Universität Augsburg
Philologisch-Historische
Fakultät

Aspects of Knowledge Representation for Discourse Relation Annotation

Christian Chiarcos

Applied Computational Linguistics (ACoLi)

University of Augsburg, Germany

Institute of Computer Science, Polish Academy of Sciences, Warsaw, Nov 21, 2024

Agenda

1 Semantic Technologies: Knowledge Graphs and Language Technology

2 Discourse and Discourse Relations

3 Formalizing Discourse Relations

4 Linking Discourse Marker Inventories

5 Inducing Discourse Marker Inventories

6 Annotation Engineering with Knowledge Graph Technologies

7 Towards a Multilingual Corpus of Discourse and Reference

Semantic Technologies

Knowledge Graphs and Language Technology

Semantic Technologies

Two main aspects

Provide and Process Structured Information

- Knowledge Representation
- technologies and protocols for sharing, accessing and inference over knowledge graphs
  - Ontologies, Property Graphs, graph technologies
- grounded on web (W3C) standards
  - HTTP, URI, RDF, SPARQL, OWL
  - **federation & interoperablity**: integrate information that is provided by independent sources with heterogeneous technologies from different locations
- The field formerly known as Semantic Web

  (What you expect at ISWC, ESWC, etc.)

Identify Information in Natural Language

- Natural Language Understanding (NLU)
- given natural language input, provide a structured representation of its information according to a specific representation formalism
- traditionally (mostly) supervised learning problems
- addressed in long-standing series of Shared Tasks devoted to individual sub-problems
  - Syntactic Parsing, Named Entity Recognition, Entity Linking, Co-Reference, Information Extraction, Semantic Role Labelling, Semantic Parsing, ...
- One of the primary concerns of the NLP community

  (What you expect at ACL, EMNLP, etc.)

Semantic Technologies

Two main aspects

Provide and Process Structured Information

- Knowledge Representation

Identify Information in Natural Language

- Natural Language Understanding (NLU)

Linguistic Data Science

bringing together both aspects/communities/worlds

use knowledge representation standards to solve interoperability problems in NLP (NLU), e.g., in the creation of training data

- The field formerly known as Semantic Web

  (What you expect at ISWC, ESWC, etc.)

- One of the primary concerns of the NLP community

  (What you expect at ACL, EMNLP, etc.)

Semantic Technologies

Two main aspects

Provide and Process Structured Information

Identify Information in Natural Language

LANGUAGE, DATA and KNOWLEDGE 2023

13-15 September in Vienna, Austria

LDK 2023 – 4th Conference on Language, Data and Knowledge

ABOUT
Find out more about LDK 2023
CALL FOR PAPERS
Submit a paper or abstract
PROGRAM
Coming soon
CALL FOR WORKSHOPS & TUTORIALS
Now open
ATTENDING
Local info about Vienna

Linguistic Data Science

bringing together both aspects/communities/worlds

use knowledge representation standards to solve interoperability problems in NLP (NLU), e.g., in the creation of training data

Nexus Linguarum

cost

EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

W3C COMMUNITY & BUSINESS GROUPS

LINKED DATA FOR LANGUAGE TECHNOLOGY

ONTOLOGY-LEXICA COMMUNITY GROUP

BEST PRACTICES FOR MULTILINGUAL LINKED OPEN DATA COMMUNITY GROUP

Semantic Technologies

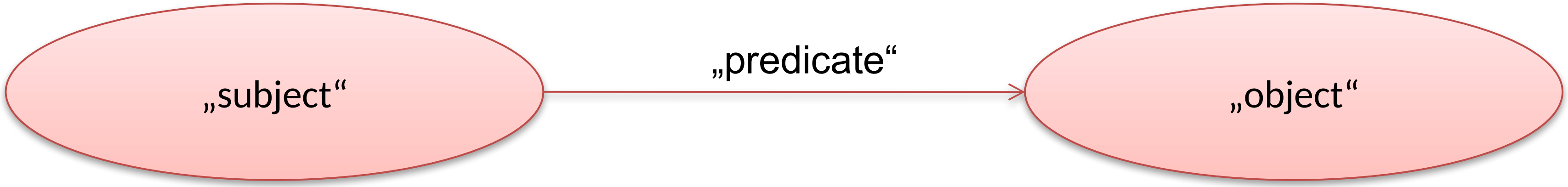
Parts of our Technology Stack (also see https://github.com/acoli-repo/)

- preprocessing
  - CoNLL-Merge: merge TSV files, normalize tokenization, merge annotations (Chiarcos & Schenk@LDK-2019)
- RDF conversion and enrichment
  - FINTAN: Flexible Integrated Transformation and Annotation eNgineering platform
    - more in a moment ;) (Fäth et al.@ LREC-2020)
  - CoNLL-RDF: FINTAN customization for CoNLL/TSV files (Chiarcos & Fäth@LDK-2017)
- selected knowledge graphs
  - ACoLi Dictionary Graph: 3000+ bilingual dictionaries (Chiarcos et al.@LREC-2020)
  - Ontologies of Linguistic Annotation: 100+ annotation schemes (Chiarcos & Sukhareva, SWJ, 2015)
- foundational standards cf. Cimiano, Chiarcos, Gracia & McCrae (2020), *Linguistic Linked Data*. Springer, Cham
  - W3C standards: URI, HTTP, RDF
  - community standards: OntoLex, NLP Interchange Format, CoNLL-RDF data model

Resource Description Framework (RDF)

https://www.w3.org/RDF/

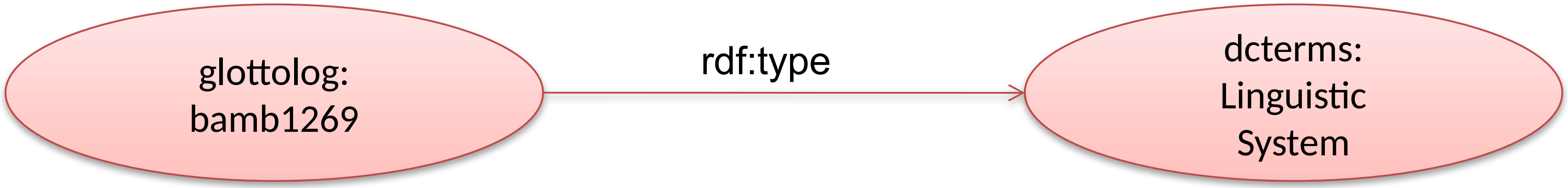
- a (labeled directed multi-) graph
  - nodes („RDF resources“)
    - anything we want to provide information about
  - edges („RDF properties“)
    - assigns a source node („subject“) a target node („object“) or a value („literal“)
  - nodes and edges are unambiguously identified
    - Uniform Resource Identifiers (URIs), e.g., URLs



Resource Description Framework (RDF)

https://www.w3.org/RDF/

glottolog:bamb1269 rdf:type dcterms:LinguisticSystem.

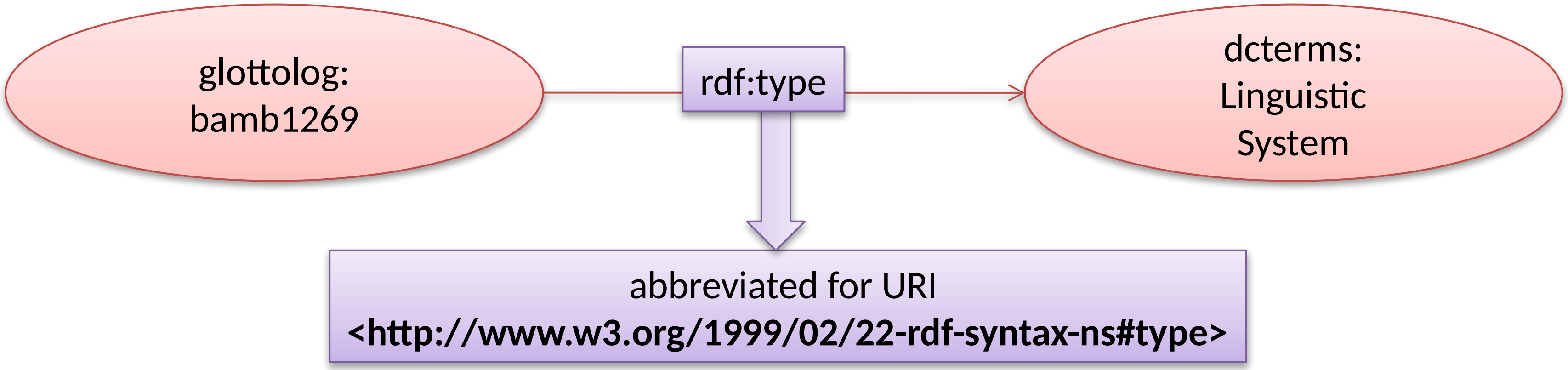


(the concept) „bamb1269“ is a(n instance of concept) „LinguisticSystem“

Resource Description Framework (RDF)

https://www.w3.org/RDF/

glottolog:bamb1269 rdf:type dcterms:LinguisticSystem.



glottolog:
bamb1269

rdf:type

dcterms:
Linguistic
System

abbreviated for URI
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>

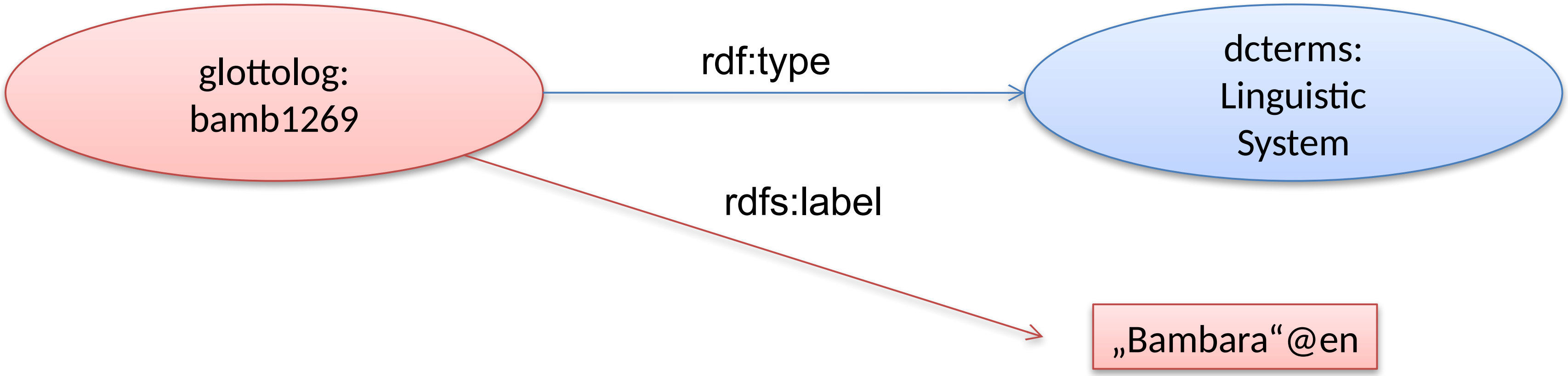
could be opened in a browser
resolvable URIs may provide further information

Resource Description Framework (RDF)

https://www.w3.org/RDF/

glottolog:bamb1269 rdf:type dcterms:LinguisticSystem.

glottolog:bamb1269 rdfs:label „Bambara"@en.



in English (en), a label for (the concept) „bamb1269“ is „*Bambara*“

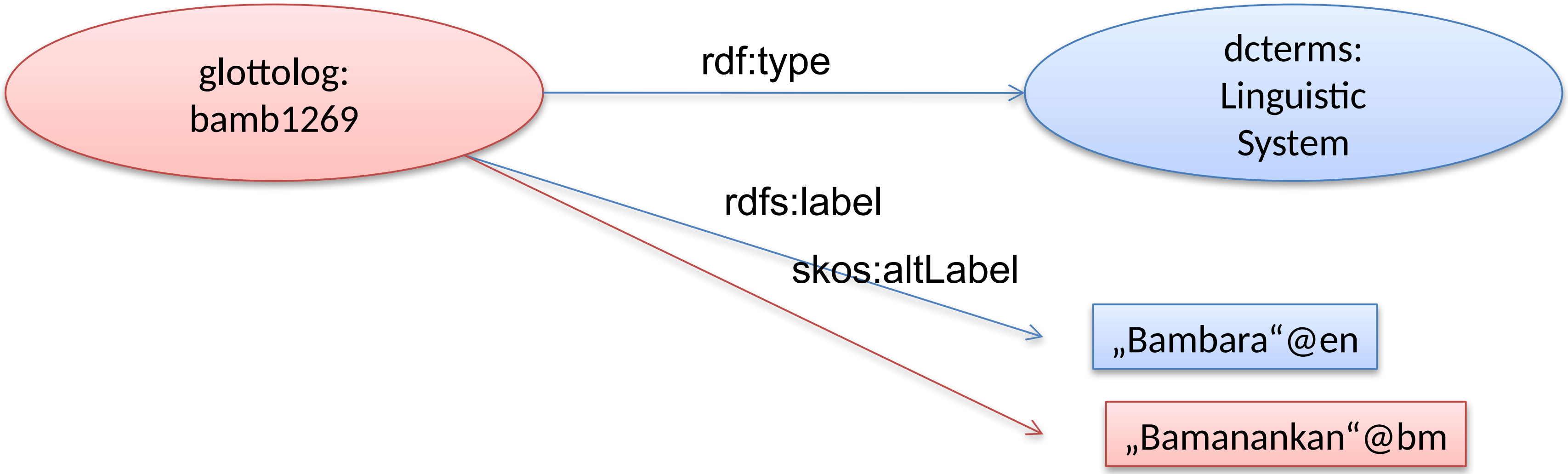
Resource Description Framework (RDF)

https://www.w3.org/RDF/

glottolog:bamb1269 rdf:type dcterms:LinguisticSystem.

glottolog:bamb1269 rdfs:label „Bambara"@en.

glottolog:bamb1269 skos:altLabel „Bamanankan“@bm.



in Bambara (bm), an alternative label for (the concept) „bamb1269“ is „*Bamanankan*“

Resource Description Framework (RDF)

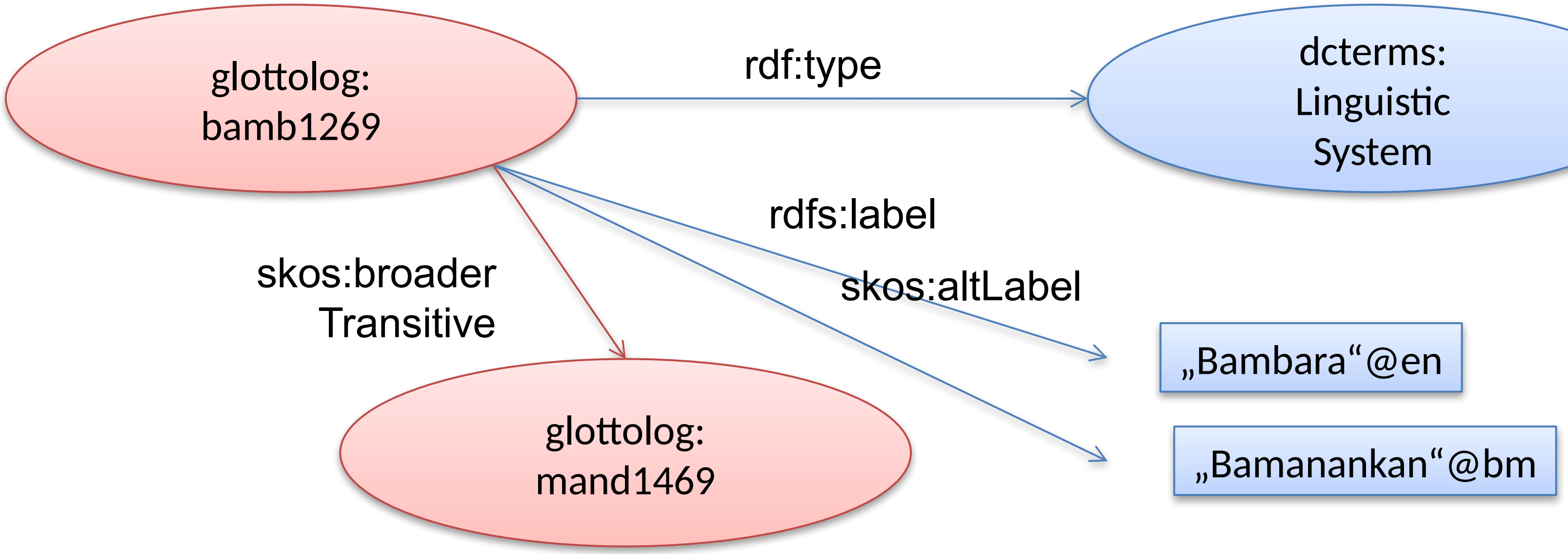
https://www.w3.org/RDF/

glottolog:bamb1269 rdf:type dcterms:LinguisticSystem.

glottolog:bamb1269 rdfs:label „Bambara"@en.

glottolog:bamb1269 skos:altLabel „Bamanankan“@bm.

glottolog:bamb1269 skos:broaderTransitive glottolog:mand1469.



