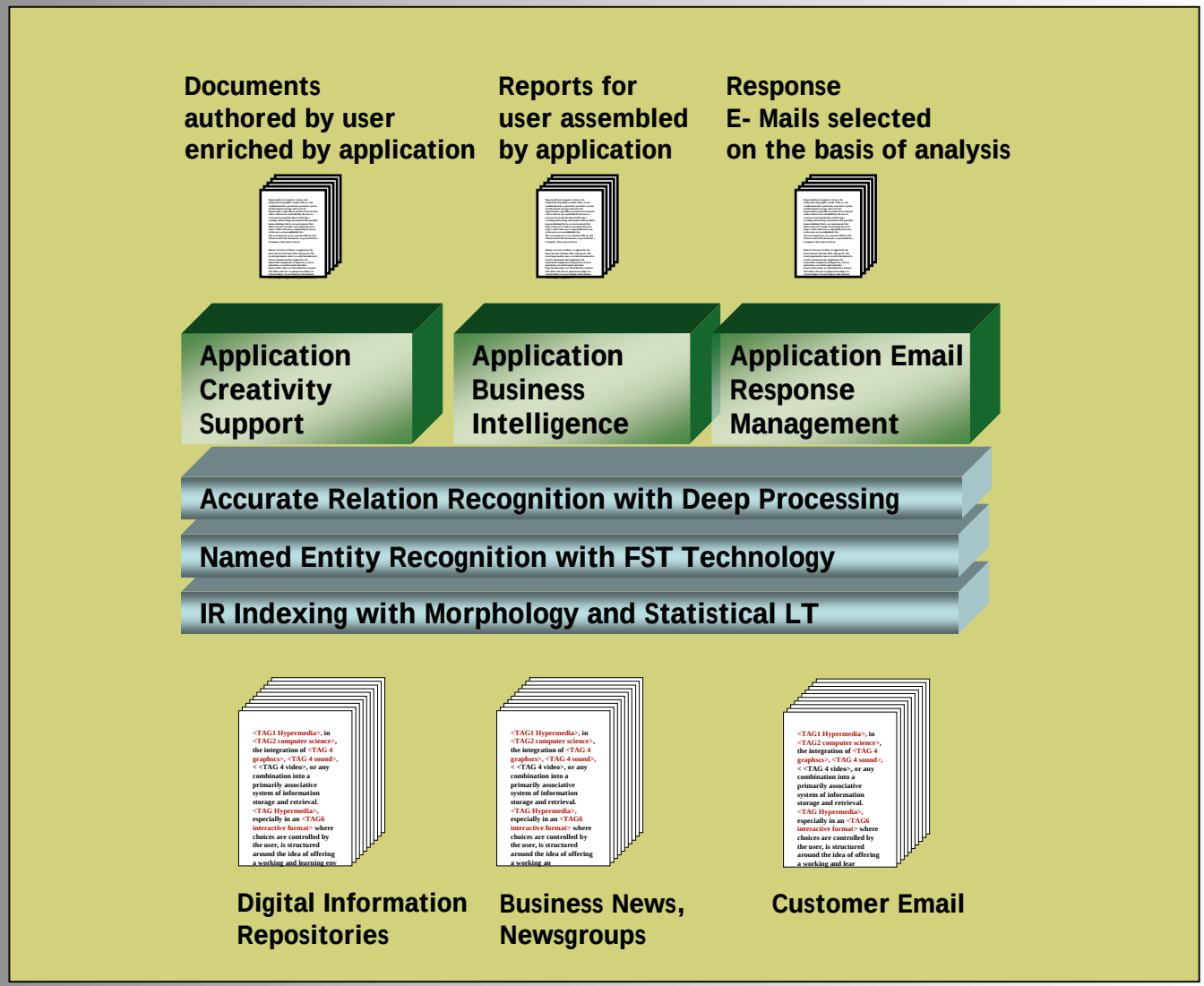
CELI Company of Torino

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

Main project results

- **RMRS provides a common representation for linguistic processing on all levels and for all languages. Shallow and deep processing produce compatible semantic representations.**
- **Different strategies for hybrid NLP have been exploited and implemented.**
- **Strategies for lexicon acquisition based on deep and statistical processing have been implemented.**
- **The theory and implementations are inherently multilingual.**
  - **The Grammar Matrix provides ways to implement new grammars for new languages based on insights from former implementations.**

DeepThought

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 RMRS 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 results

- **RMRS provides a common representation for linguistic processing on all levels and for all languages. Shallow and deep processing produce compatible semantic representations.**
- **Different strategies for hybrid NLP exploited and implemented.**
- **Strategies for lexicon acquisition based on deep and statistical processing implemented.**
- **The theory and implementations are inherently multilingual.**
  - **The Grammar Matrix provides ways to implement new grammars for new languages based on insights from former implementations.**

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.**
- **Integration of disambiguation models into a high-performance deep parser.**

