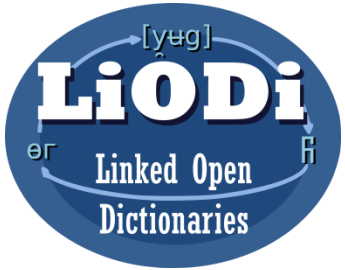
Augsburg, 13.10.2021

Maschinenlesbare Diskursmarkerinventorien

Christian Chiarcos

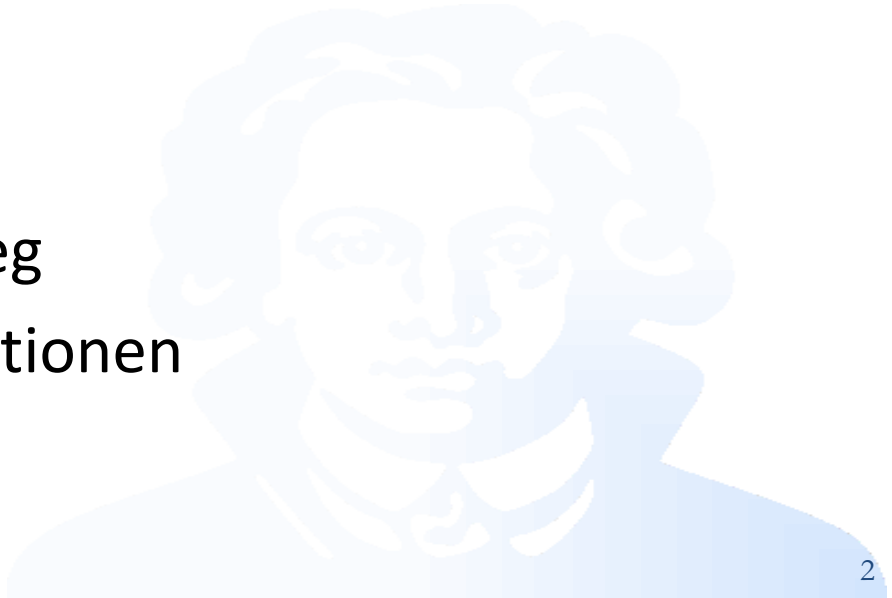
Angewandte Computerlinguistik (ACoLi)

chiarcos@informatik.uni-frankfurt.de



Maschinenlesbare Diskursmarkerinventorien

- Sprach- und theorieübergreifende Modellierung
  - OntoLex-Lemon und Webstandards
  - OLiA Discourse Extensions
- und deren Nutzen
  - Suche über Sprachen hinweg
  - Inferenz von Diskursannotationen



Diskursmarker

- Zeigen an, wie eine Äußerung mit ihrem Diskurskontext zu verbinden ist
  - lexikalische Mittel (z.B. Adverbien, Phrasen), um eine *Diskursrelation* (*Kohärenzrelation*) auszudrücken
    - Hans kann nicht gehen. ...
      - **Und** auch Maria ist es nicht möglich. *additive*
      - **Daher** ist es auch Maria nicht möglich. *kausale*
      - **Aber** auch Maria ist es nicht möglich. *kontrastive*
      - ∅ Auch Maria ist es nicht möglich. *implizite (unmarkierte)*

Relation

Diskursmarkerinventorien

- Für mehrere Sprachen wurden Diskursmarkerinventorien entwickelt
  - bilden Diskursmarker auf Funktionen (Relationen) ab
  - unterstützen Diskursparsing & dessen Anwendungen
- Verschiedene Format, verschiedene Theorien
  - ⇒ Beitrag hier: Konsolidierung, Integration, Nutzung

State of the Art: TextLink

- Cost Action *Structuring Discourse in Multilingual Europe* (2014-2018)
  - multilinguale Diskursmarkerinventorien
    - Relationen: (zumeist) Penn Discourse Treebank (PDTB)
    - Format: (zumeist) XML-Formate in Anlehnung an DimLex (Stede & Umbach 1998)

⇒ http://connective-lex.info/

Maschinenlesbarkeit in TextLink

Italienisch (LICO)

```
<entry id="1">
  <orth type="cont">
    <part type="phrasal">a causa di</part>
  </orth>
  <syn type="prepositional">
    <sem>
      <coh-relation>CONTINGENCY:Cause:reason</coh-relation>
      <example>A causa del maltempo il St. Gotthard è rimasto chiuso.</example>
      <example>Una chiusura a causa del maltempo verrà presa in considerazione.</example>
    </sem>
  </syn>
  <commento>DimLex.xml/id="k19"</commento>
  <commento>DimLex.xml/id="k21"</commento>
  <commento>DimLex.xml/id="k160"</commento>
</entry>
```

