

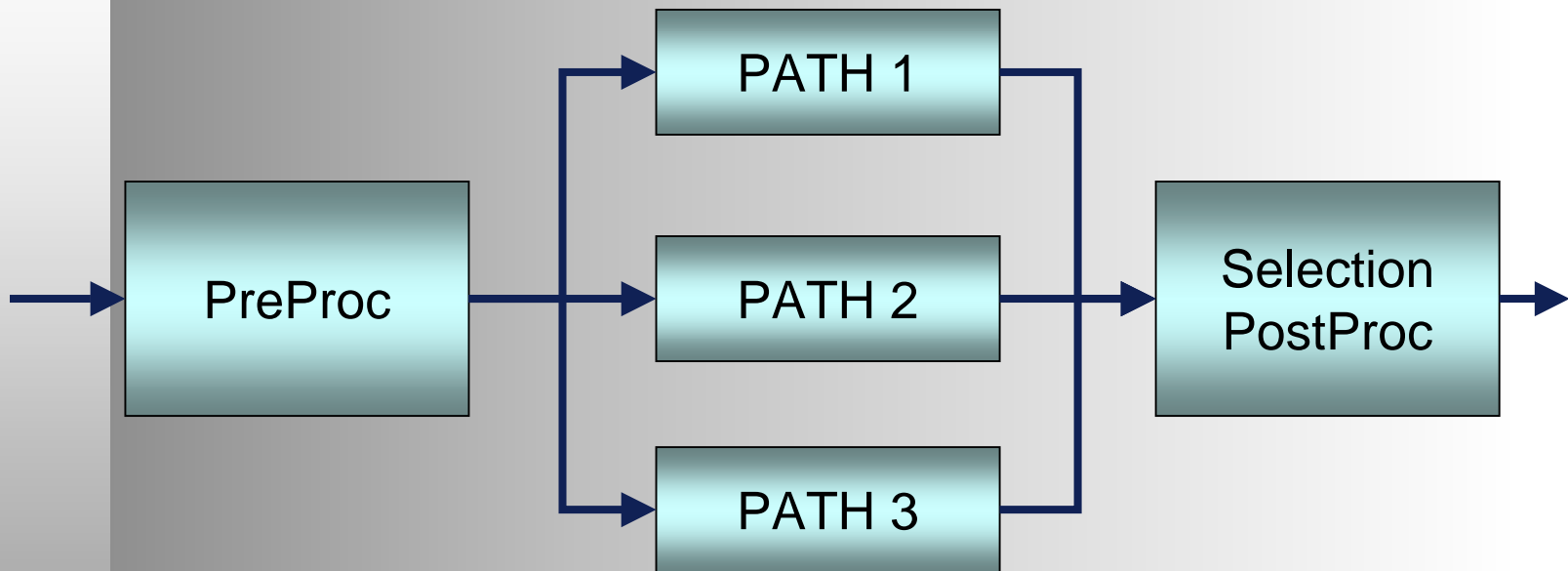
# Hybrid NLP

# 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**

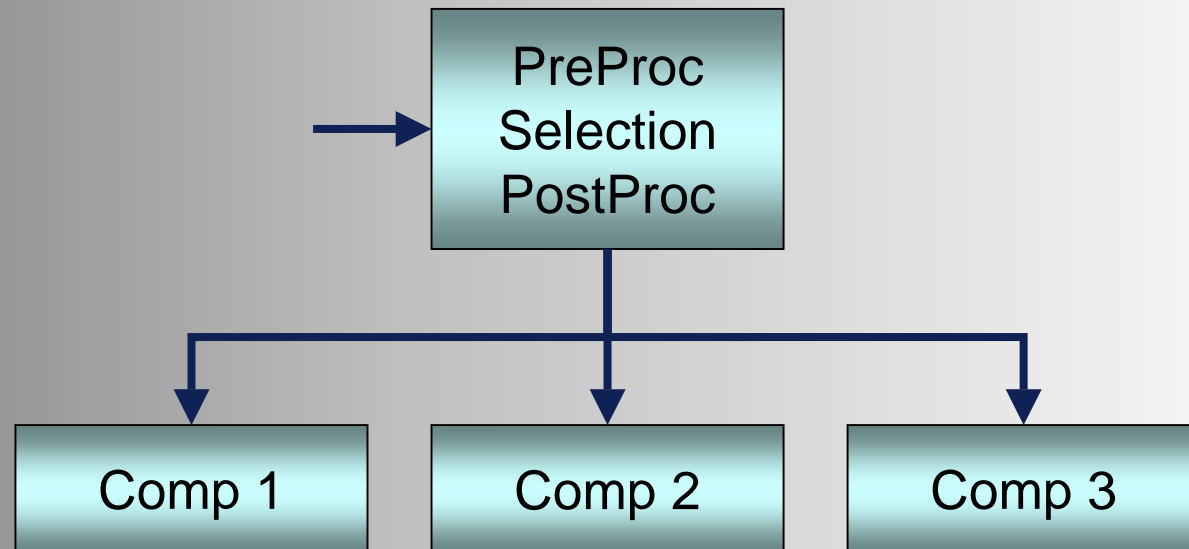