DeepThought

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**

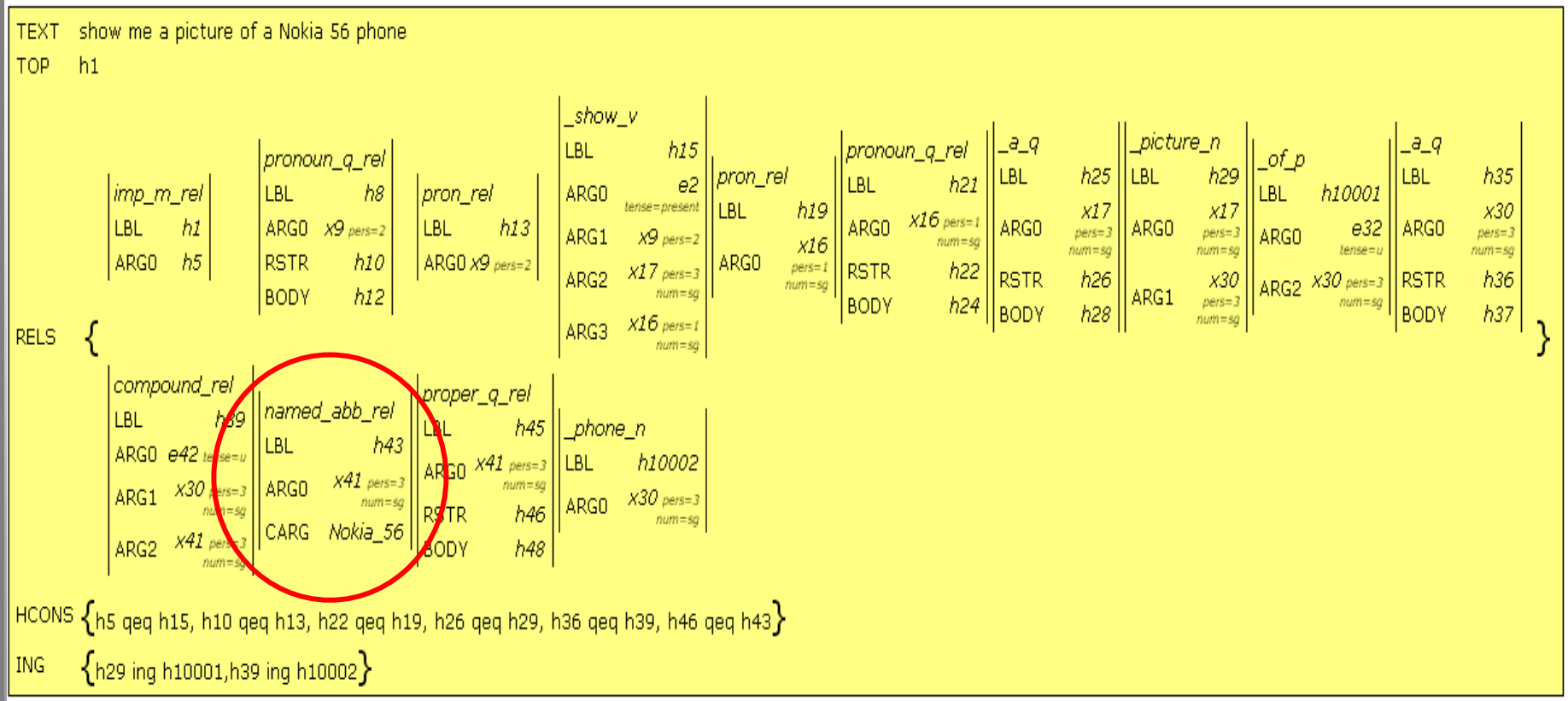
DeepThought

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.**

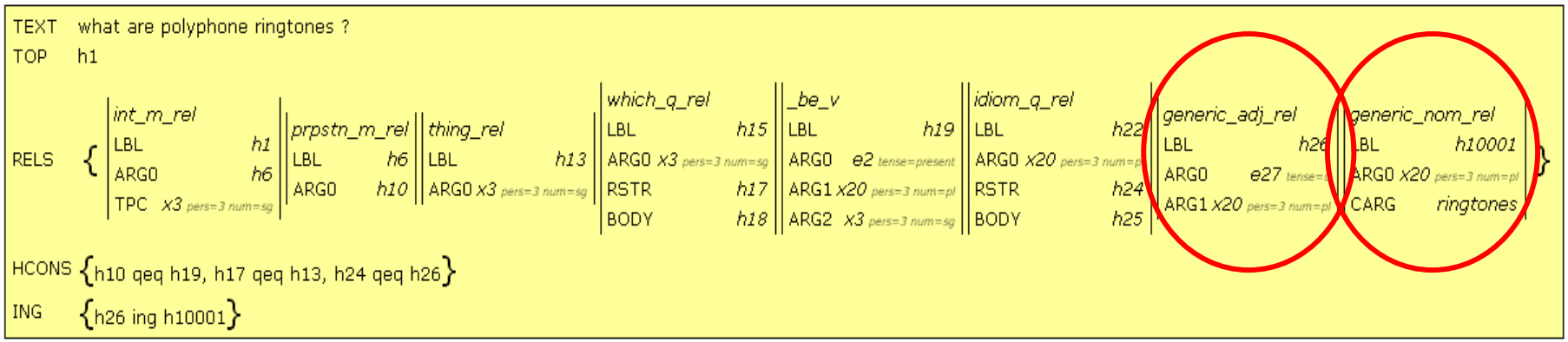
Hybrid Natural Language Processing

- **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

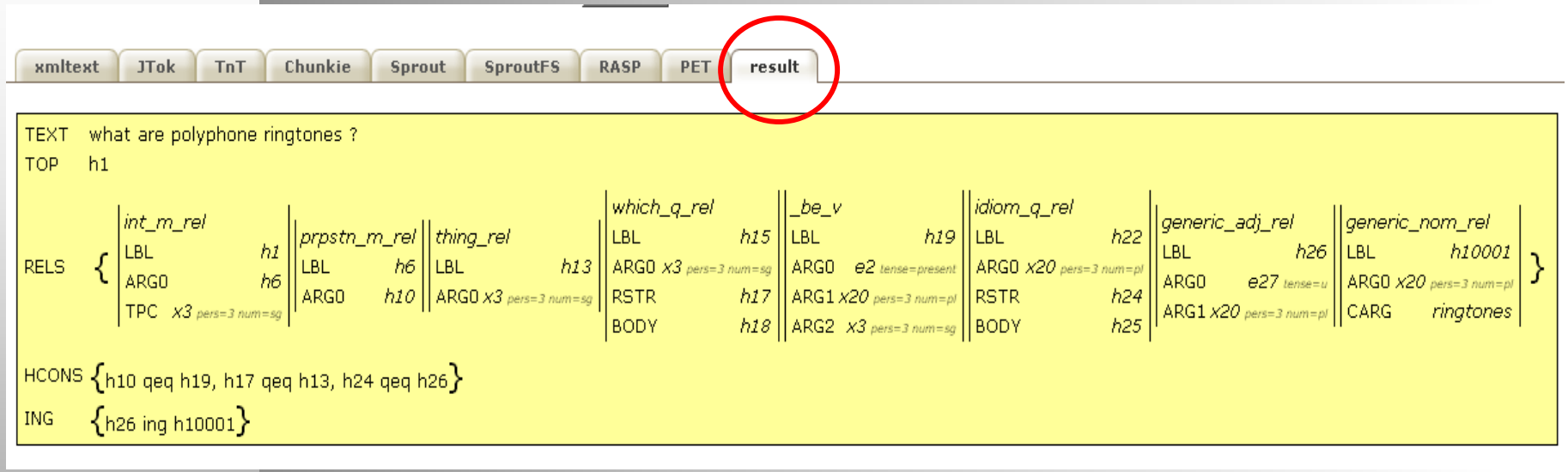
