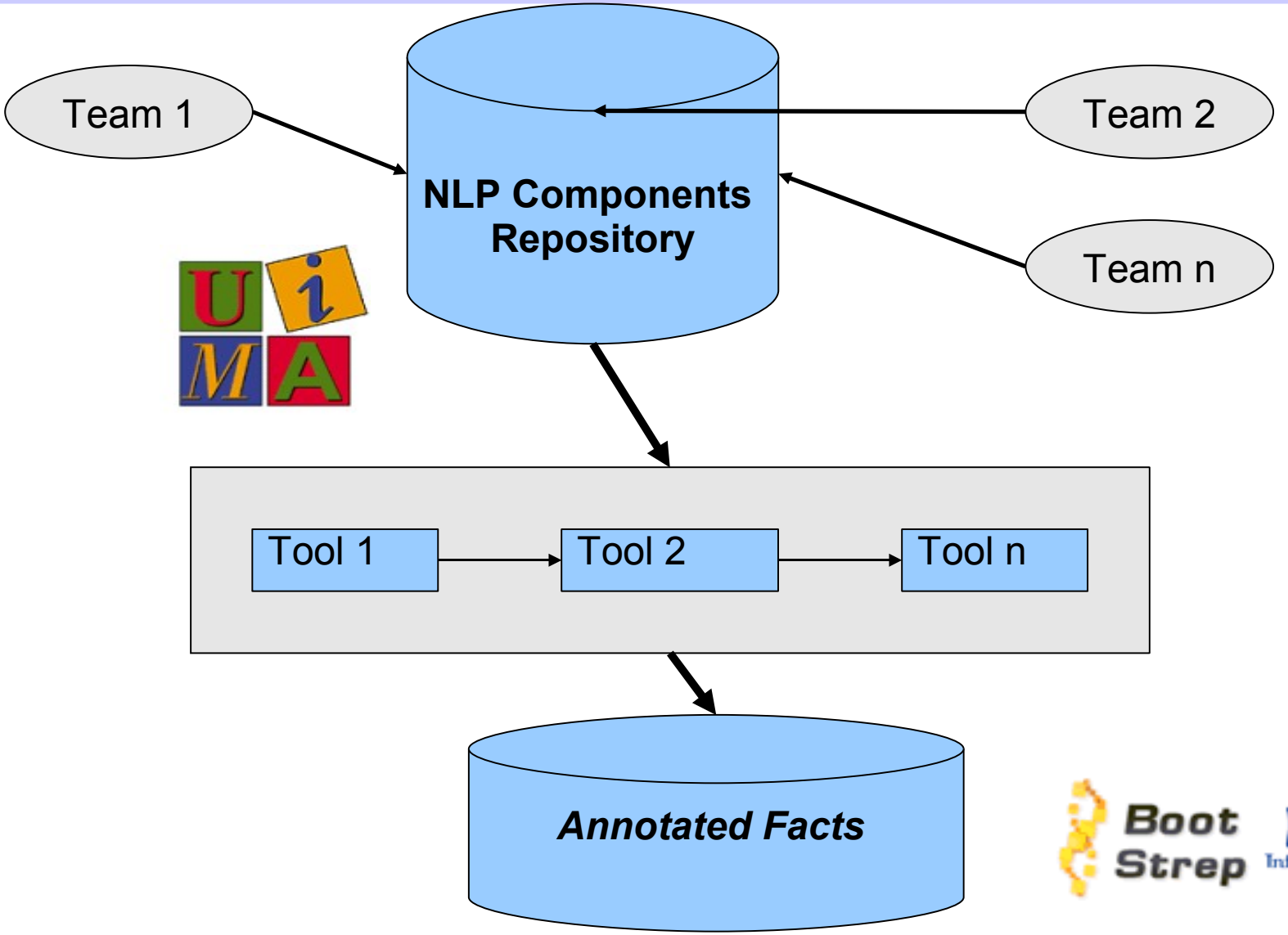
UIMA-based Annotation Type System for a Text Mining Architecture

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

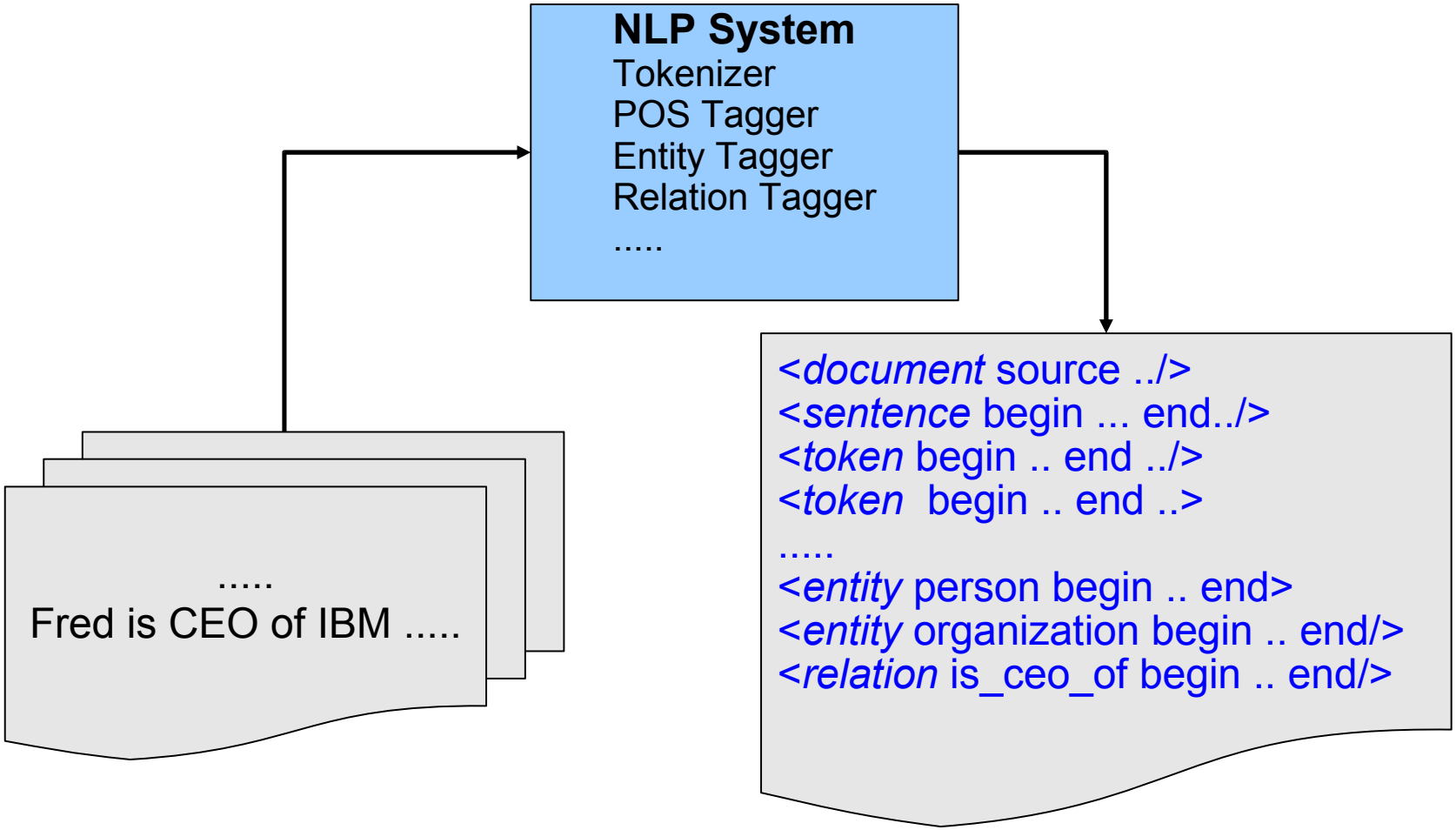
Jena University Language and Information Engineering Lab
& School of Computer Science, University of Manchester

BOOTStrep NLP Infrastructure

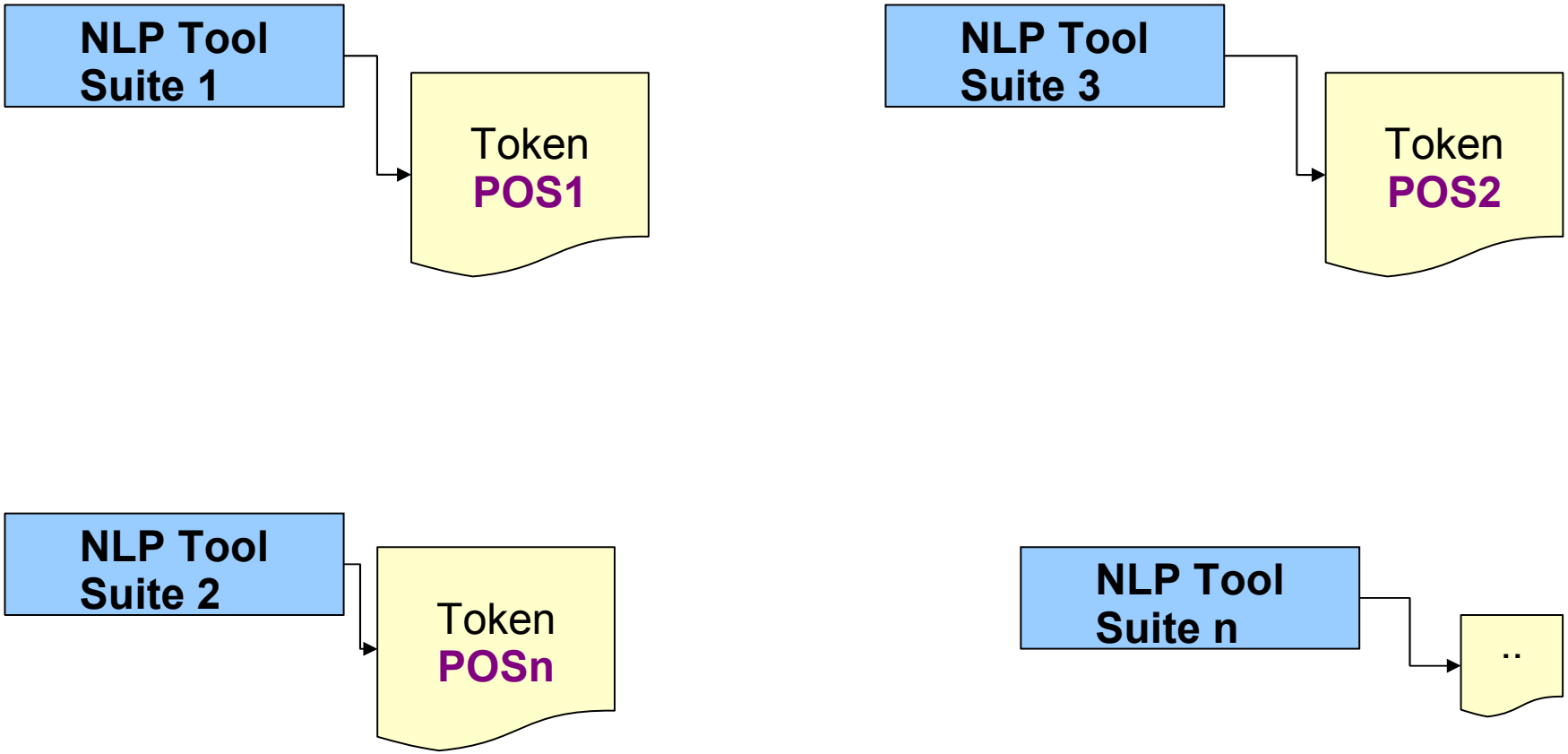
Bootstrapping Of Ontologies and Terminologies STrategic REsearch Project



