

UIMA-based Annotation Type System for a Text Mining Architecture

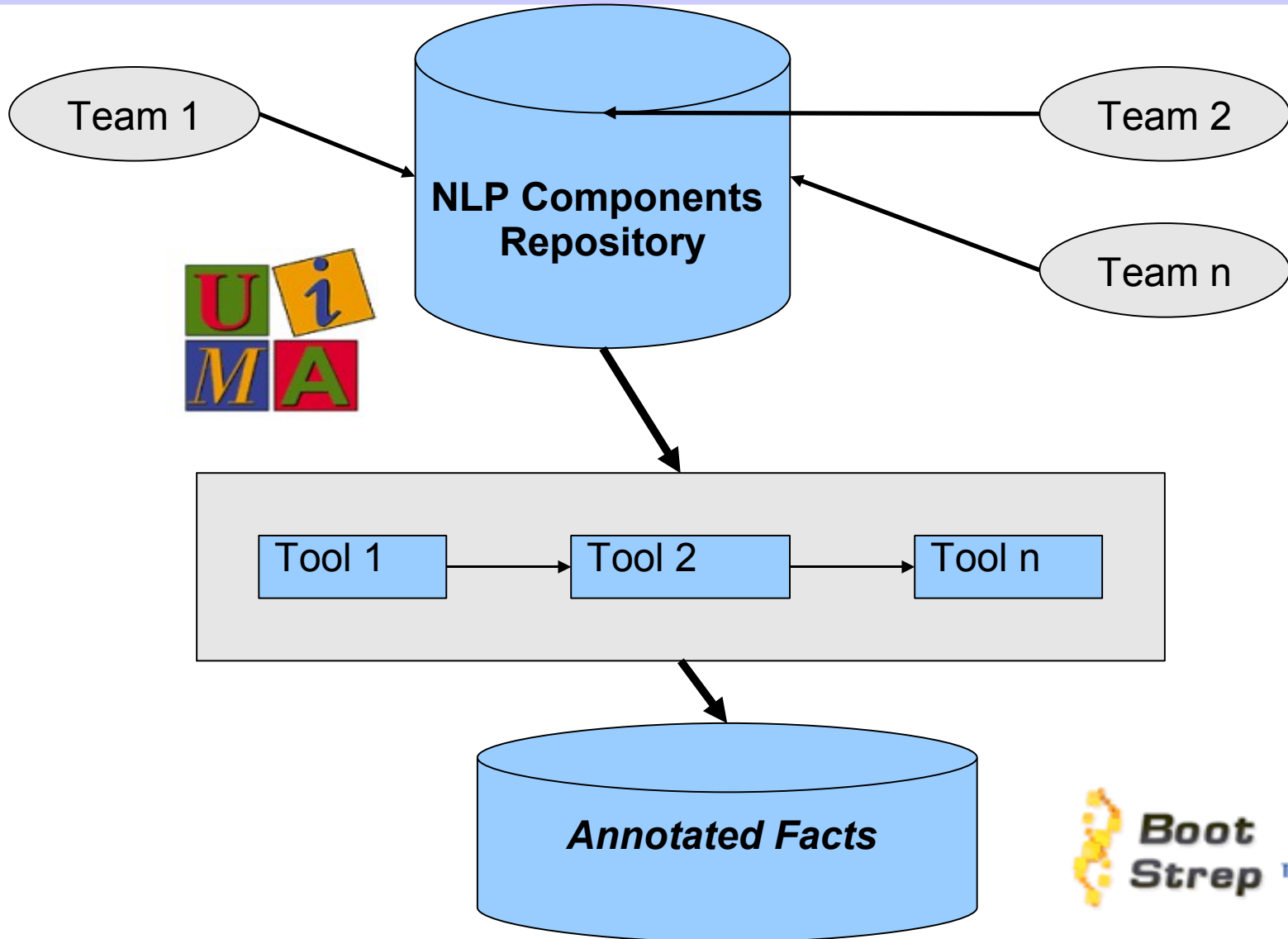
Udo Hahn, **Ekaterina Buyko**, Katrin Tomanek,
Scott Piao, Yoshimasa Tsuruoka, John McNaught, Sophia Ananiadou

*Jena University Language and Information Engineering Lab
& School of Computer Science, University of Manchester*