maschinenlesbare Syntax

- Ähnliche XML-Formate
- nicht identisch, aber transformierbar

keine maschinen-lesbare Semantik

- Relationen sind Zeichenketten, keine formal definierten Objekte

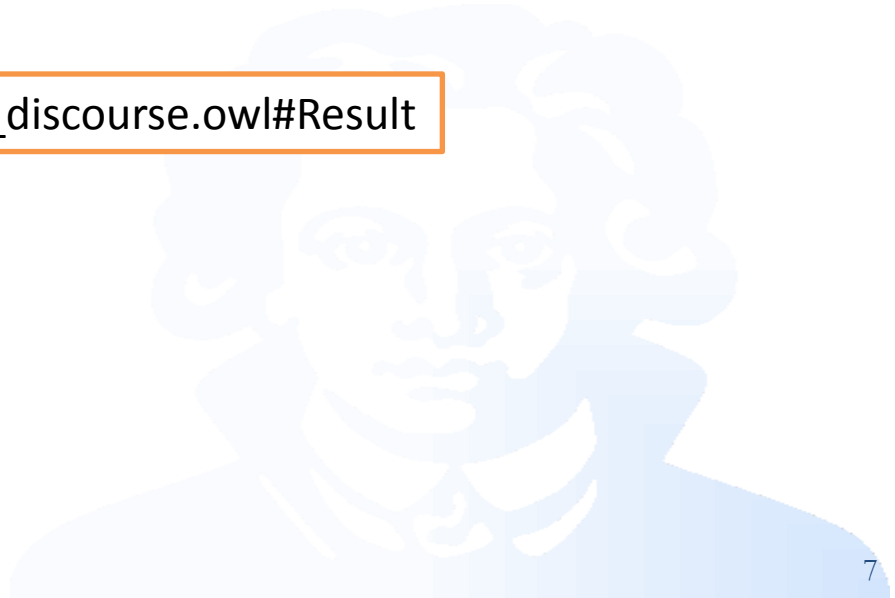
Deutsch (DimLex)

```
<entry id="k160" word="wegen">
   <orths>
      <orth type="cont" canonical="1" onr="k160o1">
         <part type="single">wegen</part>
      </orth>
   </orths>
   <non_conn_reading>
      <example>Allein schon dieses Gerede von wegen wichtigster Tag in unserem Leben.</
   </non_conn_reading>
   <syn>
      <cat>praep</cat>
      <ordering/>
      <sem>
         <pdtb3_relation sense="cause-reason" freq="5" anno_N="5"/>
      </sem>
   </syn>
</entry>
```

Maschinenlesbare Semantik
Wissensrepräsentation im Web

- eindeutige Bezeichnung von Konzepten und Objekten
  - Webstandard: Uniform Resource Identifier (URI)

http://purl.org/olia/discourse/olia_discourse.owl#Result



Maschinenlesbare Semantik
Wissensrepräsentation im Web

- eindeutige Bezeichnung von Konzepten und Objekten
  - Webstandard: Uniform Resource Identifier (URI)

```
http://purl.org/olia/discourse/olia_discourse.owl#Result
```

oder, etwas kompakter

```
PREFIX olia: <http://purl.org/olia/discourse/olia_discourse.owl#>
olia:Result
```

Maschinenlesbare Semantik
Wissensrepräsentation im Web

- eindeutige Bezeichnung von Konzepten und Objekten
  - Webstandard: Uniform Resource Identifier (URI)
- semantische Beziehungen zwischen diesen
  - Webstandard: Resource Description Framework (RDF)

```
PREFIX olia: <http://purl.org/olia/discourse/olia_discourse.owl#>
olia:Result rdfs:subClassOf olia:Cause .
```

„Result ist eine Art von Cause [Kausalbeziehung].“

Maschinenlesbare Semantik
Wissensrepräsentation im Web

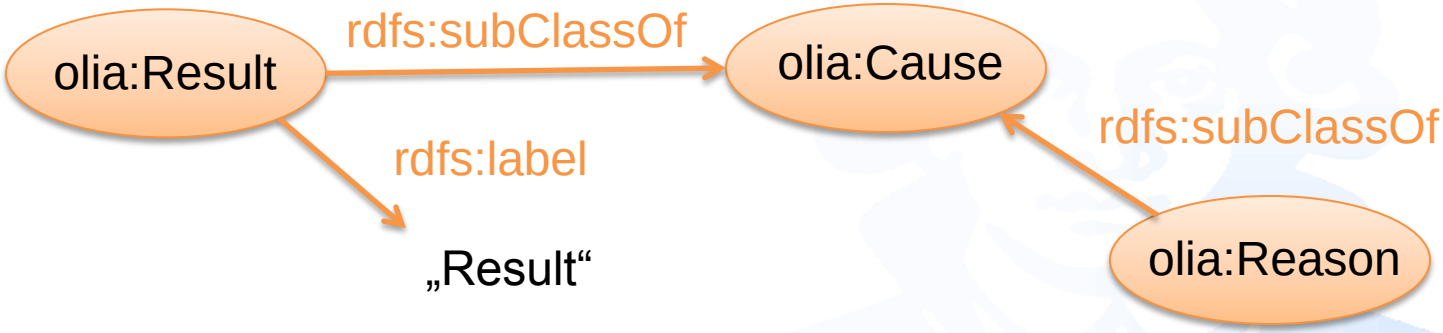
- eindeutige Bezeichnung von Konzepten und Objekten
  - Webstandard: Uniform Resource Identifier (URI)
- semantische Beziehungen zwischen diesen
  - Webstandard: Resource Description Framework (RDF)


```
PREFIX olia: <http://purl.org/olia/discourse/olia_discourse.owl#>
olia:Result rdfs:subClassOf olia:Cause .
```

„Result ist eine Art von Cause [Kausalbeziehung].“

Maschinenlesbare Semantik
Wissensrepräsentation im Web

- eindeutige Bezeichnung von Konzepten und Objekten
  - Webstandard: Uniform Resource Identifier (URI)
- semantische Beziehungen zwischen diesen
  - Webstandard: Resource Description Framework (RDF) => Graph



Maschinenlesbare Semantik
Wissensrepräsentation im Web

- eindeutige Bezeichnung von Konzepten und Objekten
  - Webstandard: Uniform Resource Identifier (URI)
- semantische Beziehungen zwischen diesen
  - Webstandard: Resource Description Framework (RDF)
- Zugriff auf Wissensresourcen im Web
  - Webstandards: HTTP (Protokoll)

```
PREFIX olia: <http://purl.org/olia/discourse/olia_discourse.owl#>
olia:Result rdfs:subClassOf olia:Cause .
```

Maschinenlesbare Semantik
Wissensrepräsentation im Web

- eindeutige Bezeichnung von Konzepten und Objekten
  - Webstandard: Uniform Resource Identifier (URI)
- semantische Beziehungen zwischen diesen
  - Webstandard: Resource Description Framework (RDF)
- Zugriff auf Wissensresourcen im Web
  - Webstandards: HTTP (Protokoll), SPARQL (Anfragesprache)

```
SELECT ?relation
{  ?relation rdfs:subClassOf olia:Cause . }
```

„Welche Arten von Cause [Kausalbeziehungen] gibt es?“

Maschinenlesbare Semantik
Wissensrepräsentation im Web

- eindeutige Bezeichnung von Konzepten und Objekten
  - Webstandard: Uniform Resource Identifier (URI)
- semantische Beziehungen zwischen diesen
  - Webstandard: Resource Description Framework (RDF)
- Zugriff auf Wissensresourcen im Web
  - Webstandards: HTTP (Protokoll), SPARQL (Anfragesprache)
- Formalisierung: Ontologien
  - Webstandard: Web Ontology Language (OWL)

olia:Result **rdfs:subClassOf** olia:Cause .