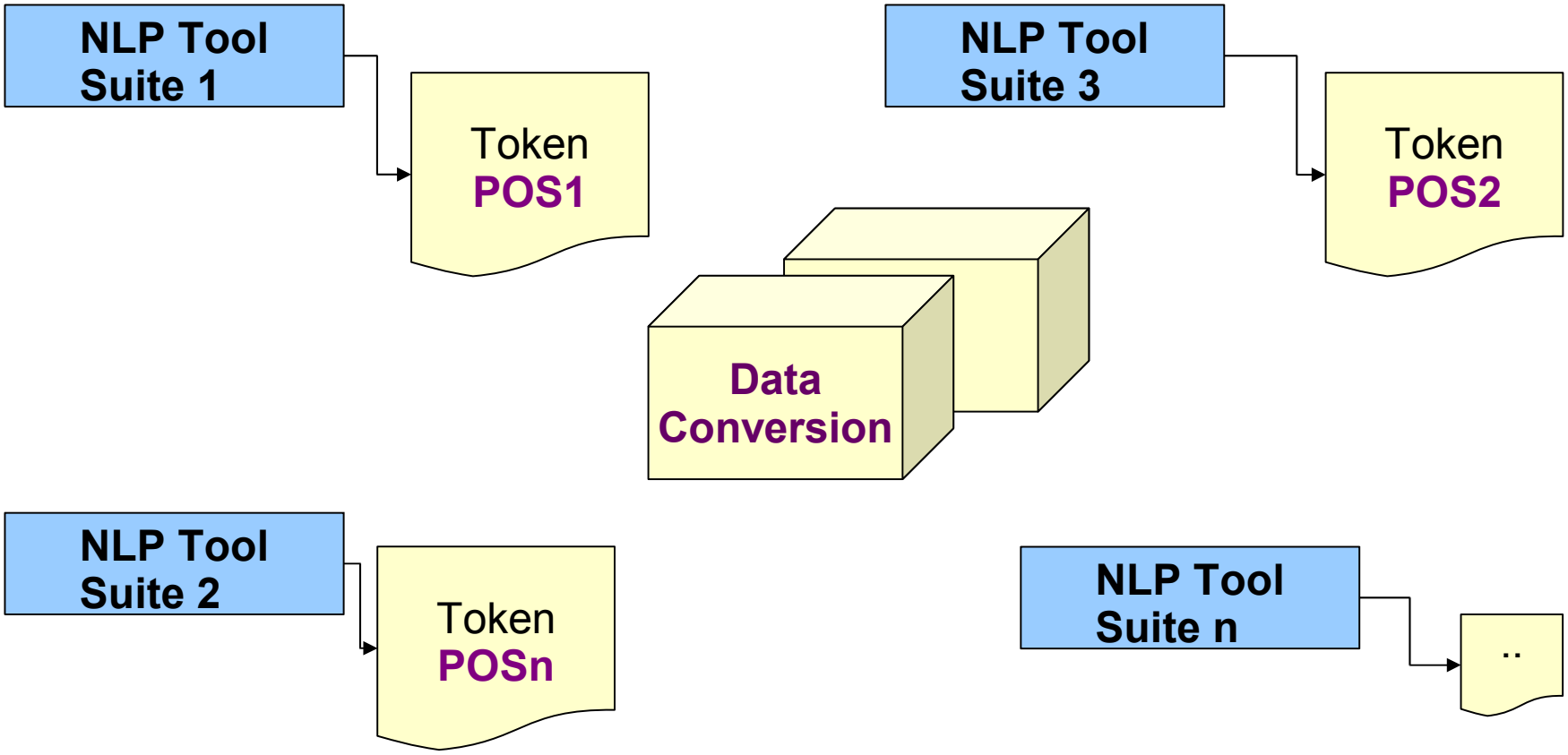
Annotation in Natural Language Processing (NLP)