- **The analysis results of NLP tools at lower processing levels can be used by components at higher levels.**
  - **For example, the deep linguistic analysis module PET uses default lexicon entries for Part-of-Speech tags that the POS tagger TnT delivers.**



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}

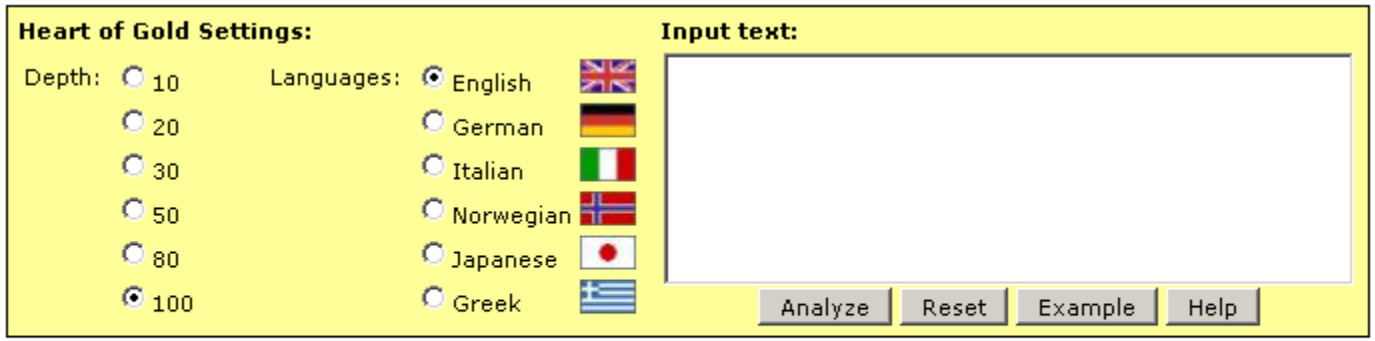
TEXT I want a new phone do you ?
TOP h1

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