# Selection Mechanism

- **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**

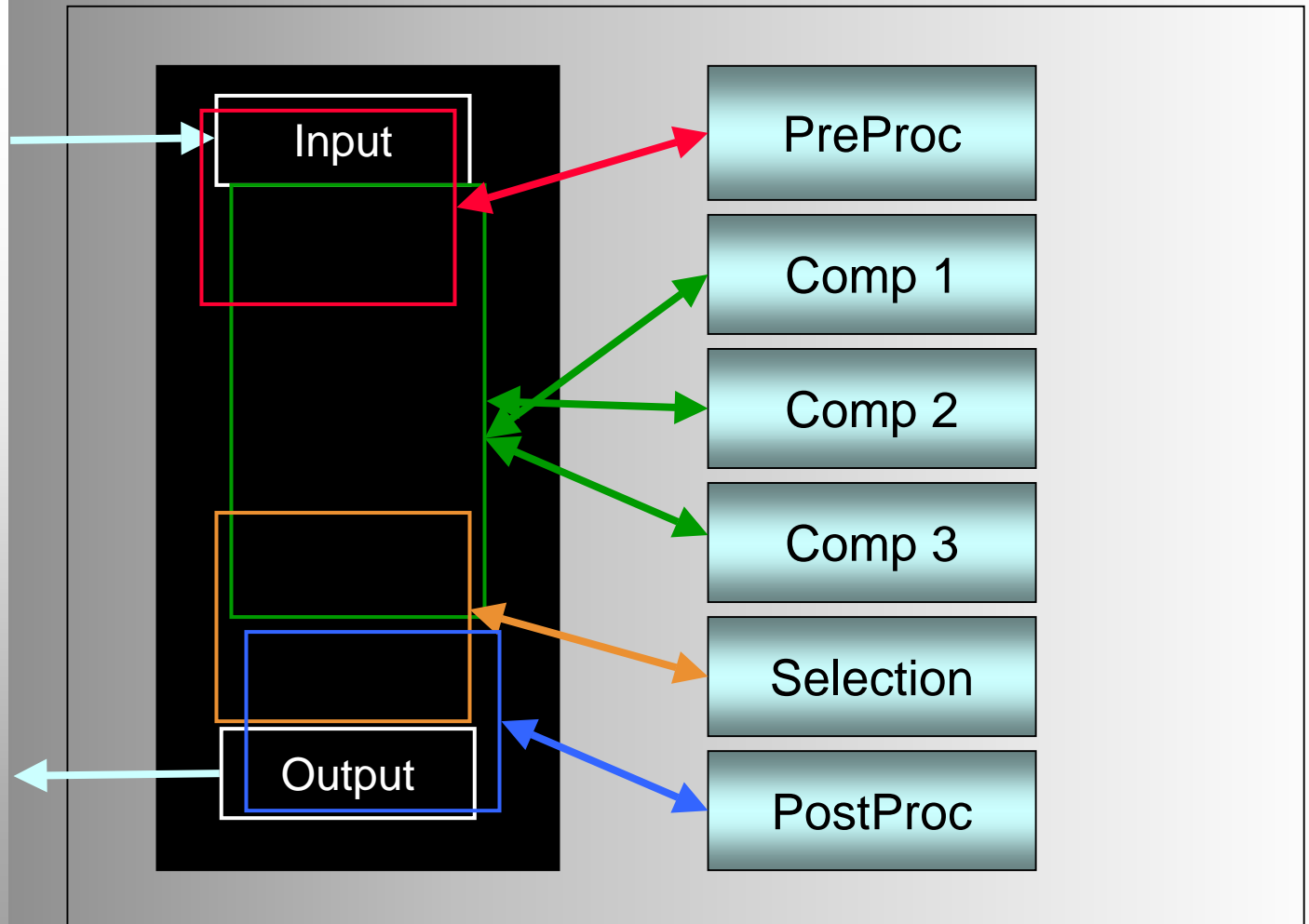
# One Integrated System

- **Manager Architecture**



# One Integrated System