Annotation in NLP Systems



Annotation in NLP Systems

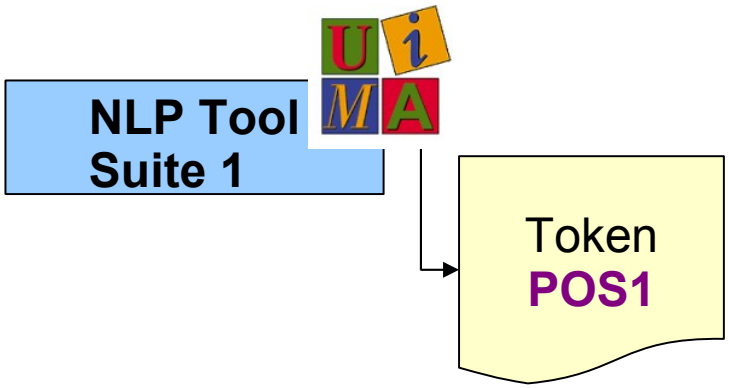


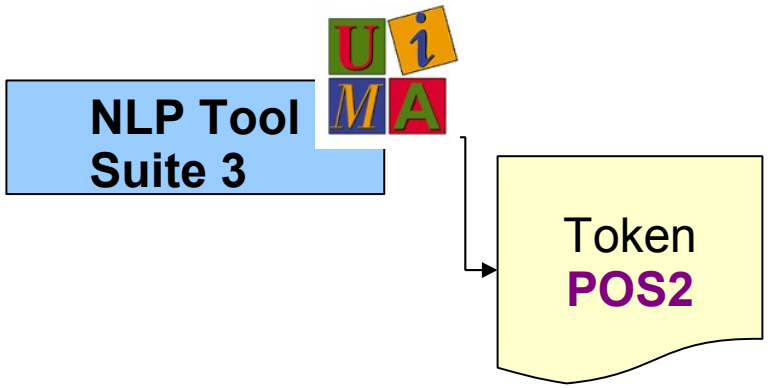
Advantages of the UIMA Framework

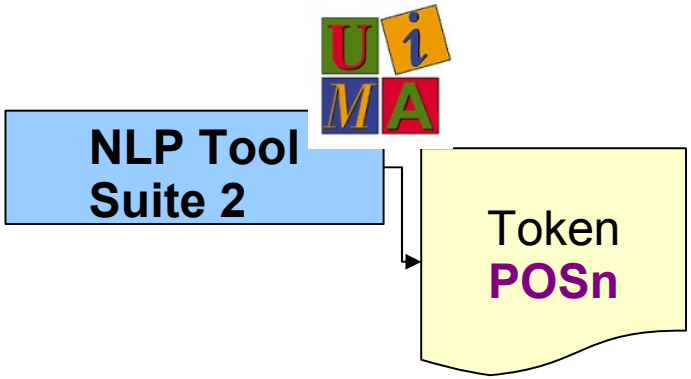
Interoperability between NLP systems

- Portability of components
- Flexible exchange of components

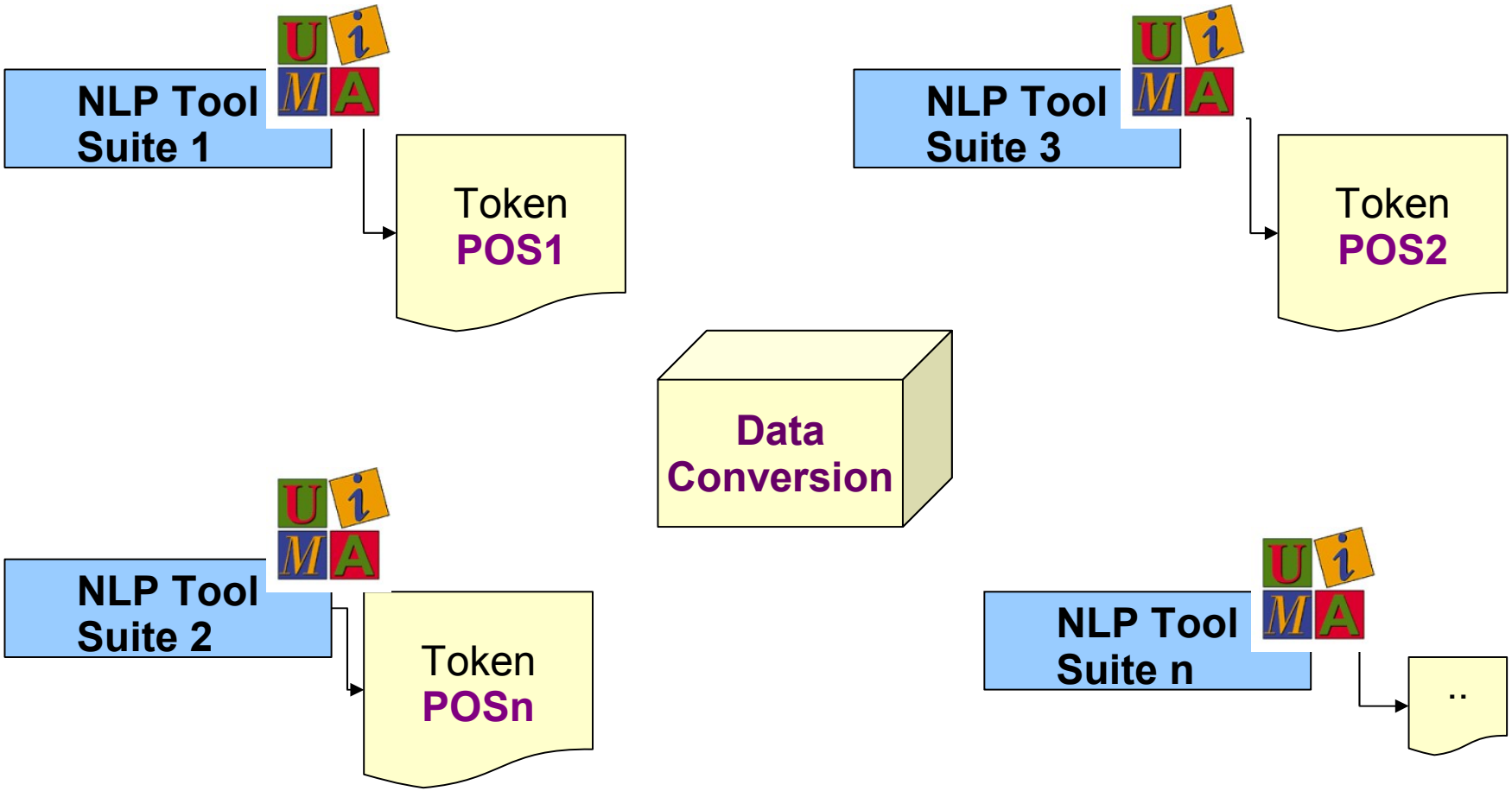
Annotation in NLP Systems



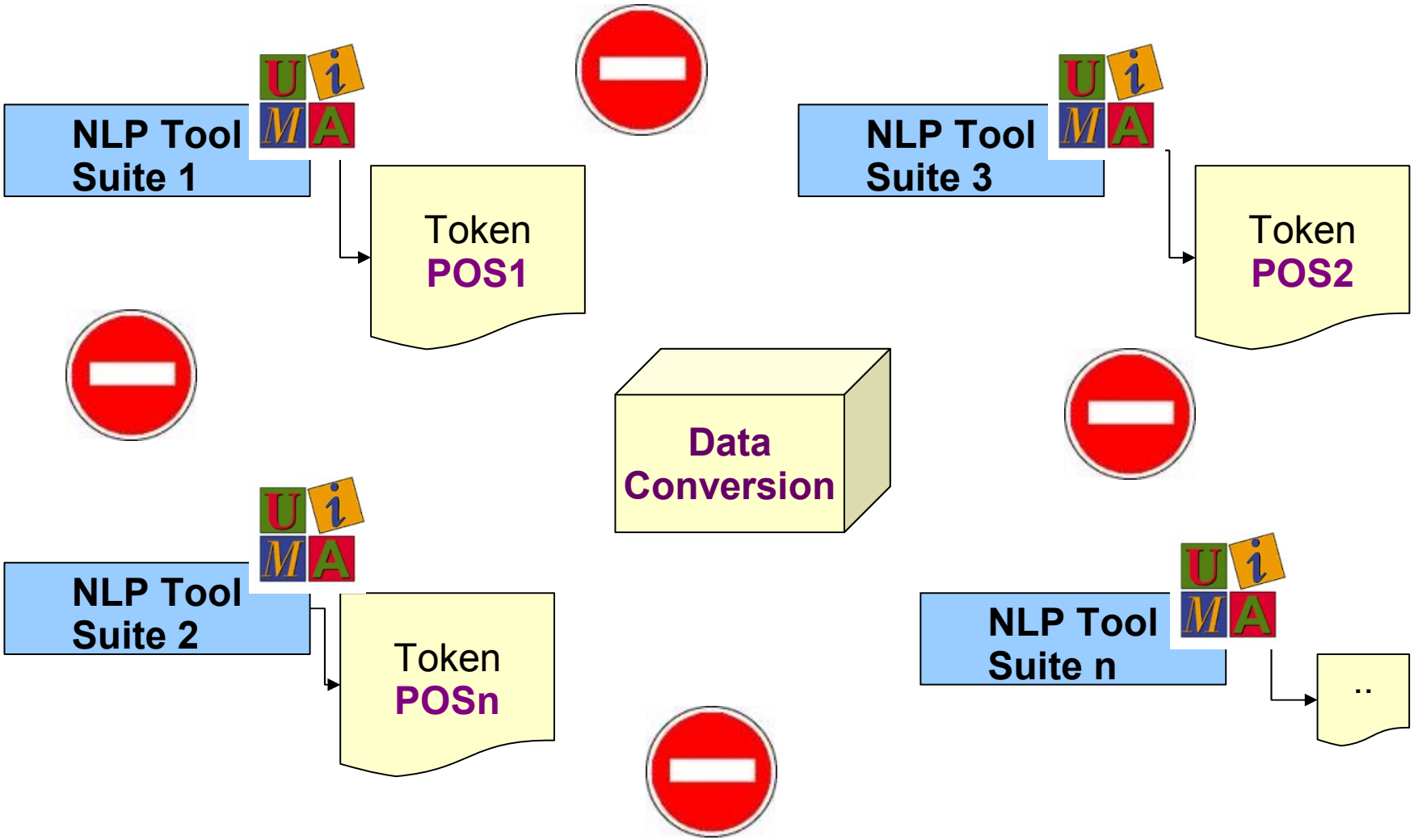




Annotation in NLP Systems



Annotation in NLP Systems



Advantages of the UIMA Framework

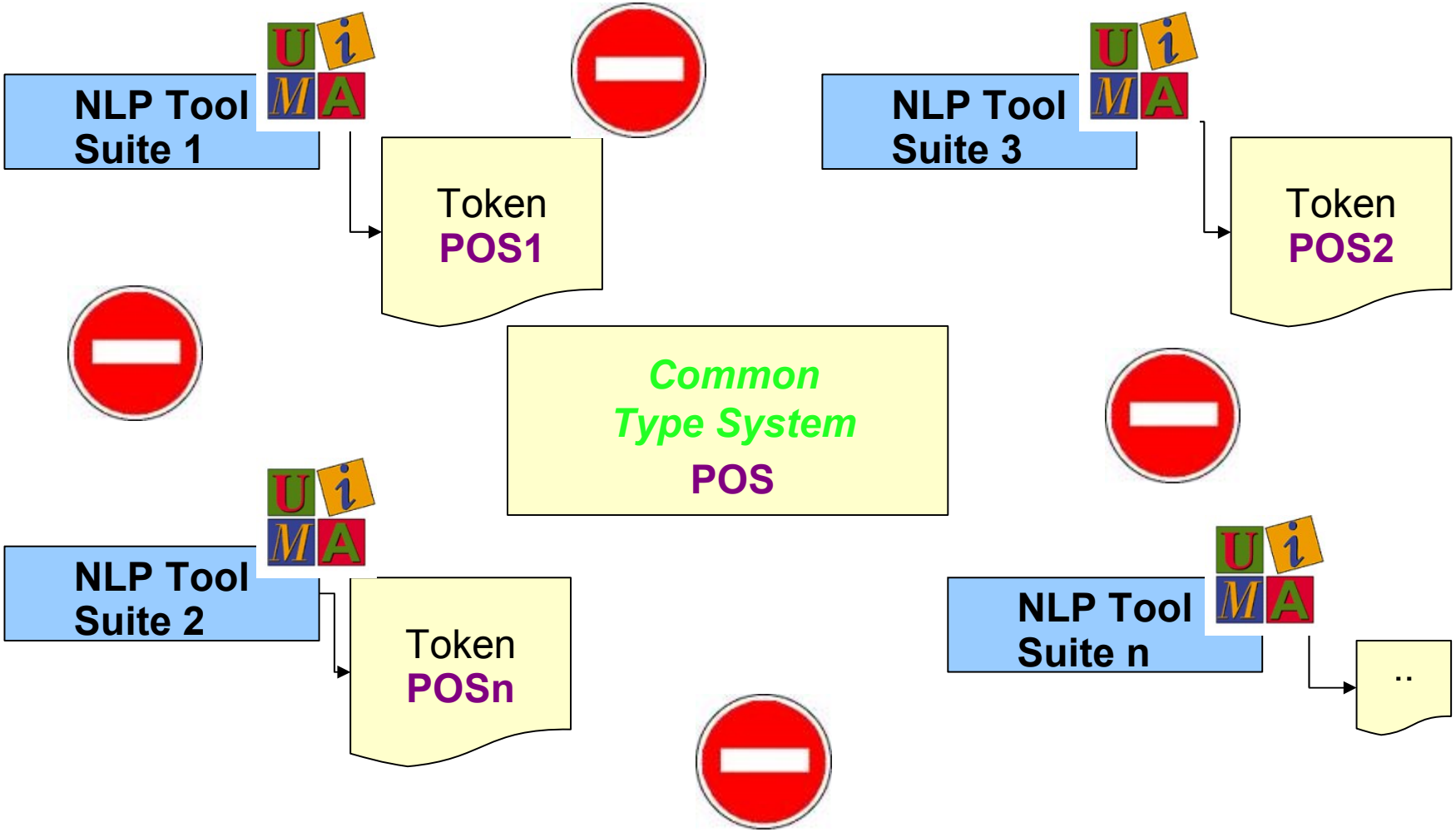
Interoperability between NLP systems

- ✔ Portability of components
- ✗ Flexible exchange of components

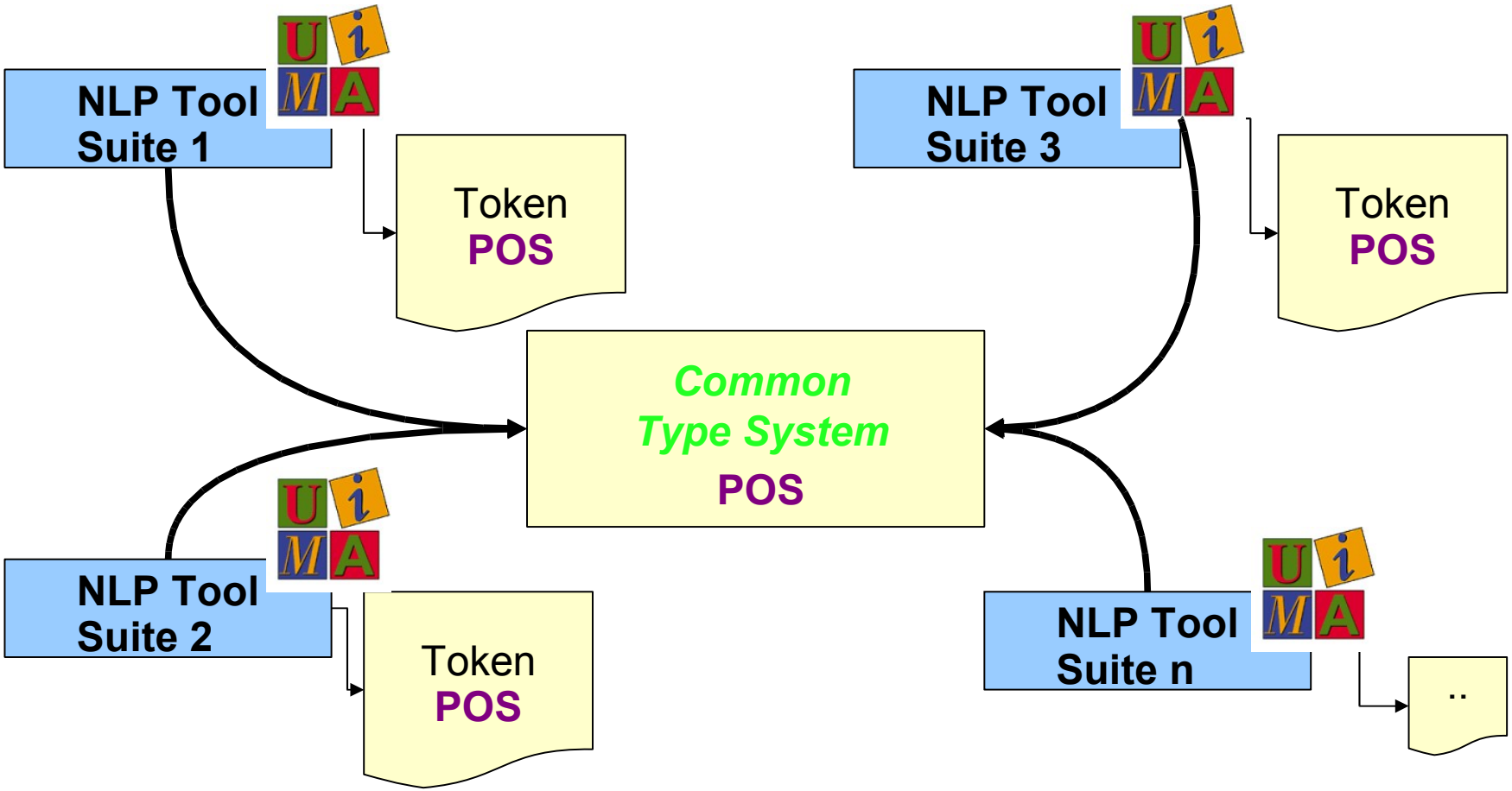
Exchange of components in UIMA

- **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