„ist eine Art von“

Maschinenlesbare Semantik: Ontologies of Linguistic Annotation (OLiA)

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

- OLiA Discourse Extensions (Chiarcos 2014)
  - 12 Annotationsschemata für
    - Informationsstruktur
    - Koreferenz
    - Diskursstruktur
    - Diskursrelationen
  - Referenzmodell

Maschinenlesbare Semantik: Ontologies of Linguistic Annotation (OLiA)

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

- OLiA Discourse Extensions (Chiarcos 2014)
  - 12 Annotationsschemata für
    - Informationsstruktur
    - Koreferenz
    - Diskursstruktur
    - Diskursrelationen
  - Referenzmodell

PDTB
RST
RST-DTB
PDGB
Knott

http://purl.org/olia/discourse/discourse.RST.owl
visualisiert mit Protégé

Maschinenlesbare Semantik: Ontologies of Linguistic Annotation (OLiA)

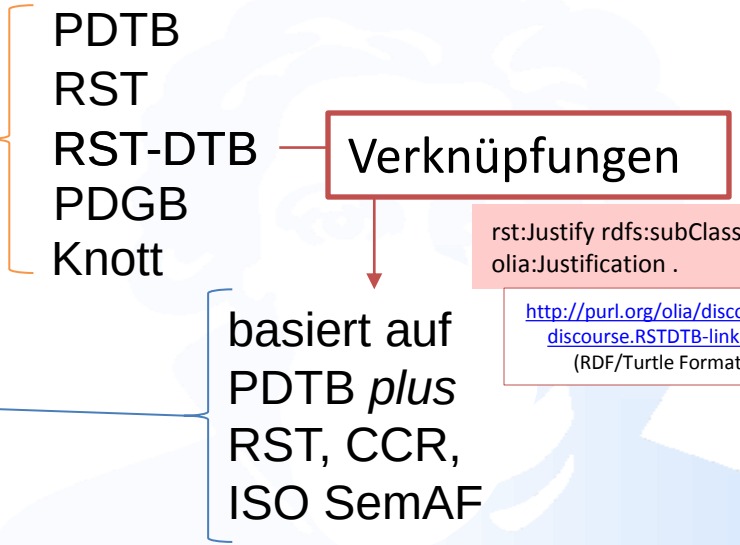
http://github.com/acoli-repo/olia

- OLiA Discourse Extensions
  - 12 Annotationsschemata für
    - Informationsstruktur
    - Koreferenz
    - Diskursstruktur
    - Diskursrelationen
  - Referenzmodell

Maschinenlesbare Semantik: Ontologies of Linguistic Annotation (OLiA)

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

- OLiA Discourse Extensions
  - 12 Annotationsschemata für
    - Informationsstruktur
    - Koreferenz
    - Diskursstruktur
    - Diskursrelationen
  - Referenzmodell



Diskursmarker in OntoLex

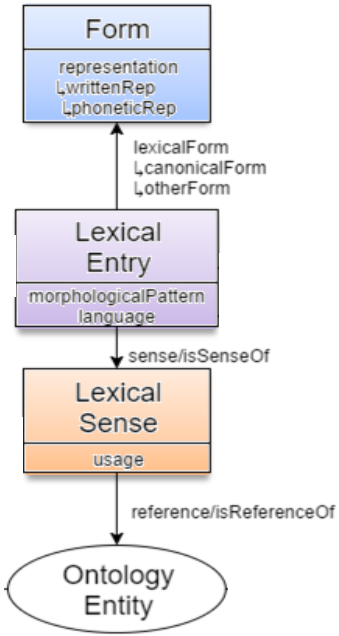
Maschinenlesbare
Diskursmarkerinventorien

OntoLex

https://www.w3.org/2016/05/ontolex

- Community-Standard für maschinenlesbare Wörterbücher im Web (of Data)
- Anwendungen u.a. in
  - Textgenerierung (v.a. Lexikalisierung)
  - Lexikographie (z.B. DitMao/Lex-0)
  - Terminologiemanagement (z.B. Terme-a-LLOD)
  - Computerlinguistik (z.B. TIAD Shared Tasks)

OntoLex (Auszug)



- LexicalEntry
  - entspricht einem Lexem (Schlagwort im Wörterbuch)
- Form
  - Schreibung (+linguistische Informationen)
- LexicalSense
  - Wortbedeutung, ggf. mit externer Ontologie/Wissensbasis verknüpft

Datengrundlage

	>10 Sprachen	7 Formate	4 Theorien	
DimLex	Deutsch	DimLex-XML	PDTB 3.0	
DisCoDict	Niederländisch	DimLex-XML	PDTB 3.0	
LICO	Italienisch	mod. DimLex	PDTB 2.0/3.0	
LDM-PT	Portugiesisch	mod. DimLex	PDTB 3.0	TextLink/ DimLex
LexConn	Französisch	mod. DimLex	SDRT	
PDTB	Englisch	PDTB-Format	PDTB 2.0	eigene Konverter
CzeDLex	Tschechisch	PML-XML	PDiT 2.0	
DiscMar	Engl., Span., Katalanisch	TSV/HTML	DiscMar	
TED-MDB	7 Sprachen*	PDTB-Format	PDTB 3.0	

* geringer Datenumfang, Konverter anwendbar auf Hindi und Chinesisch

Beispiel: DimLex Deutsch

DimLex-XML Exzerpt

```
<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>
```