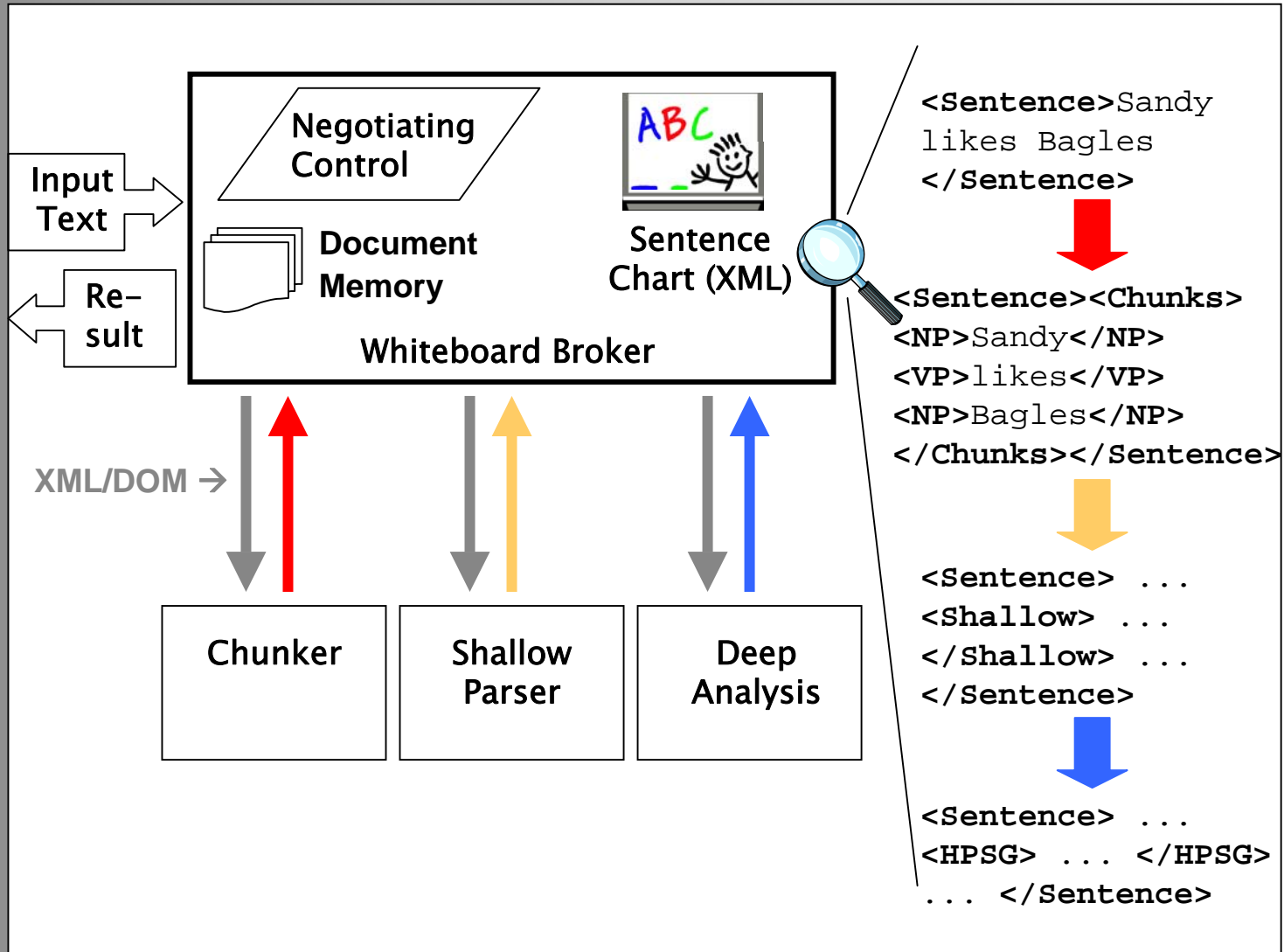
## ■ Blackboard Architecture



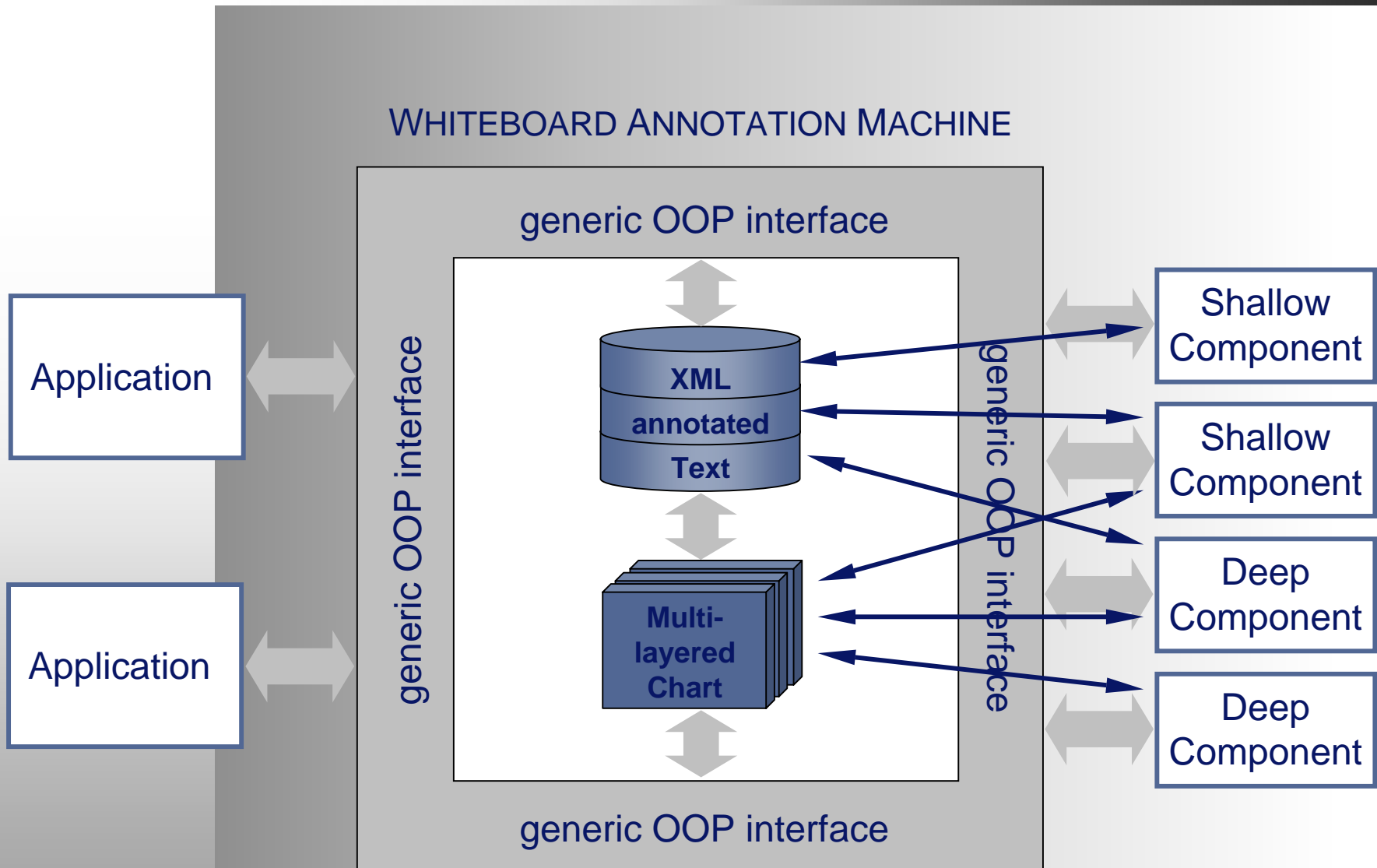
- **Combination of shallow and deep processing by incremental text enrichment**
  
- **Shallow Methods**
  - **POS Tagging -- TnT**
  - **Morphology -- Morphix and MMorph**
  - **Word Nets -- GermaNet**
  - **Information Extraction -- SPPC and SPROUT**
  