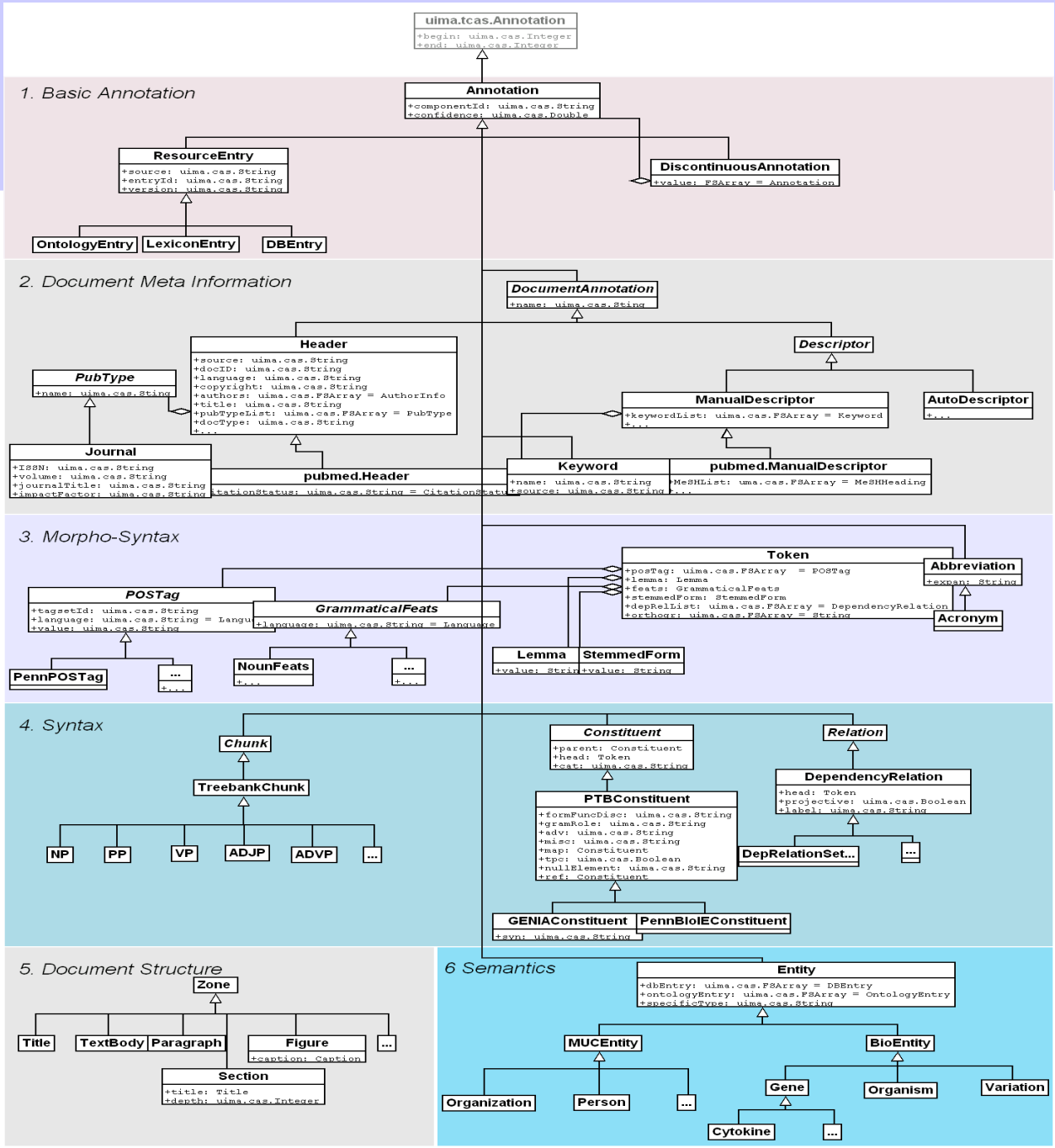
- Broad c**overage** for the information extraction
- Compatible to “standard” NLP annotation schemata
- Definition of the **core** type system which is **extensible**
- Using **UIMA specific features**
  - **Multiple** annotation of the same type
  - **Annotation control** through the restriction of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

Annotation Guidelines & Schemata
Corpus Annotation

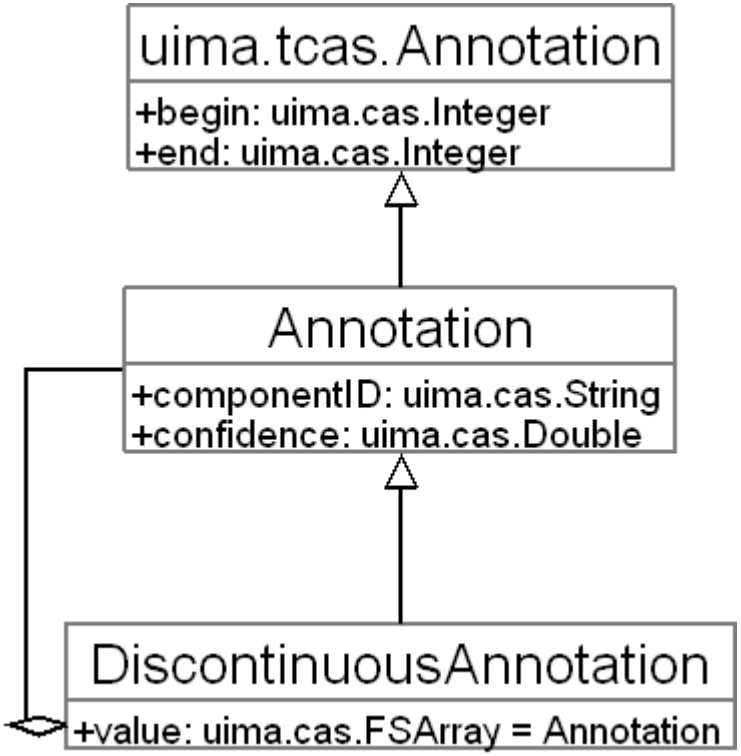
- Annotation languages (e.g. XML (in-line, stand-off))
- Annotation levels:
  - Document Meta (e.g. Dublin Core Metadata Initiative)
  - Linguistic Analysis (e.g. TEI, XCES (EAGLES), Penn Treebank)
  - Semantic Analysis (e.g. MUC, ACE, GENIA)
- NLP system annotation guidelines?