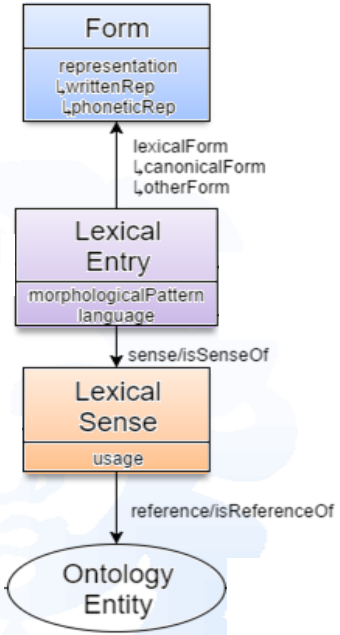
- Stede & Umbach (1998)
  - PDTB-Relationen nach Scheffler & Stede (2016)

    https://github.com/discourse-lab/dimlex

Beispiel: DimLex Deutsch

DimLex-XML $\Rightarrow_{XSLT}$ OntoLex

```
<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>
```



Beispiel: DimLex Deutsch

DimLex-XML $\Rightarrow_{XSLT}$ OntoLex + freie Ergänzungen

(markiert durch *dimlex:*)

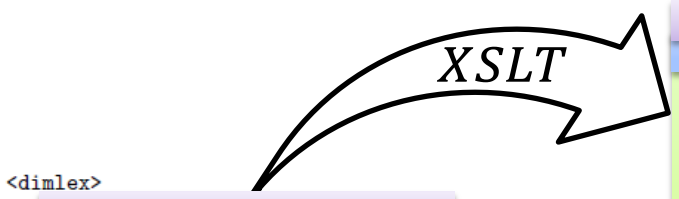
```
<dimlex>
   <entry id="k1" word="aber">
      <orths>
         <orth type="cont" canonical="1" onr="k1o1">
            <part type="single">aber</part>
         </orth>
      </orths>
      <non_conn_reading>
         <example type="ADV" tfreq="940">aber und abermals</example>
         <example type="ADV">Du 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>
```

freie Ergänzungen für *alle* XML-Elemente und -Attribute

⇒ verlustfreie* Repräsentation

* nicht standardisiert

Beispiel: DimLex Deutsch



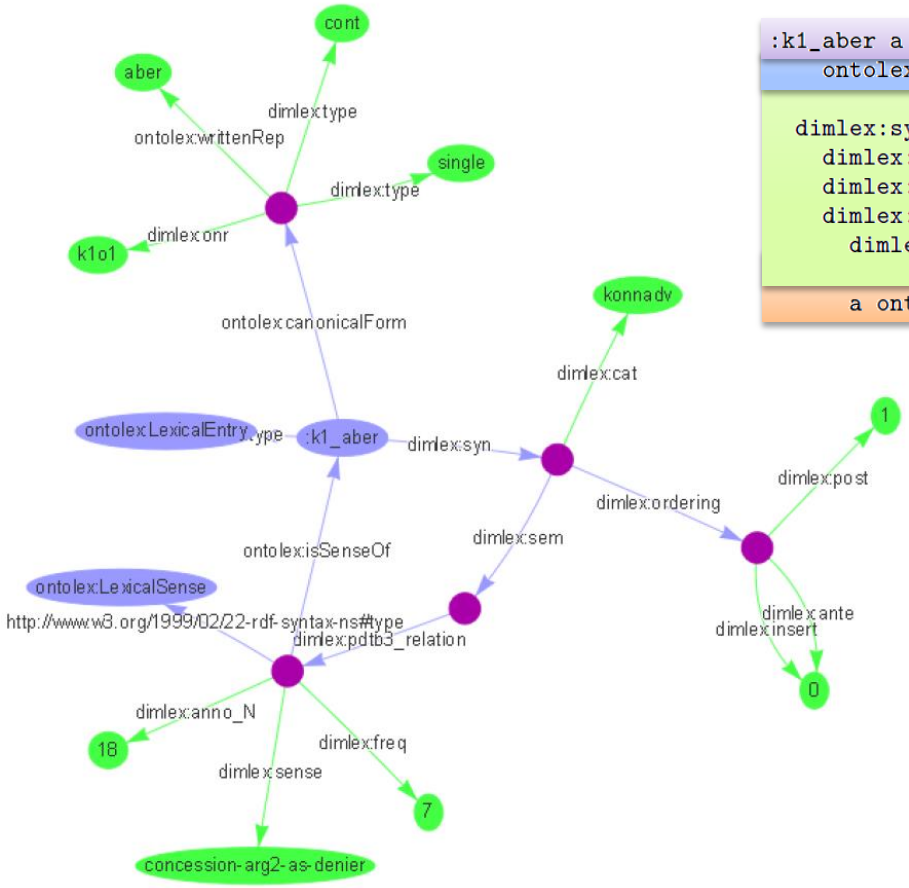

```
<dimlex>
   <entry id="k1" word="aber">
      <orths>
         <orth type="cont" canonical="1" onr="k1o1">
            <part type="single">aber</part>
         </orth>
      </orths>
      <non_conn_reading>
         <example type="ADV" tfreq="940">aber und abermals</example>
         <example type="ADV">Du 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>
```

```
: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-RDF

- OntoLex-Konzepte
- gleiche Struktur
- verlustfrei

Beispiel: DimLex Deutsch als RDF Graph




```
: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-RDF

- OntoLex-Konzepte
- gleiche Struktur
- verlustfrei

Verknüpfung mit OLiA (SPARQL)

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

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

lade die PDTB-Ontologie (aus dem Web)

- Die PDTB-Ontologie ist ihrerseits mit OLiA verknüpft, usw.

Verknüpfung mit OLiA (SPARQL)

```
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.
}
```

lade die PDTB-Ontologie (aus dem Web)

erzeuge Verknüpfung

- Die PDTB-Ontologie ist ihrerseits mit OLiA verknüpft, usw.