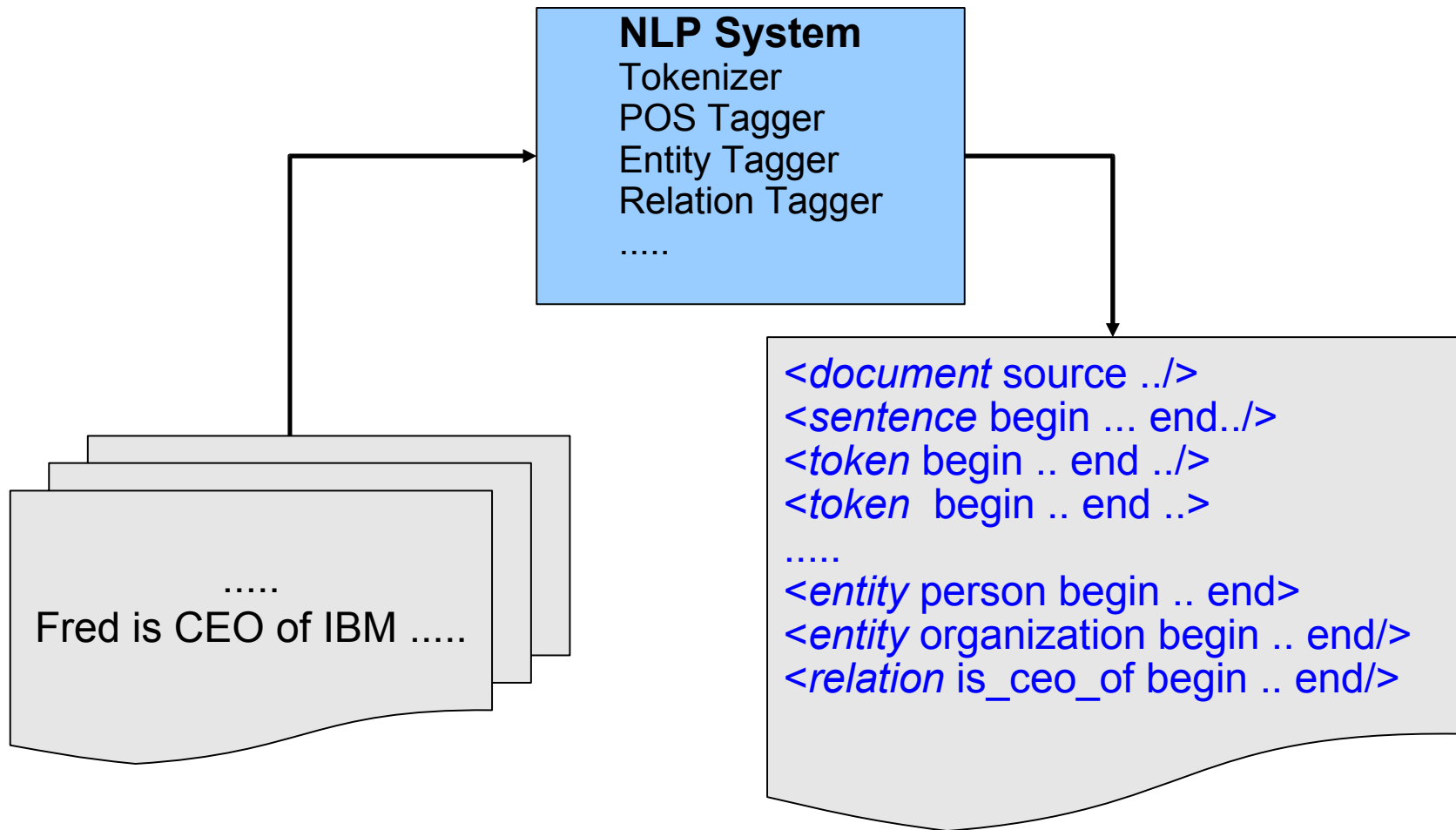


BOOTStrep NLP Infrastructure

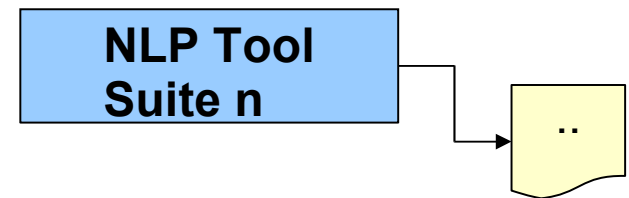
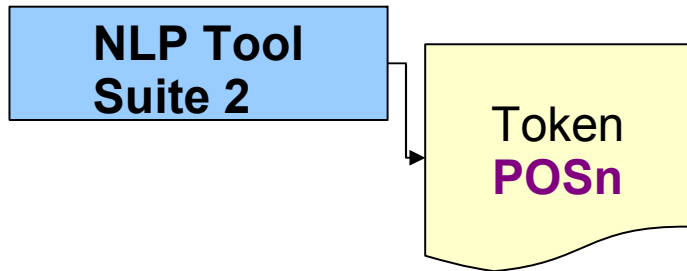
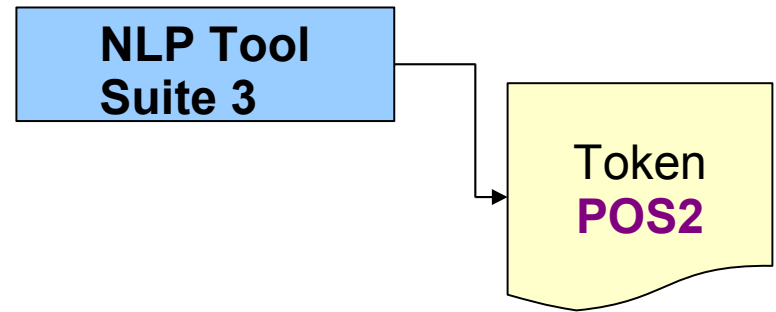
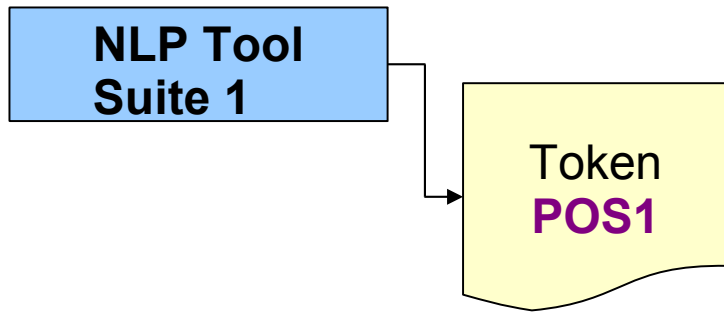
Bootstrapping Of Ontologies and Terminologies Strategic REsearch Project



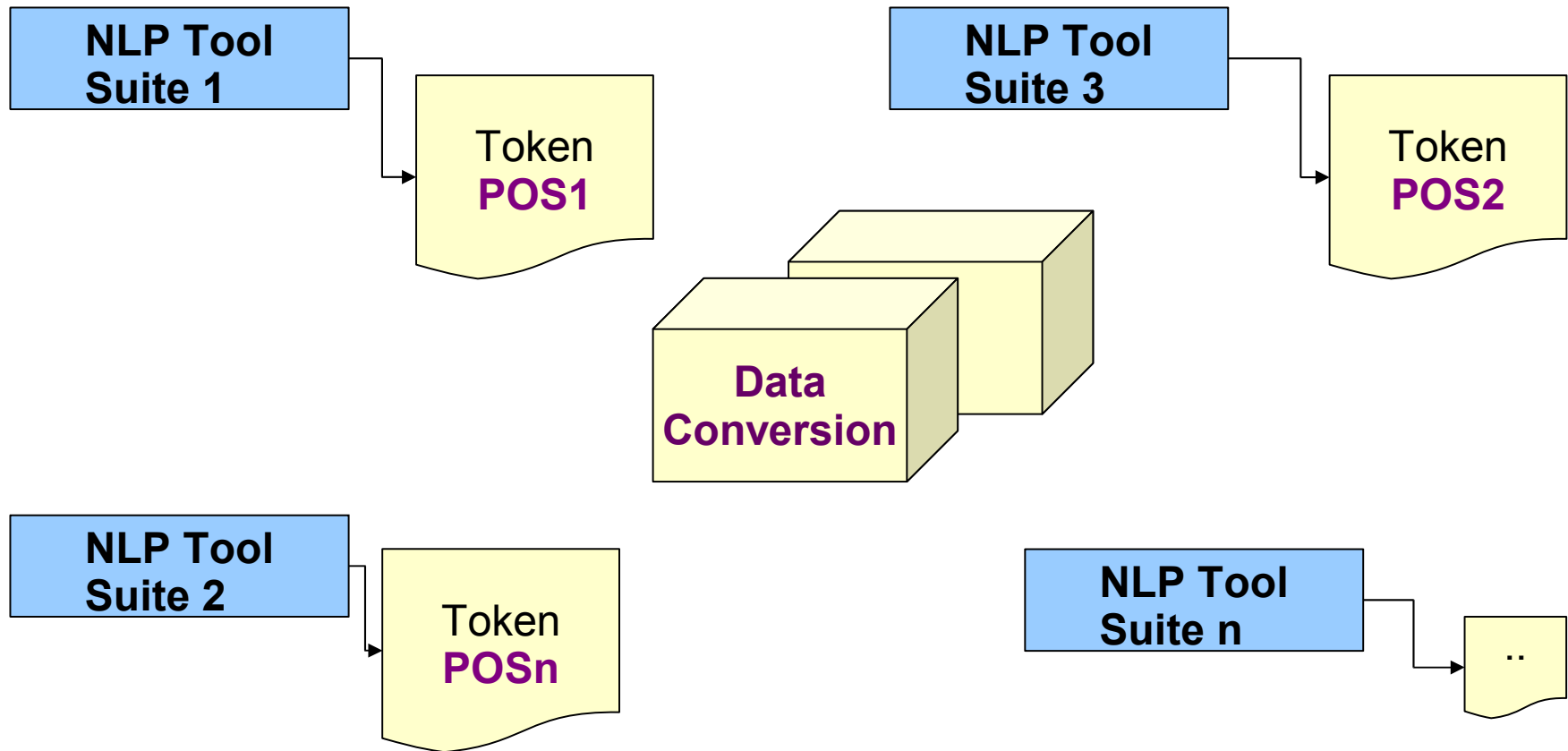
Annotation in Natural Language Processing (NLP)



Annotation in NLP Systems



Annotation in NLP Systems



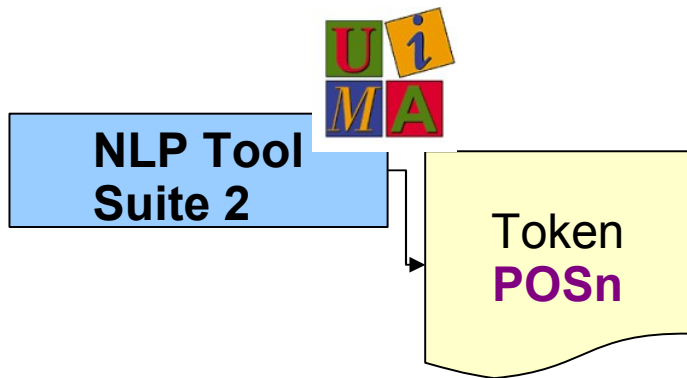
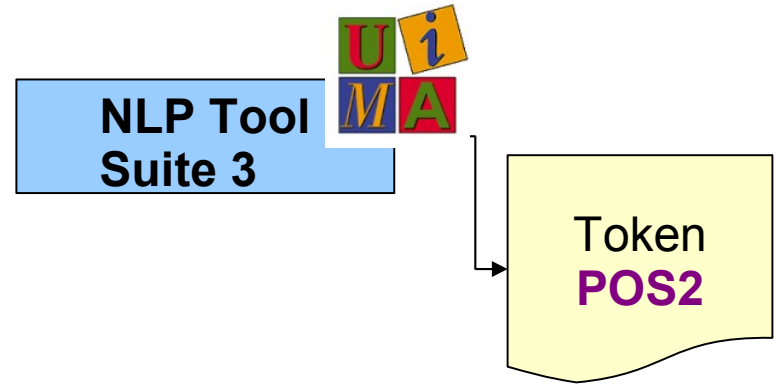
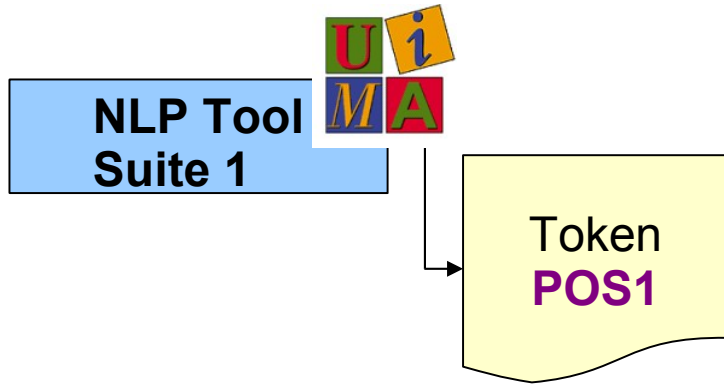
Advantages of the UIMA Framework