„bamb1269“ pertains to a subgroup of „mand1469“ (= Mande language family)

Resource Description Framework (RDF)

https://www.w3.org/RDF/

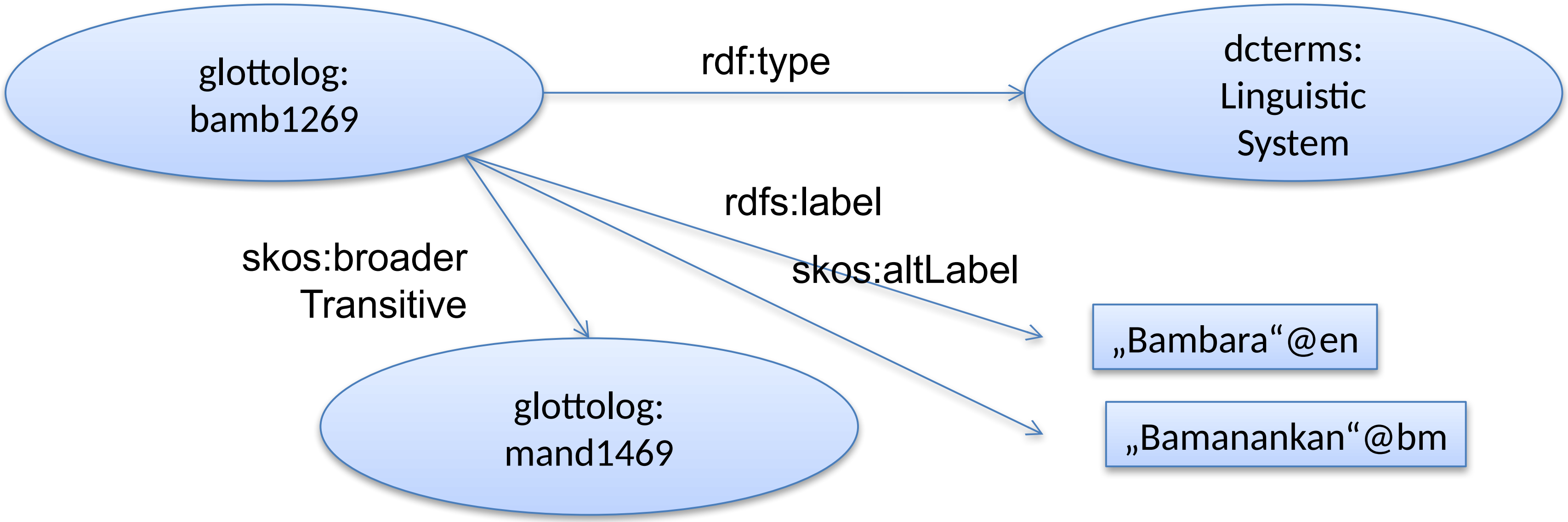
glottolog:bamb1269 rdf:type dcterms:LinguisticSystem.

glottolog:bamb1269 rdfs:label „Bambara"@en.

glottolog:bamb1269 skos:altLabel „Bamanankan“@bm.

glottolog:bamb1269 skos:broaderTransitive glottolog:mand1469.

triple notation (Turtle)



graphical notation

Querying it with SPARQL

https://www.w3.org/TR/sparql11-query/

glottolog:bamb1269 rdfs:label „Bambara"@en.

glottolog:bamb1269 skos:broaderTransitive glottolog:mand1469.

triple notation (Turtle)

```
SELECT ?language_name
WHERE {
      ?language rdfs:label ?language_name.
      ?language skos:broaderTransitive glottolog:mand1469.
}
```

query (SPARQL)

“give me the names of all Mande (glottolog:mand1469) languages”

Querying it with SPARQL*

https://www.w3.org/TR/sparql11-query/

glottolog:bamb1269 rdfs:label „Bambara"@en.

glottolog:bamb1269 skos:broaderTransitive glottolog:mand1469.

triple notation (Turtle)

```
SELECT ?language_name
WHERE {
      ?language rdfs:label ?language_name.
      ?language skos:broaderTransitive glottolog:mand1469.
}
```

query (SPARQL)

“give me the names of all Mande (glottolog:mand1469) languages”

* with SELECT, we perform queries
  with DELETE and INSERT, we perform updates

Linked Data

Rules of best practice for publishing data on the web

- use URIs as names for things (1)

  links to external URIs retrieve more information
- **if** they can be resolved via HTTP (2)
- **and** provide information as RDF, SPARQL, etc. (3)
- **and** they include links to other URIs (4)

⇒ **then**, this is Linked Data (informally)

http://www.w3.org/DesignIssues/LinkedData.html

Rules of best practice

=> Information integration

- ❑ Interoperability

  => the same query to query different datasets

- ❑ Federation
  - data published on the web
    - ❑ with a query interface (SPARQL end point)

  => a single query to query different datasets simultaneously

a formalism to

„build bridges“

=> more (re-)usable resources and technologies

coupled with the dynamics of the

open source / open data movement

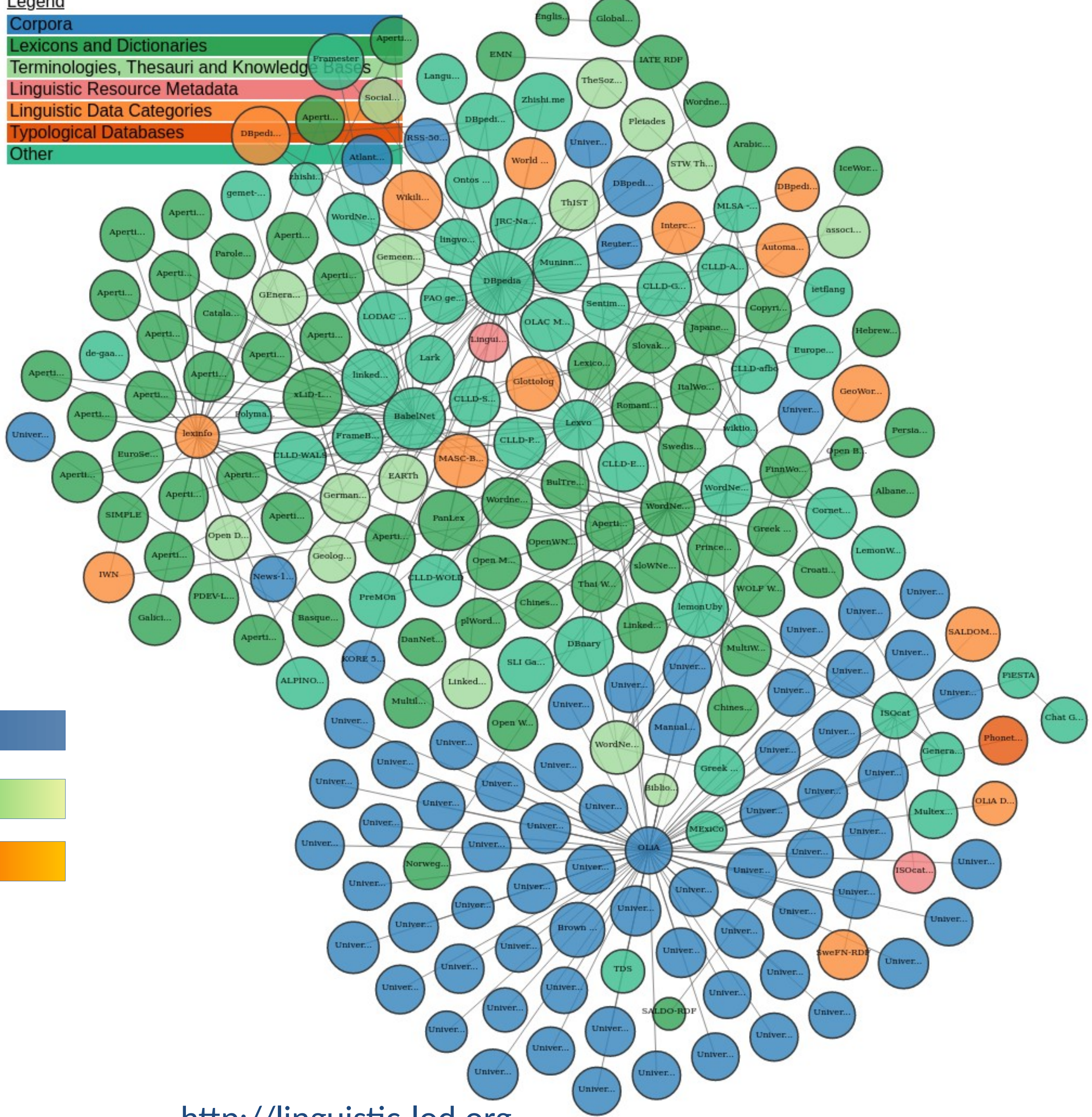
Linked **Open** Data

Linguistic Linked Open Data (LLOD)

LLOD cloud diagram

sub-diagram of the Linked Open Data (LOD) cloud diagram

open resources for
- linguistic annotation 🟦
- lexical/conceptual knowledge 🟩
- linguistically relevant metadata 🟥



http://linguistic-lod.org

Limitations and Potential

- Linked Data and RDF tech are not (and probably won‘t ever be) sufficiently user-friendly for end users (say, a linguist)

- BUT
  - Most users won‘t have to work with it directly, but only mediated through software tools.
- IF
  - The data can be prepared by/for them

Illustrated here for aspects of discourse annotation

Discourse and Discourse Relations

Some Theoretical Background

Discourse

Some non-trivial aspects of Natural Language Understanding

- In Natural Language Understanding, the semantic analysis of individual sentences is an established field of research (and to a large extent, solved).
- But there is meaning between the lines (resp., sentences) ...

  - Peter pushed John.
  - He was hurt badly.

Who was hurt?

Discourse

Some non-trivial aspects of Natural Language Understanding

- In Natural Language Understanding, the semantic analysis of individual sentences is an established field of research (and to a large extent, solved).
- But there is meaning between the lines (resp., sentences) ...
  - Peter pushed John.
  - He was hurt badly.

Probably

John was hurt

and this is the result of Peter pushing him

Discourse

Some non-trivial aspects of Natural Language Understanding

- In Natural Language Understanding, the semantic analysis of individual sentences is an established field of research (and to a large extent, solved).
- But there is meaning between the lines (resp., sentences) ...

  - Peter pushed John.
  - He was hurt badly.

Probably

John was hurt

and this is the result of Peter pushing him

  - Peter pushed John.
  - He was hurt badly.
  - And so, the retaliation felt good, for a while.

Could also be

Peter was hurt

and he pushed John for retaliation

Discourse Markers

How to make the meaning between the lines visible

- Make explicit how an utterance (clause, sentence, etc.) is linked to its discourse context
  - lexical expressions, mostly conjunctions, adverbs and PPs
    - John can‘t go. …

And	Mary can‘t go either.	*additive*
Therefore,	Mary can‘t go either.	*causal*
However,	Mary can‘t go either.	*contrastive*
	Mary can‘t go either.	*implicit (unmarked)*

relation

Discourse Relations

The meaning of discourse markers

- Semantic, pragmatic or conversational relation holding between an utterance and its discourse context.
  - different theories and annotation frameworks
    - Coherence relations (Hobbs, 1979)
    - RST (Mann and Thompson, 1987)
    - SDRT (Asher & Lascarides, 2003)
    - PDTB (Prasad et al., 2008)

overlapping in intent and content, but not compatible with each other

We focus on RST and PDTB here, these provide the major corpora.

Discourse Parsing

- Identify relations to assess how utterances are connected

  for information extraction, text summarization, machine translation, etc.

  off-the-shelf LLMs are still rather poor at such context-dependencies ;)

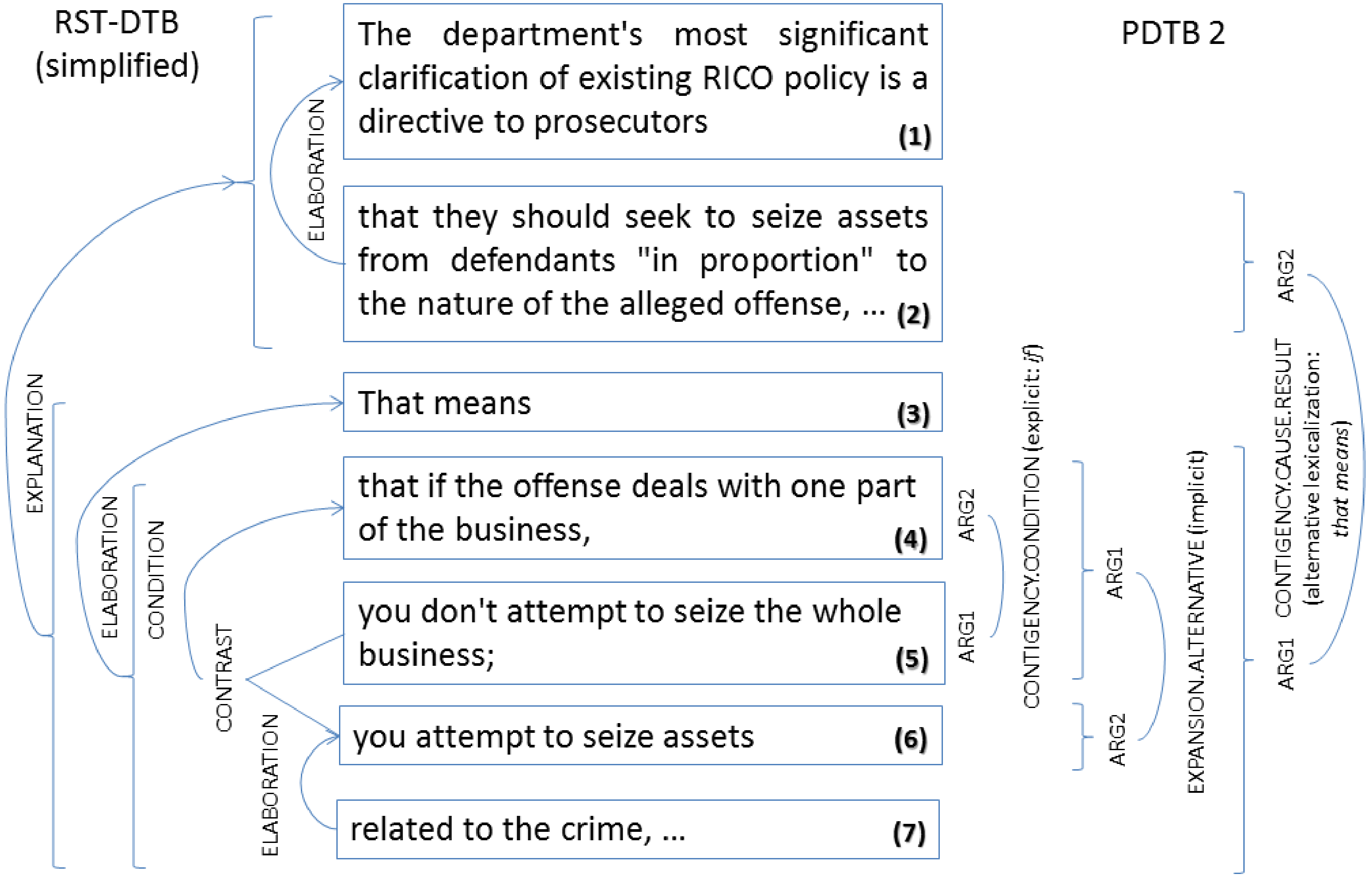
- Two primary (and incompatible) frameworks

Rhetorical Structure Theory (RST)	Penn Discourse Treebank (PDTB)
discourse relations constitute a tree structure that encompasses all utterances of a coherent discourse (“deep” discourse parsing)	forget about the tree annotate any discourse relation you see in the local context (shallow discourse parsing)

Comparing RST and PDTB Annotations

RST-DTB (simplified)

PDTB 2



The department's most significant clarification of existing RICO policy is a directive to prosecutors **(1)**

that they should seek to seize assets from defendants "in proportion" to the nature of the alleged offense, … **(2)**

That means **(3)**

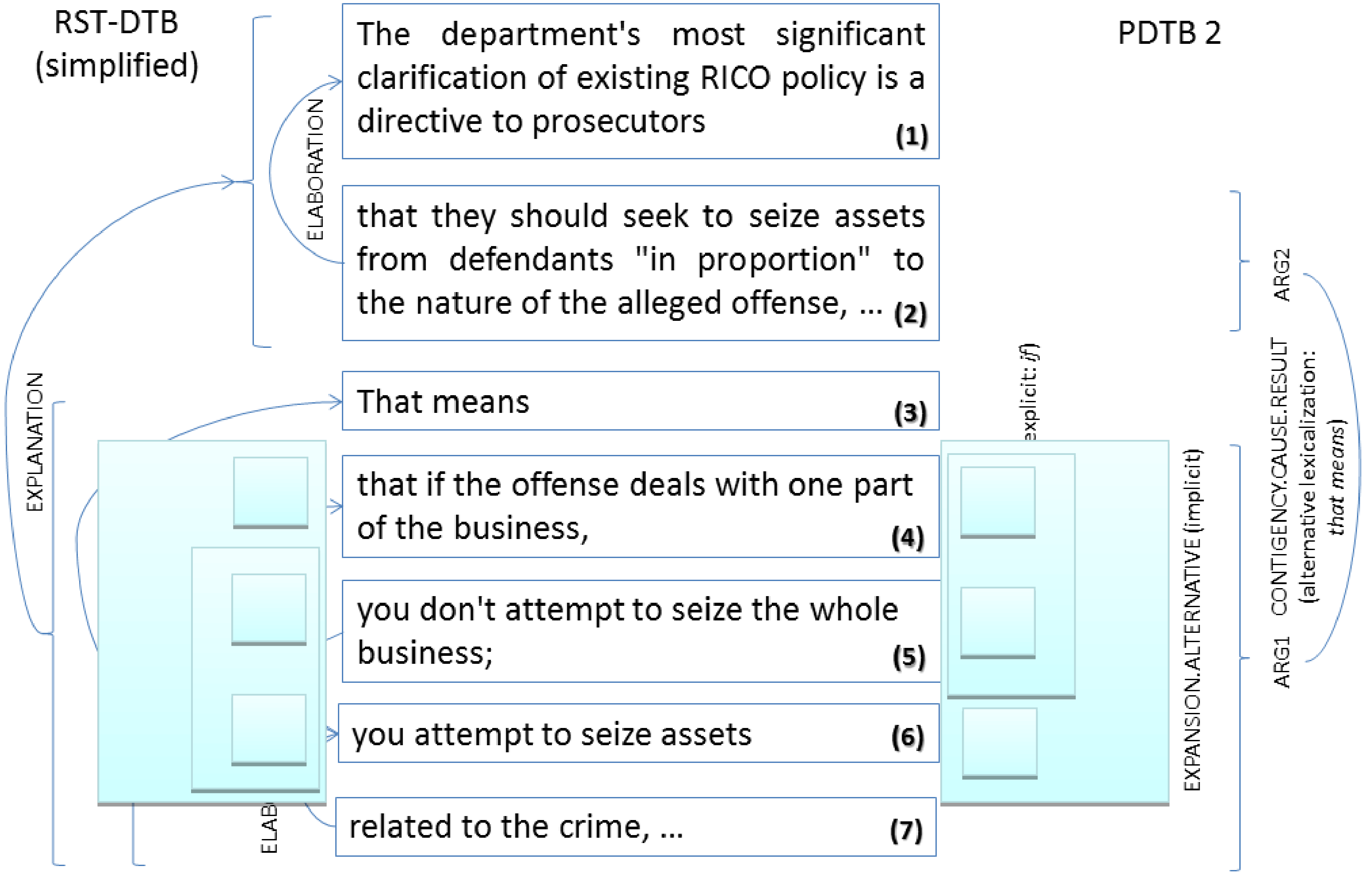
that if the offense deals with one part of the business, **(4)**

you don't attempt to seize the whole business; **(5)**

you attempt to seize assets **(6)**

related to the crime, … **(7)**

Comparing RST and PDTB: Structural Differences



RST-DTB (simplified)

PDTB 2

The department's most significant clarification of existing RICO policy is a directive to prosecutors **(1)**

that they should seek to seize assets from defendants "in proportion" to the nature of the alleged offense, ... **(2)**

That means **(3)**

that if the offense deals with one part of the business, **(4)**

you don't attempt to seize the whole business; **(5)**

you attempt to seize assets **(6)**

related to the crime, ... **(7)**

Goals: Consolidate and Integrate Existing Data Sources

- across languages
  - multilingual discourse markers
- across frameworks
  - RST, PDTB, etc.
- across formats
  - various CSV, XML and special-purpose formats
- machine-readable semantics
  - knowledge graph(s)

Formalizing Discourse Relations

Discourse in the Ontologies of Linguistic Annotation (OLiA)

(Chiarcos@LREC-2014)

What is an Ontology?