Coverage
Multi-Layered Annotation Type System

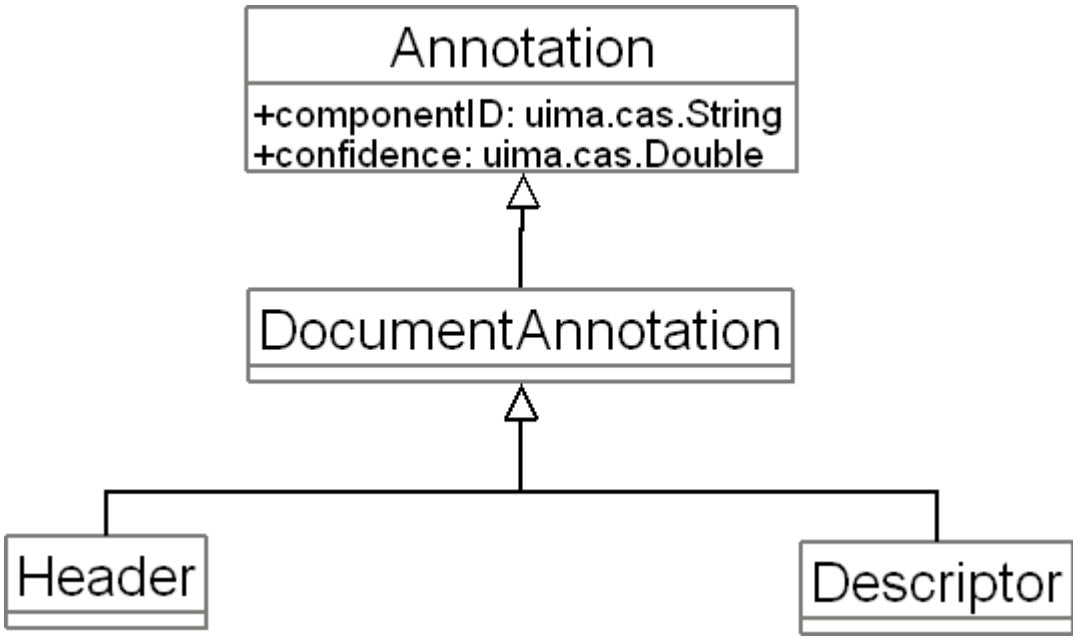
1. **Document Meta**: *author, publication data, source*
2. **Document Structure & Style** : *title, sections, text bold*
3. **Morpho-Syntax**: *token, part-of speech, lemma*
4. **Syntax**: *chunks, constituents, dependency relations*
5. **Semantics**: *entities, relations, events*
6. ***Discourse**: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


uima.tcas.Annotation
+begin: uima.cas.Integer
+end: uima.cas.Integer
1. Basic Annotation
Annotation
+componentId: uima.cas.String
+confidence: uima.cas.Double
ResourceEntry
+source: uima.cas.String
+entryId: uima.cas.String
+version: uima.cas.String
DiscontinuousAnnotation
+value: FSArray = Annotation
OntologyEntry
LexiconEntry
DBEntry
2. Document Meta Information
DocumentAnnotation
+name: uima.cas.String
Header
+source: uima.cas.String
+docID: uima.cas.String
+language: uima.cas.String
+copyright: uima.cas.String
+authors: uima.cas.FSArray = AuthorInfo
+title: uima.cas.String
+pubTypeList: uima.cas.FSArray = PubType
+docType: uima.cas.String
+...
PubType
+name: uima.cas.String
Descriptor
ManualDescriptor
+keywordList: uima.cas.FSArray = Keyword
+...
AutoDescriptor
+...
Journal
+ISSN: uima.cas.String
+volume: uima.cas.String
+journalTitle: uima.cas.String
+impactFactor: uima.cas.String
pubmed.Header
+citationStatus: uima.cas.String = CitationStatu
Keyword
+name: uima.cas.String
+source: uima.cas.String
pubmed.ManualDescriptor
+MeSHList: uima.cas.FSArray = MeSHHeading
+...
3. Morpho-Syntax
Token
+posTag: uima.cas.FSArray = POSTag
+lemma: Lemma
+feats: GrammaticalFeats
+stemmedForm: StemmedForm
+depRelList: uima.cas.FSArray = DependencyRelation
+orthogr: uima.cas.FSArray = String
Abbreviation
+expan: String
Acronym
POSTag
+tagsetId: uima.cas.String
+language: uima.cas.String = Langu
+value: uima.cas.String
GrammaticalFeats
+language: uima.cas.String = Language
PennPOSTag
...
+...
NounFeats
+...
...
+...
Lemma
+value: Strin
StemmedForm
+value: String
4. Syntax
Chunk
TreebankChunk
NP
PP
VP
ADJP
ADVP
...
Constituent
+parent: Constituent
+head: Token
+cat: uima.cas.String
PTBConstituent
+formFuncDisc: uima.cas.String
+gramRole: uima.cas.String
+adv: uima.cas.String
+misc: uima.cas.String
+map: Constituent
+tpc: uima.cas.Boolean
+nullElement: uima.cas.String
+ref: Constituent
GENIAConstituent
PennBioIEConstituent
+syn: uima.cas.String
Relation
DependencyRelation
+head: Token
+projective: uima.cas.Boolean
+label: uima.cas.String
DepRelationSet...
...
5. Document Structure
Zone
Title
TextBody
Paragraph
Figure
+caption: Caption
...
Section
+title: Title
+depth: uima.cas.Integer
6 Semantics
Entity
+dbEntry: uima.cas.FSArray = DBEntry
+ontologyEntry: uima.cas.FSArray = OntologyEntry
+specificType: uima.cas.String
MUCEntity
BioEntity
Organization
Person
...
Gene
Organism
Variation
Cytokine
...