- **Deep Methods**
  - **HPSG Parsing -- PET**

# WHITEBOARD

Application



# POOL ARCHITECTURE



# POOL ARCHITECTURE



**POS Tagging  
by TnT**

**Morphology  
by Morphix  
or M Morph**

**Word Senses  
by GermaNet**

**Topological  
Parsing by  
TOPOPARSER**

**Information  
Extraction by  
SPPC - SPROU**

**HPSG Parsing  
by PET**

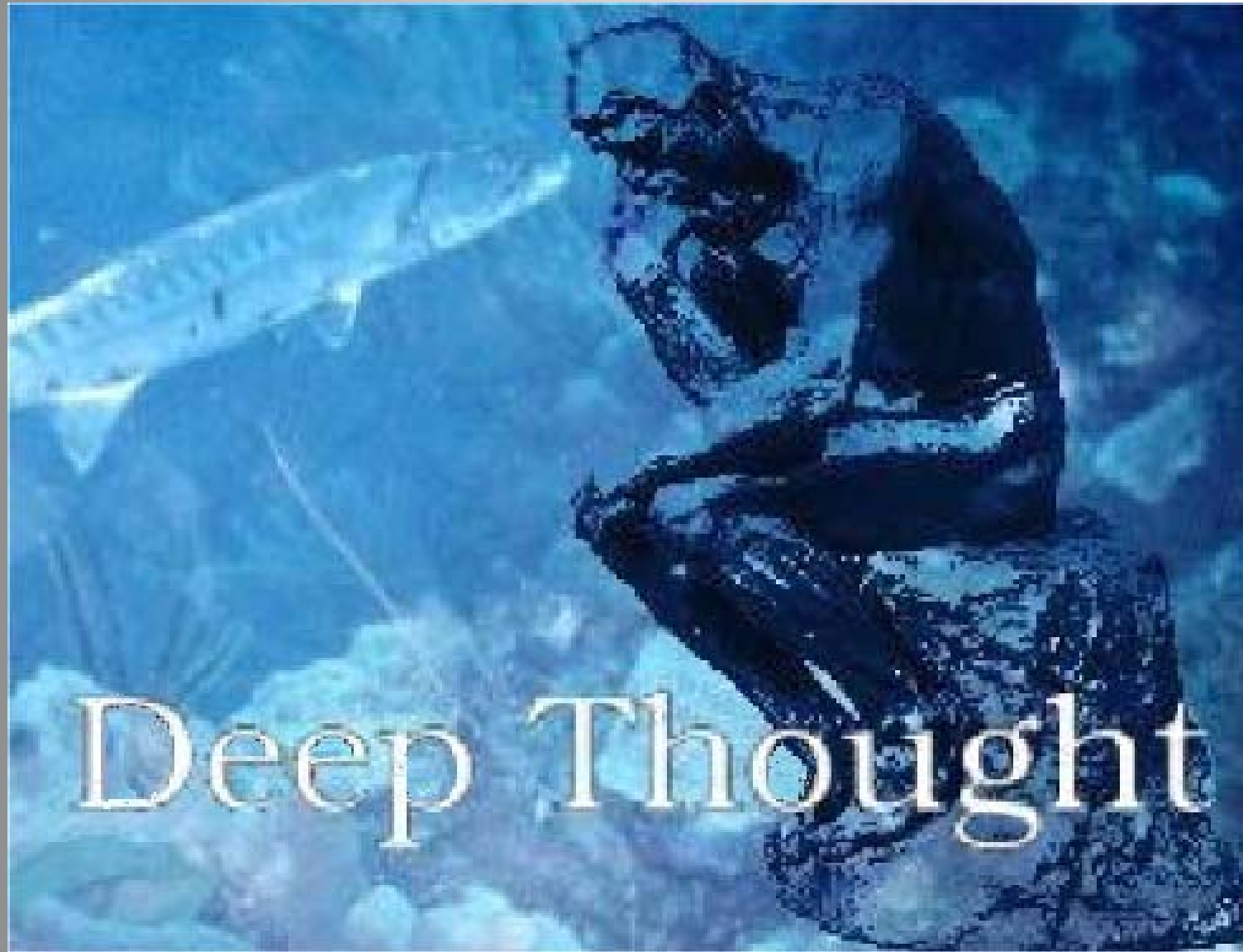
**B. Crysmann, A. Frank,  
B. Kiefer, S. Müller,  
G. Neumann, J. Piskorski,  
U. Schäfer, M. Siegel,  
H. Uszkoreit, F. Xu,  
M. Becker, H.-U. Krieger**

