Hybrid NLP

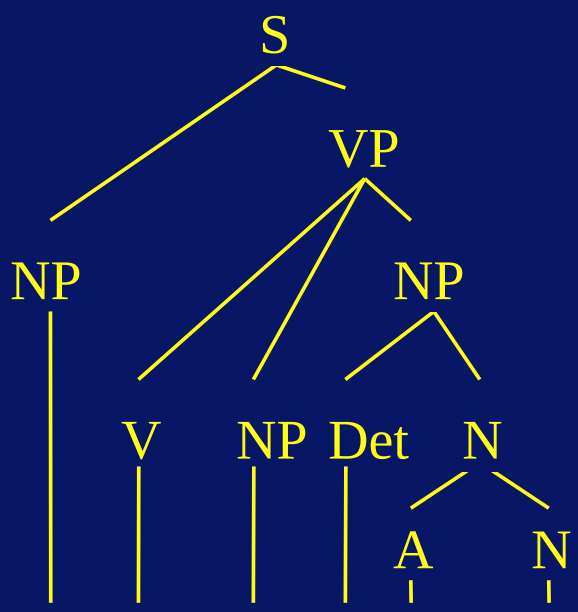
OUTLINE

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

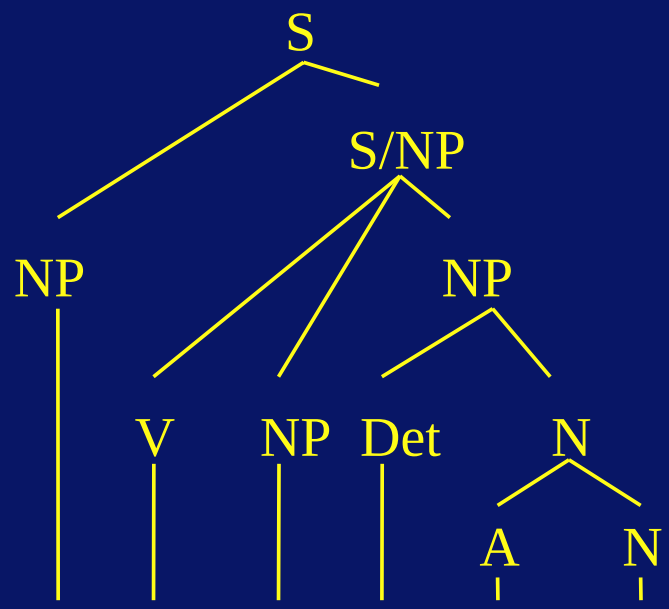
DEEP & SHALLOW PROCESSING

- **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.

APPLICATIONS

- Machine Translation

  e.g. Systran, Logos, METAL-Comprendium, IBM PT

- Access to Databases

  e.g. Core Language Engine

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 a 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.

REAL GRAMMARS

- **LinGO - English Resource Grammar**
  - **8.000 types**
  - **100.000 lines of code**
  - **average feature structure > 300 nodes**
- **German Grammar of equal size**
- **Japanese and Norwegian grammars are getting close**

International Collaboration

Tokyo

- **Tsujii Lab at the University of Tokyo**
- Tsujii, Torisawa, Ninomiya, Taura, Yoshida, Mitsoishi,...