Interoperability between NLP systems

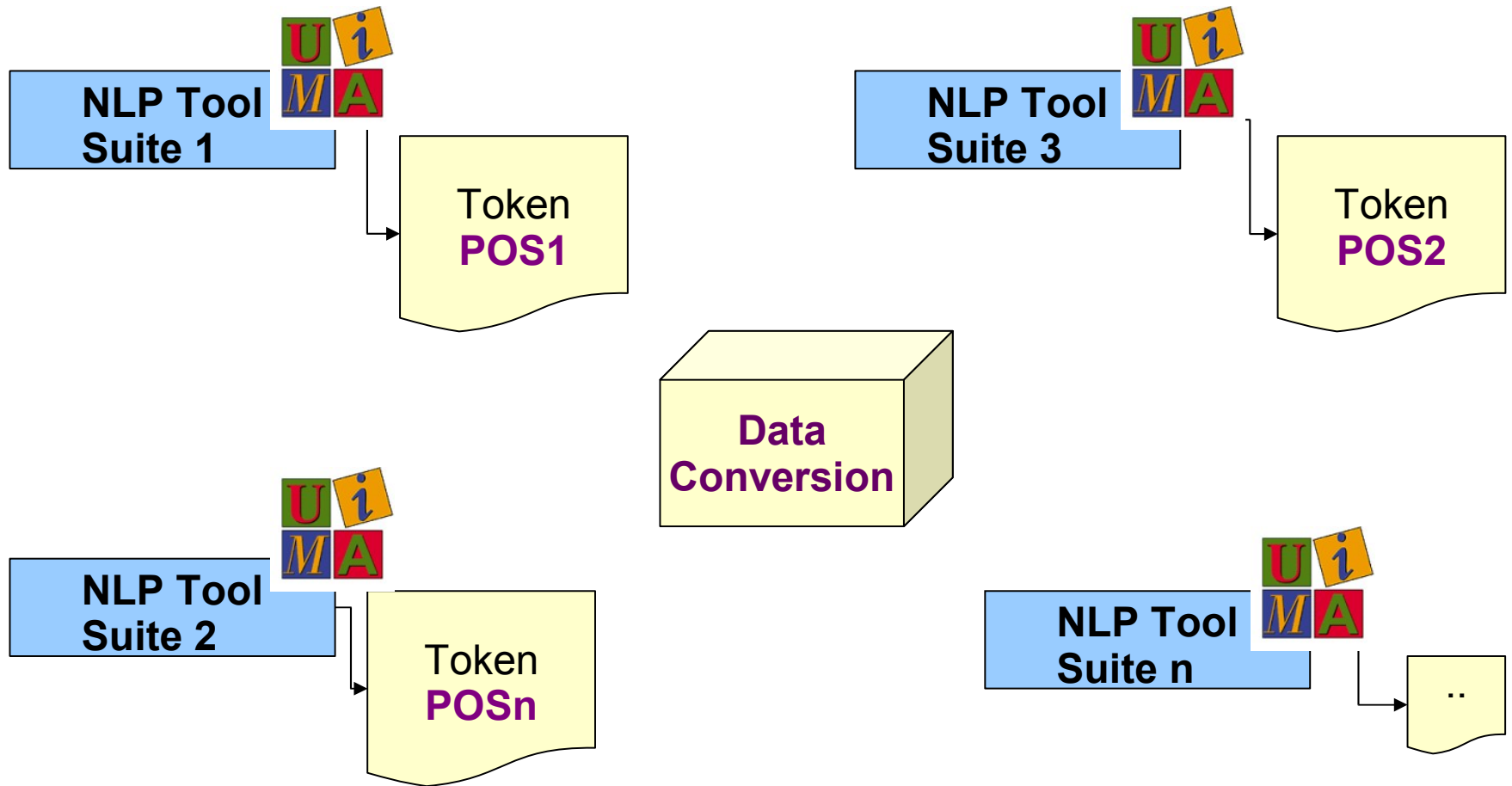
- Portability of components
- Flexible exchange of components