Verknüpfung mit OLiA (SPARQL)

```
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))
};
```

lade die PDTB-Ontologie (aus dem Web)

erzeuge Verknüpfung

wenn PDTB-Label = dimlex:sense

- Die PDTB-Ontologie ist ihrerseits mit OLiA verknüpft, usw.

Verknüpfte OntoLex-Inventorien

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

lan-guage	dataset http://purl.org/acoli/dimlex/...	license	PDTB links	markers (canonical)	granu-larity
ar	.../ar/arabic.ttl	t.b.d.	505	505	14
bn	.../bn/dimlex-bangla.ttl	CC-BY-NC-SA 4.0	107	122 (101)	16
ca	.../ca/discmar.ca.ttl	CC-BY-NC 3.0	97	93	5
cs	.../cs/czedlex0.6.ttl	CC-BY-NC-SA 4.0	1883	1459 (204)	20
de	.../de/DimLex.ttl	CC-BY-NC-SA 4.0	411	763 (274)	18
de	.../de/ted-mdb-german.ttl	CC-BY 4.0	27	31	15
en	.../en/discmar.en.ttl	CC-BY-NC 3.0	90	98	5
en	.../en/pdtb2.ttl	CC-BY-NC-SA 4.0	535	186 (92)	21
en	.../en/ted-mdb-english.ttl	CC-BY 4.0	23	24	11
es	.../es/discmar.es.ttl	CC-BY-NC 3.0	93	97	5
fr	.../fr/lexconn.ttl	CC-BY-NC 3.0	416	603	13
it	.../it/LICO-v.1.0.ttl	CC-BY 4.0	174	204	19
lt	.../lt/ted-mdb-lithuanian.ttl	CC-BY 4.0	27	24	13
nl	.../nl/discodict.ttl	CC-BY-NC-SA 4.0	244	473 (207)	21
pl	.../pl/ted-mdb-polish.ttl	CC-BY 4.0	4	12	3
pt	.../pt/LDM-v.1.3.ttl	CC-BY-NC-SA 4.0	663	254	22
pt	.../pt/ted-mdb-portuguese.ttl	CC-BY 4.0	21	22	9
ru	.../ru/ted-mdb-russian.ttl	CC-BY 4.0	21	21	11
tr	.../tr/ted-mdb-turkish.ttl	CC-BY 4.0	28	31	11

Nutzung: Anfragen

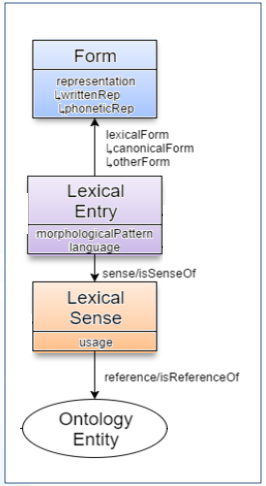
Diskursmarker ↦ PDTB-Relation ↦ Diskursmarker

- für einen gegebenen Diskursmarker, bestimme bedeutungsgleiche Marker (z.B. in einer anderen Sprache)

Diskursmarker ↦ Relation (DiscMar, Englisch)

```
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>
PREFIX pdtb: <http://purl.org/olia/discourse/discourse.PDTB.owl#>

SELECT distinct ?en ?pdtb
FROM <http://purl.org/acoli/dimlex/en/discmar.en.ttl>
WHERE {
  ?form ontolex:writtenRep ?en.
  filter(lang(?en) = "en")
  ?entry (ontolex:lexicalForm|ontolex:canonicalForm) ?form.
  ?sense ontolex:isSenseOf ?entry.
  ?sense ontolex:reference ?pdtb.
} ORDER BY ?en ?pdtb
```



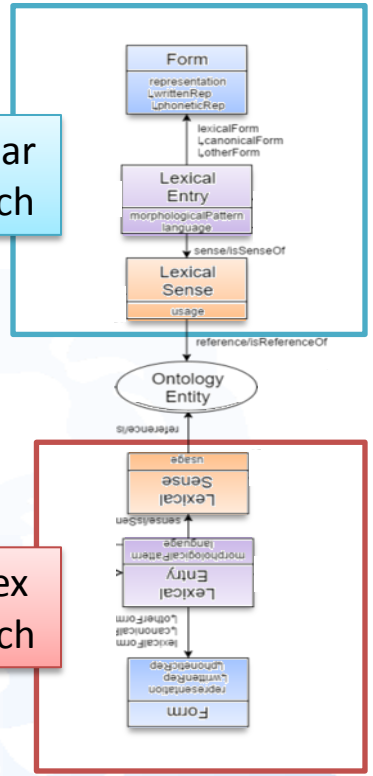
Diese Anfrage und alle folgenden können online über Webservices (z.B. http://sparql.org) ausgeführt werden. Keine Datenbankinstallation notwendig.

Englisch ↦ Relation ↦ Deutsch

```
SELECT distinct ?en ?pdtb ?de
FROM <http://purl.org/acoli/dimlex/en/discmar.en.ttl>
FROM <http://purl.org/acoli/dimlex/de/DimLex.ttl>
WHERE {
   ?pdtb ^ontolex:reference/ontolex:isSenseOf/
         (ontolex:lexicalForm|ontolex:canonicalForm)/
         ontolex:writtenRep ?en.
   filter(lang(?en) = "en")
   ?pdtb ^ontolex:reference/ontolex:isSenseOf/
         (ontolex:lexicalForm|ontolex:canonicalForm)/
         ontolex:writtenRep ?de.
   filter(lang(?de) = "de")
} ORDER BY ?en ?pdtb ?de
```

DiscMar Englisch

DimLex Deutsch



Einbeziehung der Ontologie

- Anfragen nach *identischen* Diskursrelationen sind auch in TextLink machbar
- ABER: DiscMar and DimLex haben nur wenige 1:1-Entsprechungen
  - DiscMar (englisch): 5 Diskursrelationen
  - DimLex (deutsch): 18 Diskursrelationen