... in Knowledge Representation

An ontology is the formalization of concepts and their relations for a particular domain

- Formalized in terms of the Web Ontology Language (OWL)
  - i.e., an RDF vocabulary for classes (concepts), properties (relations) and axioms

- selected properties

rdf:type	(*a*)	assign a class (type) to an object	
rdfs:subClassOf	(⊑)	subclass relation	(cf. logical →)
owl:intersectionOf	(⊓)	intersection between two classes	(cf. logical ∧)
owl:unionOf	(⊔)	union between two classes	(cf. logical ∨)
owl:complementOf	(¬)	complement of a class	(cf. logical ¬)

Ontologies of Linguistic Annotation (OLiA)

http://purl.org/olia

Chiarcos and Sukhareva, SWJ 2015

- one ontology per annotation schema
  - OLiA Annotation Model
- one ontology that defines common terminology
  - OLiA Reference Model
- one RDF file with rdfs:subClassOf statements
  - OLiA Linking Model: Annotation Model => Reference Model
- annotation schemas for 100+ languages
  - mostly morphosyntax, inflectional features and syntax

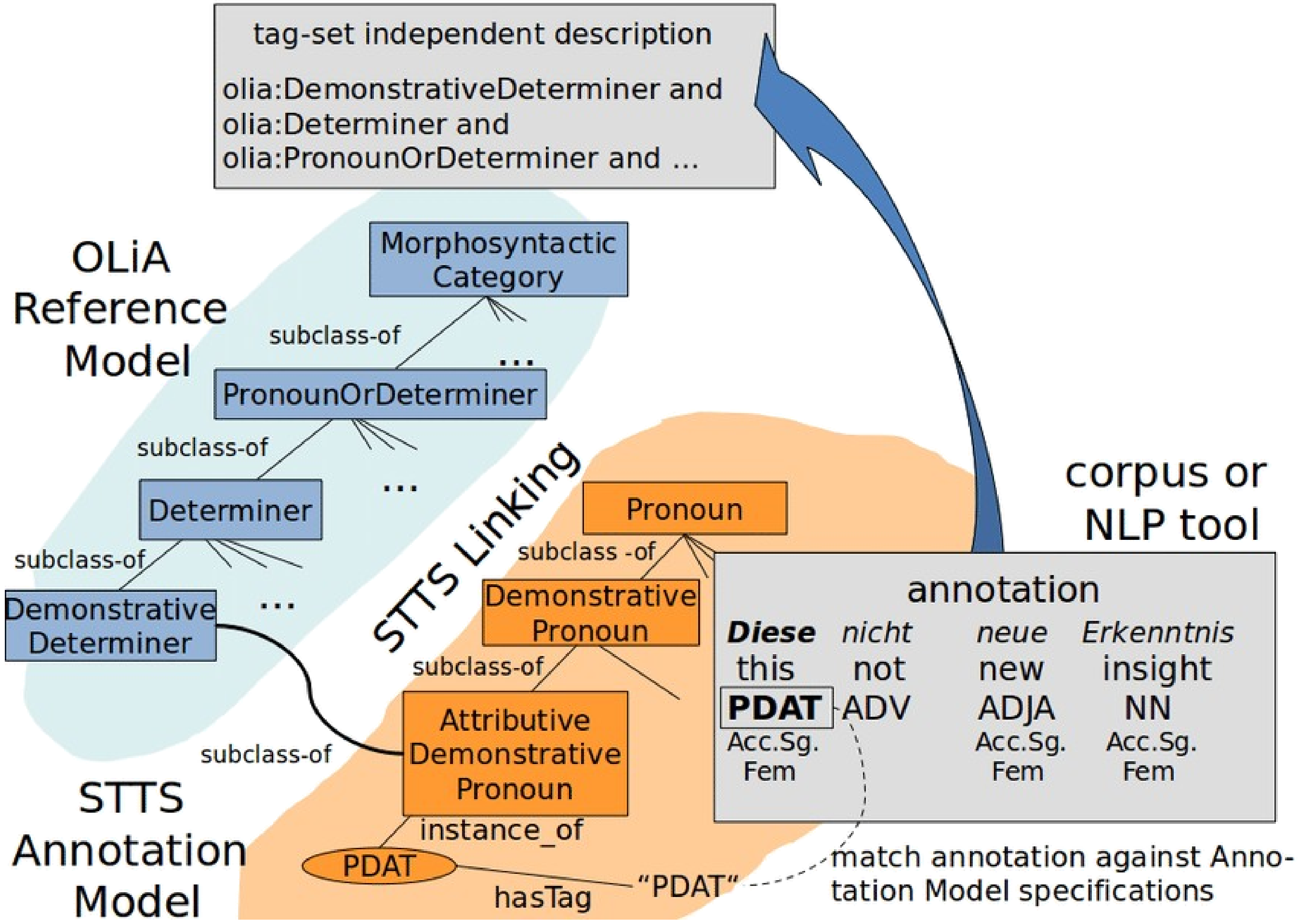
Ontologies of Linguistic Annotation (OLiA)

http://purl.org/olia

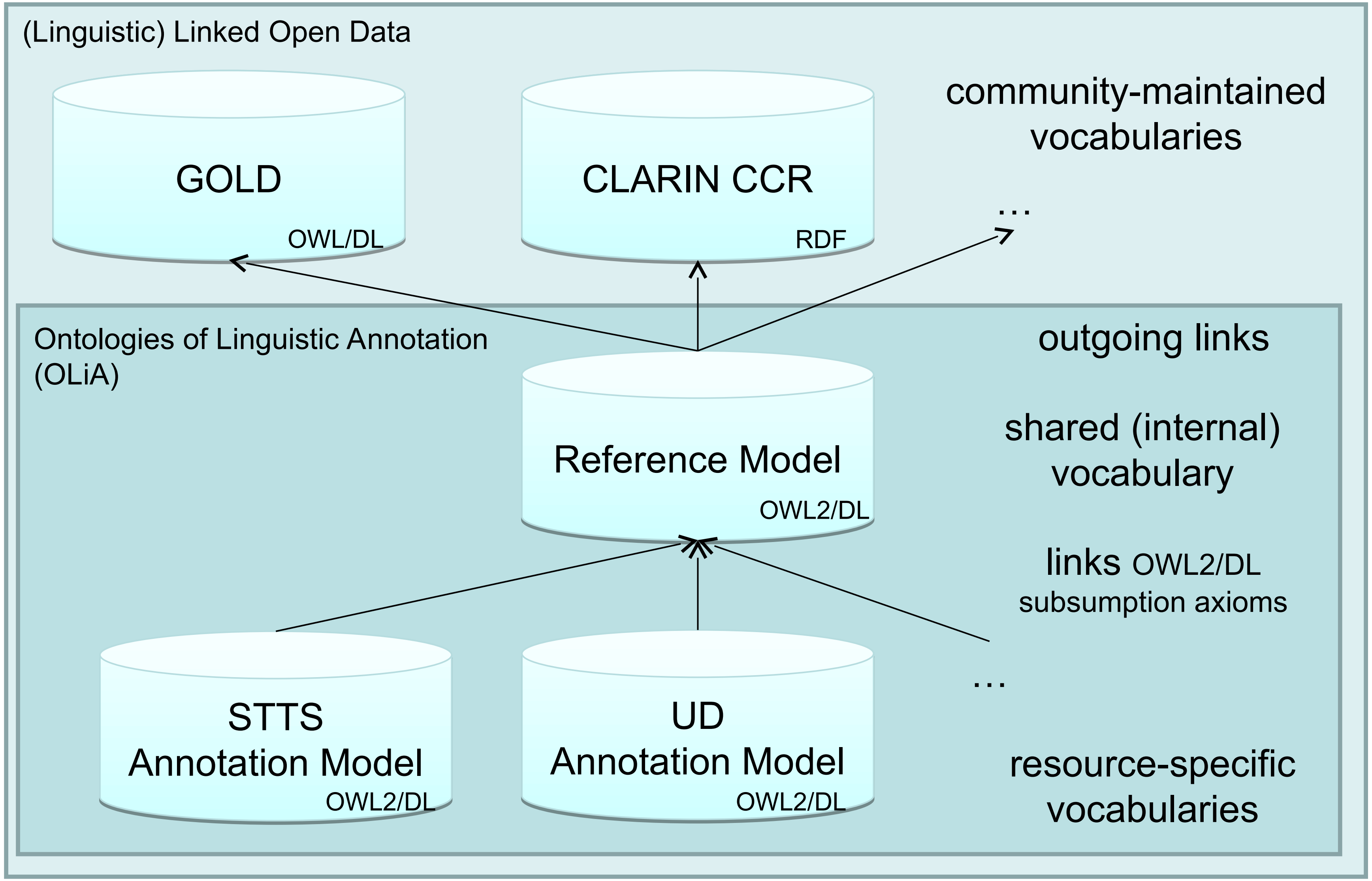
German

parts of

speech



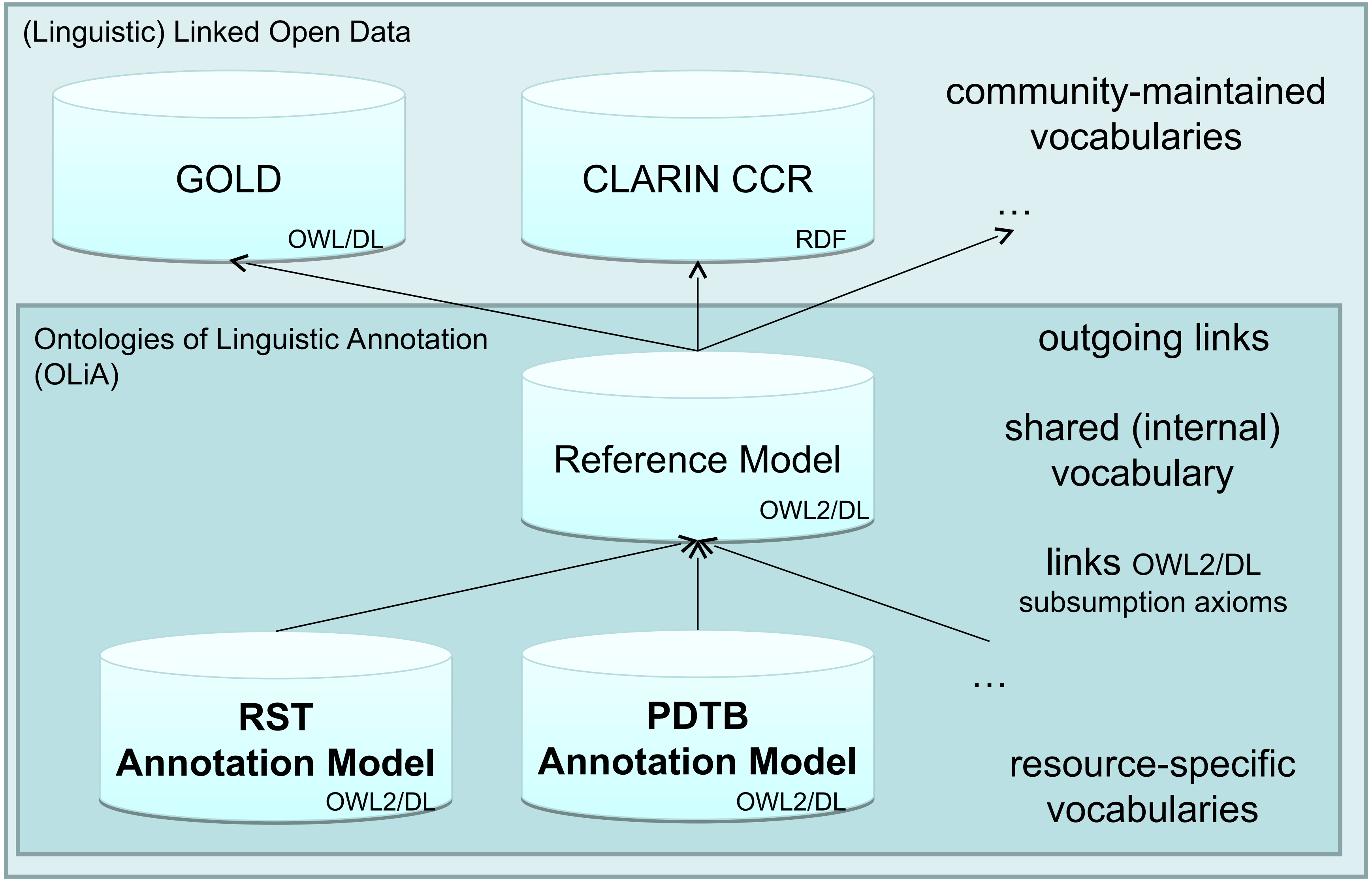
Ontologies of Linguistic Annotation (OLiA)


(Linguistic) Linked Open Data
GOLD
OWL/DL
CLARIN CCR
RDF
community-maintained vocabularies
…
Ontologies of Linguistic Annotation (OLiA)
outgoing links
Reference Model
OWL2/DL
shared (internal) vocabulary
links OWL2/DL subsumption axioms
STTS Annotation Model
OWL2/DL
UD Annotation Model
OWL2/DL
…
resource-specific vocabularies

OLiA Discourse Extension

(Chiarcos@LREC-2014)

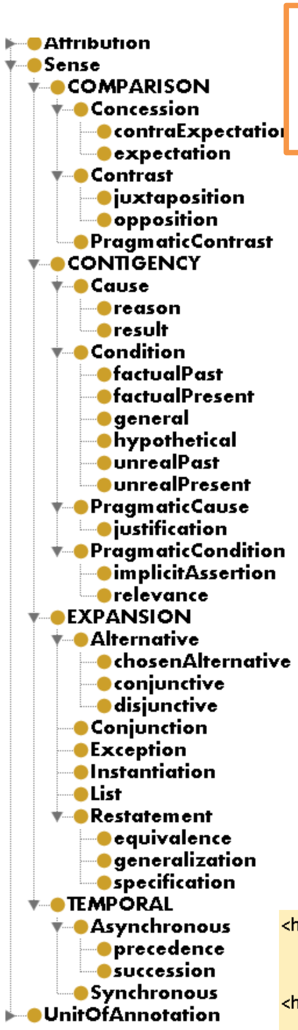
Annotation models and reference definitions for RST, PDTB, etc.



(Linguistic) Linked Open Data
GOLD
OWL/DL
CLARIN CCR
RDF
community-maintained vocabularies
…
Ontologies of Linguistic Annotation (OLiA)
outgoing links
Reference Model
OWL2/DL
shared (internal) vocabulary
links OWL2/DL subsumption axioms
RST Annotation Model
OWL2/DL
PDTB Annotation Model
OWL2/DL
…
resource-specific vocabularies

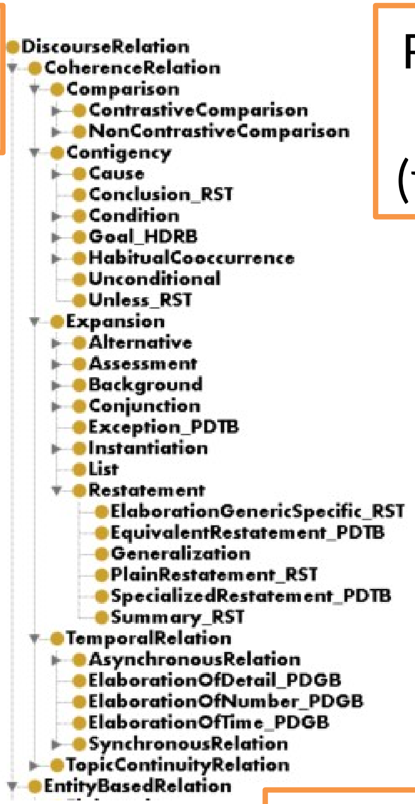
OLiA Discourse Extension

(Chiarcos@LREC-2014)



- Attribution
- Sense
  - COMPARISON
    - Concession
      - contraExpectatio
      - expectation
    - Contrast
      - juxtaposition
      - opposition
    - PragmaticContrast
  - CONTIGENCY
    - Cause
      - reason
      - result
    - Condition
      - factualPast
      - factualPresent
      - general
      - hypothetical
      - unrealPast
      - unrealPresent
    - PragmaticCause
      - justification
    - PragmaticCondition
      - implicitAssertion
      - relevance
  - EXPANSION
    - Alternative
      - chosenAlternative
      - conjunctive
      - disjunctive
    - Conjunction
    - Exception
    - Instantiation
    - List
    - Restatement
      - equivalence
      - generalization
      - specification
  - TEMPORAL
    - Asynchronous
      - precedence
      - succession
    - Synchronous
- UnitOfAnnotation

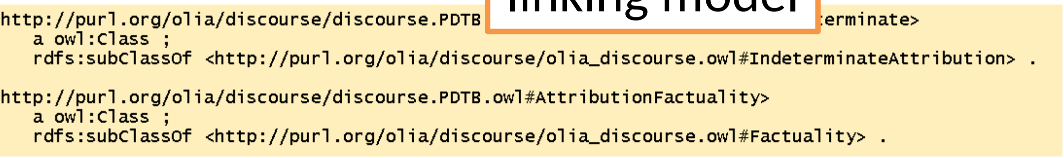
PDTB ontology



- DiscourseRelation
  - CoherenceRelation
    - Comparison
      - ContrastiveComparison
      - NonContrastiveComparison
    - Contigency
      - Cause
      - Conclusion_RST
      - Condition
      - Goal_HDRB
      - HabitualCooccurrence
      - Unconditional
      - Unless_RST
    - Expansion
      - Alternative
      - Assessment
      - Background
      - Conjunction
      - Exception_PDTB
      - Instantiation
      - List
      - Restatement
        - ElaborationGenericSpecific_RST
        - EquivalentRestatement_PDTB
        - Generalization
        - PlainRestatement_RST
        - SpecializedRestatement_PDTB
        - Summary_RST
    - TemporalRelation
      - AsynchronousRelation
      - ElaborationOfDetail_PDGB
      - ElaborationOfNumber_PDGB
      - ElaborationOfTime_PDGB
      - SynchronousRelation
    - TopicContinuityRelation
  - EntityBasedRelation

Reference Model (fragment)