- **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.**





**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

Documents authored by user enriched by application



Application Creativity Support

Reports for user assembled by application



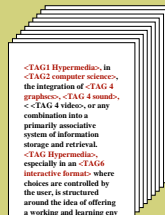
Application Business Intelligence

Response E- Mails selected on the basis of analysis

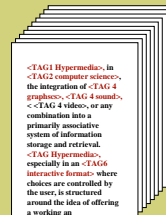


Application Email Response Management

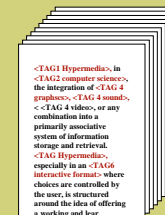
Accurate Relation Recognition with Deep Processing  
 Named Entity Recognition with FST Technology  
 IR Indexing with Morphology and Statistical LT



Digital Information Repositories



Business News, Newsgroups



Customer Email

- 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.

## NL Input

Every experienced vice president served on a committee.

## Event-Based LF

$\text{every}(x), \text{experienced}(x), \text{VP}(x), \text{serve1}(e, x), \text{spast}(e), \text{on2}(e', e, y),$   
 $\text{some}(y), \text{committee}(y).$

## Robust Minimal Recursion Semantics (RMRS)

$\text{every}(x1), \text{experienced}(x2), \text{VP}(x3), \text{serve1}(e4), \text{spast}(e4), \text{on2}(e5),$   
 $\text{some}(x6), \text{committee}(x7)$

$\text{ARG1}(e4, x1), \text{ARG1}(e5, e4), \text{ARG2}(e5, x6),$

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

## 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.

# 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.

# 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.

TEXT show me a picture of a Nokia 56 phone

TOP h1

|        |                                  |                                  |                                  |                                  |                                  |                                  |                                  |                                  |                                  |                                  |
|--------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
|        | <i>imp_m_rel</i>                 | <i>pronoun_q_rel</i>             | <i>pron_rel</i>                  | <i>_show_v</i>                   | <i>pron_rel</i>                  | <i>pronoun_q_rel</i>             | <i>_a_q</i>                      | <i>_picture_n</i>                | <i>_of_p</i>                     | <i>_a_q</i>                      |
|        | LBL h1                           | LBL h8                           | LBL h13                          | LBL h15                          | LBL h19                          | LBL h21                          | LBL h25                          | LBL h29                          | LBL h10001                       | LBL h35                          |
|        | ARGO h5                          | ARGO x9 pers=2                   | ARGO x9 pers=2                   | ARGO e2<br><i>tense=present</i>  | ARGO h19                         | ARGO x16 pers=1<br><i>num=sg</i> | ARGO x17<br><i>pers=3 num=sg</i> | ARGO x17<br><i>pers=3 num=sg</i> | ARGO e32<br><i>tense=u</i>       | ARGO x30<br><i>pers=3 num=sg</i> |
|        |                                  | RSTR h10                         | ARGO x9 pers=2                   | ARG1 x9 pers=2                   | ARGO x16<br><i>pers=1 num=sg</i> | RSTR h22                         | RSTR h26                         | ARG1 x30<br><i>pers=3 num=sg</i> | ARG2 x30 pers=3<br><i>num=sg</i> | RSTR h36                         |
|        |                                  | BODY h12                         |                                  | ARG2 x17 pers=3<br><i>num=sg</i> | BODY h24                         | BODY h28                         |                                  |                                  |                                  | BODY h37                         |
| RELS { |                                  |                                  |                                  | ARG3 x16 pers=1<br><i>num=sg</i> |                                  |                                  |                                  |                                  |                                  |                                  |
|        | <i>compound_rel</i>              | <i>named_abb_rel</i>             | <i>proper_q_rel</i>              |                                  | <i>_phone_n</i>                  |                                  |                                  |                                  |                                  |                                  |
|        | LBL h39                          | LBL h43                          | LBL h45                          |                                  | LBL h10002                       |                                  |                                  |                                  |                                  |                                  |
|        | ARGO e42<br><i>tense=u</i>       | ARGO x41 pers=3<br><i>num=sg</i> | ARGO x41 pers=3<br><i>num=sg</i> |                                  | ARGO x30 pers=3<br><i>num=sg</i> |                                  |                                  |                                  |                                  |                                  |
|        | ARG1 x30<br><i>num=sg</i>        | ARGO x41 pers=3<br><i>num=sg</i> | RSTR h46                         |                                  |                                  |                                  |                                  |                                  |                                  |                                  |
|        | ARG2 x41 pers=3<br><i>num=sg</i> | CARG Nokia_56                    | BODY h48                         |                                  |                                  |                                  |                                  |                                  |                                  |                                  |