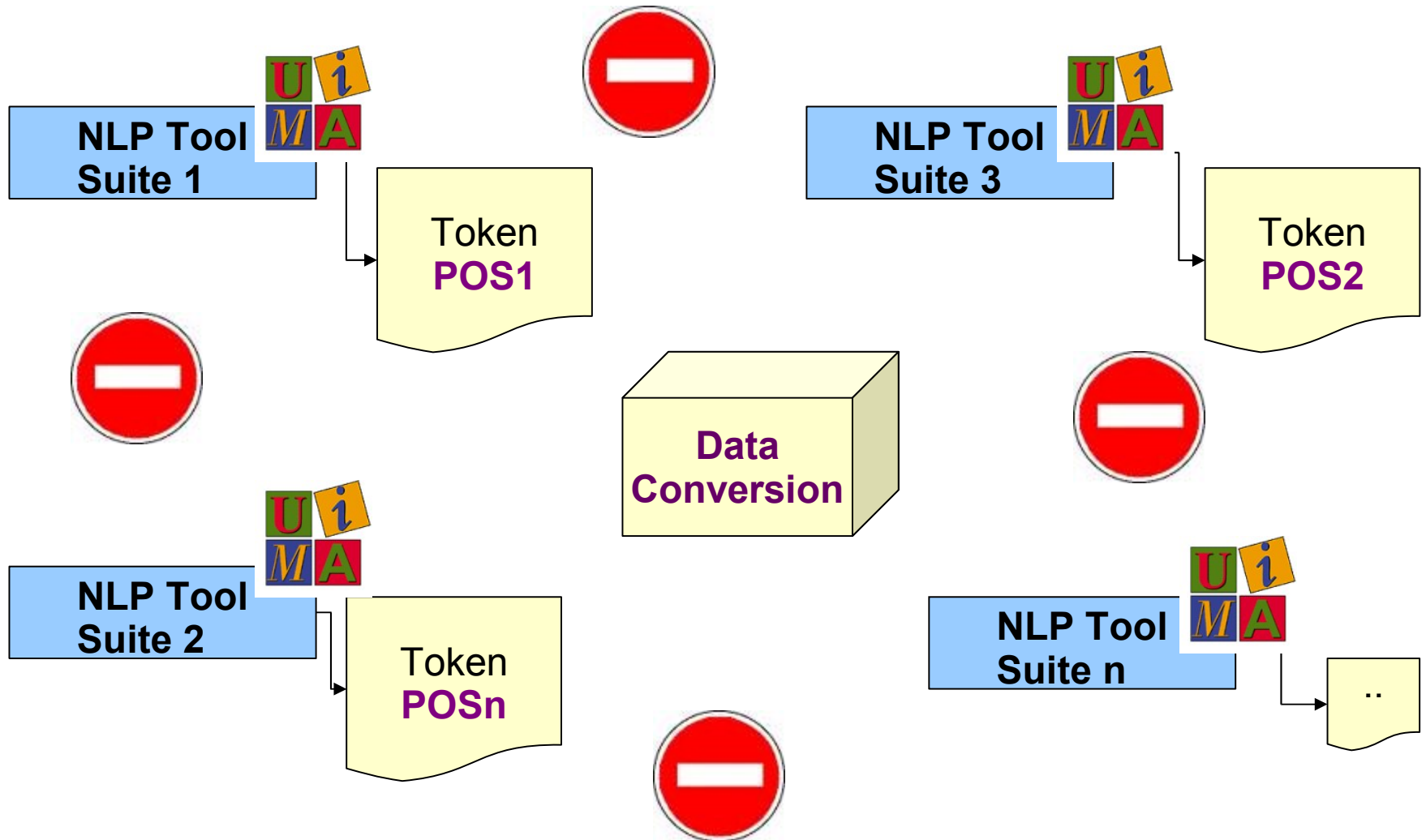
Annotation in NLP Systems



Annotation in NLP Systems



Annotation in NLP Systems



Advantages of the UIMA Framework

Interoperability between NLP systems

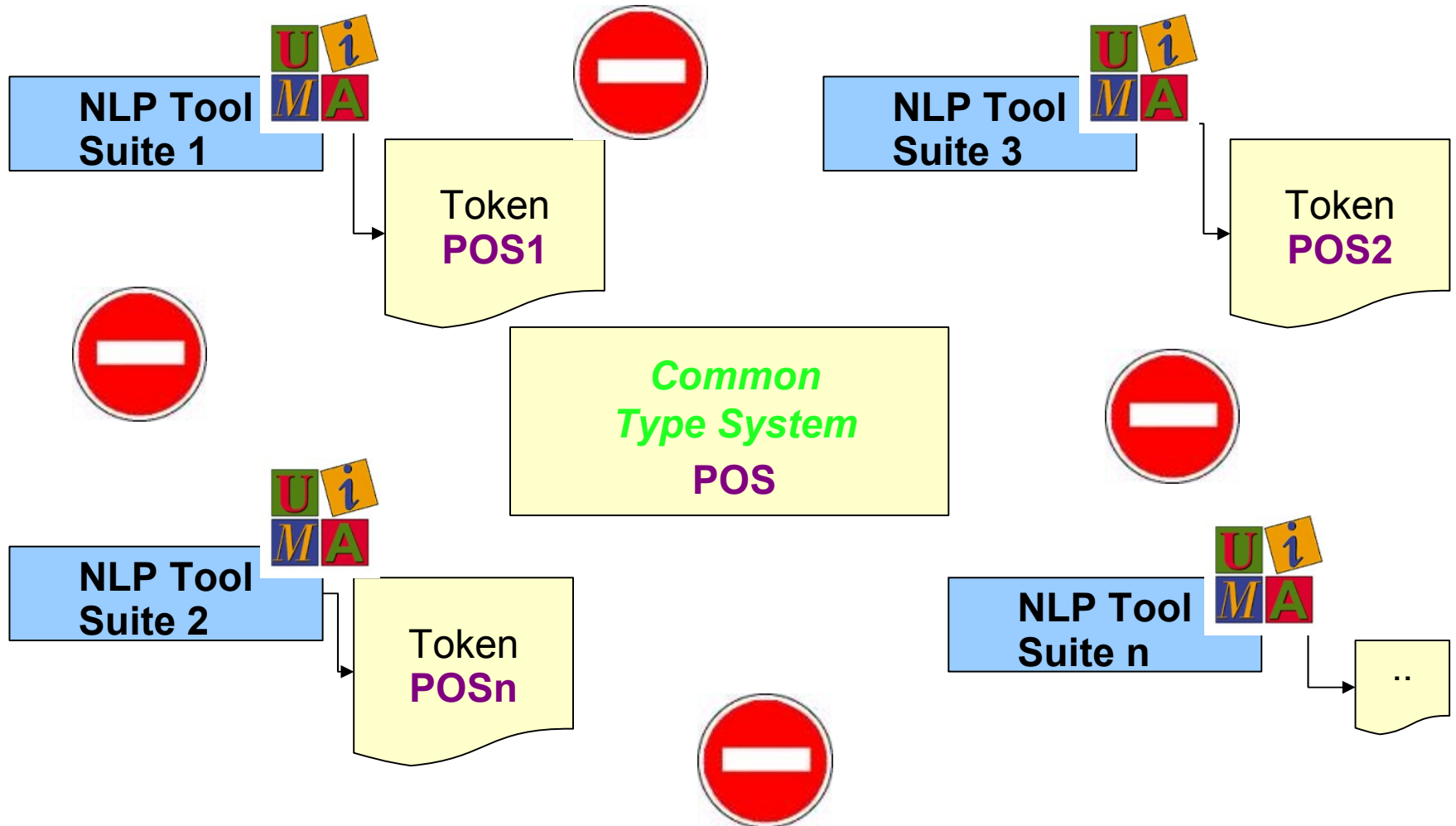
✓ Portability of components

✗ Flexible exchange of components

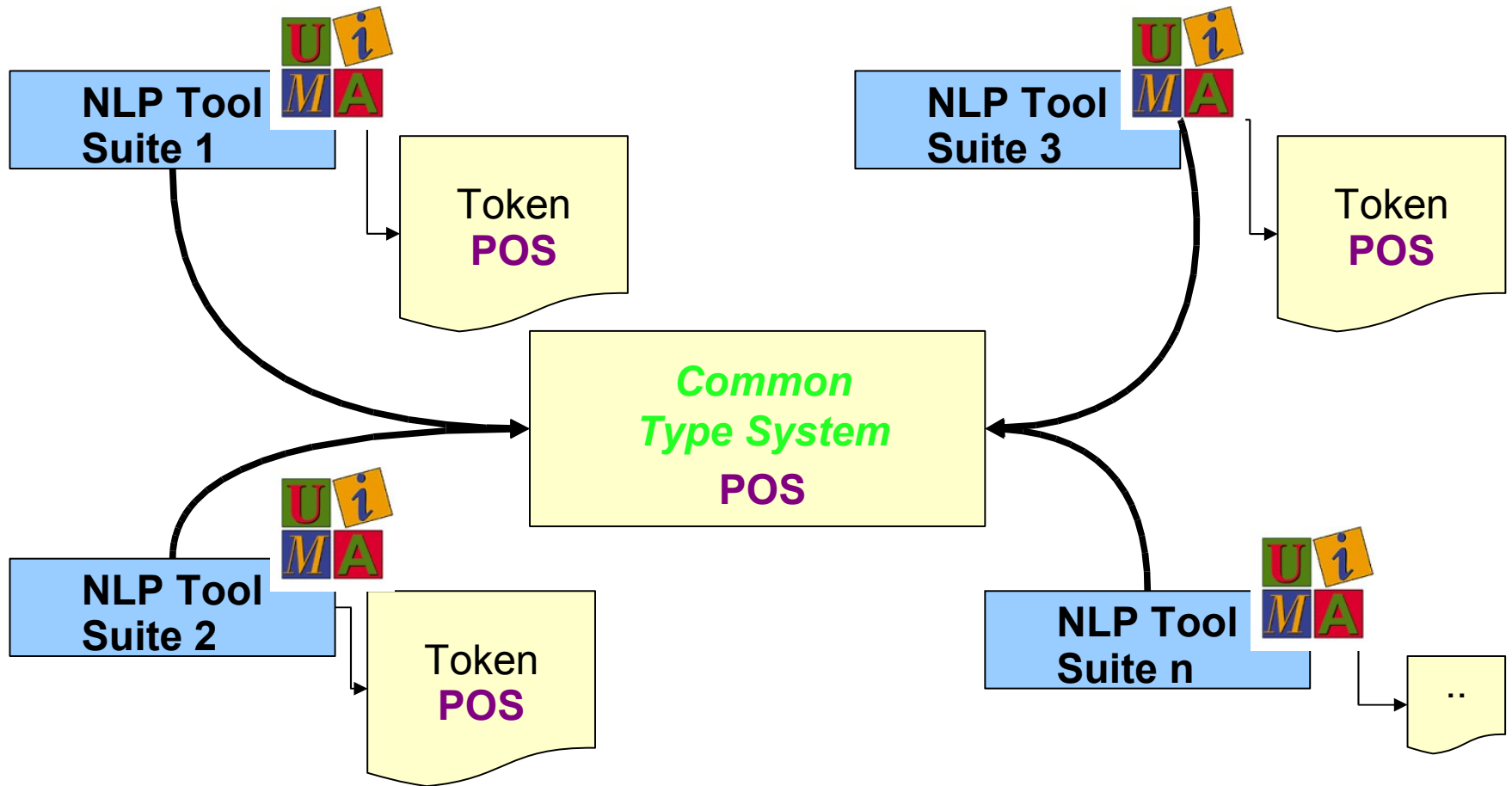
Exchange of components in UIMA

- **Adaptation Efforts**
 - **Over-write Wrappers**
 - **Create Matching Files**
- **Define a Common Annotation Type System** in advance

Annotation in NLP Systems



Annotation in NLP Systems



Advantages of the UIMA Framework

Interoperability between NLP systems

- ✓ Portability of components
- ✓ Flexible exchange of components

Design of an Annotation Type System

- Requirements from various NLP teams
- Annotation guidelines and schemata

Requirements for an Annotation Type System

- Broad **coverage** for the information extraction
- Compatible to “standard” NLP annotation schemata
- Definition of the **core** type system which is **extensible**
- Using **UIMA specific features**
 - **Multiple** annotation of the same type
 - **Annotation control** through the restriction of values

Annotation Guidelines & Schemata

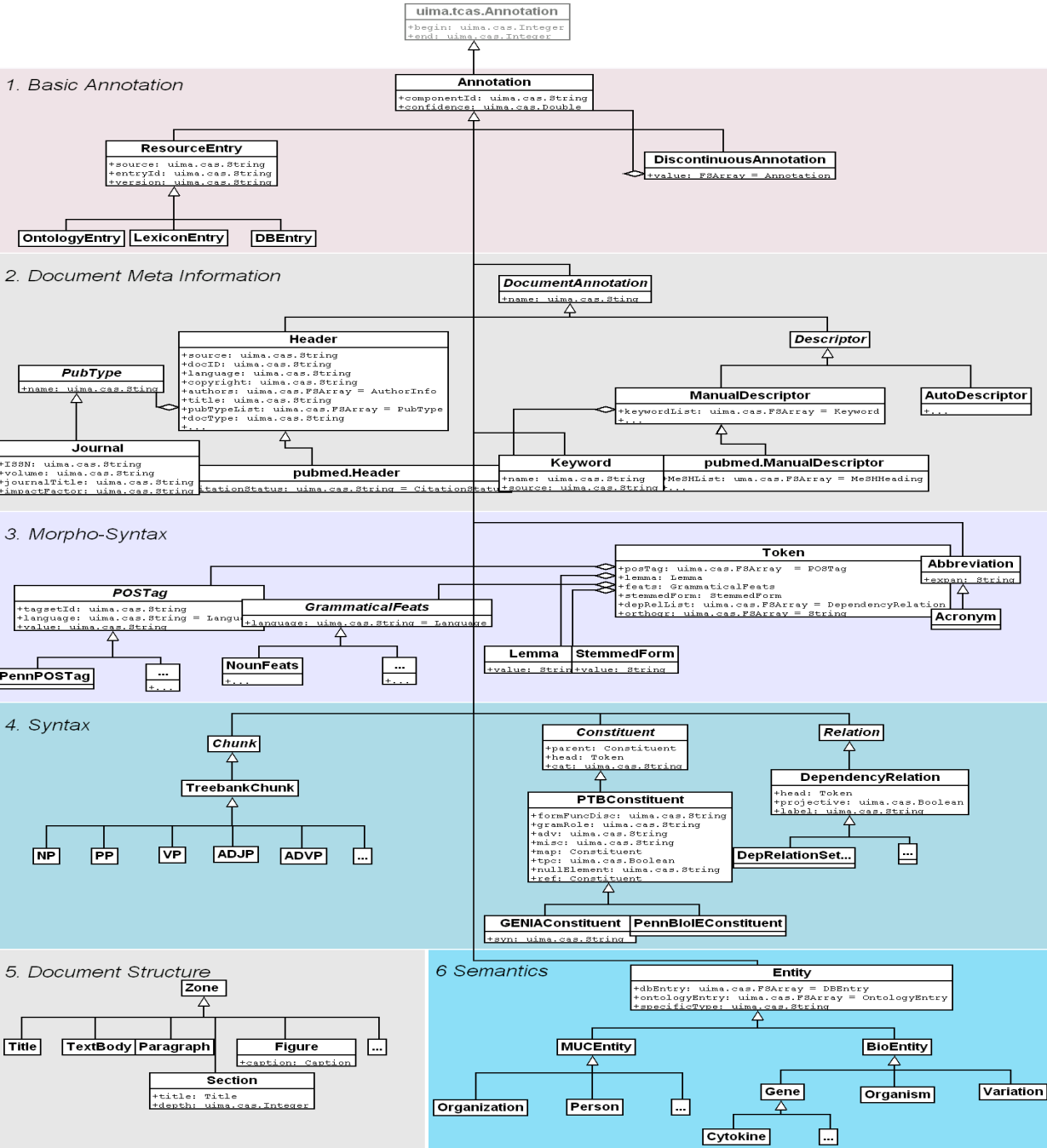
Corpus Annotation

- Annotation languages (e.g. XML (in-line, stand-off))
- Annotation levels:
 - Document Meta (e.g. Dublin Core Metadata Initiative)
 - Linguistic Analysis (e.g. TEI, XCES (EAGLES), Penn Treebank)
 - Semantic Analysis (e.g. MUC, ACE, GENIA)
- NLP system annotation guidelines?