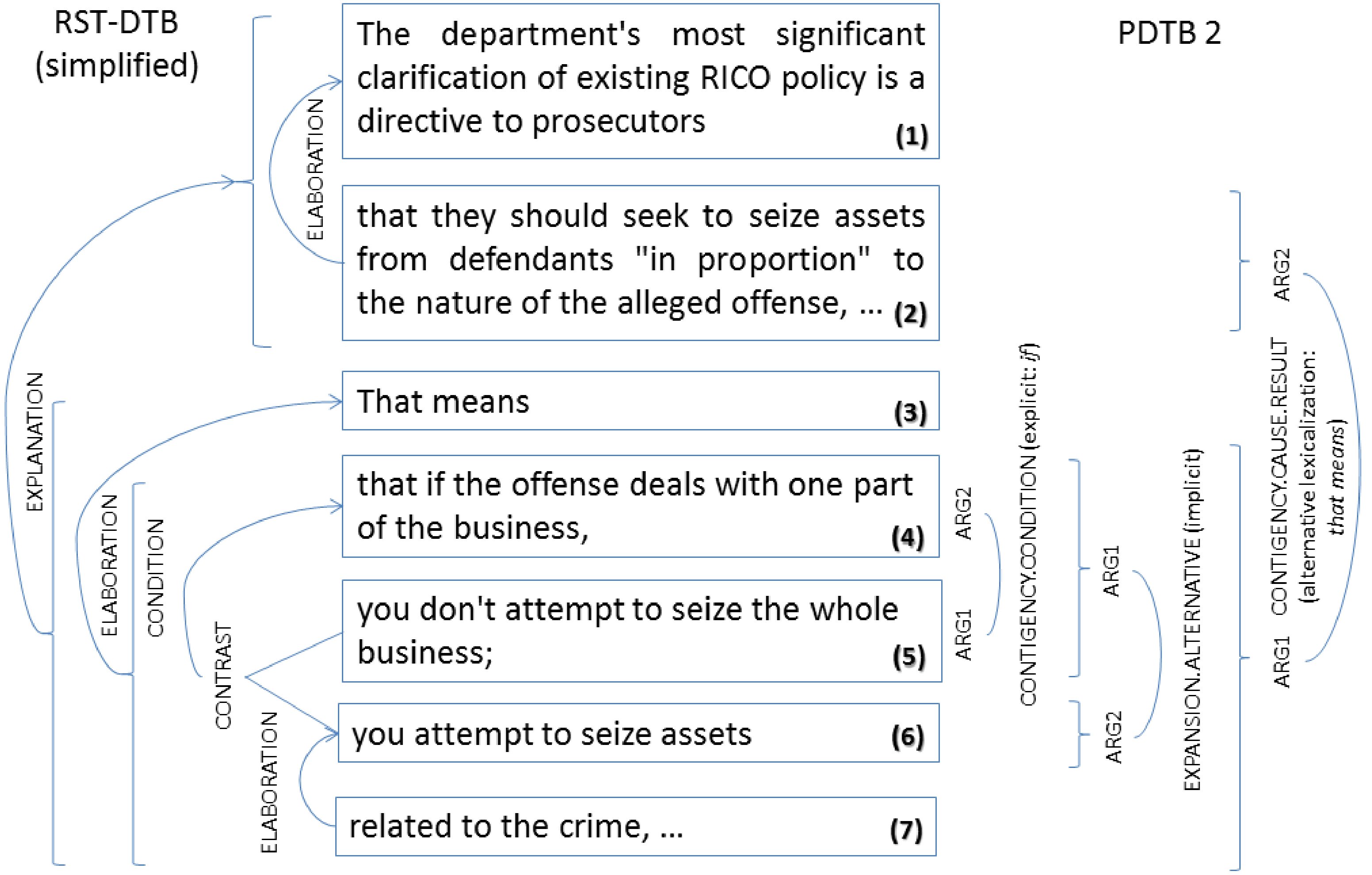
PDTB linking model



```
<http://purl.org/olia/discourse/discourse.PDTB                          erminate>
    a owl:Class ;
    rdfs:subClassOf <http://purl.org/olia/discourse/olia_discourse.owl#IndeterminateAttribution> .

<http://purl.org/olia/discourse/discourse.PDTB.owl#AttributionFactuality>
    a owl:Class ;
    rdfs:subClassOf <http://purl.org/olia/discourse/olia_discourse.owl#Factuality> .
```

- top-level structure based on PDTB
  - enriched to cover RST and other corpora
  - linked with meta-vocabularies (CCR, ISO SemAF Core DRs)

Comparing Across Frameworks



RST-DTB (simplified)

PDTB 2

The department's most significant clarification of existing RICO policy is a directive to prosecutors (1)

that they should seek to seize assets from defendants "in proportion" to the nature of the alleged offense, ... (2)

That means (3)

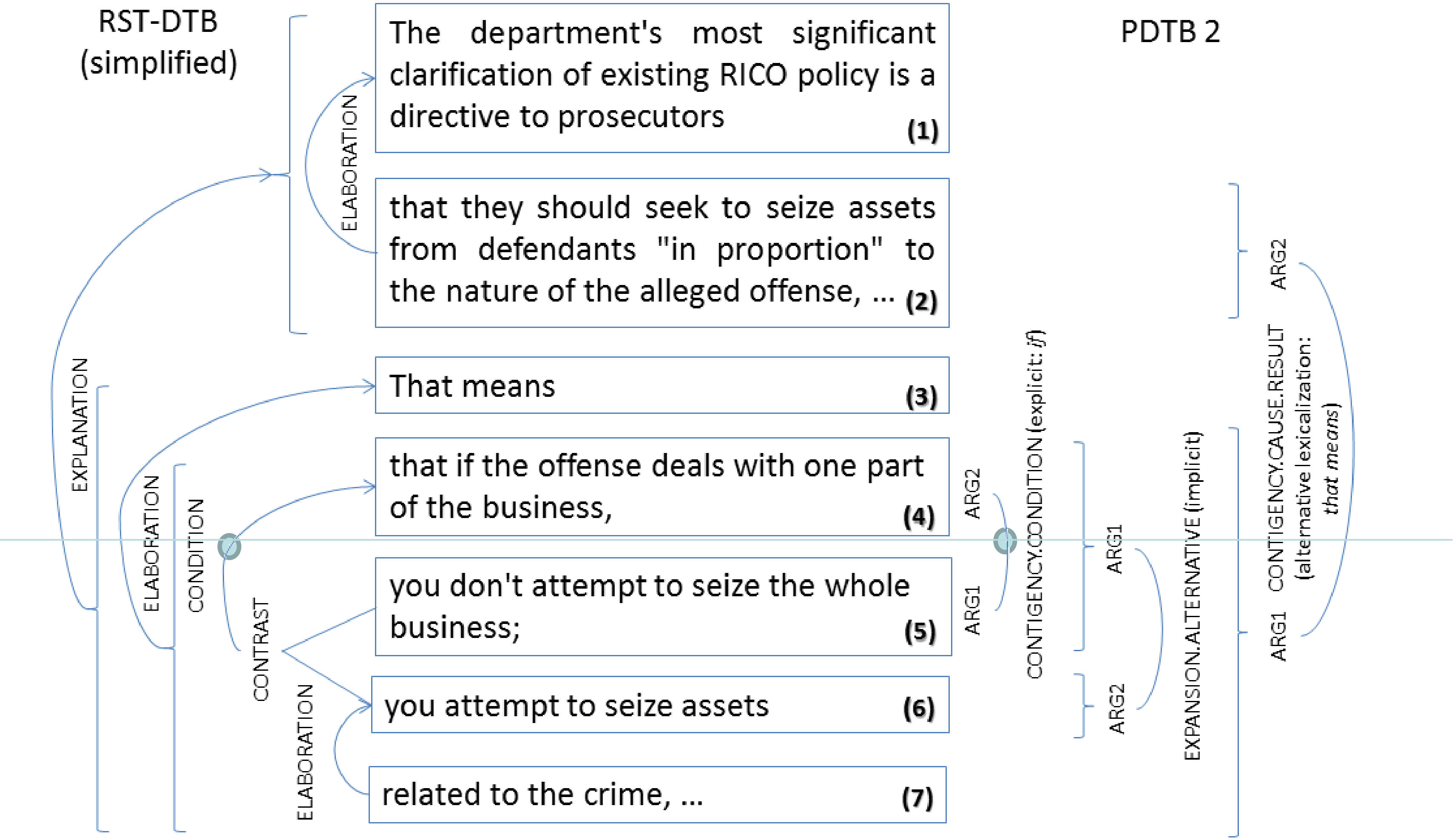
that if the offense deals with one part of the business, (4)

you don't attempt to seize the whole business; (5)

you attempt to seize assets (6)

related to the crime, ... (7)

Comparing Across Frameworks: What links (4) and (5)?



RST-DTB (simplified)

PDTB 2

The department's most significant clarification of existing RICO policy is a directive to prosecutors (1)

that they should seek to seize assets from defendants "in proportion" to the nature of the alleged offense, … (2)

That means (3)

that if the offense deals with one part of the business, (4)

you don't attempt to seize the whole business; (5)

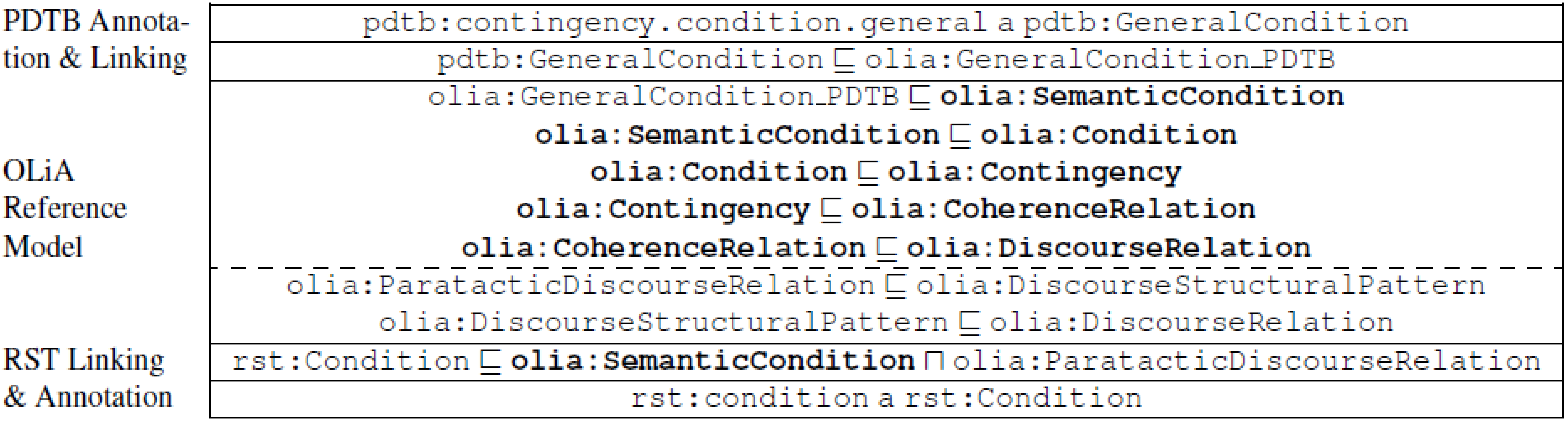
you attempt to seize assets (6)

related to the crime, … (7)

CONDITION

CONTINGENCY.CONDITION.GENERAL

Comparing Across Frameworks: What links (4) and (5)?



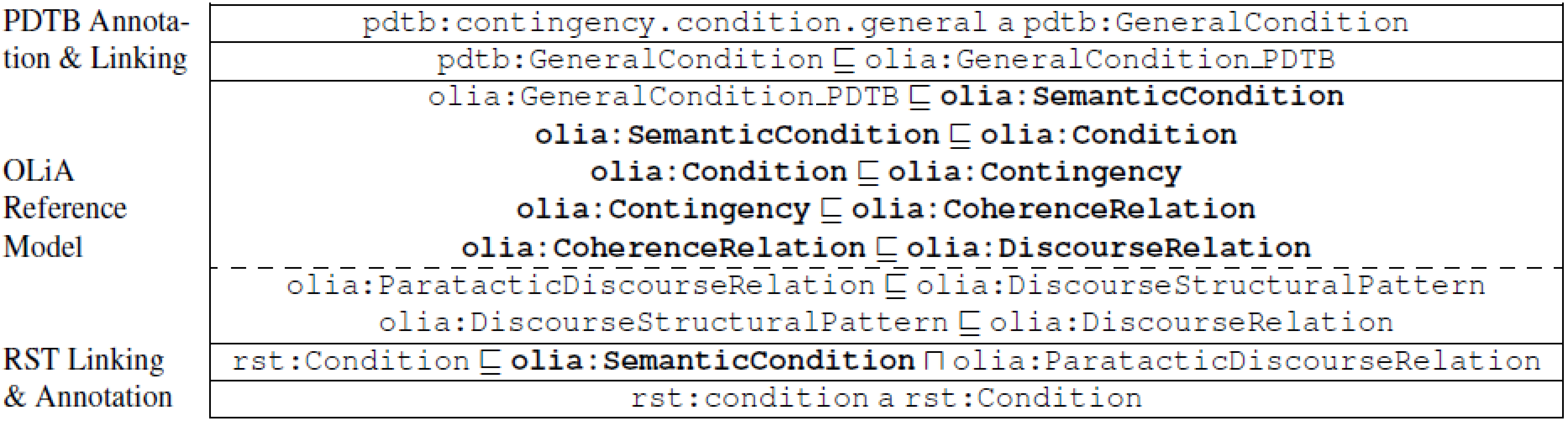
PDTB Annotation & Linking	pdtb:contingency.condition.general a pdtb:GeneralCondition
	pdtb:GeneralCondition ⊑ olia:GeneralCondition_PDTB
OLiA Reference Model	olia:GeneralCondition_PDTB ⊑ **olia:SemanticCondition**
	olia:SemanticCondition ⊑ **olia:Condition**
	olia:Condition ⊑ **olia:Contingency**
	olia:Contingency ⊑ **olia:CoherenceRelation**
	olia:CoherenceRelation ⊑ **olia:DiscourseRelation**
	olia:ParatacticDiscourseRelation ⊑ olia:DiscourseStructuralPattern
	olia:DiscourseStructuralPattern ⊑ olia:DiscourseRelation
RST Linking & Annotation	rst:Condition ⊑ **olia:SemanticCondition** ⊓ olia:ParatacticDiscourseRelation
	rst:condition a rst:Condition

- The annotations are not quite alike, but for the utterances under consideration, they agree on the features **in bold**

CONDITION

CONTINGENCY.CONDITION.GENERAL

Comparing Across Frameworks: What links (4) and (5)?



PDTB Annotation & Linking	pdtb:contingency.condition.general a pdtb:GeneralCondition
	pdtb:GeneralCondition ⊑ olia:GeneralCondition_PDTB
OLiA Reference Model	olia:GeneralCondition_PDTB ⊑ **olia:SemanticCondition**
	olia:SemanticCondition ⊑ **olia:Condition**
	olia:Condition ⊑ **olia:Contingency**
	olia:Contingency ⊑ **olia:CoherenceRelation**
	olia:CoherenceRelation ⊑ **olia:DiscourseRelation**
	olia:ParatacticDiscourseRelation ⊑ olia:DiscourseStructuralPattern
	olia:DiscourseStructuralPattern ⊑ olia:DiscourseRelation
RST Linking & Annotation	rst:Condition ⊑ **olia:SemanticCondition** ⊓ olia:ParatacticDiscourseRelation
	rst:condition a rst:Condition

- The annotations are not quite alike, but for the utterances under consideration, they agree on the features **in bold**
- We can now compare across frameworks
  - and we can derive a mapping between them
    - SPARQL ⇒ the shortest path of rdf:type (*a*) and rdfs:subClassOf (⊑) statements

Linking Discourse Marker Inventories

From Discourse Marker Inventories to OntoLex (RDF)

(Chiarcos & Ionov@LDK-2021)

Discourse Marker Inventories

- The most elementary step in discourse annotation is to identify discourse markers and their respective relations
- For a considerable number of languages, discourse marker inventories have been developed
  - ❑ to facilitate discourse parsing and downstream tasks
  - ❑ map discourse markers to (possible) discourse relations
- Different formats, different theoretical frameworks
  - ⇒ Our contribution: consolidation and integration

Related Research: TextLink

- TextLink
  - Cost Action *Structuring Discourse in Multilingual Europe* (2014-2018)
  - multilingual discourse marker inventories
    - (mostly) providing PDTB relations as senses
    - (mostly) following a consistent XML format (DimLex, Stede & Umbach 1998)
  - http://connective-lex.info/

Beyond TextLink: We aimed to improve ...

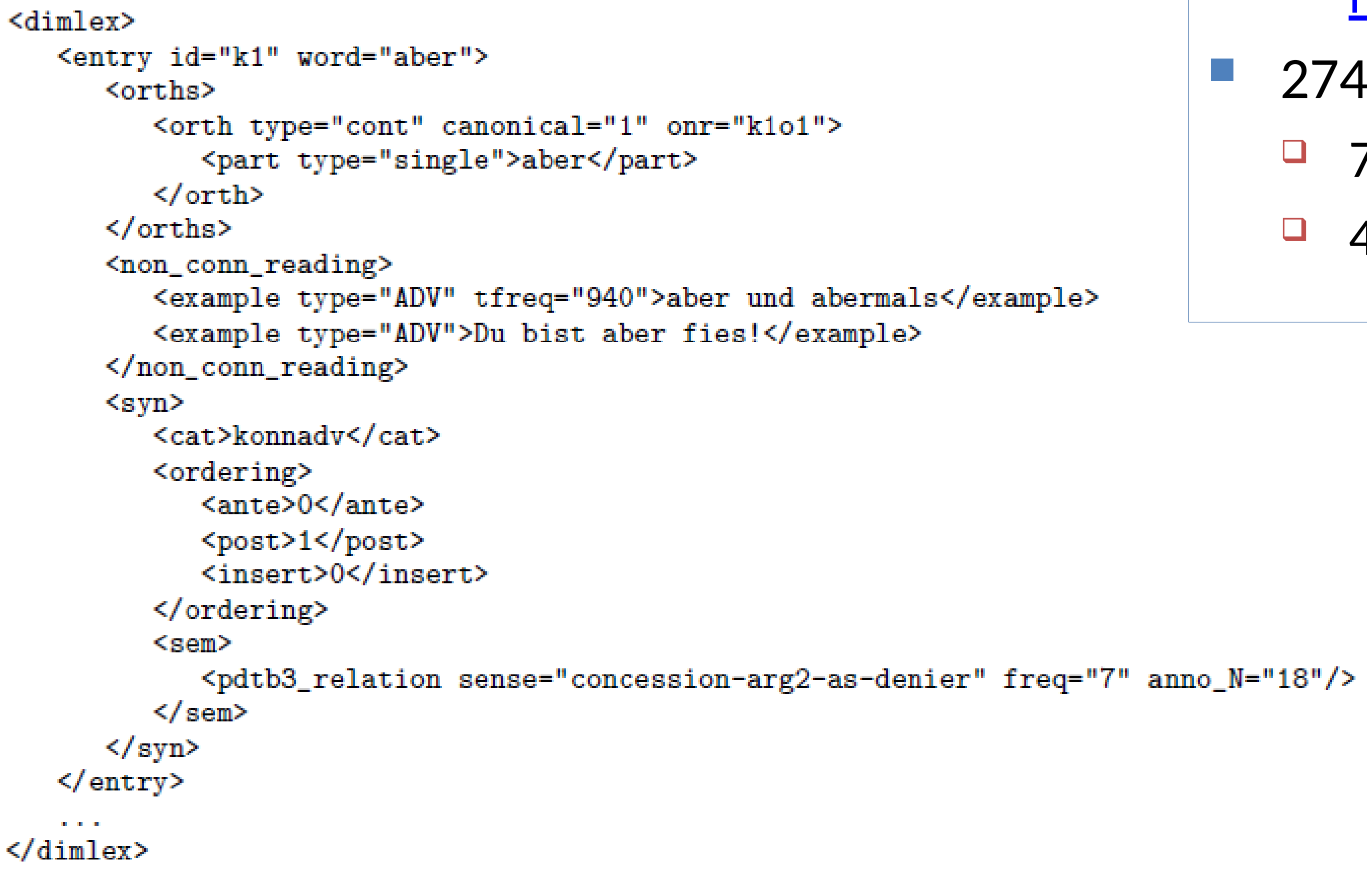
- coverage
  - extend the range of languages and resources
- semantics
  - provide machine-readable semantics
  - preserve the original sense definitions
- usability
  - enable cross-framework comparison and search
  - link and query across languages

Beyond TextLink

- An RDF edition of TextLink and other discourse marker inventories
  - using the RDF vocabulary OntoLex for machine-readable dictionaries
- Maintain original sense definitions (discourse relations)
  - link with OLiA annotation models (PDTB, RST, …)
- Map flexibly between frameworks
  - traverse with SPARQL
    - PDTB -> OLiA reference model -> RST (or ISO SemAF, CCR, etc.)