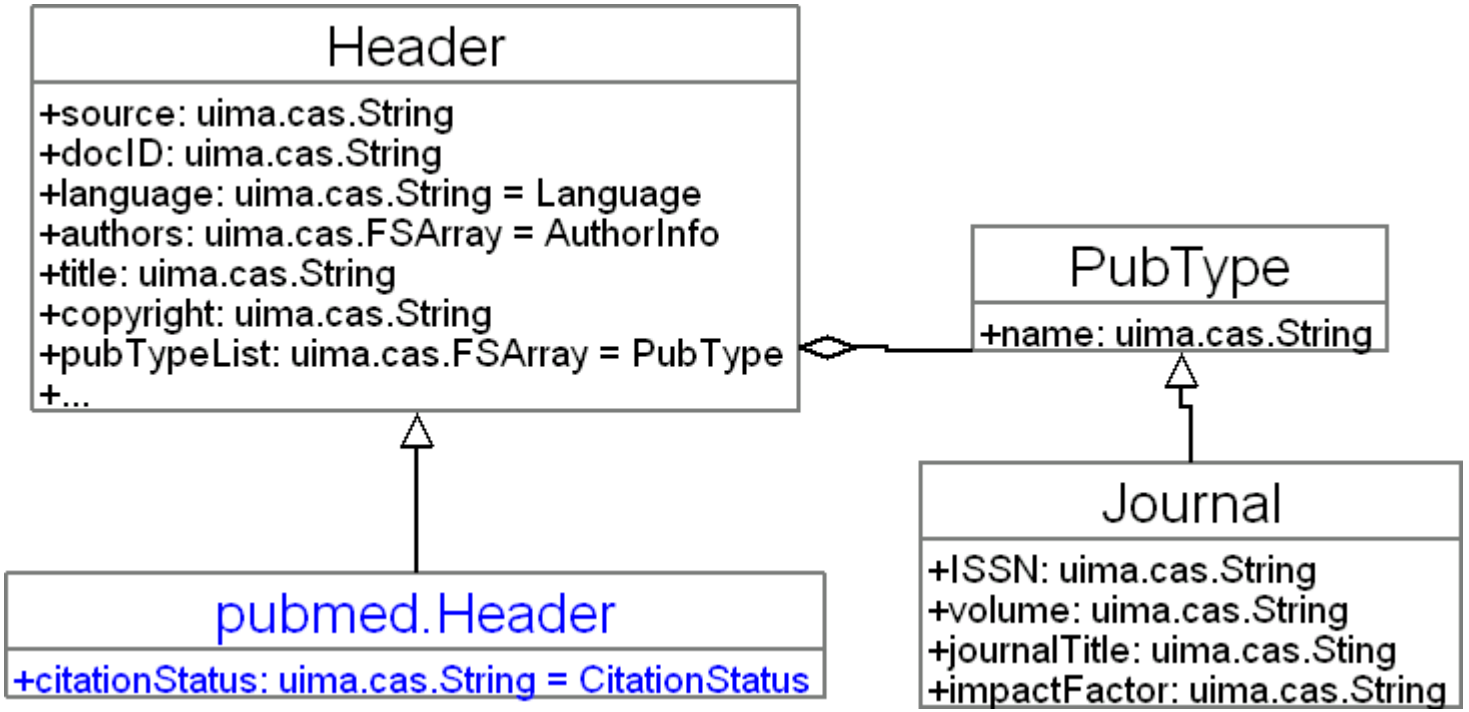
Basic Annotation Type



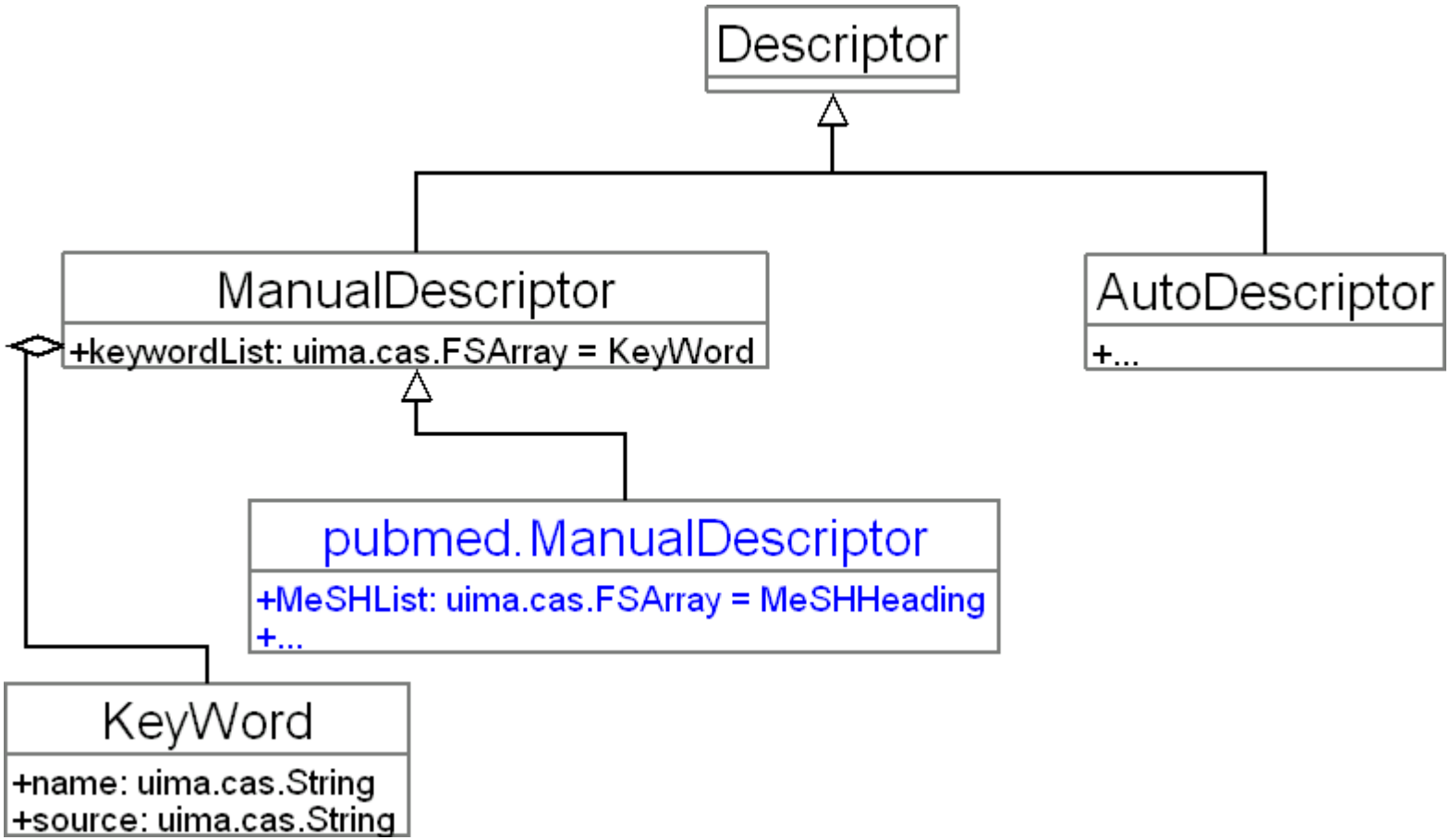
Document Meta



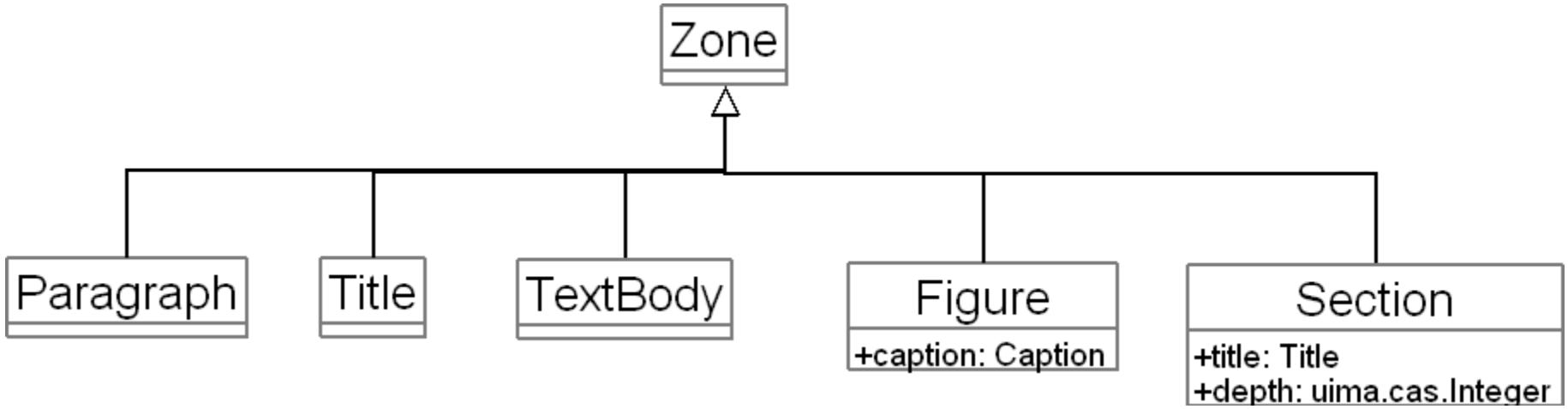
Document Meta Information I



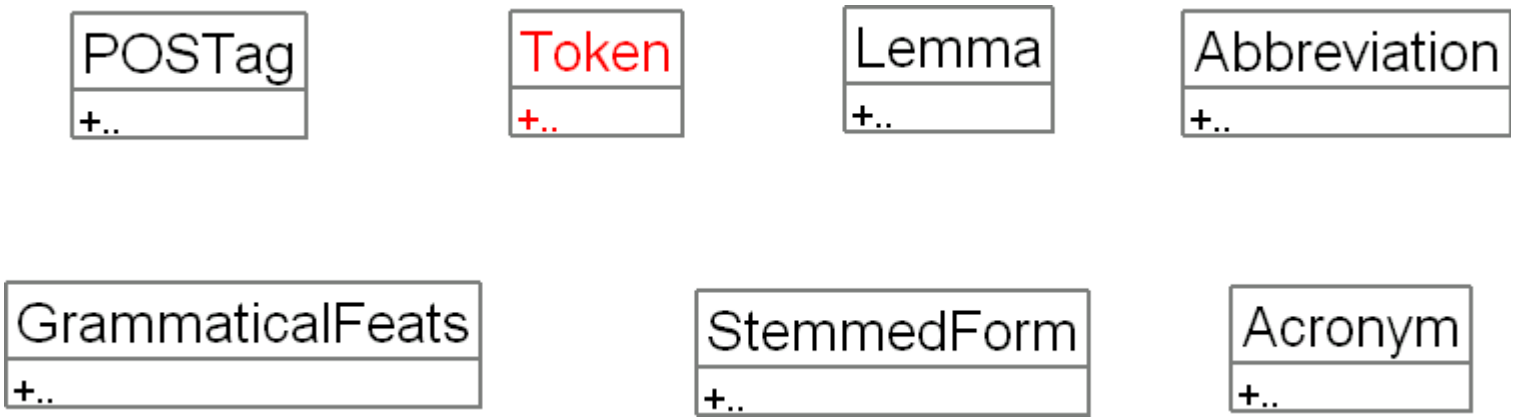
Document Meta Information II



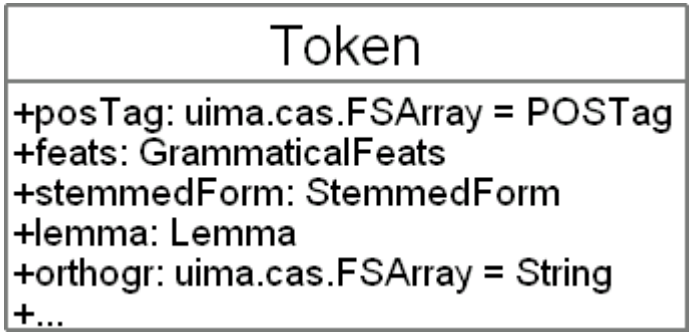
Document Structure



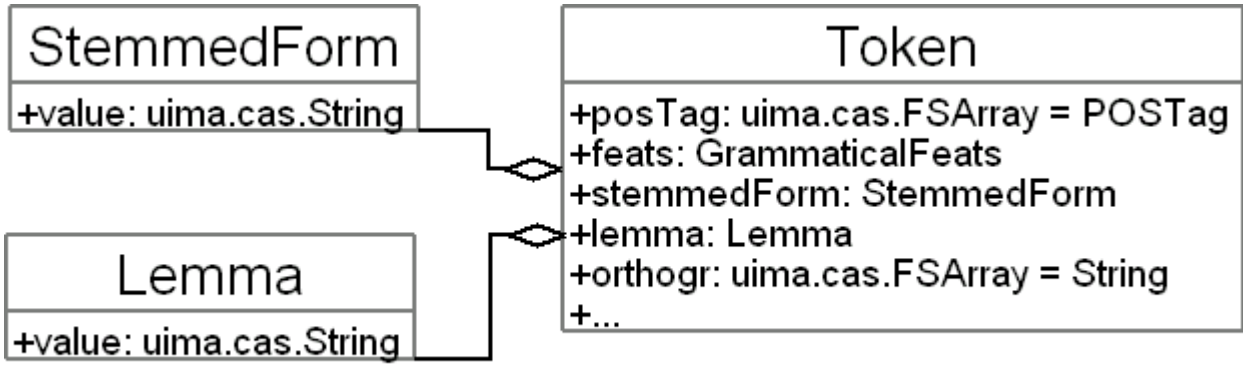