Coverage

Multi-Layered Annotation Type System

- 1. Document Meta:** *author, publication data, source*
- 2. Document Structure & Style :** *title, sections, text bold*
- 3. Morpho-Syntax:** *token, part-of speech, lemma*
- 4. Syntax:** *chunks, constituents, dependency relations*
- 5. Semantics:** *entities, relations, events*
- 6. Discourse:** *anaphora*



1. Basic Annotation

2. Document Meta Information

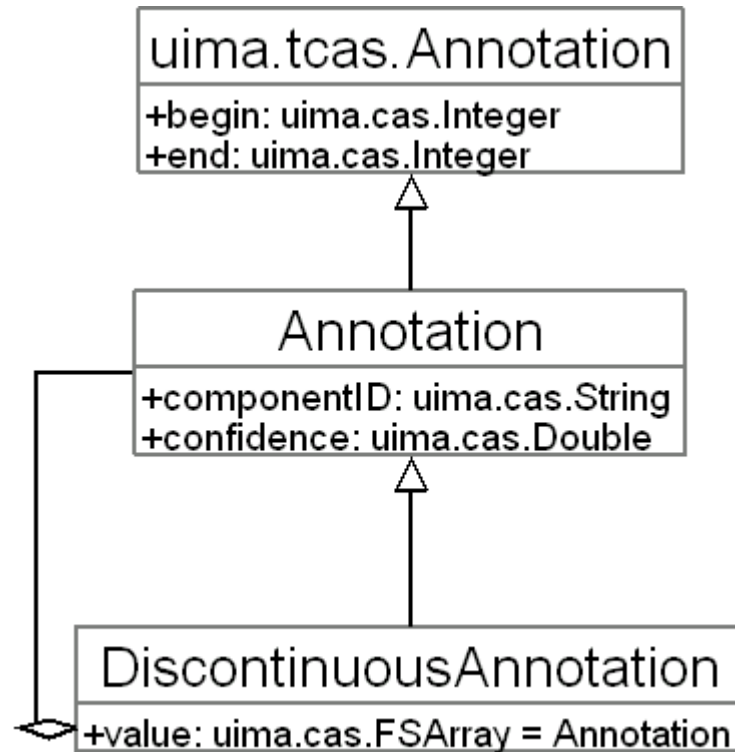
3. Morpho-Syntax

4. Syntax

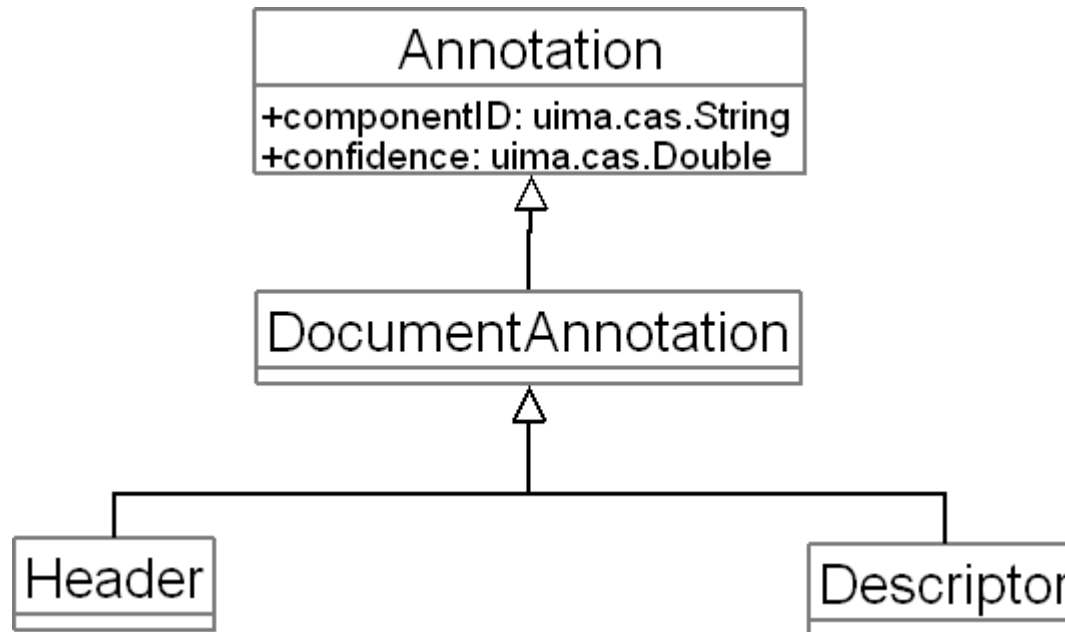
5. Document Structure

6 Semantics

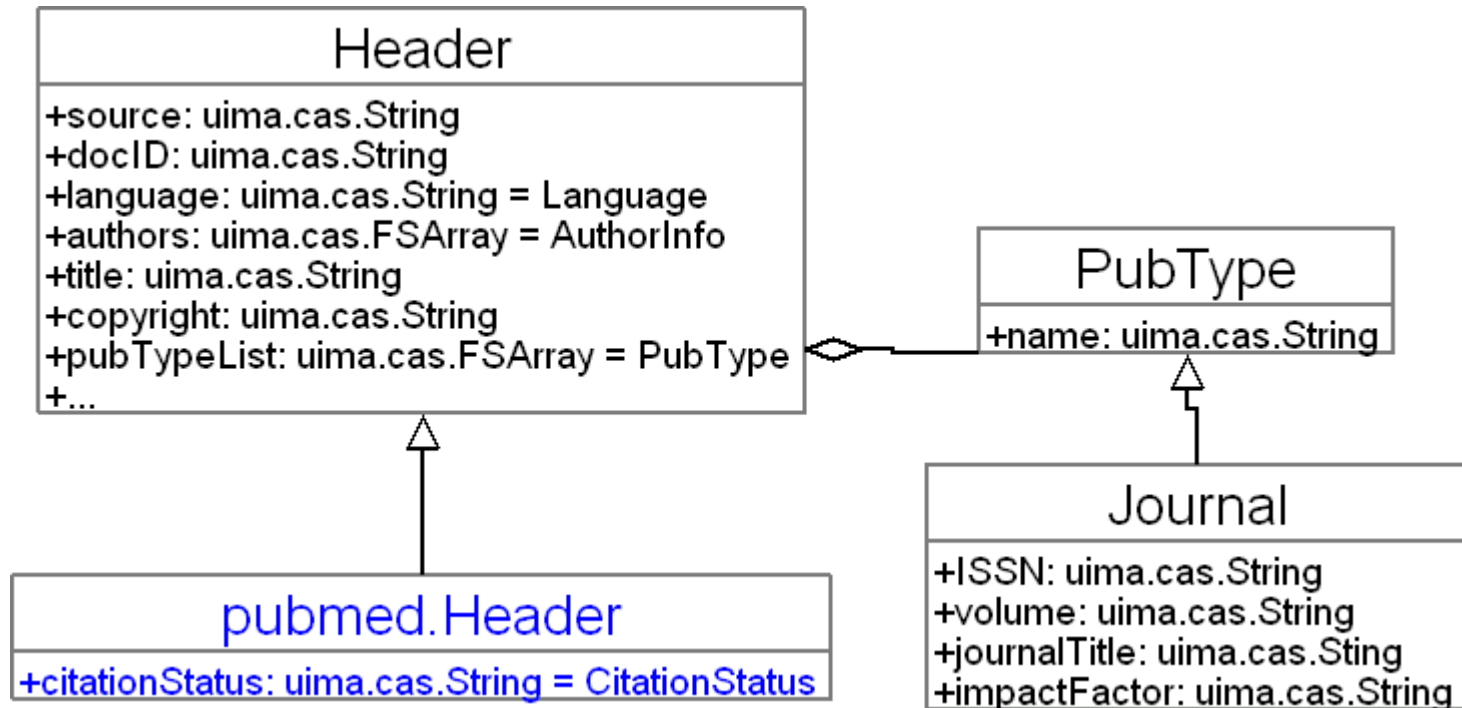