Example: German DimLex-XML

- Format: DimLex-XML

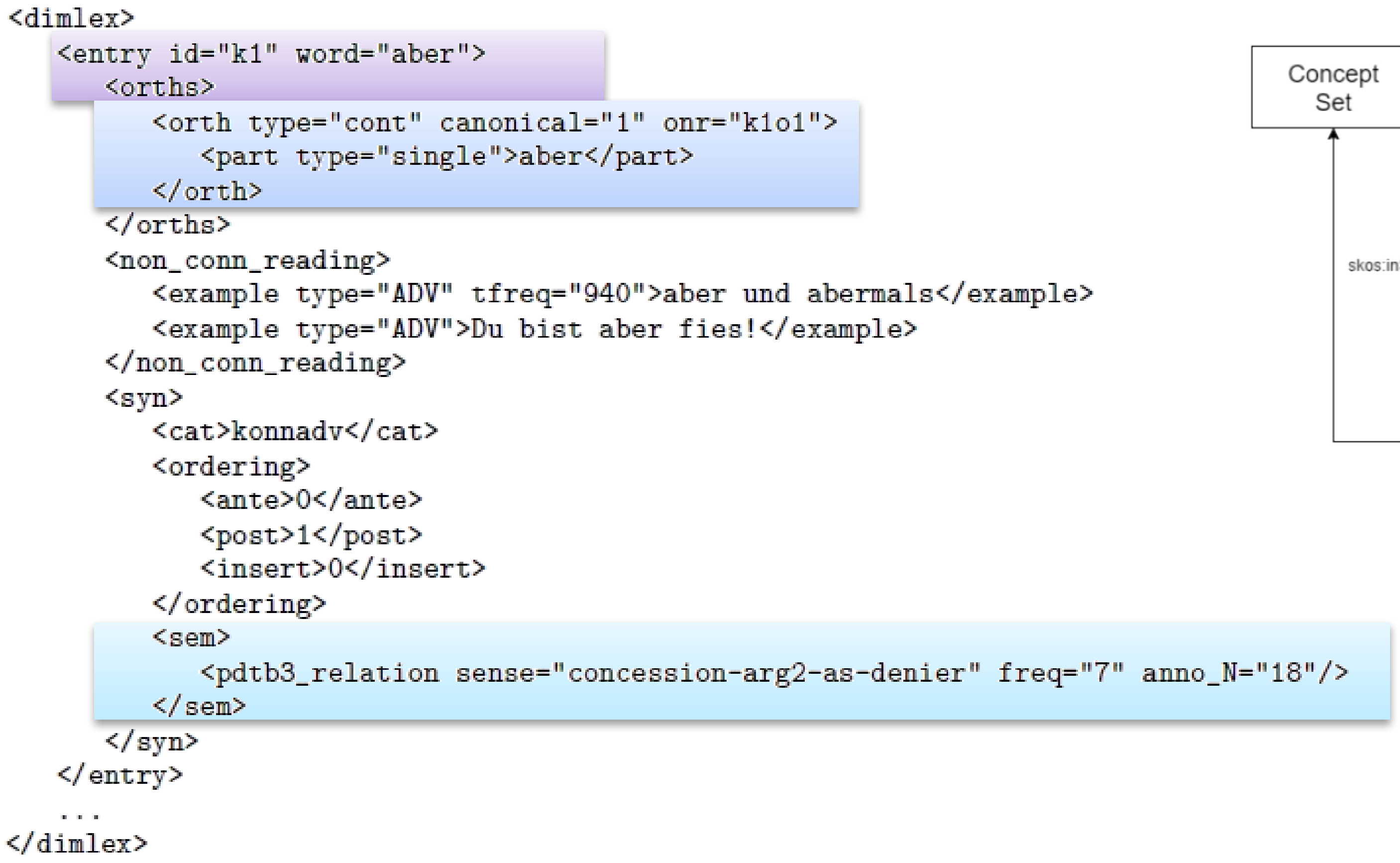


```
<dimlex>
   <entry id="k1" word="aber">
      <orths>
         <orth type="cont" canonical="1" onr="k1o1">
            <part type="single">aber</part>
         </orth>
      </orths>
      <non_conn_reading>
         <example type="ADV" tfreq="940">aber und abermals</example>
         <example type="ADV">Du bist aber fies!</example>
      </non_conn_reading>
      <syn>
         <cat>konnadv</cat>
         <ordering>
            <ante>0</ante>
            <post>1</post>
            <insert>0</insert>
         </ordering>
         <sem>
            <pdtb3_relation sense="concession-arg2-as-denier" freq="7" anno_N="18"/>
         </sem>
      </syn>
   </entry>
   ...
</dimlex>
```

- Scheffler & Stede (2016)
  - CC-BY-NC-SA 4.0
  - https://github.com/discourse-lab/dimlex
- 274 entries
  - 763 forms
  - 432 sense links (28 PDTB 3.0 relations)

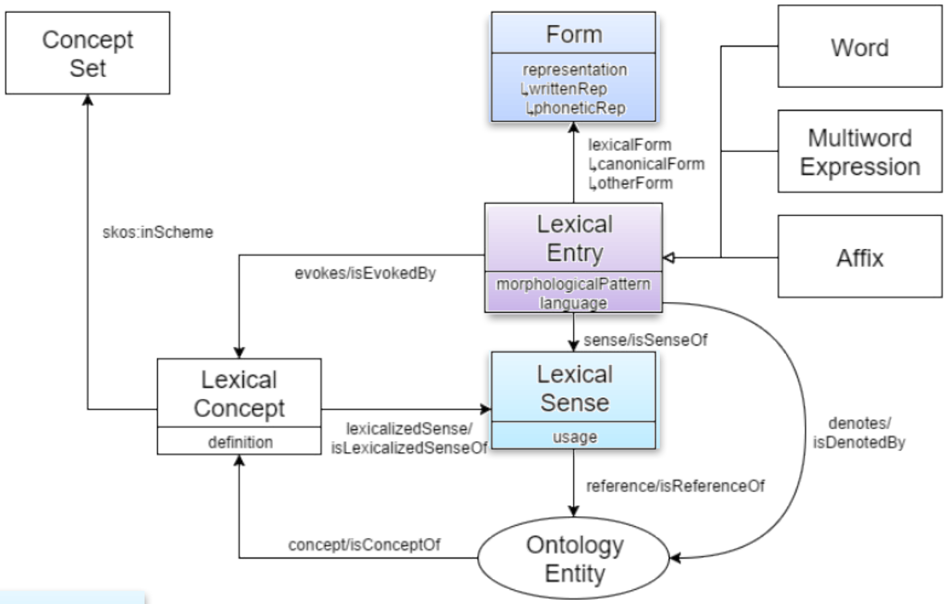
Example: German DimLex → RDF (OntoLex)

- Format: DimLex-XML



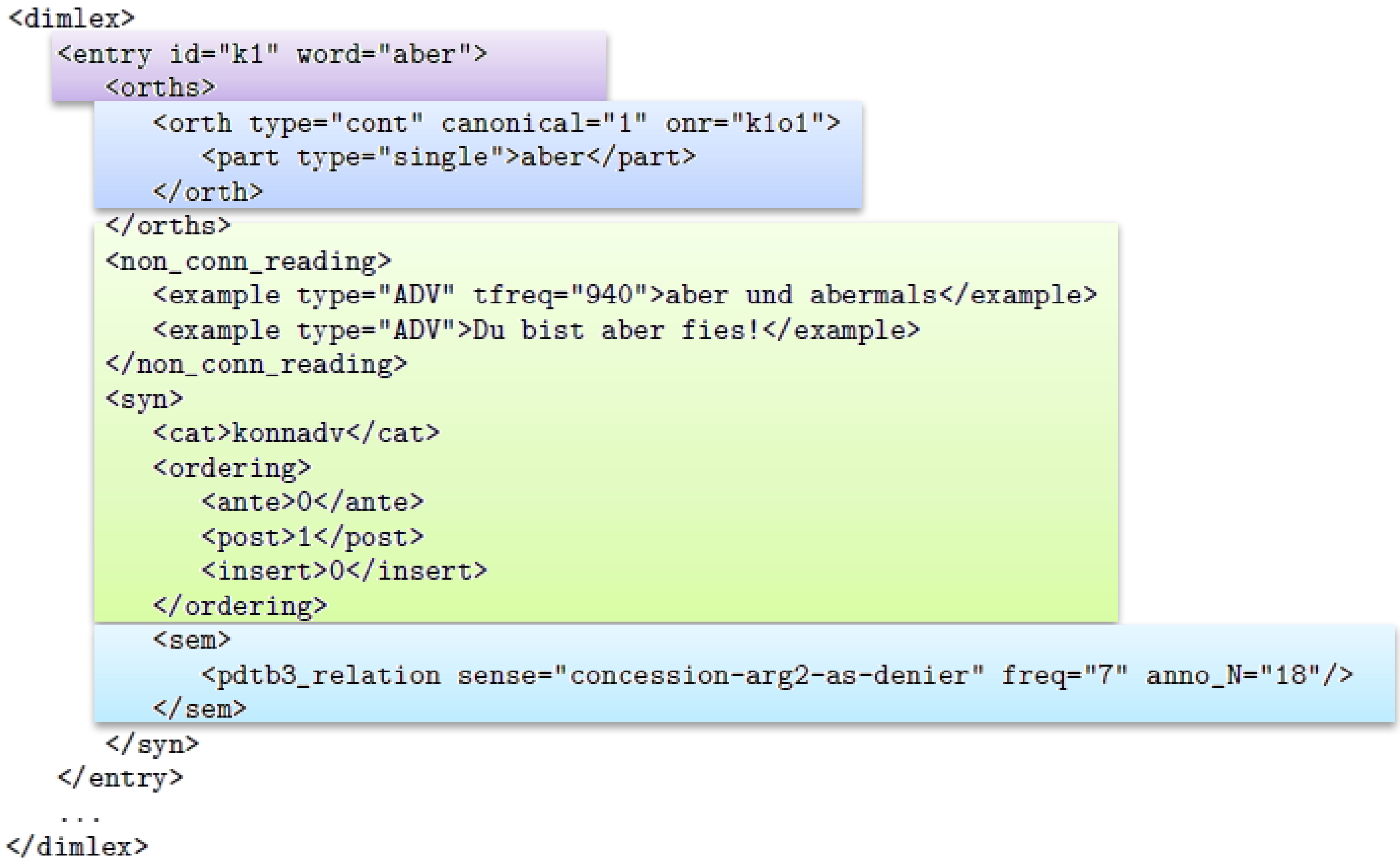
```
<dimlex>
   <entry id="k1" word="aber">
      <orths>
         <orth type="cont" canonical="1" onr="k1o1">
            <part type="single">aber</part>
         </orth>
      </orths>
      <non_conn_reading>
         <example type="ADV" tfreq="940">aber und abermals</example>
         <example type="ADV">Du bist aber fies!</example>
      </non_conn_reading>
      <syn>
         <cat>konnadv</cat>
         <ordering>
            <ante>0</ante>
            <post>1</post>
            <insert>0</insert>
         </ordering>
         <sem>
            <pdtb3_relation sense="concession-arg2-as-denier" freq="7" anno_N="18"/>
         </sem>
      </syn>
   </entry>
   ...
</dimlex>
```

OntoLex vocabulary



Example: German DimLex → RDF (OntoLex)

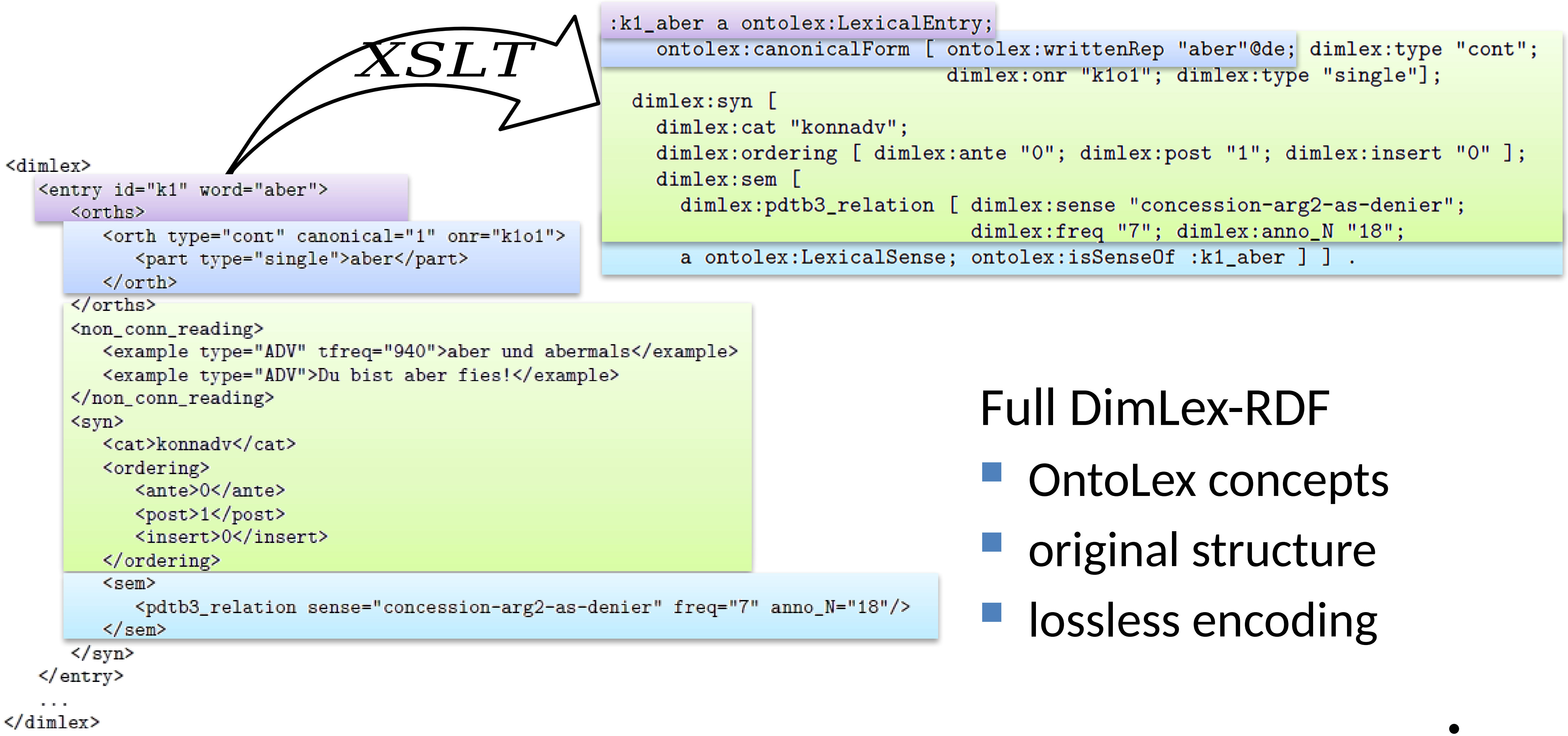
- Format: DimLex-XML OntoLex + custom properties (namespace *dimlex:*)



```
<dimlex>
   <entry id="k1" word="aber">
      <orths>
         <orth type="cont" canonical="1" onr="k1o1">
            <part type="single">aber</part>
         </orth>
      </orths>
      <non_conn_reading>
         <example type="ADV" tfreq="940">aber und abermals</example>
         <example type="ADV">Du bist aber fies!</example>
      </non_conn_reading>
      <syn>
         <cat>konnadv</cat>
         <ordering>
            <ante>0</ante>
            <post>1</post>
            <insert>0</insert>
         </ordering>
         <sem>
            <pdtb3_relation sense="concession-arg2-as-denier" freq="7" anno_N="18"/>
         </sem>
      </syn>
   </entry>
   ...
</dimlex>
```

custom properties correspond 1:1 to XML elements and attributes
⇒ different dialects represented in a lossless fashion