Morpho-Syntax



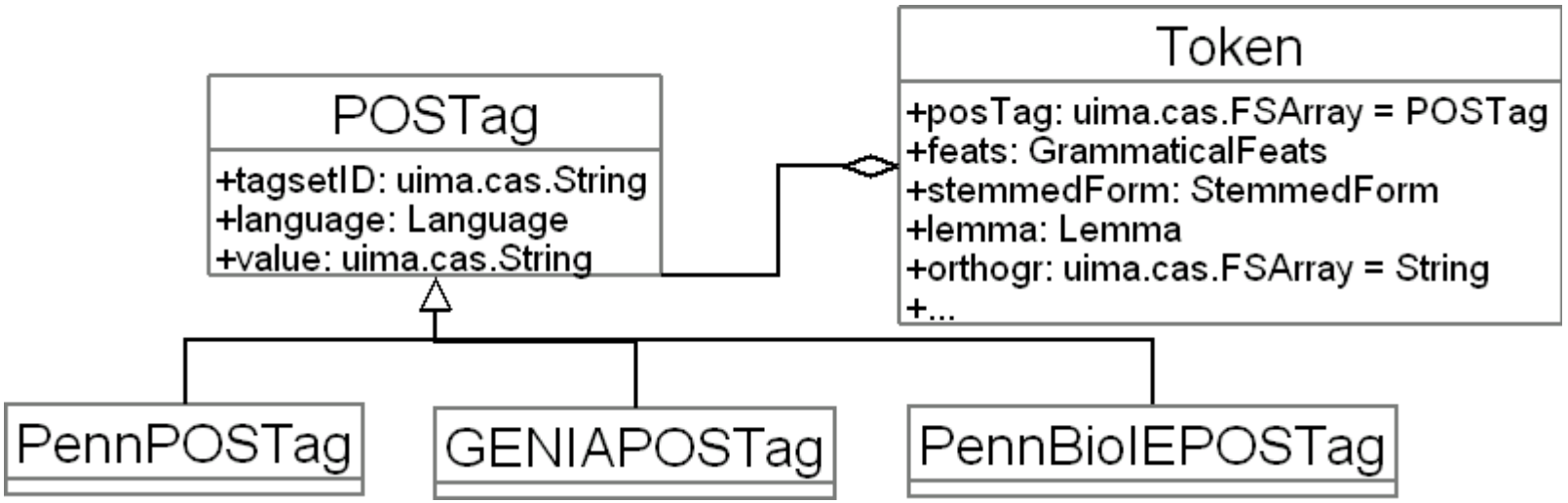
Morpho-Syntax I



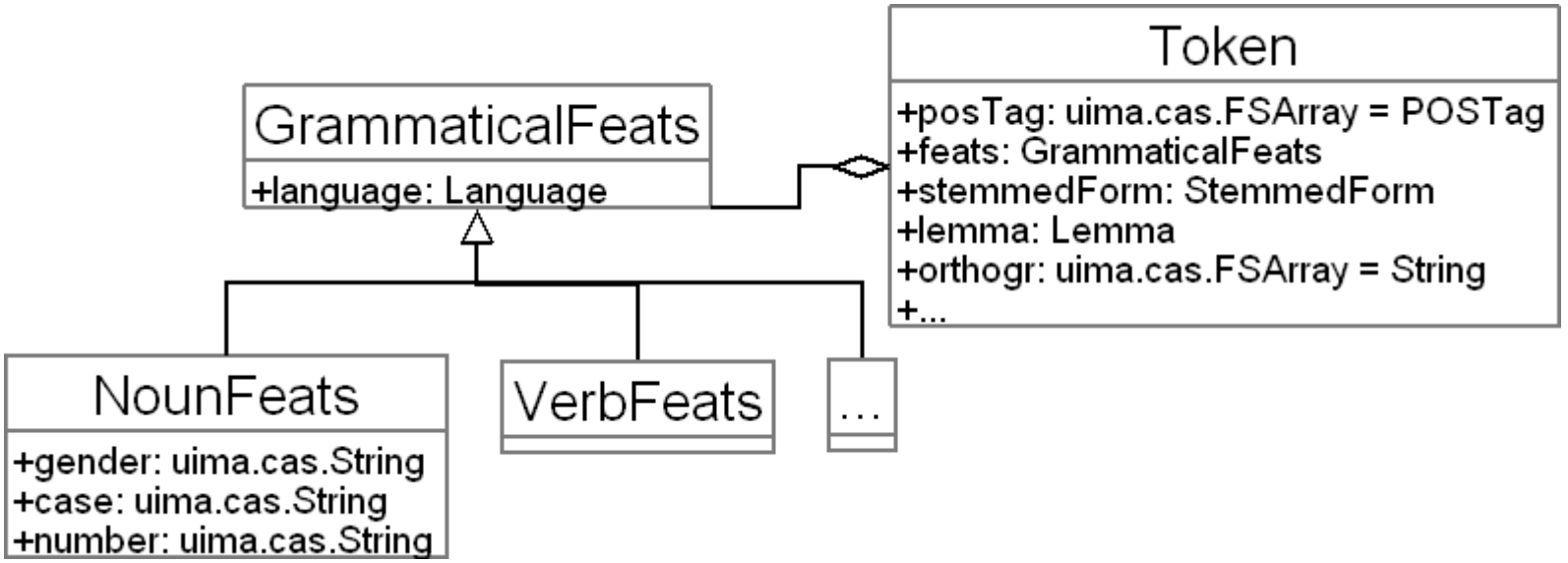
Morpho-Syntax II



Morpho-Syntax III

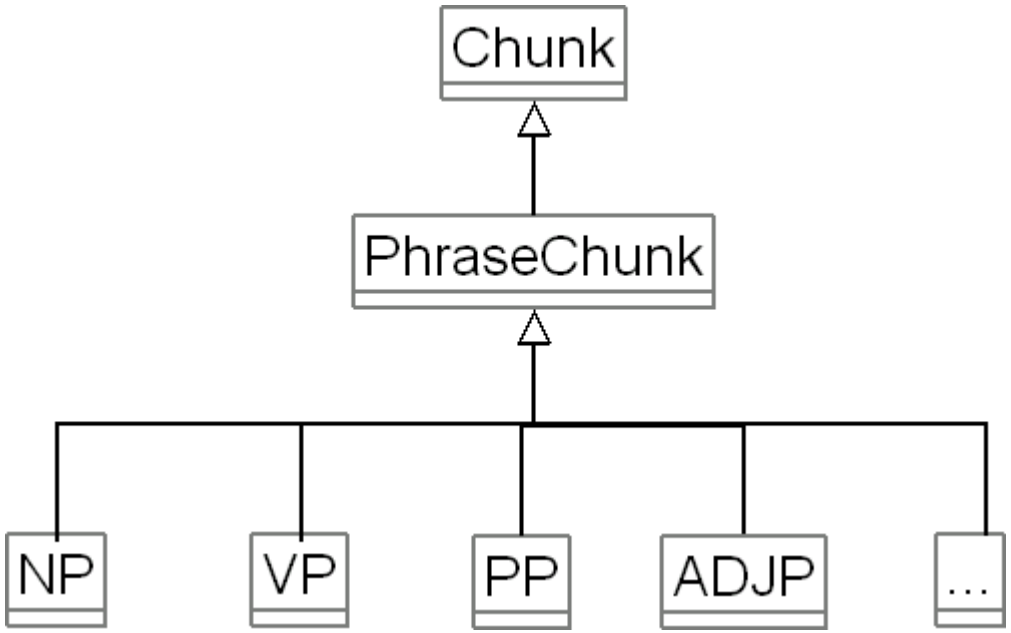


Morpho-Syntax IV

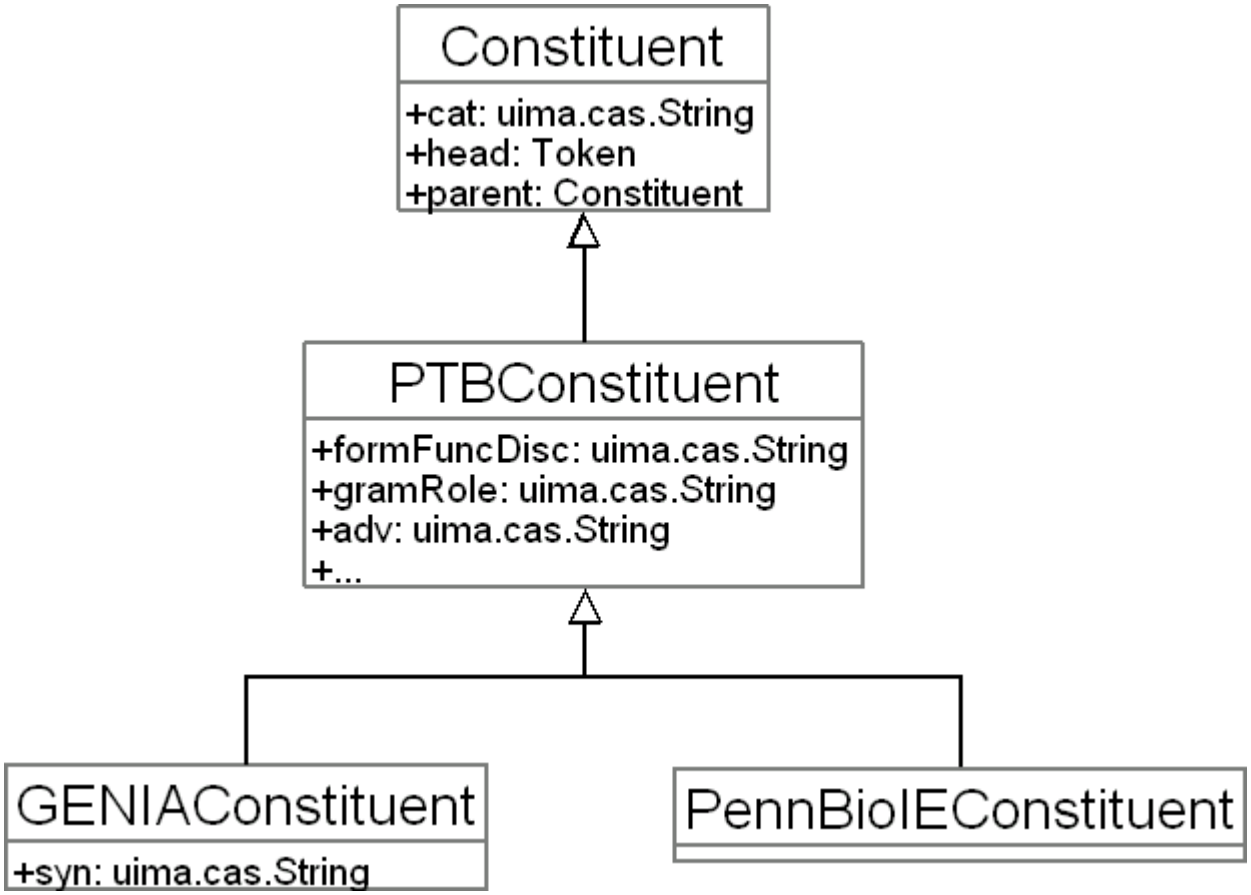


Syntax

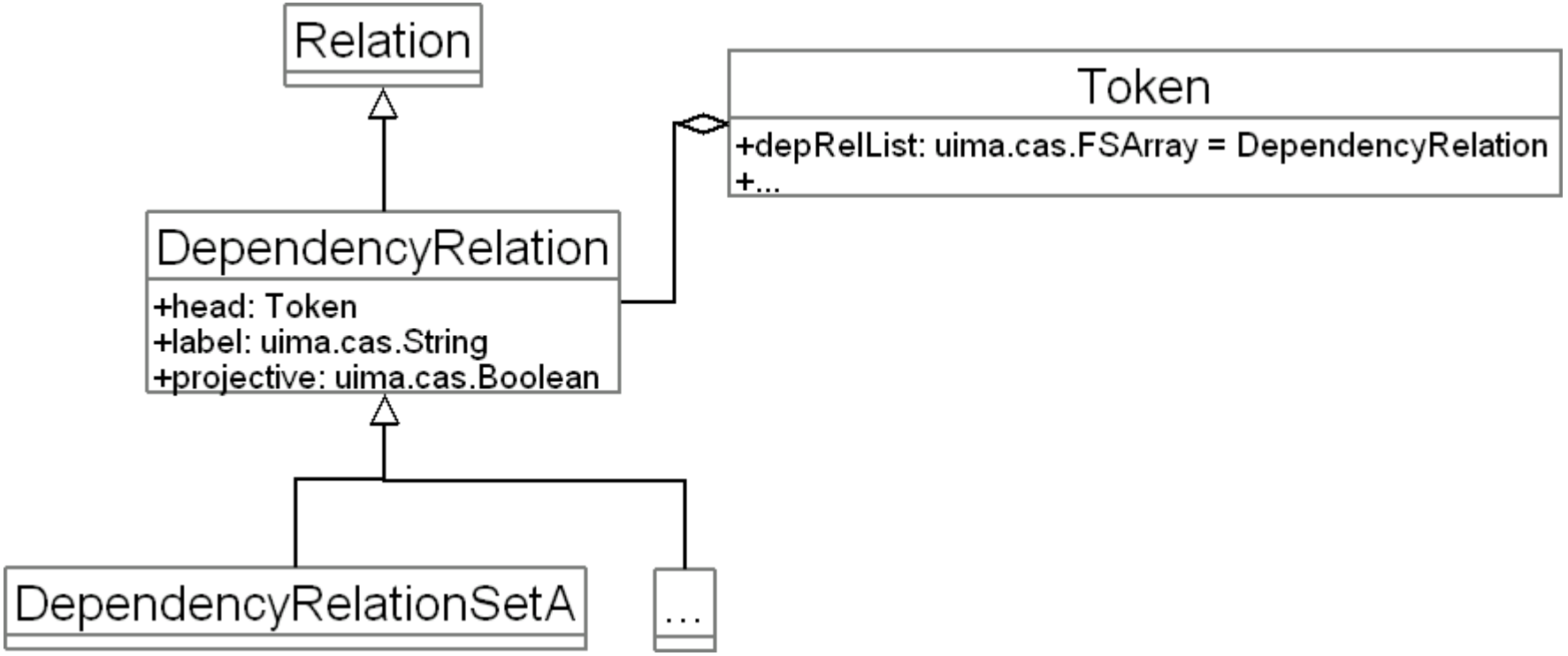
Shallow Parsing



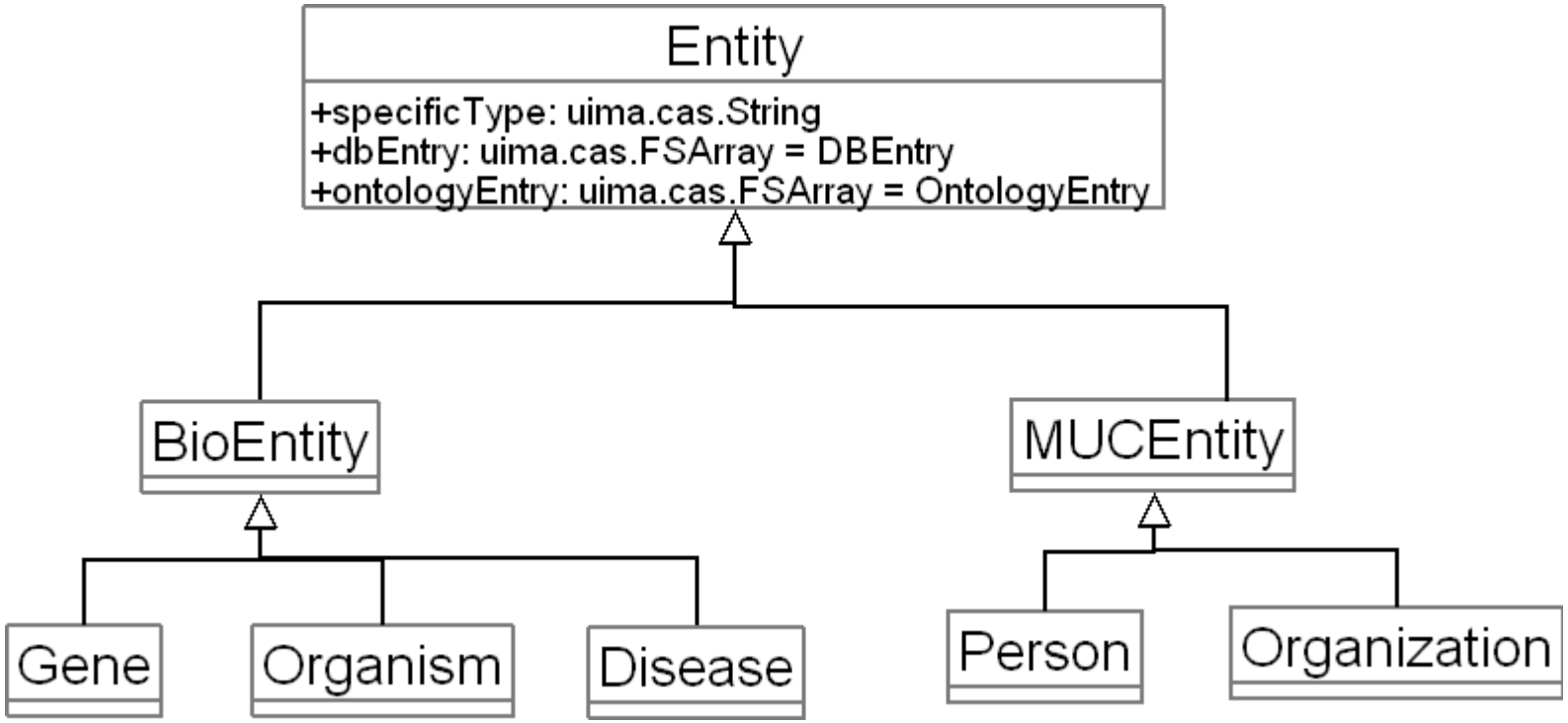