Basic Annotation Type



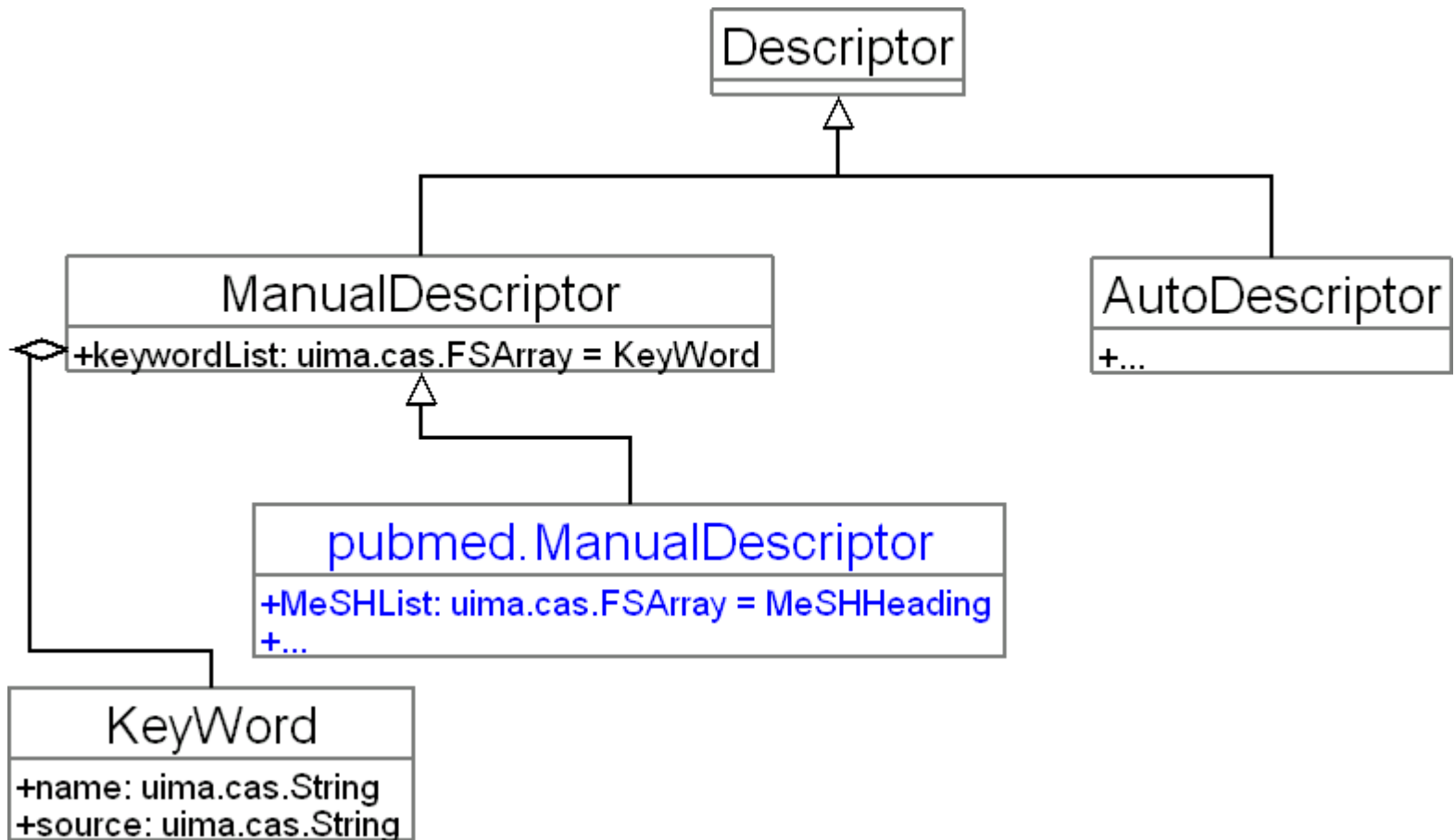
Document Meta



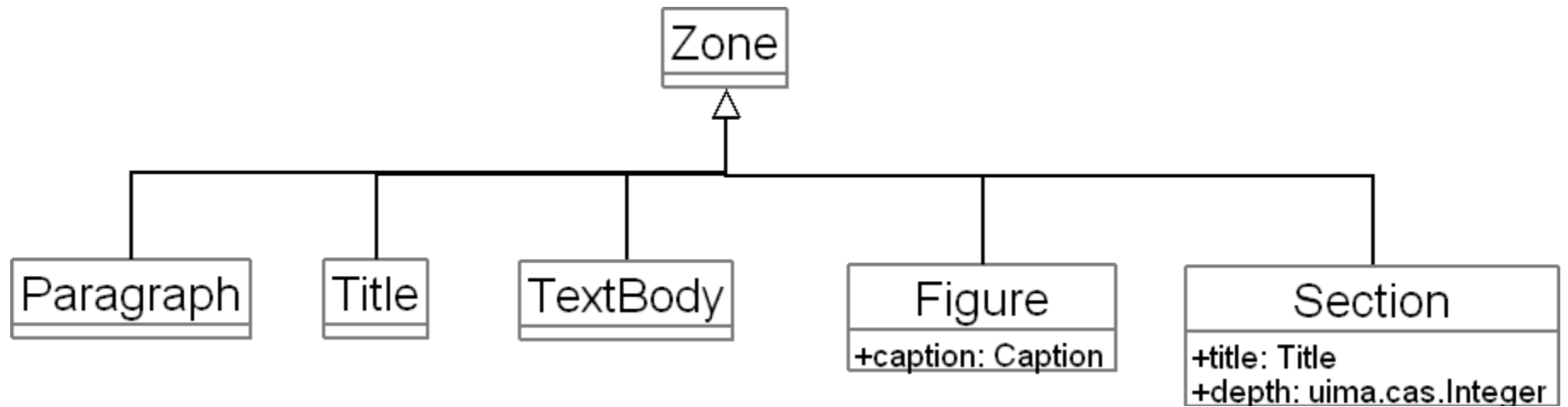
Document Meta Information I



Document Meta Information II



Document Structure



Morpho-Syntax

POSTag
+..

Token
+..

Lemma
+..

Abbreviation
+..

GrammaticalFeats
+..

StemmedForm
+..

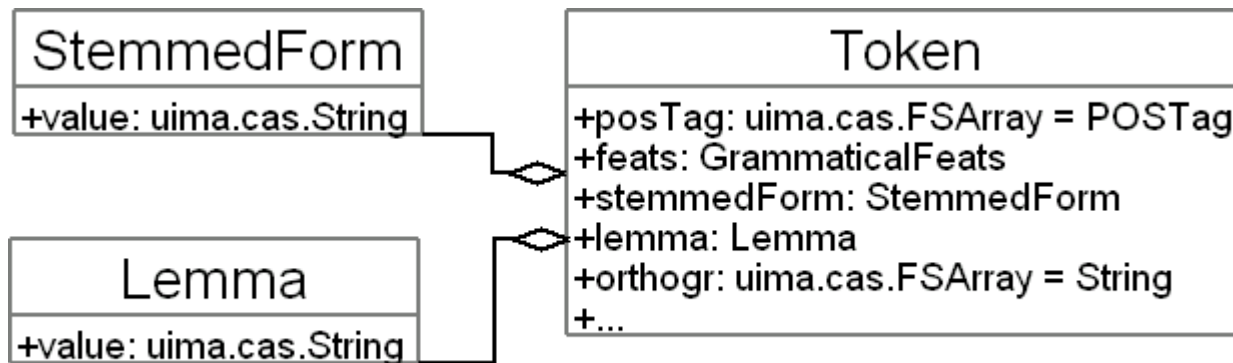
Acronym
+..

Morpho-Syntax I

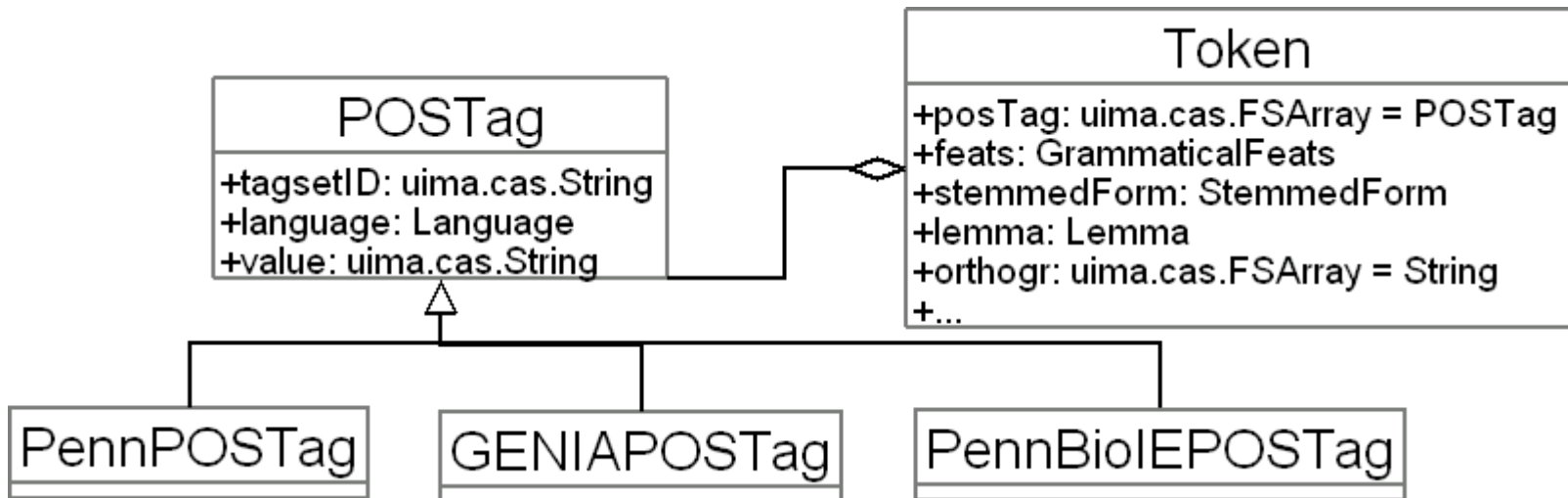
Token

+posTag: uima.cas.FSArray = POSTag
+feats: GrammaticalFeats
+stemmedForm: StemmedForm
+lemma: Lemma
+orthogr: uima.cas.FSArray = String
+...