Example: German DimLex RDF



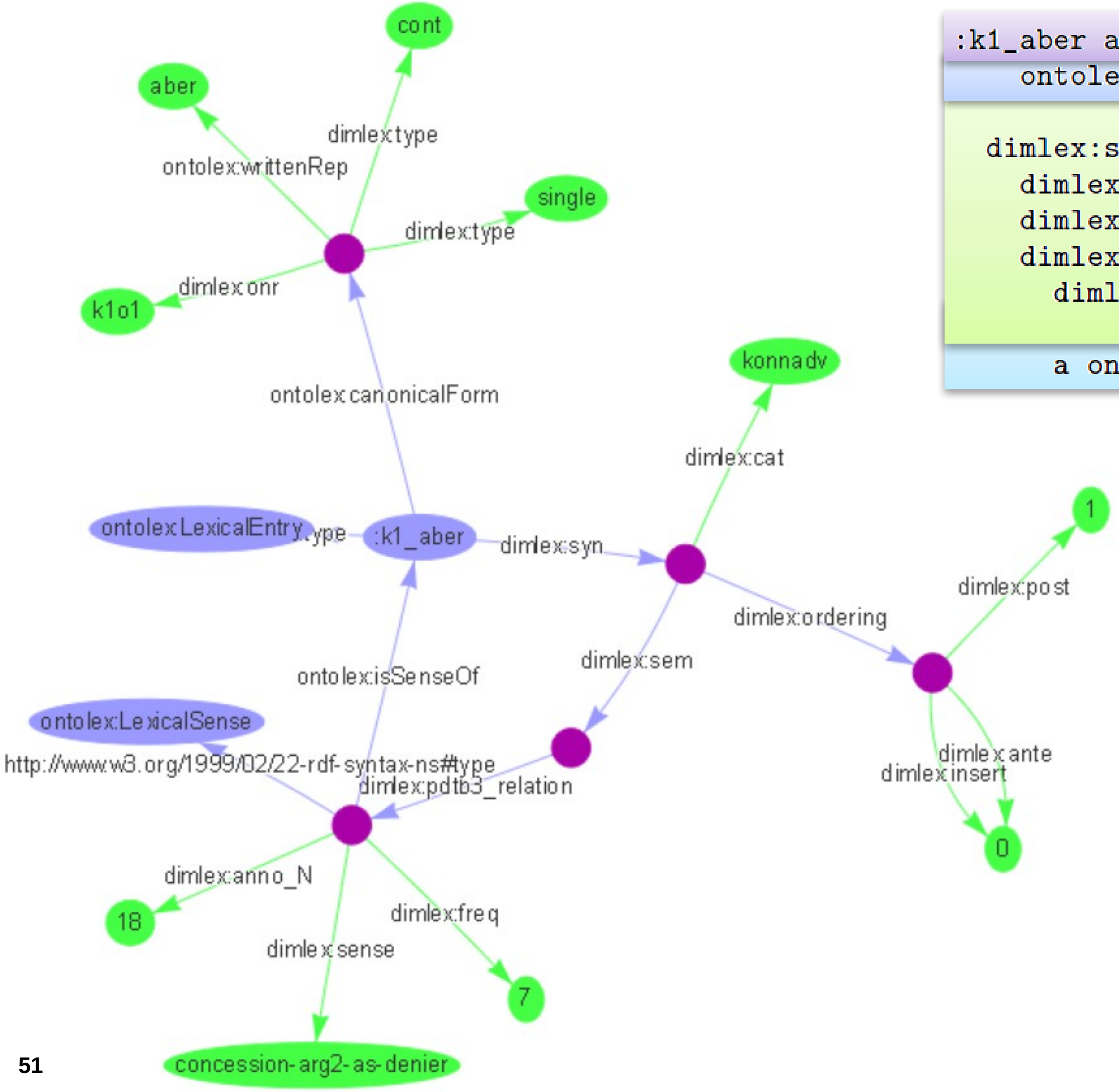
```
:k1_aber a ontolex:LexicalEntry;
  ontolex:canonicalForm [ ontolex:writtenRep "aber"@de; dimlex:type "cont";
                          dimlex:onr "k1o1"; dimlex:type "single"];
 dimlex:syn [
  dimlex:cat "konnadv";
  dimlex:ordering [ dimlex:ante "0"; dimlex:post "1"; dimlex:insert "0" ];
  dimlex:sem [
   dimlex:pdtb3_relation [ dimlex:sense "concession-arg2-as-denier";
                           dimlex:freq "7"; dimlex:anno_N "18";
   a ontolex:LexicalSense; ontolex:isSenseOf :k1_aber ] ] .
```

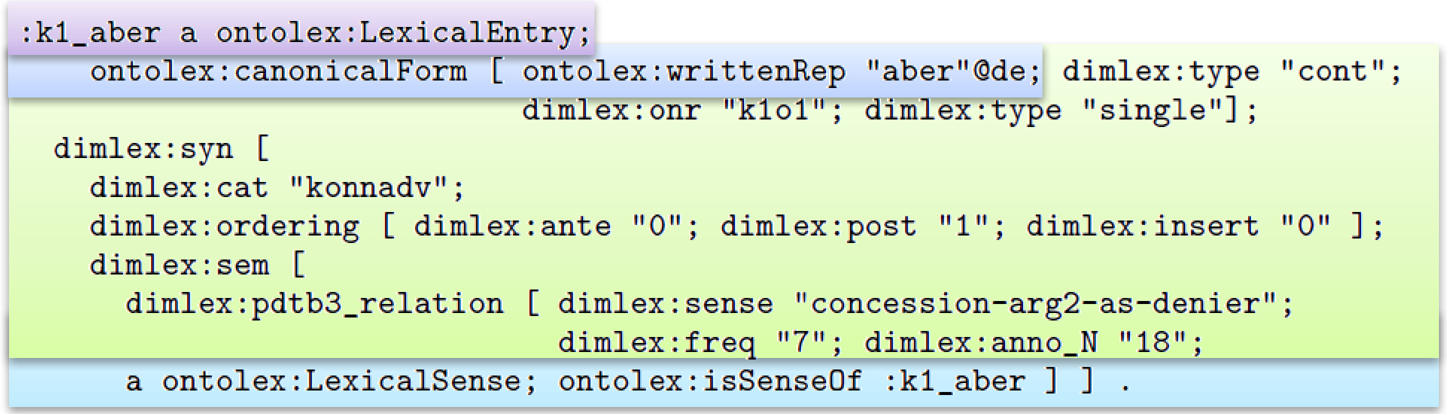
```
<dimlex>
   <entry id="k1" word="aber">
      <orths>
         <orth type="cont" canonical="1" onr="k1o1">
            <part type="single">aber</part>
         </orth>
      </orths>
      <non_conn_reading>
         <example type="ADV" tfreq="940">aber und abermals</example>
         <example type="ADV">Du bist aber fies!</example>
      </non_conn_reading>
      <syn>
         <cat>konnadv</cat>
         <ordering>
            <ante>0</ante>
            <post>1</post>
            <insert>0</insert>
         </ordering>
         <sem>
            <pdtb3_relation sense="concession-arg2-as-denier" freq="7" anno_N="18"/>
         </sem>
      </syn>
   </entry>
   ...
</dimlex>
```

Full DimLex-RDF

- OntoLex concepts
- original structure
- lossless encoding

Example: German DimLex RDF





```
:k1_aber a ontolex:LexicalEntry;
    ontolex:canonicalForm [ ontolex:writtenRep "aber"@de; dimlex:type "cont";
                            dimlex:onr "k1o1"; dimlex:type "single"];
  dimlex:syn [
    dimlex:cat "konnadv";
    dimlex:ordering [ dimlex:ante "0"; dimlex:post "1"; dimlex:insert "0" ];
    dimlex:sem [
      dimlex:pdtb3_relation [ dimlex:sense "concession-arg2-as-denier";
                              dimlex:freq "7"; dimlex:anno_N "18";
      a ontolex:LexicalSense; ontolex:isSenseOf :k1_aber ] ] .
```

Full DimLex-RDF

- OntoLex concepts
- original structure
- lossless encoding

OLiA Linking

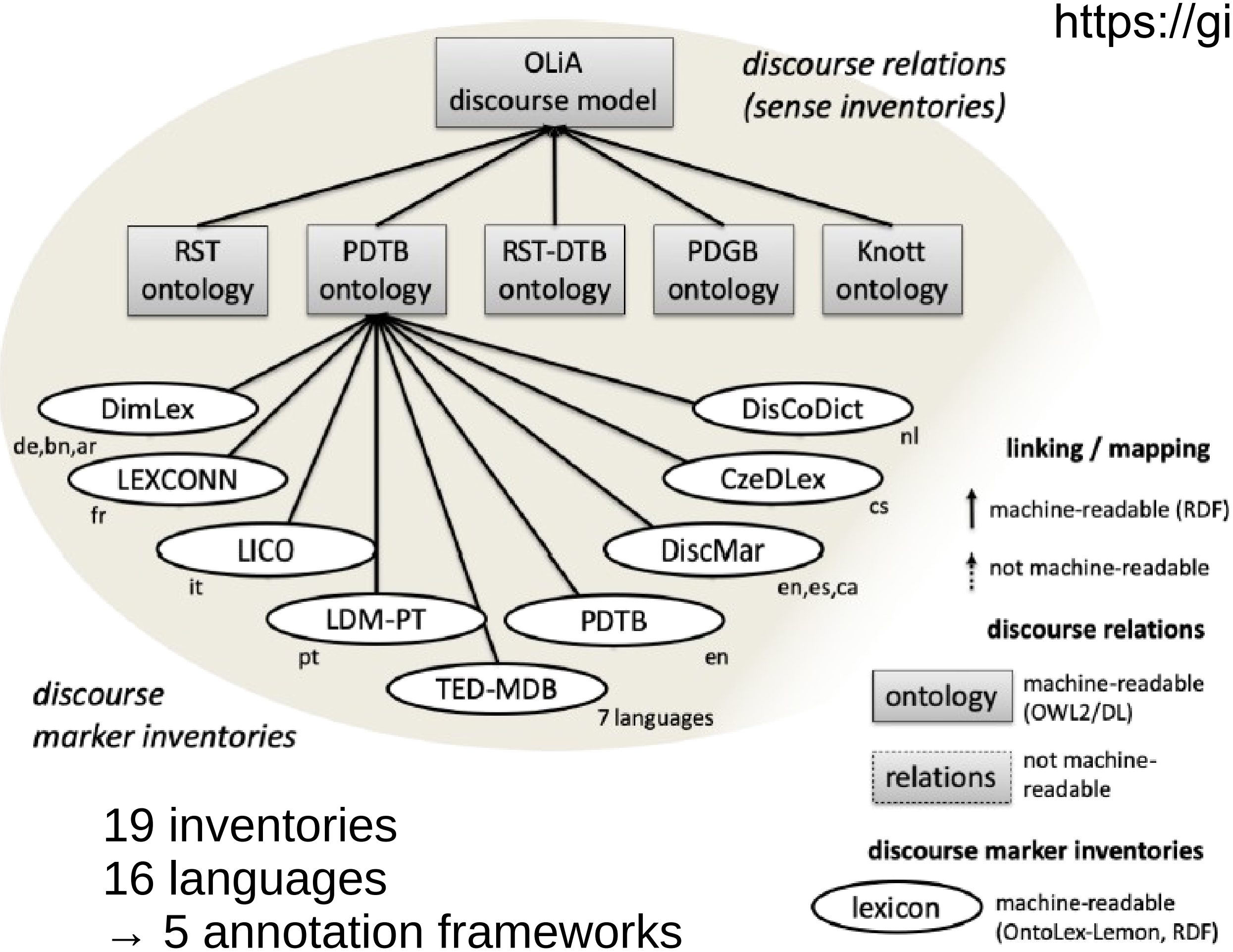
- Just use one trivial SPARQL Update

```
PREFIX dimlex: <https://github.com/discourse-lab/dimlex/blob/master/DimLex.dtd#>

LOAD <http://purl.org/olia/discourse/discourse.PDTB.owl>;

INSERT {
        ?dimlex_relation ontolex:reference ?pdtb_sense.
} WHERE {
        ?dimlex_relation dimlex:sense ?label.
        ?pdtb_sense (rdfs:label|skos:altLabel) ?sense_label.
        FILTER(lcase(?label)=lcase(?sense_label))
};
```

If a Dimlex relation has a *dimlex:sense* that matches the label of an OLiA PDTB relation, link them by *ontolex:reference*

Results: A Knowledge Graph for Discourse Markers

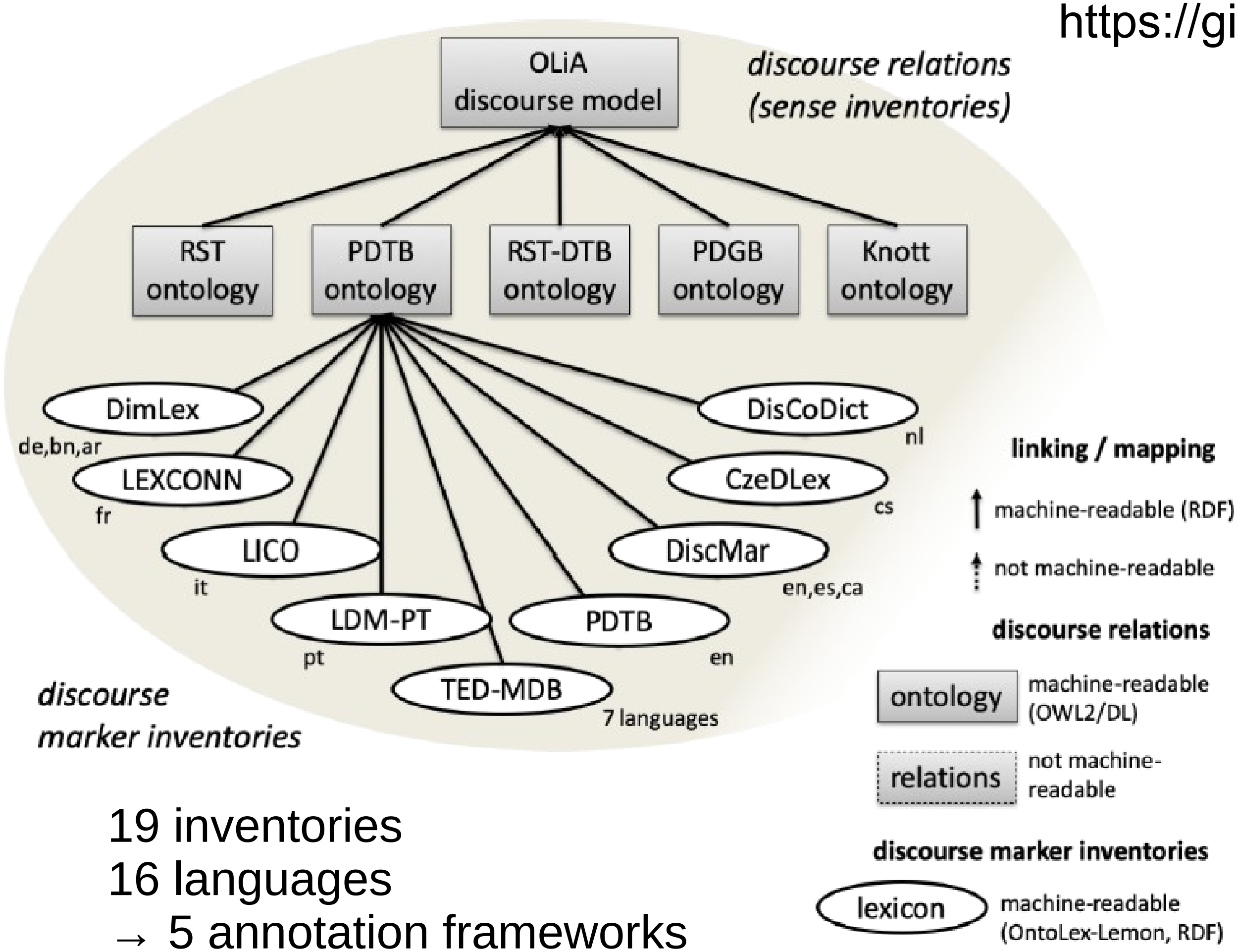
https://github.com/acoli-repo/rdf4discourse



19 inventories
16 languages
→ 5 annotation frameworks
→ 2 meta frameworks (CCR, ISO SemAF)

Results: A Knowledge Graph for Discourse Markers

https://github.com/acoli-repo/rdf4discourse



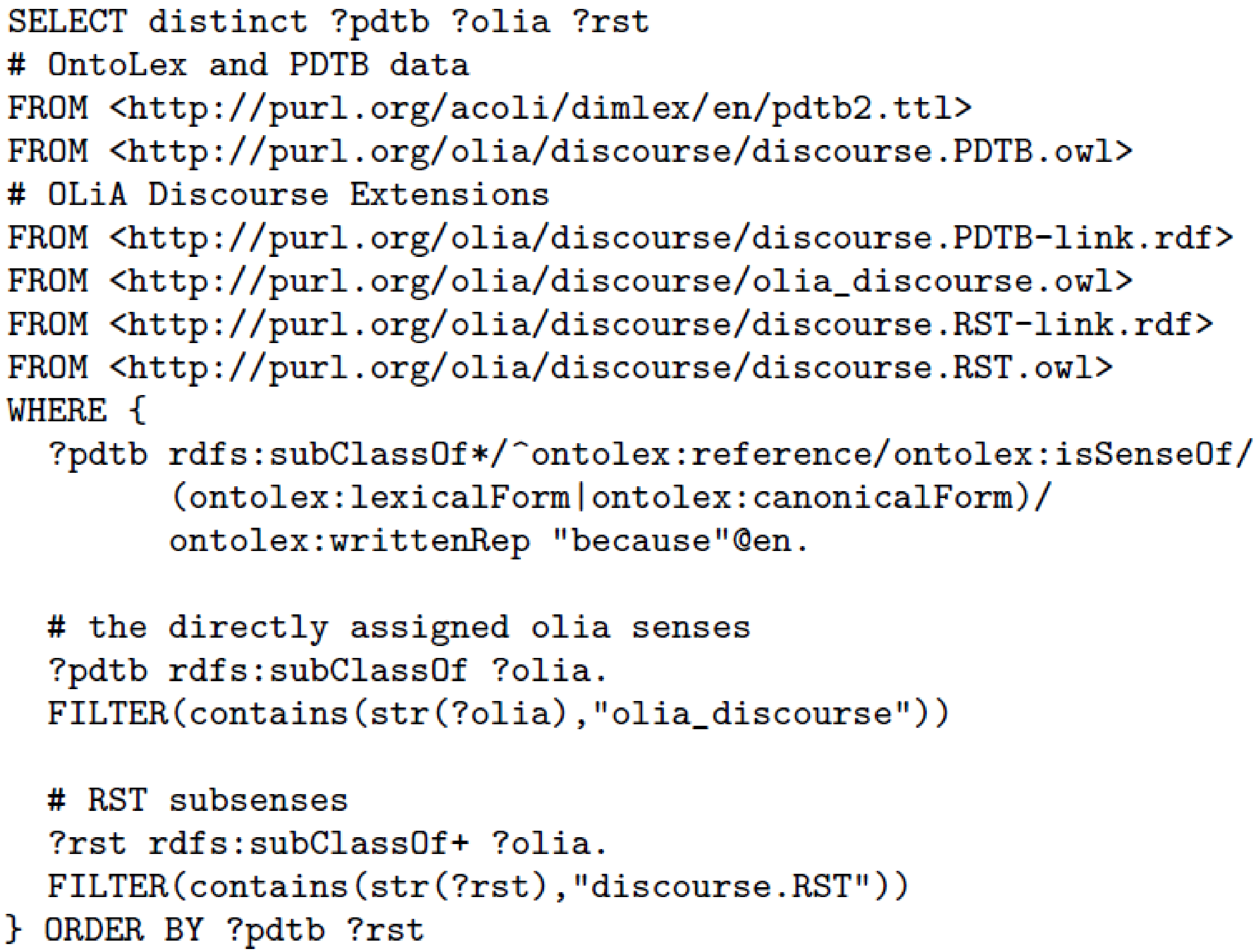
19 inventories
16 languages
→ 5 annotation frameworks
→ 2 meta frameworks (CCR, ISO SemAF)

Related research

- http://connective-lex.info/
- outcome of TextLink
- designed for human consumption
  - no machine-readable semantics
  - based on structured XML data
- PDTB senses only
- no crosslingual integration

USP: We can now much more easily retrieve information from the discourse marker inventories

Querying it with SPARQL



```
SELECT distinct ?pdtb ?olia ?rst
# OntoLex and PDTB data
FROM <http://purl.org/acoli/dimlex/en/pdtb2.ttl>
FROM <http://purl.org/olia/discourse/discourse.PDTB.owl>
# OLiA Discourse Extensions
FROM <http://purl.org/olia/discourse/discourse.PDTB-link.rdf>
FROM <http://purl.org/olia/discourse/olia_discourse.owl>
FROM <http://purl.org/olia/discourse/discourse.RST-link.rdf>
FROM <http://purl.org/olia/discourse/discourse.RST.owl>
WHERE {
  ?pdtb rdfs:subClassOf*/^ontolex:reference/ontolex:isSenseOf/
        (ontolex:lexicalForm|ontolex:canonicalForm)/
        ontolex:writtenRep "because"@en.

  # the directly assigned olia senses
  ?pdtb rdfs:subClassOf ?olia.
  FILTER(contains(str(?olia),"olia_discourse"))

  # RST subsenses
  ?rst rdfs:subClassOf+ ?olia.
  FILTER(contains(str(?rst),"discourse.RST"))
} ORDER BY ?pdtb ?rst
```