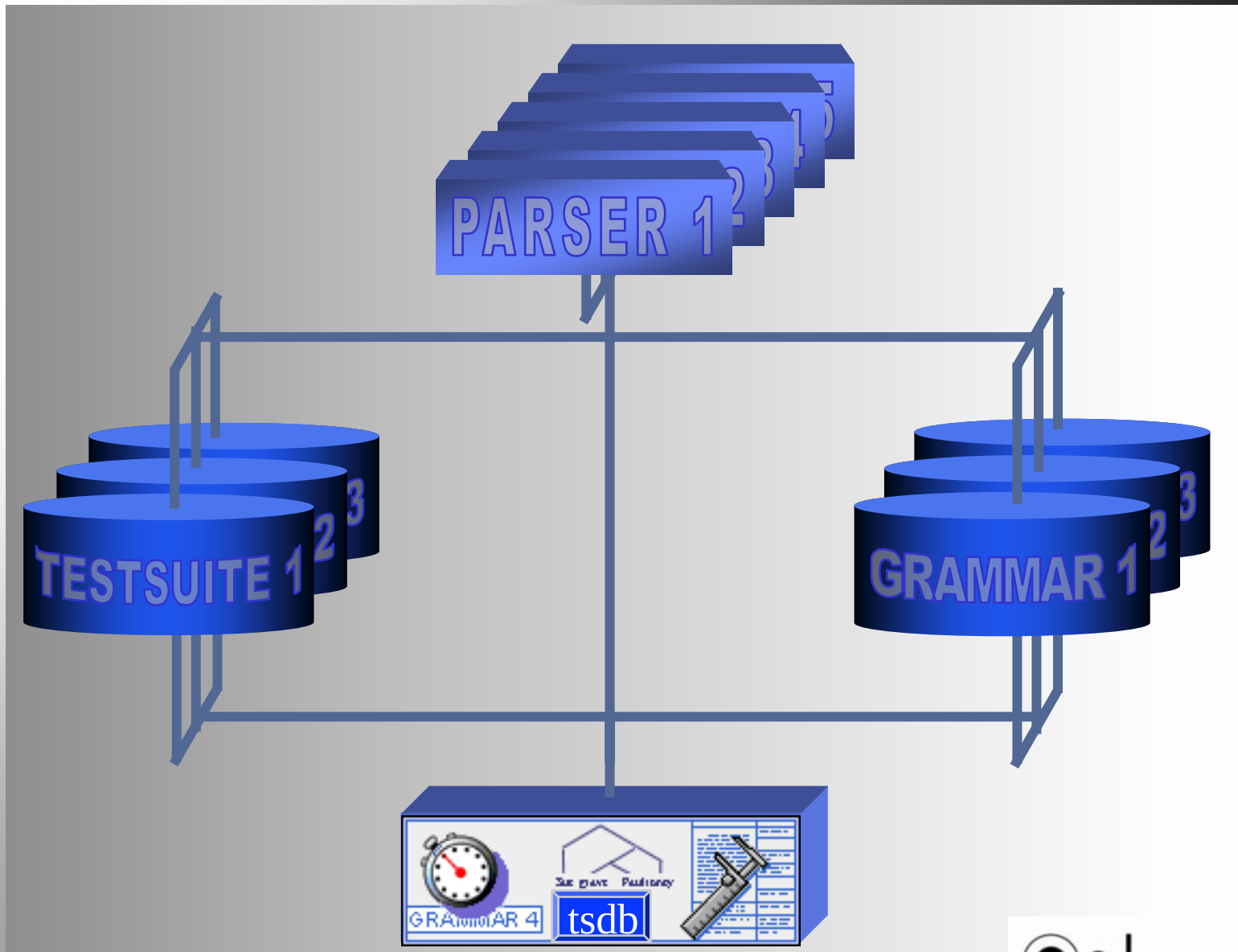
Stanford

- **HPSG Group at CSLI**
- Sag, Flickinger, Copestake, Malouf, Carroll (Brighton),...

Saarbrücken

- **LT Lab at DFKI and Dept. of CL**
- Oepen, Callmeier, Krieger, Kiefer, Ciortuz, Müller,...

The Evaluation Setup



RESULTS

- **All participating systems have benefitted from the systematic comparative evaluation**
- **Currently the fastest system is the runtime parser PET by Ulrich Callmeier (Saarbrücken)**
- **But the other parsers also improved drastically,e.g.:**
  - **LKB (Stanford, Cambridge)**
  - **LILFES (Tokyo)**
  - **PAGE (Saarbrücken)**

RESULTS

- **HPSG Parsing is now 2000 times faster than before**
- **Normal-length sentences parse in 0.1 - 1.0 seconds**
- **Steady increase in hardware efficiency will also help**