    => Subsumption: Erweitere die Suche auf Unterklassen

Englisch ↦ Relation ↦ Ontologie ↦ Deutsch

```
SELECT distinct ?en ?pdtb ?de
FROM <http://purl.org/acoli/dimlex/en/discmar.en.ttl>
FROM <http://purl.org/acoli/dimlex/de/DimLex.ttl>
FROM <http://purl.org/olia/discourse/discourse.PDTB.owl>
WHERE {
   ?pdtb rdfs:subClassOf*/^ontolex:reference/ontolex:isSenseOf/
         (ontolex:lexicalForm|ontolex:canonicalForm)/
         ontolex:writtenRep ?en.
   filter(lang(?en) = "en")
   ?pdtb ^ontolex:reference/ontolex:isSenseOf/
         (ontolex:lexicalForm|ontolex:canonicalForm)/
         ontolex:writtenRep ?de.
   filter(lang(?de) = "de")
} ORDER BY ?en ?de
```

Englisch ↦ Relation ↦ Ontologie ↦ Deutsch

```
SELECT distinct ?en ?pdtb
FROM <http://purl.org/acoli
FROM <http://purl.org/acoli/dimlex/de/DimLex.ttl>
FROM <http://purl.org/olia/discourse/discourse.PDTB.owl>
WHERE {
   ?pdtb rdfs:subClassOf*/
         (ontolex:lexicalForm|ontolex:canonicalForm)/
         ontolex:writtenRep ?en.
   filter(lang(?en) = "en")
   ?pdtb ^ontolex:reference/ontolex:isSenseOf/
         (ontolex:lexicalForm|ontolex:canonicalForm)/
         ontolex:writtenRep ?de.
   filter(lang(?de) = "de")
} ORDER BY ?en ?de
```

Lade das Annotationsmodell

Durchsuche Unterklassen

alles andere bleibt gleich

Weitere Anfragen

- über Theorien hinweg
  - Englisch ↦ PDTB-Relation ↦ Konzept (OLiA) ↦ RST-Relation
    - ganz analog
      - lade Annotationsmodelle für PDTB und RST
      - lade OLiA und dessen Verknüpfung
      - Inferenz / Suche: folge *rdfs:subClassOf*
    - exportiere das Ergebnis als Tabelle

Nutzung

Wozu brauchen wir multilingual verknüpfte Diskursmarker-Inventorien?

Induktion von Diskursannotationen Auf Basis von Parallelkorpora

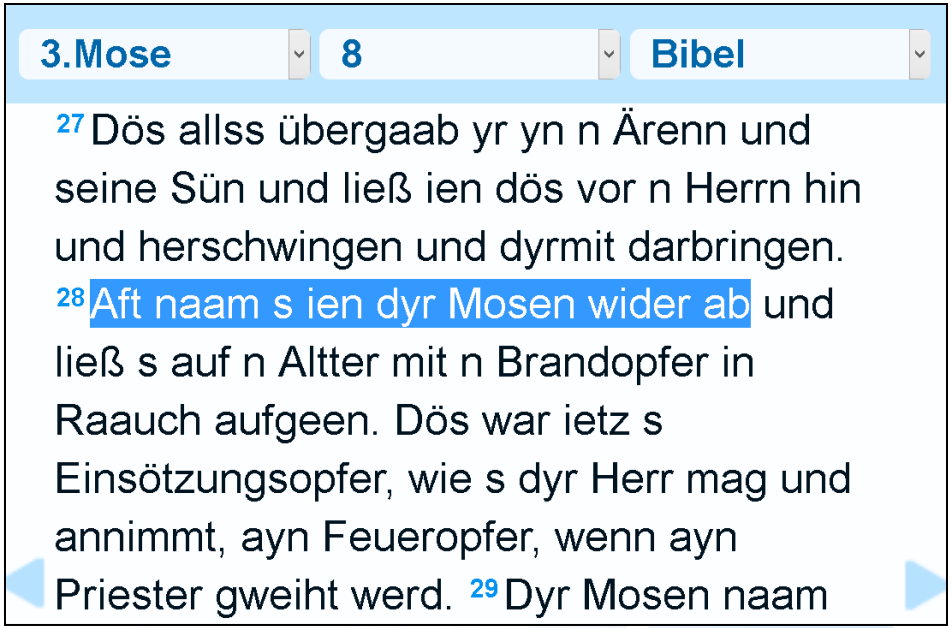
- Gegeben ein mehrfach übersetzter Text
  - Annotiere Quellsprachen nach (möglichen) Diskursmarkern und -relationen gemäß *einer* Theorie
  - *Aus der Überlappung dieser Annotationen* kann die Diskursfunktion im Kontext abgeleitet werden
- Sofern es übersetzte Texte gibt, ist das für jede beliebige Sprache möglich
  - Bayrisch?

Sturmibibl (1998)

https://bar.wikipedia.org/wiki/Sturmibibl

- vollständige Bibelübersetzung
  - 800.000 Tokens
  - “Bayrische Buchsprache”
- aligniert mit > 100 Sprachen

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



https://bibeltext.com/bairisch/leviticus/8.htm

Aft naam s ien dyr Mosen wider ab (LEV.8.28)

bar	Aft	naam s ien dyr Mosen	wider ab

Aft naam s ien dyr Mosen wider ab (LEV.8.28)

bar	**Aft**	naam s ien dyr Mosen	wider ab
de	**und**	nahm alles	wieder von ihren Händen

aft

de	COMPARISON:Contrast
de	EXPANSION:Conjunction
de	TEMPORAL:Asynchronous:Precedence

Aft naam s ien dyr Mosen wider ab (LEV.8.28)

bar	**Aft**	naam s ien dyr Mosen	wider	ab
de	**und**	nahm alles	wieder	von ihren Händen
cs	**potom**	vzvav	Ø	z rukou jejich

aft

de	COMPARISON:Contrast
de,cs	EXPANSION:Conjunction
de,cs	TEMPORAL:Asynchronous:Precedence
cs	CONTIGENCY:Condition
cs	EXPANSION:Equivalence