HCONS {h5 qeq h15, h10 qeq h13, h22 qeq h19, h26 qeq h29, h36 qeq h39, h46 qeq h43}

ING {h29 ing h10001, h39 ing h10002}

# 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.

```

TEXT  what are polyphone ringtones ?
TOP   h1

RELS  {
  int_m_rel  |  prpstn_m_rel  |  thing_rel  |  which_q_rel  |  _be_v  |  idiom_q_rel  |  generic_adj_rel  |  generic_nom_rel
  LBL        |  LBL        |  LBL        |  LBL          |  LBL    |  LBL         |  LBL             |  LBL
  ARG0       |  ARG0       |  ARG0 x3   |  ARG0 x3     |  ARG0 e2 |  ARG0 x20    |  ARG0            |  ARG0 x20
  TPC x3    |  ARG0 h10  |  ARG0 x3   |  RSTR h17    |  ARG1 x20 |  RSTR h24    |  ARG1 x20        |  ARG0 x20
  pers=3   |  num=sg    |  pers=3   |  num=sg     |  tense=present |  pers=3 num=p |  tense=...      |  pers=3 num=pl
  num=sg   |  num=sg    |  num=sg   |  num=sg     |  num=pl    |  num=pl     |  num=pl         |  num=pl
  num=pl   |  num=pl    |  num=pl   |  num=pl     |  num=sg    |  num=sg     |  num=sg         |  num=sg
}

HCONS {h10 qeq h19, h17 qeq h13, h24 qeq h26}
ING   {h26 ing h10001}
    
```

# Hybrid Natural Language Processing

***Deliver the deepest result.***

***The Email Autoresponse application follows this approach***

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

TEXT what are polyphone ringtones ?  
TOP h1

|        |                      |  |                     |  |                       |  |                       |  |                        |  |                        |  |                        |  |                        |   |
|--------|----------------------|--|---------------------|--|-----------------------|--|-----------------------|--|------------------------|--|------------------------|--|------------------------|--|------------------------|---|
| RELS { | <i>int_m_rel</i>     |  | <i>prpstn_m_rel</i> |  | <i>thing_rel</i>      |  | <i>which_q_rel</i>    |  | <i>_be_v</i>           |  | <i>idiom_q_rel</i>     |  | <i>generic_adj_rel</i> |  | <i>generic_nom_rel</i> | } |
|        | LBL h1               |  | LBL h6              |  | LBL h13               |  | ARGO x3 pers=3 num=sg |  | ARGO e2 tense=present  |  | ARGO x20 pers=3 num=pl |  | LBL h26                |  | LBL h10001             |   |
|        | ARGO h6              |  | ARGO h10            |  | ARGO x3 pers=3 num=sg |  | RSTR h17              |  | ARG1 x20 pers=3 num=pl |  | RSTR h24               |  | ARG0 e27 tense=u       |  | ARGO x20 pers=3 num=pl |   |
|        | TPC x3 pers=3 num=sg |  |                     |  |                       |  | BODY h18              |  | ARG2 x3 pers=3 num=sg  |  | BODY h25               |  | ARG1 x20 pers=3 num=pl |  | CARG ringtones         |   |

HCONS {h10 qeq h19, h17 qeq h13, h24 qeq h26}

ING {h26 ing h10001}

# 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 { | <i>prpstn_m_rel</i> | <i>pron_rel</i>       | <i>pronoun_q_rel</i>  | <i>_want_v</i>         | <i>_a_q</i>            | <i>_new_a</i>          | <i>_phone_n</i>        |
|        | LBL h1              | LBL h8                | LBL h11               | LBL h15                | LBL h18                | LBL h22                | LBL h10001             |
|        | ARG0 h5             | ARG0 x9 pers=1 num=sg | ARG0 x9 pers=1 num=sg | ARG0 e2 tense=present  | ARG0 x16 pers=3 num=sg | ARG0 e24 tense=u       | ARG0 x16 pers=3 num=sg |
|        |                     | RSTR h12              | RSTR h12              | ARG1 x9 pers=1 num=sg  | RSTR h19               | ARG1 x16 pers=3 num=sg | ARG0 x16 pers=3 num=sg |
|        |                     | BODY h14              | BODY h14              | ARG2 x16 pers=3 num=sg | BODY h20               |                        |                        |