REFERENCES

- D. Flickinger, S. Oepen, H. Uszkoreit, and J. Tsujii (eds.). 2000. *Journal of Natural Language Engineering* 6 (2000) 1. Special Issue on Efficient Processing with HPSG: Methods, Systems, Evaluation. Cambridge University Press. Cambridge.
- A. Copestake. 2002. Implementing Typed Feature Structure Grammars. CSLI Publications, Stanford. Building a Large Annotated Corpus of English:
- S. Oepen, D. Flickinger, J. Tsujii, and H. Uszkoreit. 2002. *Collaborative Language Engineering. A Case Study in Efficient Grammar-based Processing.* CSLI Publications, Stanford.

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**

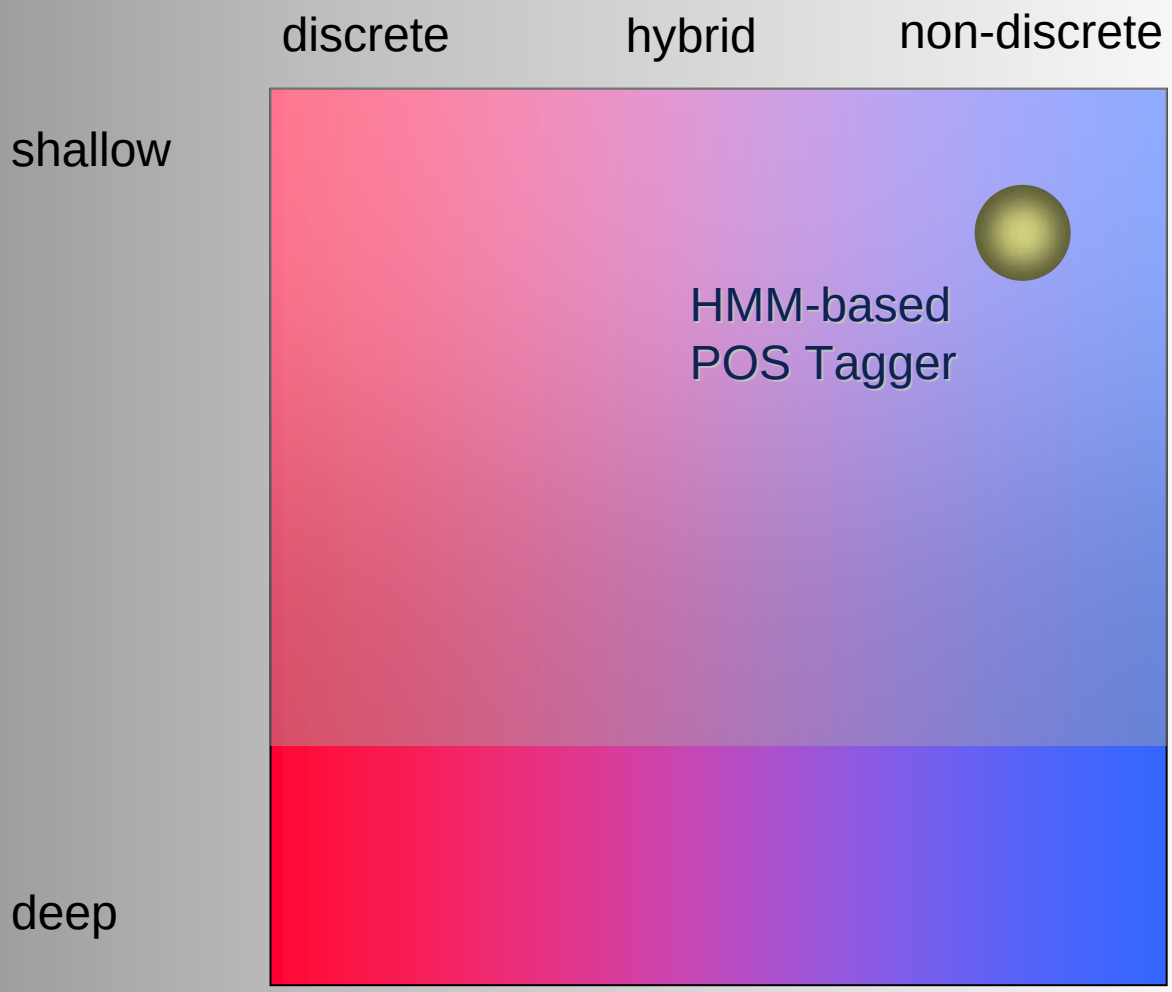
ASSUMPTIONS

- **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?**

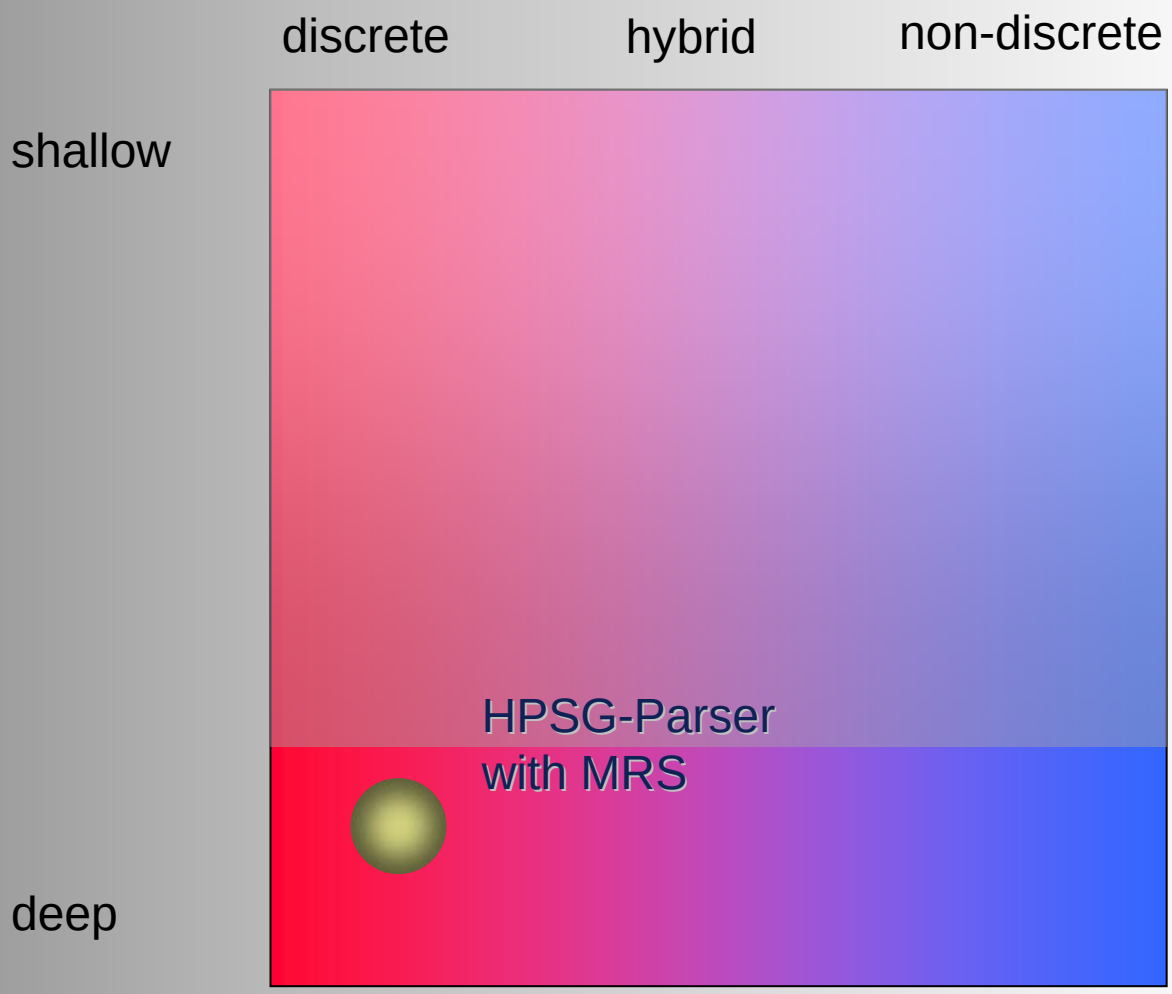
APPROACH