(1)
load the relevant knowledge graphs with FROM

Not a trivial query,

but not that hard to adapt

(2)
“because” → PDTB

(from discourse marker inventory)

(3)
PDTB → OLiA

(from OLiA PDTB model)

(4)
OLiA → RST

(from OLiA RST + reference model)

Given the English PDTB 2.0 discourse marker lexicon, retrieve all possible RST relations for “because”

Querying it with SPARQL

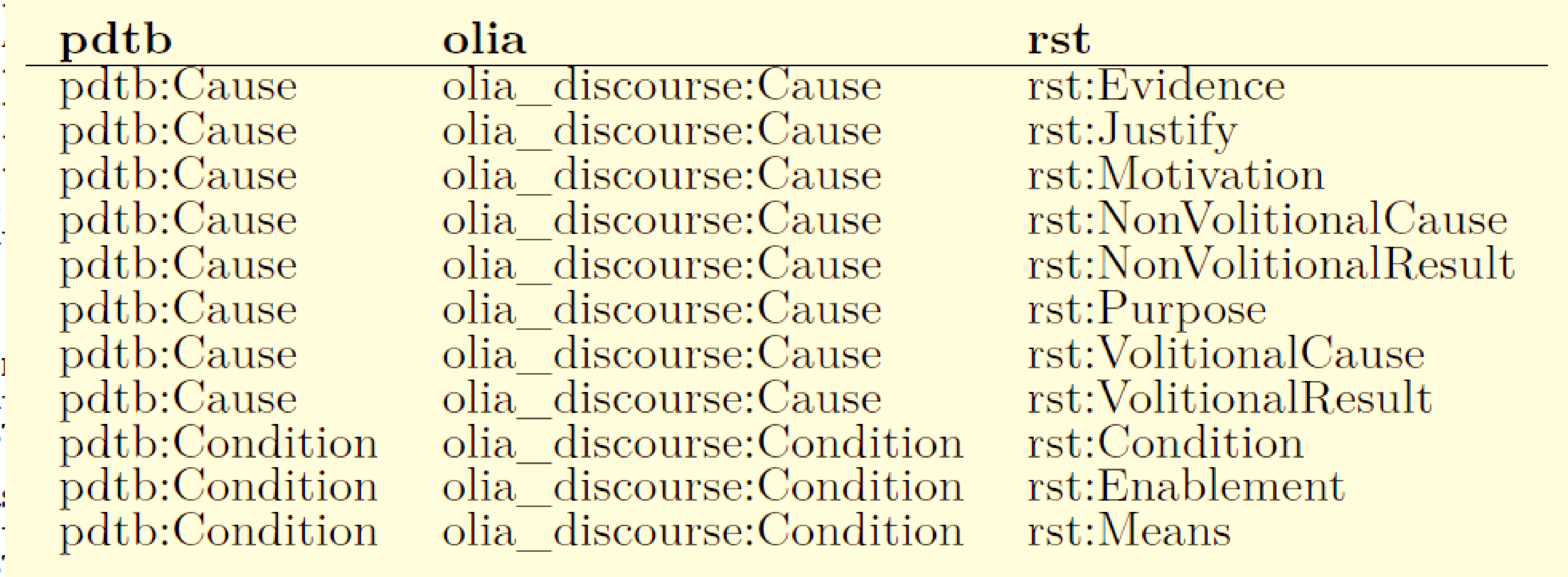
```
SELECT distinct ?pdtb ?olia ?rst
# OntoLex and PDTB data
FROM <http://purl.org/acoli/dimlex/en/pdtb2.ttl>
FROM
# OLi
FROM
FROM
FROM
FROM
WHERE
  ?pd



  # t
  ?pd
  FIL

  # R
  ?rs
  FIL
} ORDER BY ?pdtb ?rst
```

(1)

Not a trivial query,

hat hard

apt



pdtb	olia	rst
pdtb:Cause	olia_discourse:Cause	rst:Evidence
pdtb:Cause	olia_discourse:Cause	rst:Justify
pdtb:Cause	olia_discourse:Cause	rst:Motivation
pdtb:Cause	olia_discourse:Cause	rst:NonVolitionalCause
pdtb:Cause	olia_discourse:Cause	rst:NonVolitionalResult
pdtb:Cause	olia_discourse:Cause	rst:Purpose
pdtb:Cause	olia_discourse:Cause	rst:VolitionalCause
pdtb:Cause	olia_discourse:Cause	rst:VolitionalResult
pdtb:Condition	olia_discourse:Condition	rst:Condition
pdtb:Condition	olia_discourse:Condition	rst:Enablement
pdtb:Condition	olia_discourse:Condition	rst:Means

Given the English PDTB 2.0 discourse marker lexicon, retrieve all possible RST relations for “because”

Inducing Discourse Marker Inventories

from machine-readable dictionaries

(Chiarcos@LREC-2022)

Lexical Induction with the ACoLi Dictionary Graph

Discourse Marker Inventories + interlinked dictionaries → induction for other languages

Machine-readable dictionaries

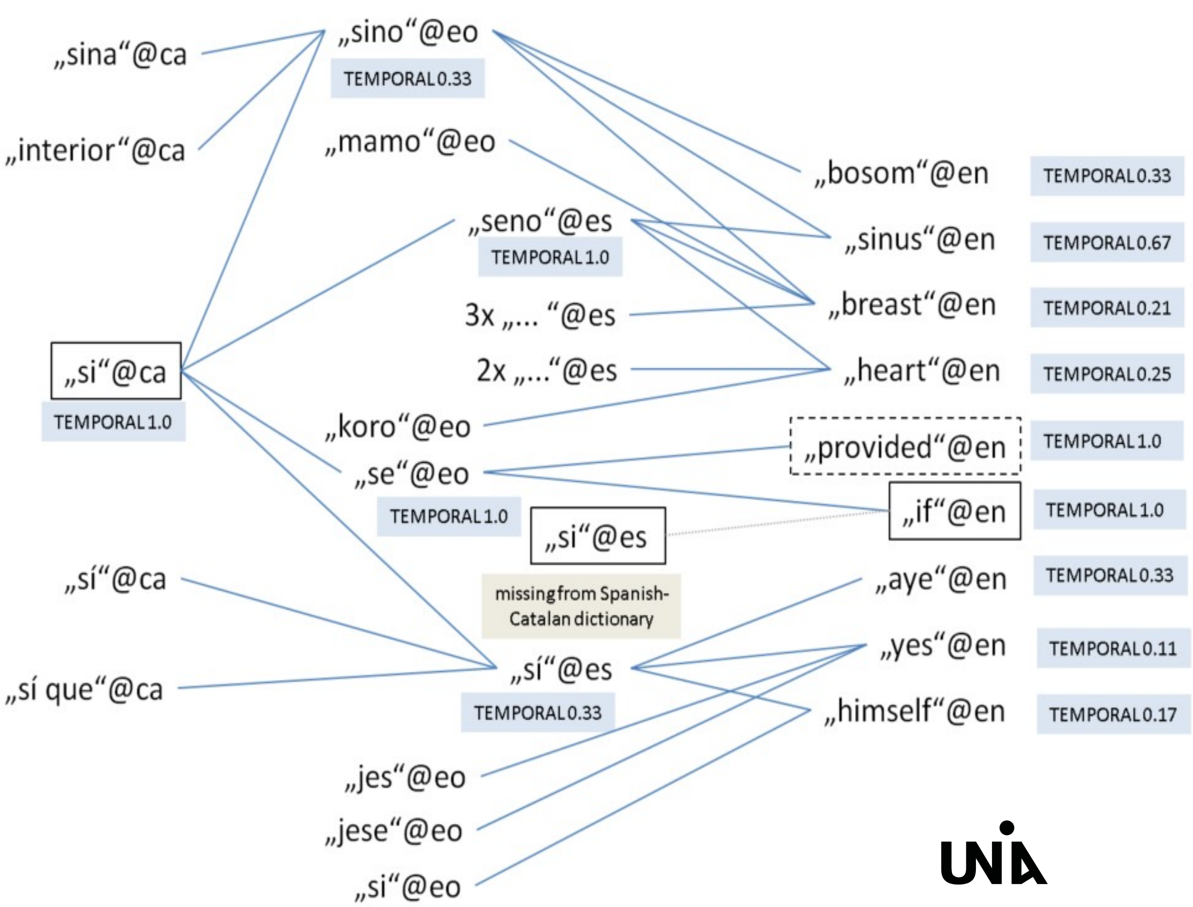
http://github.com/acoli-repo/acoli-dicts

- 430+ languages, 3000+ bilingual dictionaries (Chiarcos et al. 2020)
- RDF layer over PanLex, Apertium, FreeDict, MUSE, etc.
- **Data model**: OntoLex
- **Formats**: RDF (=> TSV, with SPARQL)
- **Selected subsets**
  - **Apertium** 53 dictionaries for MT, mostly Romance
  - **FreeDict** 145 dictionaries, heterogeneous
  - **MUSE** 108 dictionaries, machine-generated

Constrained Induction

- Operate over confidence scores for discourse relations
- Initialize word *w* with 1/(number of senses)
- Propagate relation score to word *v*:
  average over relation scores for translations (w. score)

- Constraints: (optionally) filter by
  min result score
  min pivots (translations)
  min pivot languages (of translations)
  max senses (top *k* relations, only)



Experimental Setup

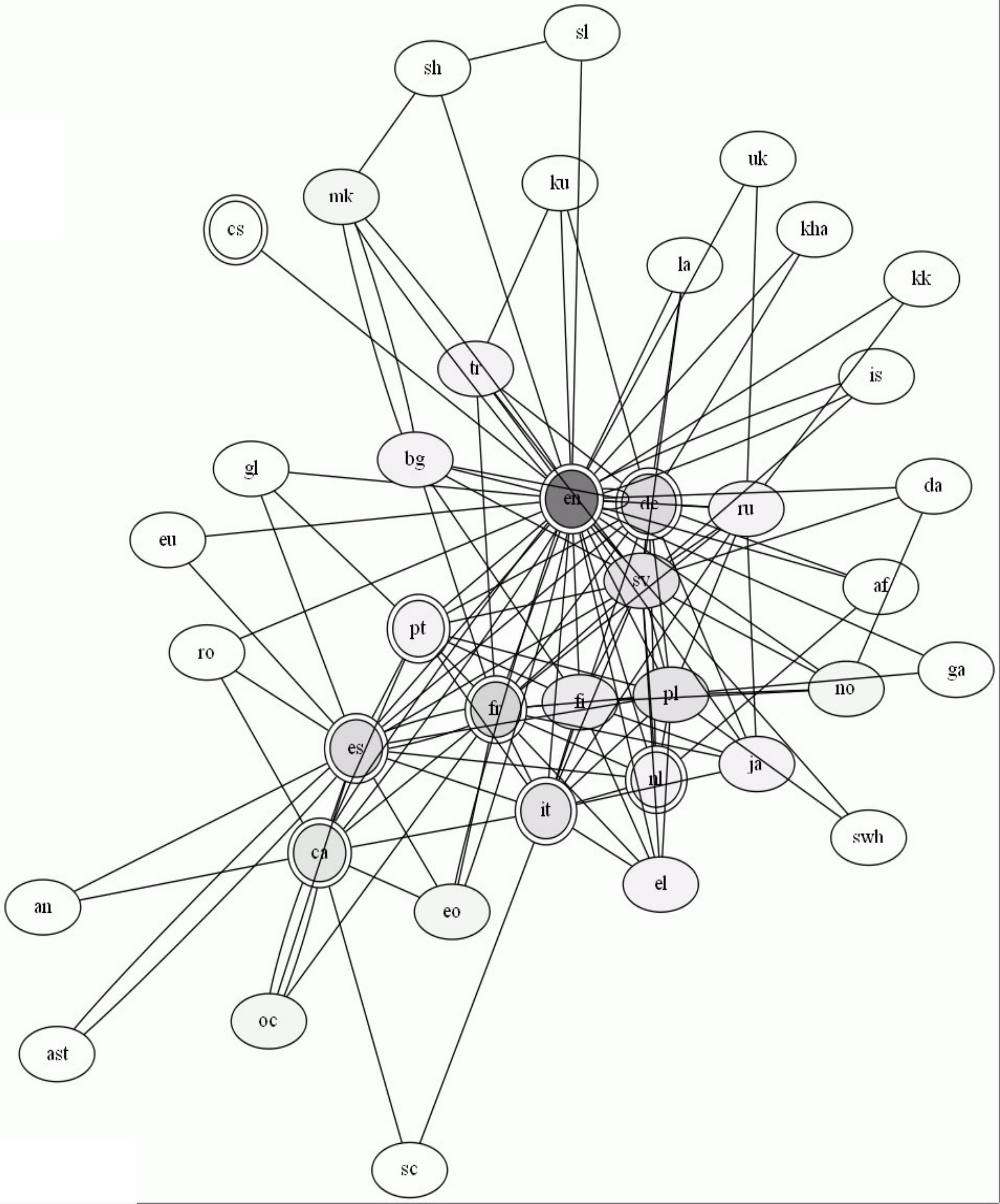
http://github.com/acoli-repo/rdf4discourse/lexical-induction

- 11 inventories, 9 languages
- mapped to PDTB and CCR
- evaluate prec, rec, f against target inventories

- Publish 10 induced inventories

(Bulgarian, Greek, Esperanto, Finnish, Japanese, Norwegian, Polish, Russian, Swedish and Turkish)

lang	language
◎	language with discourse marker inventory
╲	>= 1 dictionary

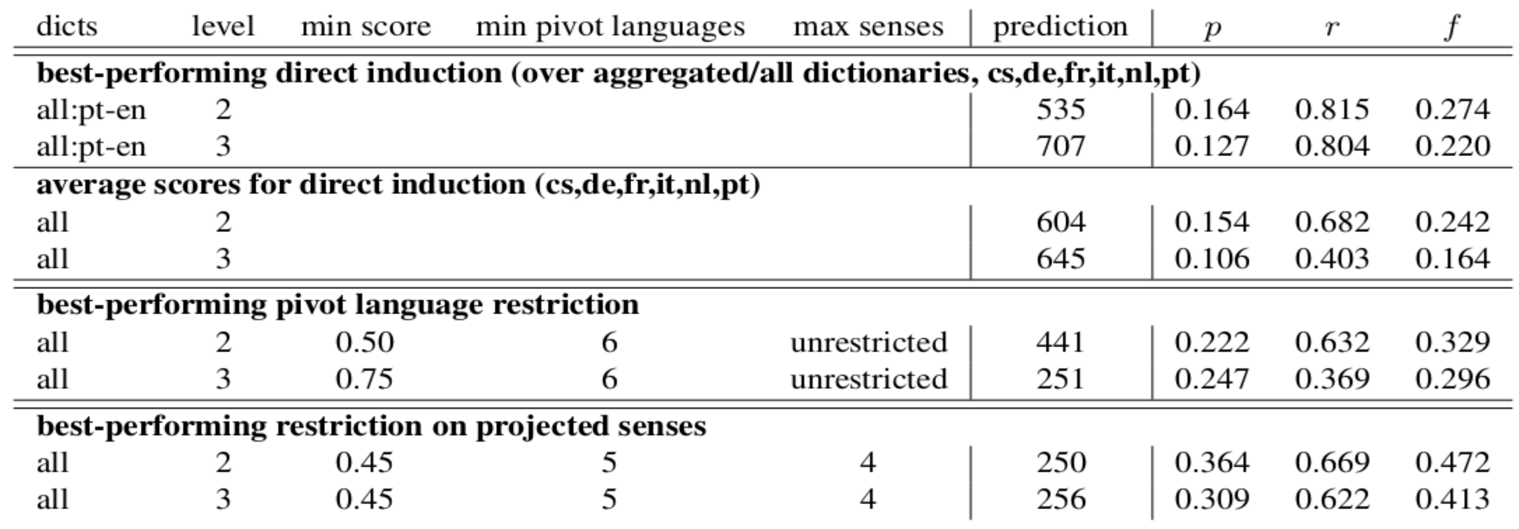


Lexical Induction with the ACoLi Dictionary Graph

Discourse Marker Inventories + interlinked dictionaries → induction for other languages

- Direct induction (e.g., from/to English) yields best results, but depends on dictionary quality (Apertium > FreeDict/MUSE)

- Constrained indirect induction is a feasible fallback-strategy



dicts	level	min score	min pivot languages	max senses	prediction	p	r	f
best-performing direct induction (over aggregated/all dictionaries, cs,de,fr,it,nl,pt)								
all:pt-en	2				535	0.164	0.815	0.274
all:pt-en	3				707	0.127	0.804	0.220
average scores for direct induction (cs,de,fr,it,nl,pt)								
all	2				604	0.154	0.682	0.242
all	3				645	0.106	0.403	0.164
best-performing pivot language restriction								
all	2	0.50	6	unrestricted	441	0.222	0.632	0.329
all	3	0.75	6	unrestricted	251	0.247	0.369	0.296
best-performing restriction on projected senses								
all	2	0.45	5	4	250	0.364	0.669	0.472
all	3	0.45	5	4	256	0.309	0.622	0.413

Precision is dissatisfying, but recall is reasonable
=> Baseline

Generated inventories can be a basis for **manual pruning**
(note that discourse marker inventories are small, < 1000 entries)

Lexical Induction with the ACoLi Dictionary Graph

DM discourse marker	DM score	PDTB relation	relation score
"dlatego"@pl	0.96	CONTIGENCY	0.958
"dlatego"@pl	0.96	CONTIGENCY.Cause	0.949
"dlatego"@pl	0.96	CONTIGENCY.Cause.result	0.931
"dlatego"@pl	0.96	CONTIGENCY.Cause.reason	0.019
"dlatego"@pl	0.96	CONTIGENCY.Condition	0.009
"dlatego"@pl	0.96	TEMPORAL.Asynchronous.precedence	0.005
"dlatego"@pl	0.96	TEMPORAL.Asynchronous	0.005
"dlatego"@pl	0.96	TEMPORAL	0.005
"zatem"@pl	0.95	CONTIGENCY	0.699
"zatem"@pl	0.95	CONTIGENCY.Cause	0.671
"zatem"@pl	0.95	CONTIGENCY.Cause.result	0.435
"zatem"@pl	0.95	CONTIGENCY.Cause.reason	0.256
"zatem"@pl	0.95	TEMPORAL	0.199
"zatem"@pl	0.95	TEMPORAL.Asynchronous	0.180
"zatem"@pl	0.95	TEMPORAL.Asynchronous.precedence	0.176
"zatem"@pl	0.95	EXPANSION	0.048
"zatem"@pl	0.95	CONTIGENCY.Condition	0.038
"zatem"@pl	0.95	TEMPORAL.Synchronous	0.029
"zatem"@pl	0.95	EXPANSION.Alternative	0.029
"zatem"@pl	0.95	EXPANSION.Alternative.disjunctive	0.024
"gdy"@pl	0.84	TEMPORAL	0.461
"gdy"@pl	0.84	CONTIGENCY	0.361
"gdy"@pl	0.84	TEMPORAL.Synchronous	0.310