Aft naam s ien dyr Mosen wider ab (LEV.8.28)

bar	**Aft**	naam s ien dyr Mosen	wider	ab
de	**und**	nahm alles	wieder	von ihren Händen
cs	**potom**	vzvav	Ø	z rukou jejich
en	**Ø**	Moses took them	Ø	from their hands

aft

de	COMPARISON:Contrast
de,cs	EXPANSION:Conjunction
de,cs	TEMPORAL:Asynchronous:Precedence
cs	CONTIGENCY:Condition
cs	EXPANSION:Equivalence
en	no translation/alignment

Aft naam s ien dyr Mosen wider ab (LEV.8.28)

bar	**Aft**	naam s ien dyr Mosen	**wider** ab
de	**und**	nahm alles	**wieder** von ihren Händen
cs	**potom**	vzvav	**Ø** z rukou jejich
en	**Ø**	Moses took them	**Ø** from their hands
fr	**puis**	Moïse les ôta	**(puis)** de leurs mains

aft

de	COMPARISON:Contrast
de,cs	EXPANSION:Conjunction
de,cs,fr	TEMPORAL:Asynchronous:Precedence
cs	CONTIGENCY:Condition
cs	EXPANSION:Equivalence
en	no translation/alignment

wider

de	no annotation
cs,en	no translation/alignment
fr	TEMPORAL:Asynchronous:Precedence

Aft naam s ien dyr Mosen wider ab (LEV.8.28)

bar	**Aft**	naam s ien dyr Mosen	**wider**	ab
de	**und**	nahm alles	**wieder**	von ihren Händen
cs	**potom**	vzvav	**Ø**	z rukou jejich
en	**Ø**	Moses took them	**Ø**	from their hands
fr	**puis**	Moïse les ôta	**(puis)**	de leurs mains
it	**poi**	Mosè prese quelle cose	**(poi)**	d'in su le lor mani

aft

de	COMPARISON:Contrast
de,cs	EXPANSION:Conjunction
de,cs,fr,it	TEMPORAL:Asynchronous:Precedence
cs	CONTIGENCY:Condition
cs	EXPANSION:Equivalence
en	no translation/alignment

wider

de	no annotation
cs,en	no translation/alignment
fr,it	TEMPORAL:Asynchronous:Precedence

Aft naam s ien dyr Mosen wider ab (LEV.8.28)

bar		**Aft**	naam s ien dyr Mosen	**wider**	ab
de		**und**	nahm alles	**wieder**	von ihren Händen
cs		**potom**	vzvav	**Ø**	z rukou jejich
en		**Ø**	Moses took them	**Ø**	from their hands
fr		**puis**	Moïse les ôta	**(puis)**	de leurs mains
it		**poi**	Mosè prese quelle cose	**(poi)**	d'in su le lor mani
nl	Mozes nam het	**daarna**		**weer**	van hen over

aft

de	COMPARISON:Contrast
de,cs	EXPANSION:Conjunction
de,cs,fr,it,nl	TEMPORAL:Asynchronous:Precedence
cs	CONTIGENCY:Condition
cs	EXPANSION:Equivalence
en	no translation/alignment

wider

de,nl	no annotation
cs,en	no translation/alignment
fr,it	TEMPORAL:Asynchronous:Precedence

Aft naam s ien dyr Mosen wider ab (LEV.8.28)

bar		**Aft**	naam s ien dyr Mosen	**wider**	ab
de		**und**	nahm alles	**wieder**	von ihren Händen
cs		**potom**	vzvav	**Ø**	z rukou jejich
en		**Ø**	Moses took them	**Ø**	from their hands
fr		**puis**	Moïse les ôta	**(puis)**	de leurs mains
it		**poi**	Mosè prese quelle cose	**(poi)**	d'in su le lor mani
nl	Mozes nam het	**daarna**		**weer**	van hen over
pt		**Ø**		**então**	Moisés os tomou das mãos deles

aft

de	COMPARISON:Contrast
de,cs	EXPANSION:Conjunction
de,cs,fr,it,nl	TEMPORAL:Asynchronous:Precedence
cs	CONTIGENCY:Condition
cs	EXPANSION:Equivalence
en,pt	no translation/alignment

wider

de,nl	no annotation
cs,en	no translation/alignment
fr,it,pt	TEMPORAL:Asynchronous:Precedence

Ensemble-Architektur

vgl. Chiarcos (2010), Towards robust multi-tool tagging. An OWL/DL-based approach, Proc. ACL-2010, Uppsala

- (Annotationen der) Übersetzungen ergeben eine Gewichtung von möglichen Diskursfunktionen
  - Baseline: einfache Mehrheit

aft

de	COMPARISON:Contrast
de,cs	EXPANSION:Conjunction
de,cs,fr,it,nl	TEMPORAL:Asynchronous:Precedence
cs	CONTIGENCY:Condition
cs	EXPANSION:Equivalence
en,pt	no translation/alignment

wider

de,nl	no annotation
cs,en	no translation/alignment
fr,it,pt	TEMPORAL:Asynchronous:Precedence

Ensemble-Architektur

vgl. Chiarcos (2010), Towards robust multi-tool tagging. An OWL/DL-based approach, Proc. ACL-2010, Uppsala

- (Annotationen der) Übersetzungen ergeben eine Gewichtung von möglichen Diskursfunktionen
  - Baseline: einfache Mehrheit

		durch Übersetzungen	
aft		vorausgesagt	nicht vorausgesagt
de	COMPARISON:Contrast	1	4
de,cs	EXPANSION:Conjunction	2	3
de,cs,fr,it,nl	TEMPORAL:Asynchronous:Precedence	5	0
cs	CONTIGENCY:Condition	1	4
cs	EXPANSION:Equivalence	1	4
en,pt	no translation/alignment	n/a	n/a