- **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.**

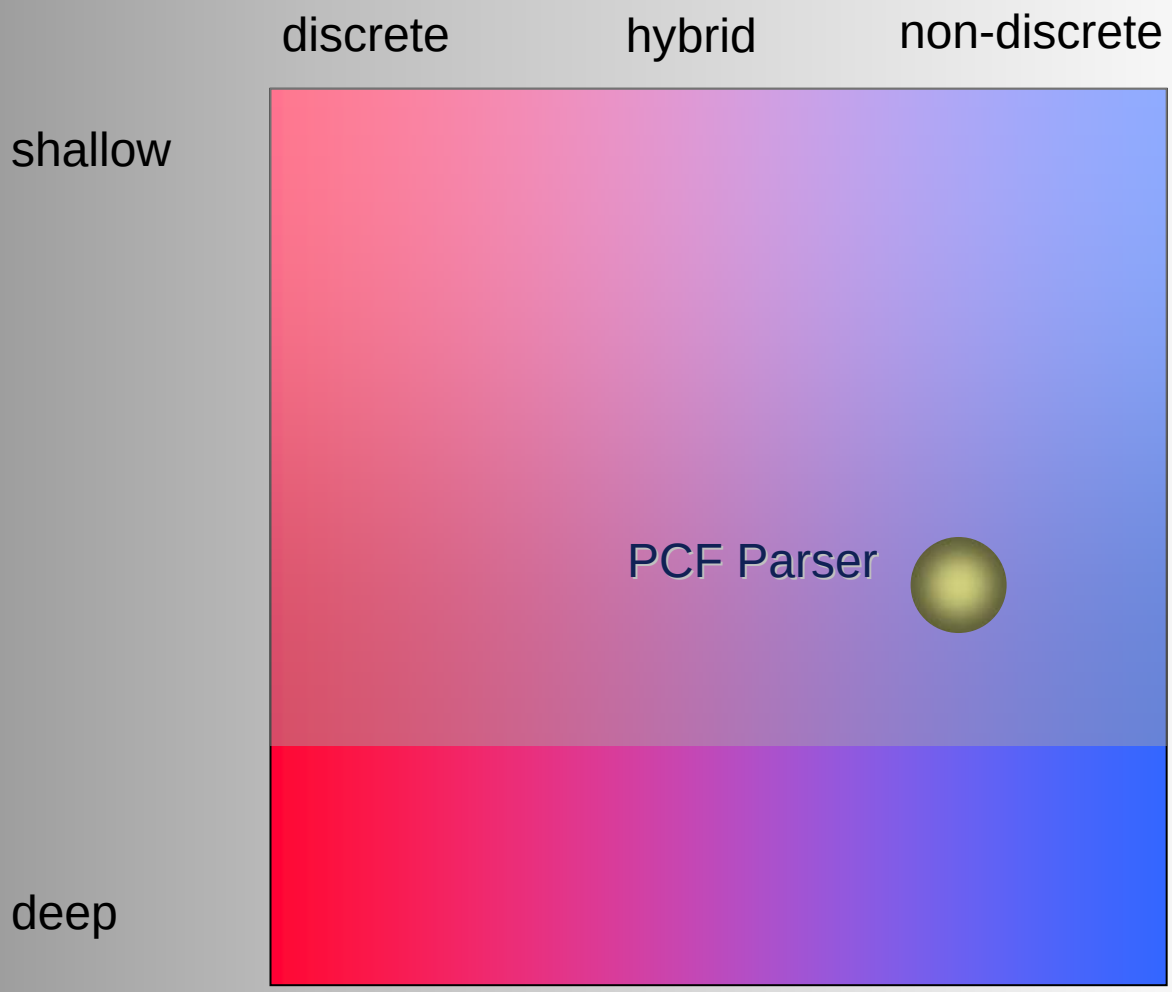
LT METHODS


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

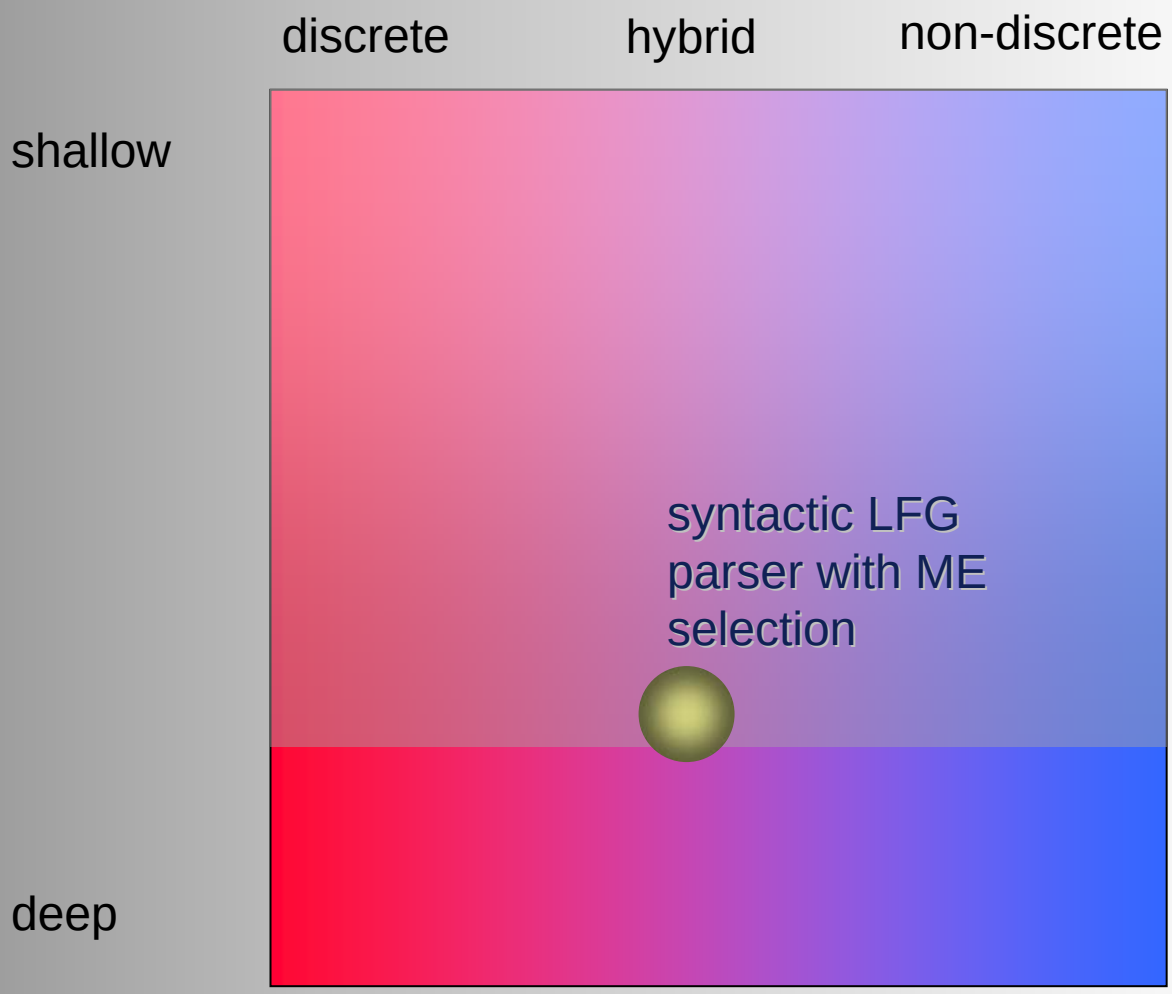
LT METHODS


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

Combination of Methods