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 { | <i>_do_v</i>    | <i>pron_rel</i> | <i>pronoun_q_rel</i> |
|        | LBL h1          | LBL h7          | LBL h9               |
|        | ARG0 e2 tense=u | ARG0 x3 pers=2  | ARG0 x3 pers=2       |
|        | ARG1 u4         | RSTR h10        | RSTR h10             |
|        | ARG2 u5         | BODY h12        | BODY h12             |
|        | ARG3 x3 pers=2  |                 |                      |


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.**


**Heart of Gold Settings:**


Depth:  10      Languages:  English 

20                       German 

30                       Italian 

50                       Norwegian 

80                       Japanese 

100                       Greek 

**Input text:**

Analyze    Reset    Example    Help

Please note that JavaScript™ is required.

# Hybrid Natural Language Processing

## The different modules use a compatible output formalism, RMRS.

In case of shallower modules, this robust semantic

xmltext JTok TnT Chunkie Sprout SproutFS RASP PET result

TEXT Kim intended to buy the phone  
TOP h31

RELS {

|                     |     |                     |    |                  |     |                  |               |                     |     |               |     |               |     |
|---------------------|-----|---------------------|----|------------------|-----|------------------|---------------|---------------------|-----|---------------|-----|---------------|-----|
| <i>prpstn_m_rel</i> |     | <i>proper_q_rel</i> |    | <i>named_rel</i> |     | <i>_intend_v</i> |               | <i>prpstn_m_rel</i> |     | <i>_buy_v</i> |     | <i>_the_q</i> |     |
| LBL                 | h31 | LBL                 | h1 | LBL              | h3  | LBL              | h6            | LBL                 | h24 | LBL           | h10 | LBL           | h12 |
| ARG0                | h33 | ARG0                | x2 | ARG0             | x2  | ARG0             | e7 tense=past | ARG0                | h26 | ARG0          | e11 | ARG0          | x13 |
|                     |     | RSTR                | h4 | ARG0             | x2  | ARG1             | x2            | ARG0                | h26 | ARG0          | x13 | RSTR          | h19 |
|                     |     | BODY                | h5 | CARG             | kim | ARG2             | h24           |                     |     | ARG2          | x13 | BODY          | h16 |

ARGO h14

HCONS {h33 qeq h6, h4 qeq h3, h26 qeq h10, h19 qeq h14}

ING {}

xmltext JTok TnT Chunkie Sprout SproutFS RASP PET result

TEXT Kim intended to buy the phone  
TOP h1

RELS {

|                     |    |                     |            |                  |            |                  |               |                     |     |               |             |               |            |
|---------------------|----|---------------------|------------|------------------|------------|------------------|---------------|---------------------|-----|---------------|-------------|---------------|------------|
| <i>prpstn_m_rel</i> |    | <i>proper_q_rel</i> |            | <i>named_rel</i> |            | <i>_intend_v</i> |               | <i>prpstn_m_rel</i> |     | <i>_buy_v</i> |             | <i>_the_q</i> |            |
| LBL                 | h1 | LBL                 | h8         | LBL              | h13        | LBL              | h15           | LBL                 | h17 | LBL           | h23         | LBL           | h27        |
| ARG0                | h5 | ARG0                | x10 num=sq | ARG0             | x10 num=sq | ARG0             | e2 tense=past | ARG0                | h20 | ARG0          | e25 tense=u | ARG0          | x24 num=sq |
|                     |    | RSTR                | h9         | ARG0             | x10 num=sq | ARG1             | x10 num=sq    | ARG0                | h20 | ARG1          | x10 num=sq  | RSTR          | h28        |
|                     |    | BODY                | h12        | CARG             | kim        | ARG2             | h17           |                     |     | ARG2          | x24 num=sq  | BODY          | h30        |

ARGO h31

HCONS {h5 qeq h15, h9 qeq h13, h20 qeq h23, h28 qeq h31}

ING {}