PDTB induction, Polish inventory, 4 pivot languages

538 potential discourse markers

not evaluated, but ranked according to confidence scores for being a discourse marker and for each possible relation

can be a seed for a discourse marker inventory, requires manual pruning

limitations:

- only if in dictionary, mostly single word translations, no phrasal expressions

- some potential discourse markers might actually not be discourse markers after all

Towards a Multilingual Corpus of Discourse and Reference

Bringing it all together

Interim Summary

- We now have
  - a number of multilingual discourse marker inventories
  - a technology to induce discourse marker inventories for hundreds of languages
  - and we can flexibly switch between theory-specific relation inventories
- This can be used to
  - create discourse marker pre-annotation for a novel languages
  - convert an RST corpus into a PDTB or ISO SemAF corpus, say, to increase the amount of training data (if the framework-specific data structures can be transformed, as well)
- This has not been done yet, but we have the right technology in place
  - Take a corpus, transform it into an (RDF) graph
  - Apply SPARQL updates for enrichment and transformation
  - Serialize into target format

Interim Summary

Doing that with off-the-shelf RDF technology sounds like a performance nightmare
But we provide special tooling

The Flexible Integrated Transformation and Annotation eNgineering platform

- NLP formats ↔ RDF graphs
- one sentence (and its local context) at a time
- parallel processing
- streaming

FinTan

- This has not been done yet, but we have the right technology in place
  - Take a corpus, transform it into an (RDF) graph
  - Apply SPARQL updates for enrichment and transformation
  - Serialize into target format

FINTAN: Transforming heterogeneous data in a unified way

Fäth et al.@LREC-2020

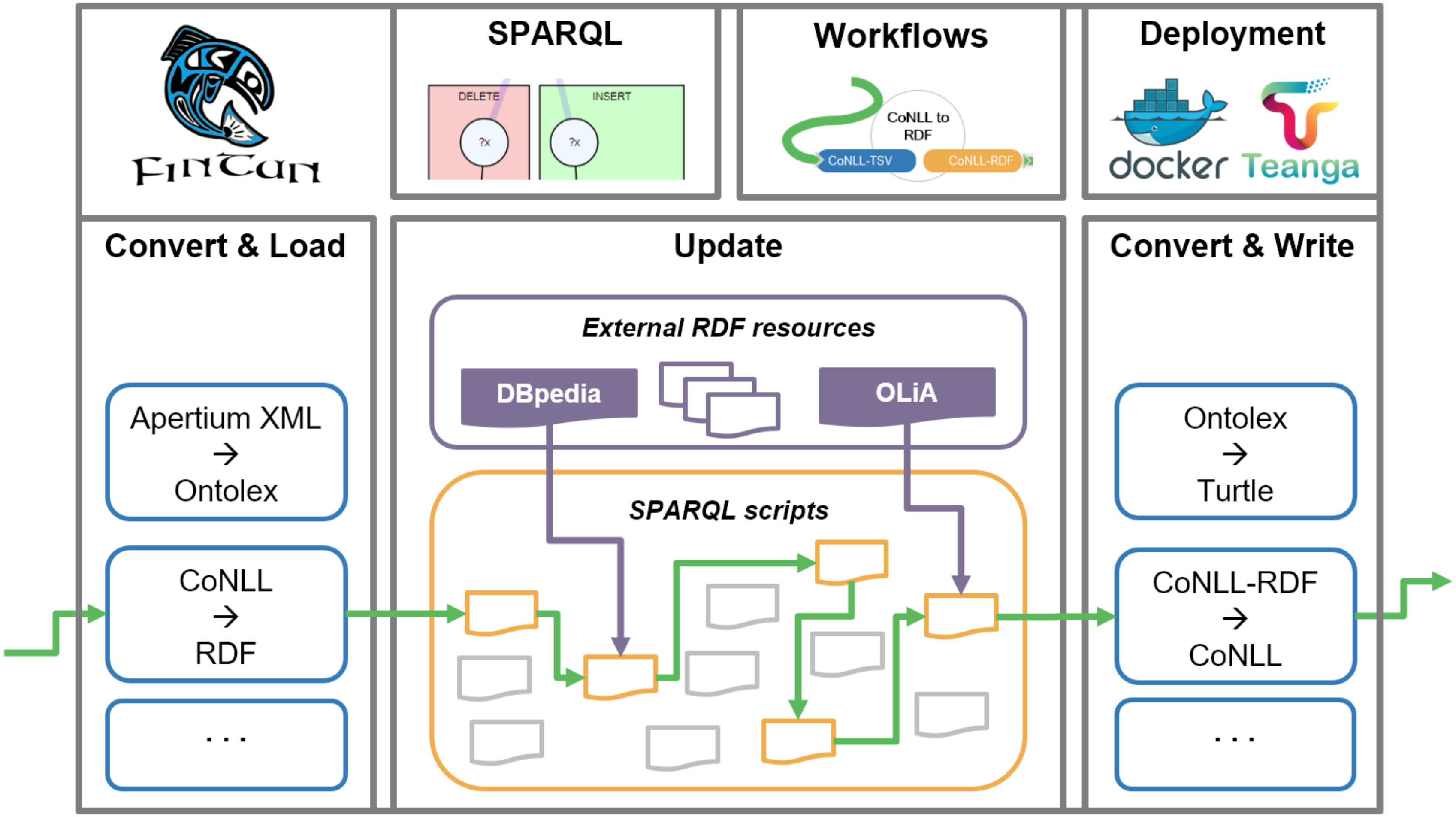
- Convert *any* kind of language resource to RDF graphs.
- Manipulate/link/transform graphs with SPARQL.
- Serialize as RDF or in conventional NLP formats

Modular: Pipelines composed of small, reusable pieces
Reusable: Same RDF vocabulary => same modules
Extensible: Add your own (SPARQL, Docker, Java, ...)
Scalable: Stream processing & parallelization

https://github.com/Pret-a-LLOD/Fintan (wrapper repo)
https://github.com/acoli-repo/conll-rdf (CoNLL customization)

FINTAN: Transforming heterogeneous data in a unified way

Fäth et al.@LREC-2020



FINTAN: Transforming heterogeneous data in a unified way

Fäth et al.@LREC-2020

- previously, FINTAN has been used for
  - various conversion and enrichment/linking tasks
  - rule-based post-processing of automated annotation tasks
  - unified querying of heterogeneously annotated corpora
  - creating a semantically annotated treebank by transforming, decomposing and combining information from PropBank and UD
    - for Role and Reference Grammar
  - pre-annotation of the Augsburg Corpus for Reference and Information Structure
    - automated pre-annotations for discourse markers and the language-specific classification of referring expressions
    - converting existing annotations for coreference (Disco-MT) and discourse (TED-MDB) to the AURIS schema

Augsburg Corpus for Reference and Information Structure (AURIS)

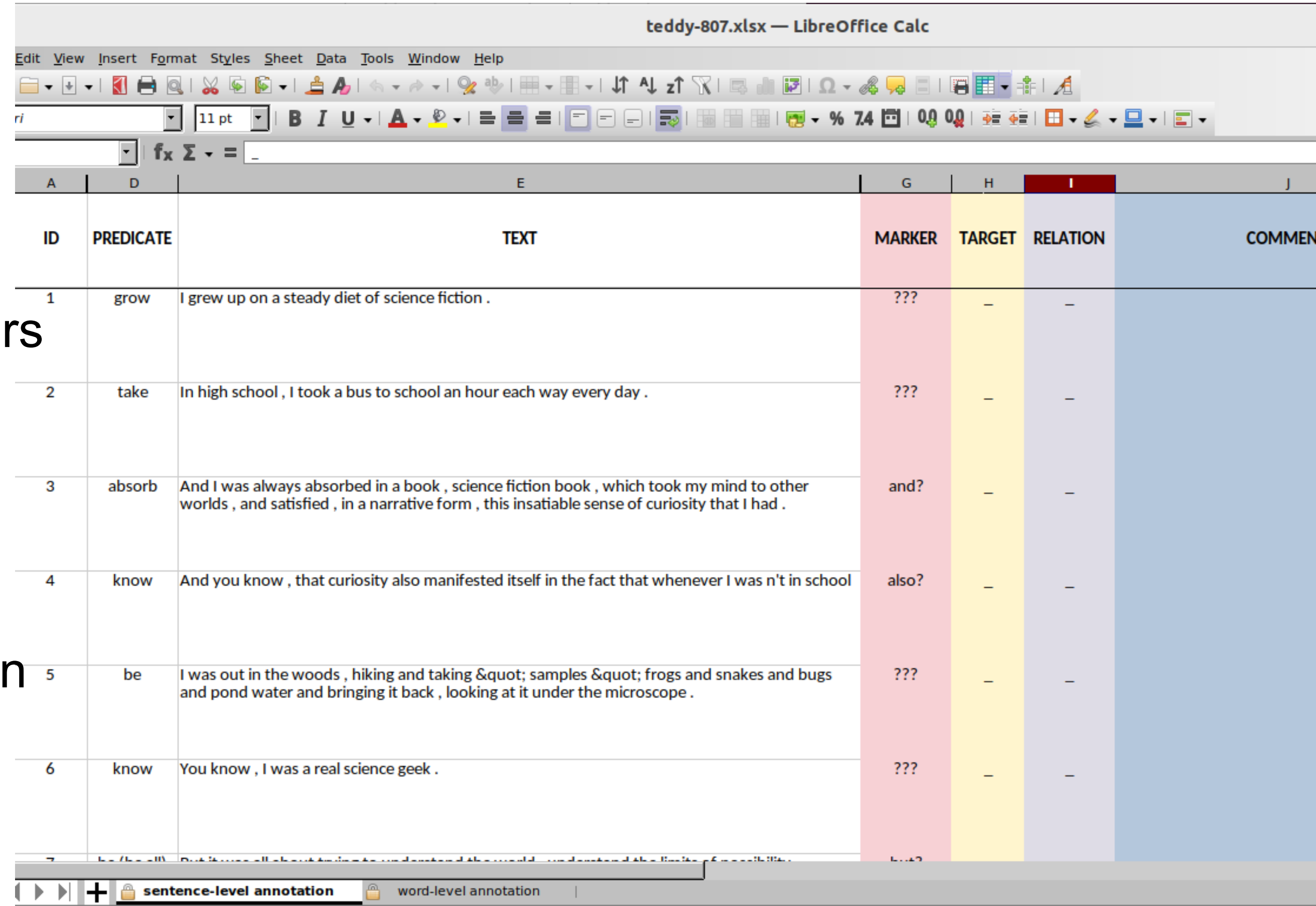
- general lack of language resources for discourse, pragmatics and semantics beyond the sentence
- build such resources together with students (department of philology and history)
  - seminars in German, English & Romance studies, translation science since 2023
- requires / benefits from
  - minimal technical entry barrier
  - offline and online editing
  - multilingual data (students must be able to work on the language of their studies)
  - existing annotations to evaluate students

Augsburg Corpus for Reference and Information Structure (AURIS)

- general lack of language resources for discourse, pragmatics and semantics beyond the sentence
- build such resources together with students (department of philology and history)
  - seminars in German, English & Romance studies, translation science since 2023
- requires / benefits from
  - minimal technical entry barrier **(pre-annotation => spreadsheets)**
  - offline and online editing
  - multilingual data **(parallel text in 5-750 languages, literature, religious, news, TED)**
  - existing annotations to evaluate students **(bootstrapping AURIS annotations from TED MDB, DiscoMT, OntoNotes, FrameNet)**

Augsburg Corpus for Reference and Information Structure (AURIS)

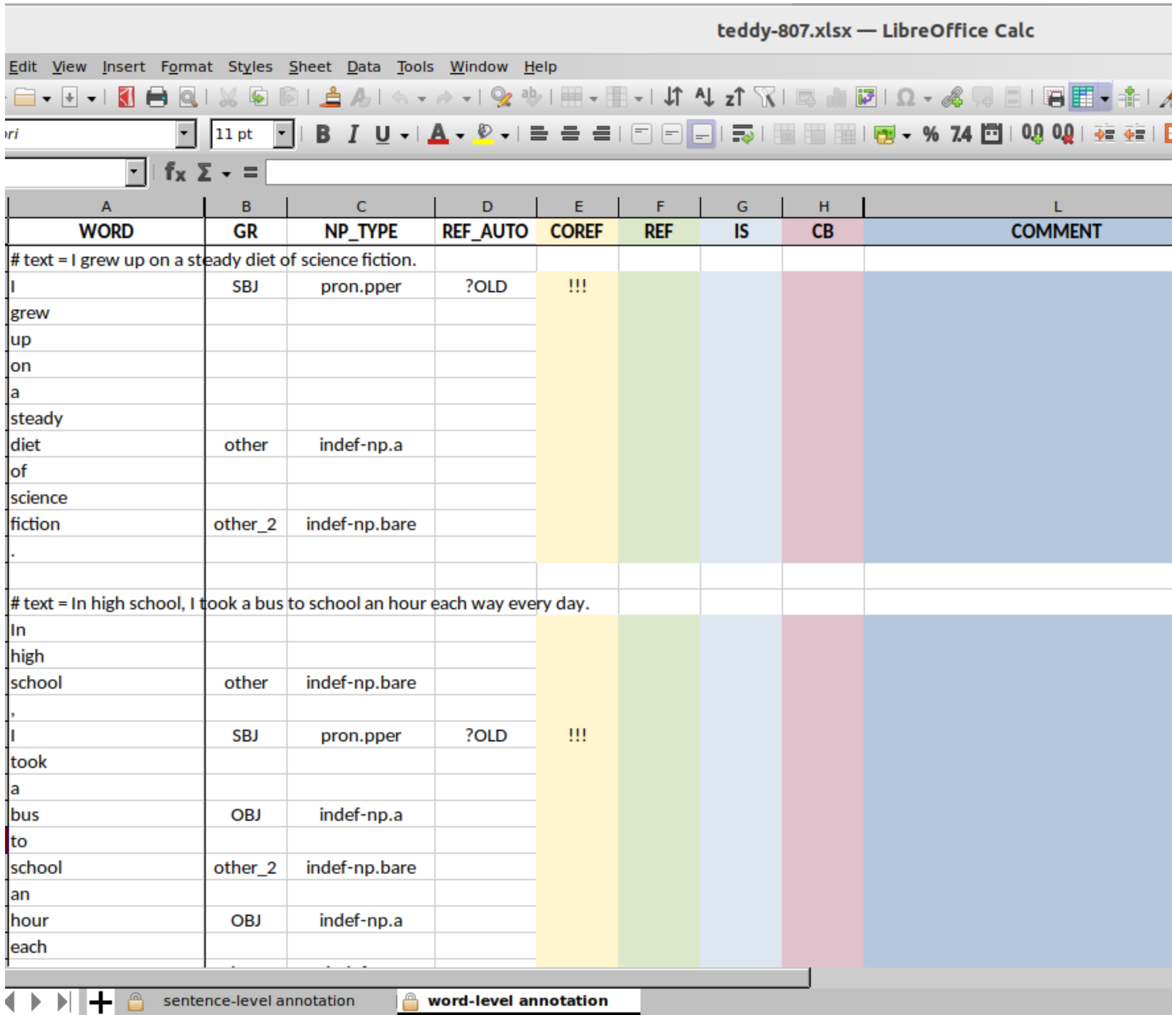
- general lack of language resources for discourse, pragmatics and semantics beyond the sentence
- UD parser + FINTAN => spreadsheet
- **discourse-level sheet**
  - segmented by sentence
  - pre-annotation for discourse markers
  - annotate discourse relations
    - target / external argument
    - relation
  - formulas for dynamic pre-annotation
  - sheet protection


teddy-807.xlsx — LibreOffice Calc
Edit View Insert Format Styles Sheet Data Tools Window Help
11 pt
A D E G H I J
sentence-level annotation
word-level annotation

ID	PREDICATE	TEXT	MARKER	TARGET	RELATION	COMMEN
1	grow	I grew up on a steady diet of science fiction .	???	_	_	
2	take	In high school , I took a bus to school an hour each way every day .	???	_	_	
3	absorb	And I was always absorbed in a book , science fiction book , which took my mind to other worlds , and satisfied , in a narrative form , this insatiable sense of curiosity that I had .	and?	_	_	
4	know	And you know , that curiosity also manifested itself in the fact that whenever I was n't in school	also?	_	_	
5	be	I was out in the woods , hiking and taking &quot; samples &quot; frogs and snakes and bugs and pond water and bringing it back , looking at it under the microscope .	???	_	_	
6	know	You know , I was a real science geek .	???	_	_	
7	be (be all)	But it was all about trying to understand the world , understand the limits of possibility .	but?			

Augsburg Corpus for Reference and Information Structure (AURIS)

- general lack of language resources for discourse, pragmatics and semantics beyond the sentence
- UD parser + FINTAN => spreadsheet
- **word-level sheet**
  - segmented by token and sentence
  - grammatical roles and syntactic embedding
  - automatically classify referring expressions
  - COREF: annotate referent ID, manually
  - REF: referentiality, predicted from COREF
  - IS: information status, -“-
  - CB: topic annotation, -“-


teddy-807.xlsx — LibreOffice Calc
Edit View Insert Format Styles Sheet Data Tools Window Help
11 pt
sentence-level annotation
word-level annotation

A	B	C	D	E	F	G	H	L
WORD	GR	NP_TYPE	REF_AUTO	COREF	REF	IS	CB	COMMENT
# text = I grew up on a steady diet of science fiction.								
I	SBJ	pron.pper	?OLD	!!!				
grew								
up								
on								
a								
steady								
diet	other	indef-np.a						
of								
science								
fiction	other_2	indef-np.bare						
.								
# text = In high school, I took a bus to school an hour each way every day.								
In								
high								
school	other	indef-np.bare						
,								
I	SBJ	pron.pper	?OLD	!!!				
took								
a								
bus	OBJ	indef-np.a						
to								
school	other_2	indef-np.bare						
an								
hour	OBJ	indef-np.a						
each								

Epilogue

wrap up ;)

Summing up

I presented a number of technologies and resources designed to support aspects of discourse processing and discourse annotation, in particular

- RDF technologies and Linked Open Data, and their application to
  - establish a level of interoperability over theory-specific inventories of discourse relations
  - access discourse marker inventories as a knowledge graph,
  - link them them with these inventories and map their,
  - link them with a lexical knowledge graph in order to induce discourse marker inventories in other langages, and
  - convert (or, pre-annotate) annotations for discourse and co-reference

Overall, the main contribution of this technology is its versatility, in discourse studies, in proving training data for NLU, or beyond

If you want to learn more, please consider to participate in our MOOCs on Linguistic Linked Data (see QR Codes)

Thank you very much!