Full Parsing (constituent-based)



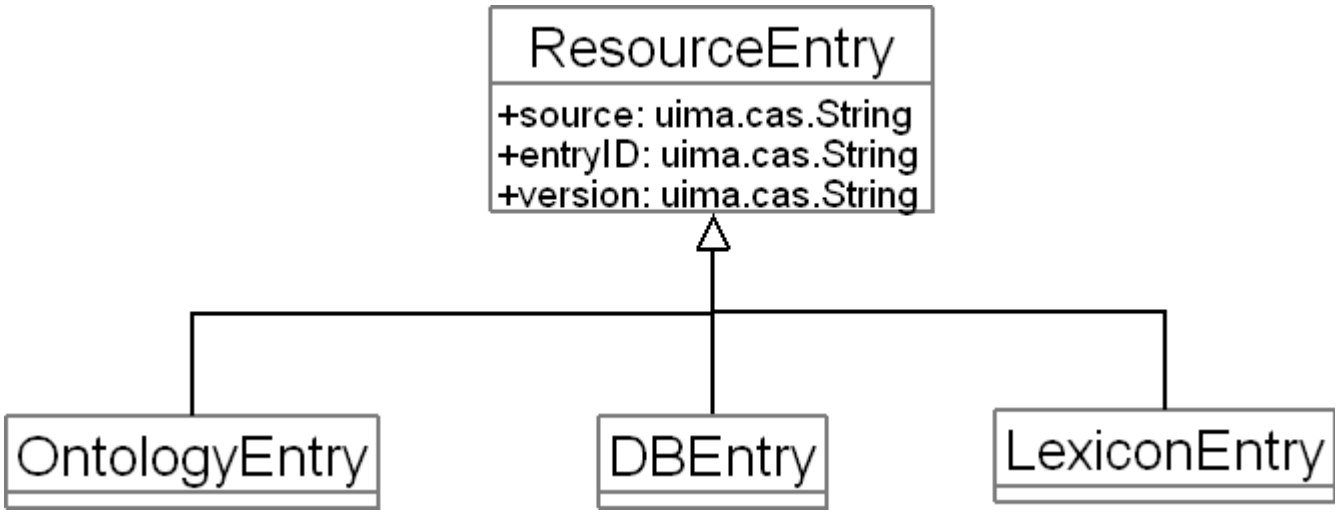
Full Parsing (dependency-based)



Semantics


Entity
+specificType: uima.cas.String
+dbEntry: uima.cas.FSArray = DBEntry
+ontologyEntry: uima.cas.FSArray = OntologyEntry
BioEntity
MUCEntity
Gene
Organism
Disease
Person
Organization

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