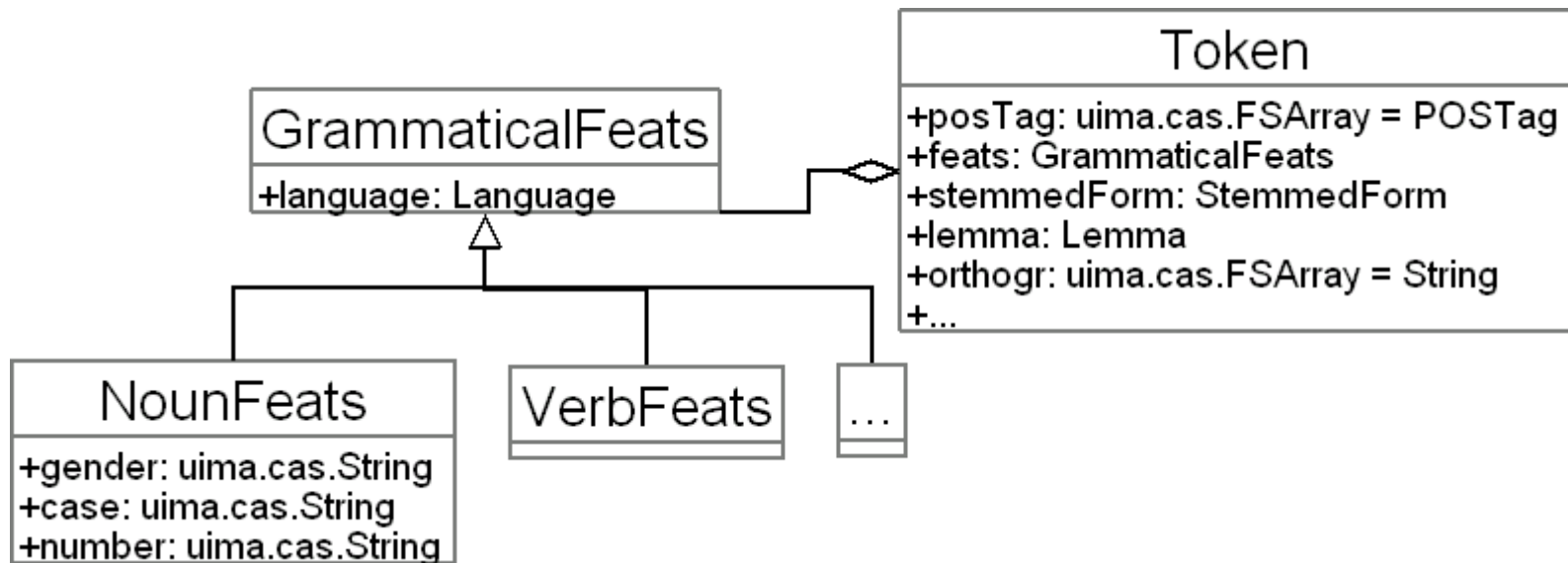
Morpho-Syntax II



Morpho-Syntax III



Morpho-Syntax IV



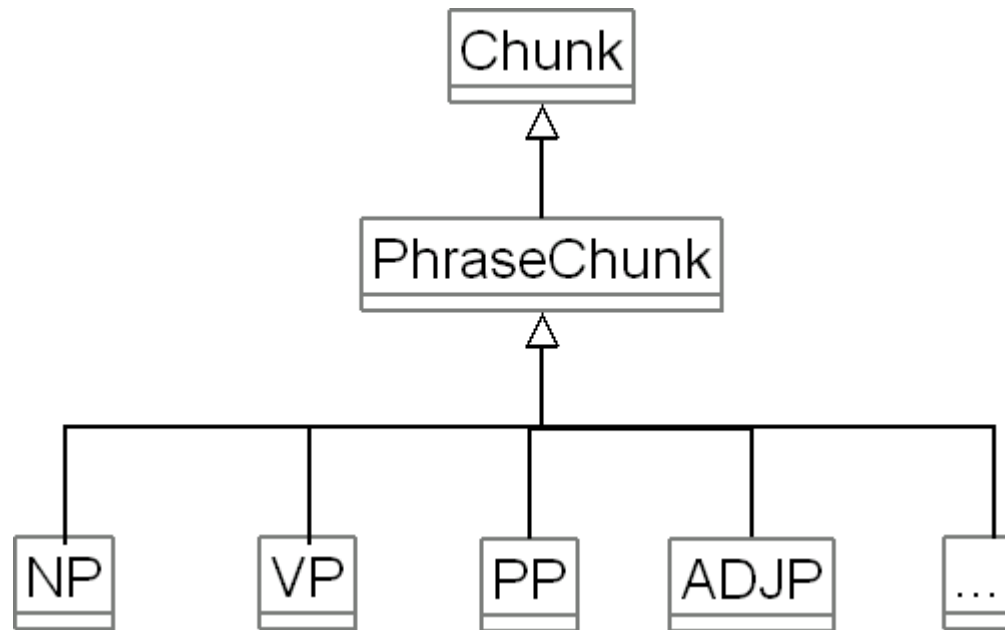
Syntax

Chunk

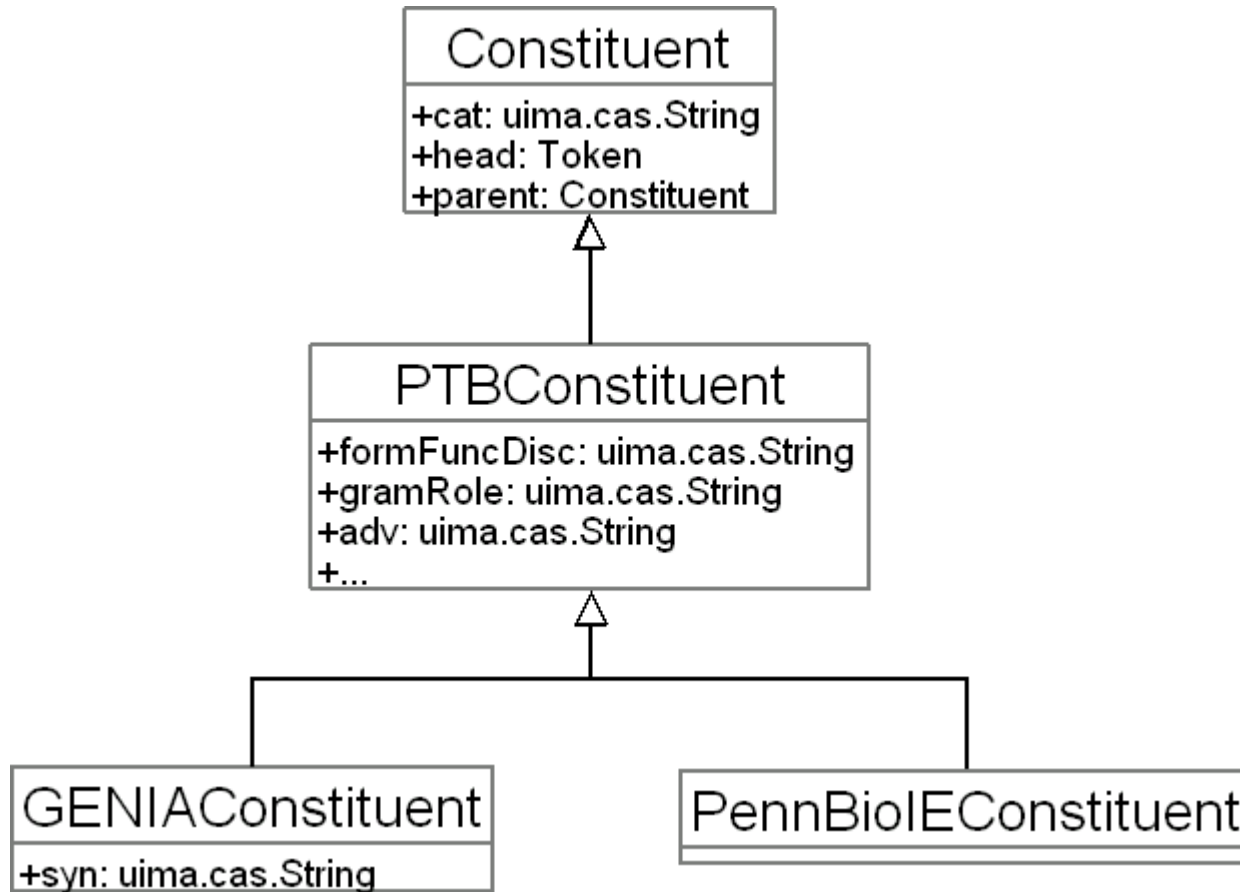
Constituent

DependencyRelation

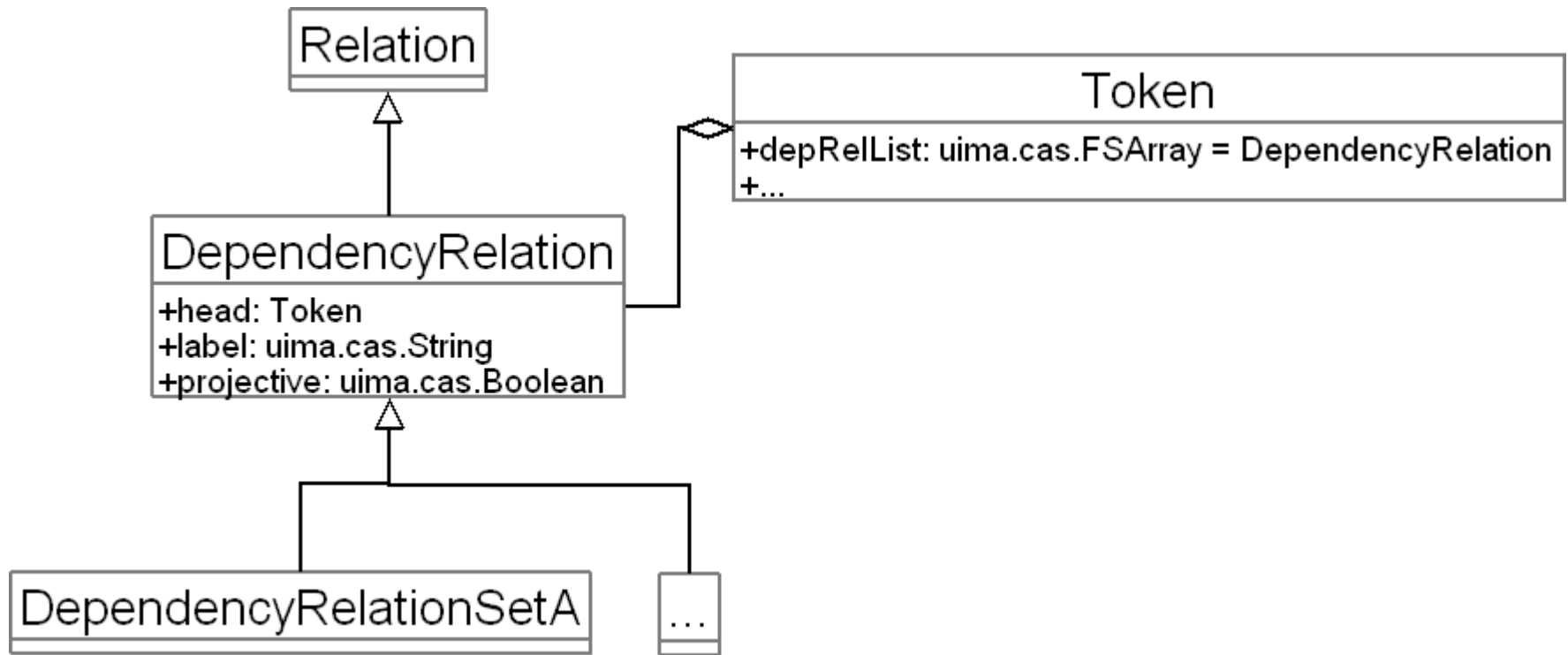
Shallow Parsing



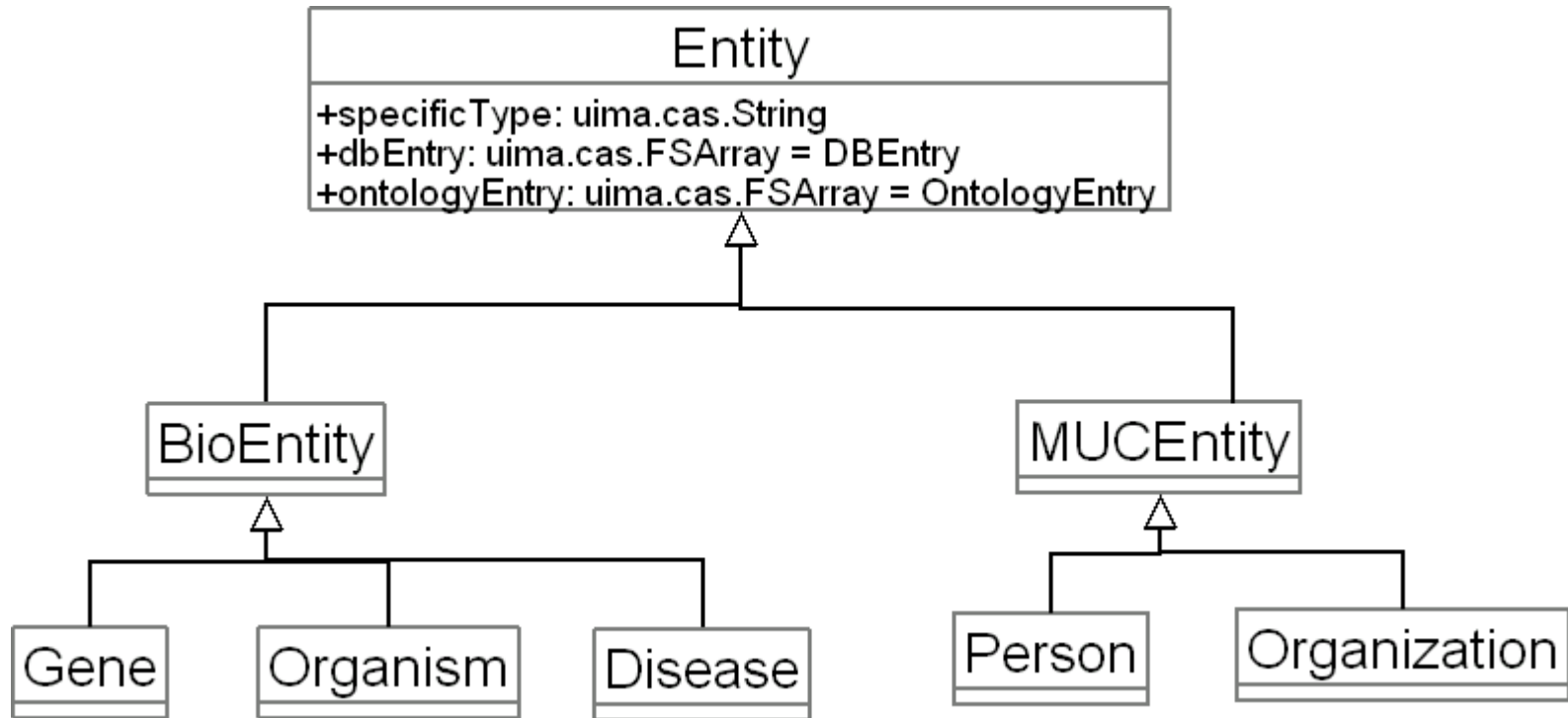
Full Parsing (constituent-based)



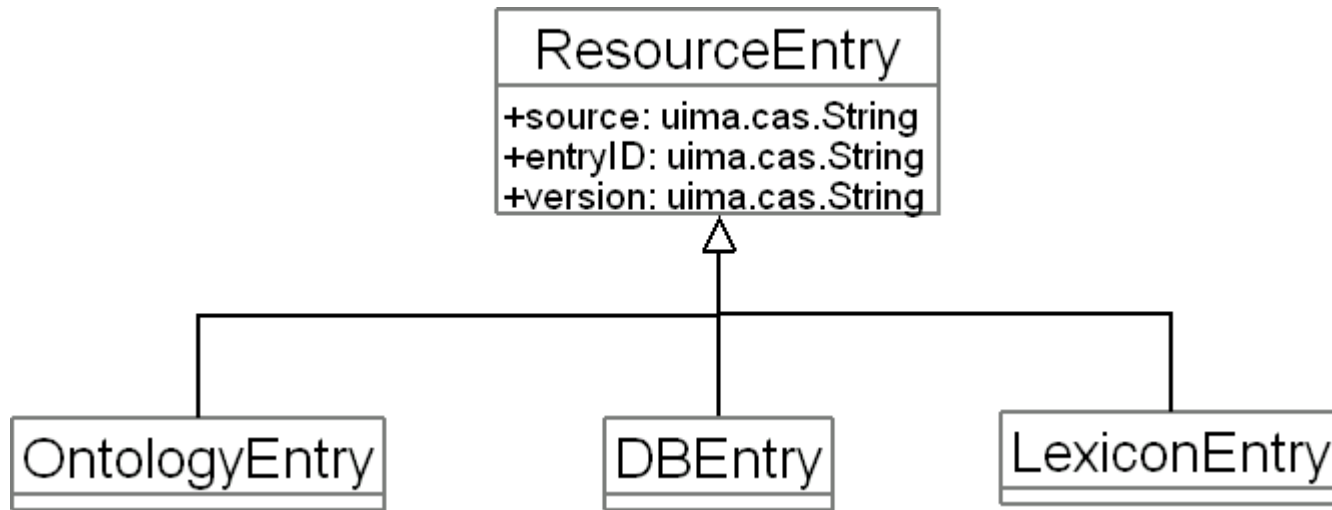
Full Parsing (dependency-based)



Semantics



Resource Connection



To wrap up ..

- **Multi-layered annotation**
- **Core** annotation type system
- Extended for the **biomedical** domain
- **Can easily be extended** for other domains
- **Restriction of values** for the annotation **control**
- Sub-Types for **multiple** annotation (e.g. POS, Chunk)
- **Connection** to external resources

Open Issues

- **Performance** measure of the type system
- **Definitions:**
 - Semantics (Relation, Event)
 - Discourse (Anaphora)

UIMA Annotation Type System Working Group?

Download: <http://www.julielab.de/>

Contact: buyko@coling-uni-jena.de

Sponsored by