Ensemble-Architektur
⇒ Annotiertes Korpus

```
# b.LEV.8.28
# Aft naam s ien dyr Mosen wider ab und ließ s auf
ester gweiht werd .
1       Aft     TEMPORAL:Asynchronous:Precedence
2       naam    _
3       s       _
4       ien     _
5       dyr     _
6       Mosen   _
7       wider   TEMPORAL:Asynchronous:Precedence
8       ab      _
9       und     _
10      ließ    _
11      s       _
12      auf     _
13      n       _
14      Altter  _
15      mit     _
16      n       _
17      Brandopfer      _
18      in      _
19      Raauch  _
20      aufgeen _
21      .       _
```

Grundlage weiterer Studien

- Trainingsdaten für maschinelles Lernen
- Einspeisung in Korpusmanage-mentsystem und Suche
- Verbesserte Ensemble-Architektur
- Lexikographie / Diskursmarker-inventorium

Ensemble-Architektur
Evaluation

- für Diskursannotation auf Bayrisch gibt es keine Golddaten
  - Evaluation gegen eine der Projektionen (hier Englisch)

    tp true positive: *alle* vorausgesagten Relationen werden auch aus dem Englischen projiziert

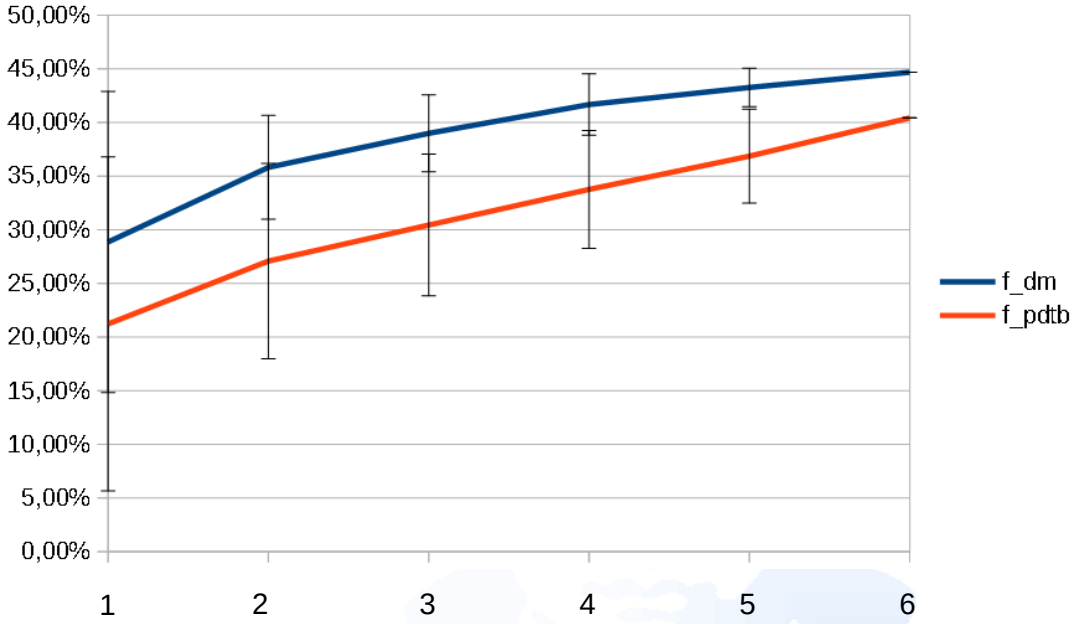
$$precision = {tp}/{(tp+fp)}$$

$$recall = {tp}/{(tp+fn)}$$

$$f = 2\frac{precision \cdot recall}{precision + recall}$$

Ensemble-Architektur Evaluation

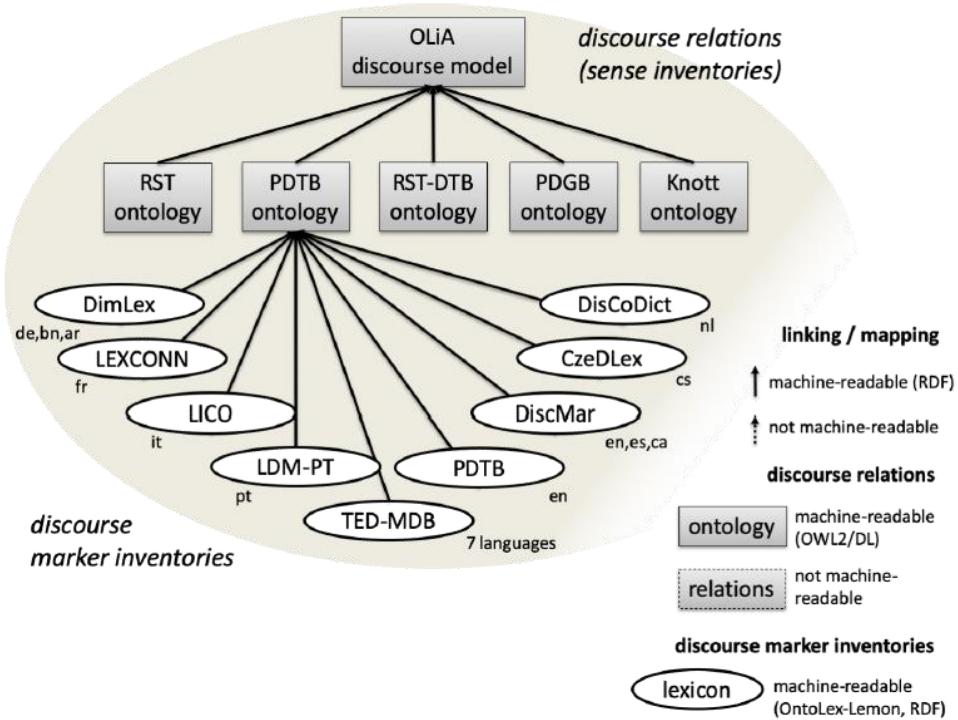
- Je größer das Ensemble, um so zuverlässiger ist die Voraussage
  - Bestätigt* für *f* (*precision & recall*)



* Hinweis: Ohne Gold-Daten ist diese Evaluation heuristisch und *unter*schätzt die Performance systematisch.

Zusammenfassung

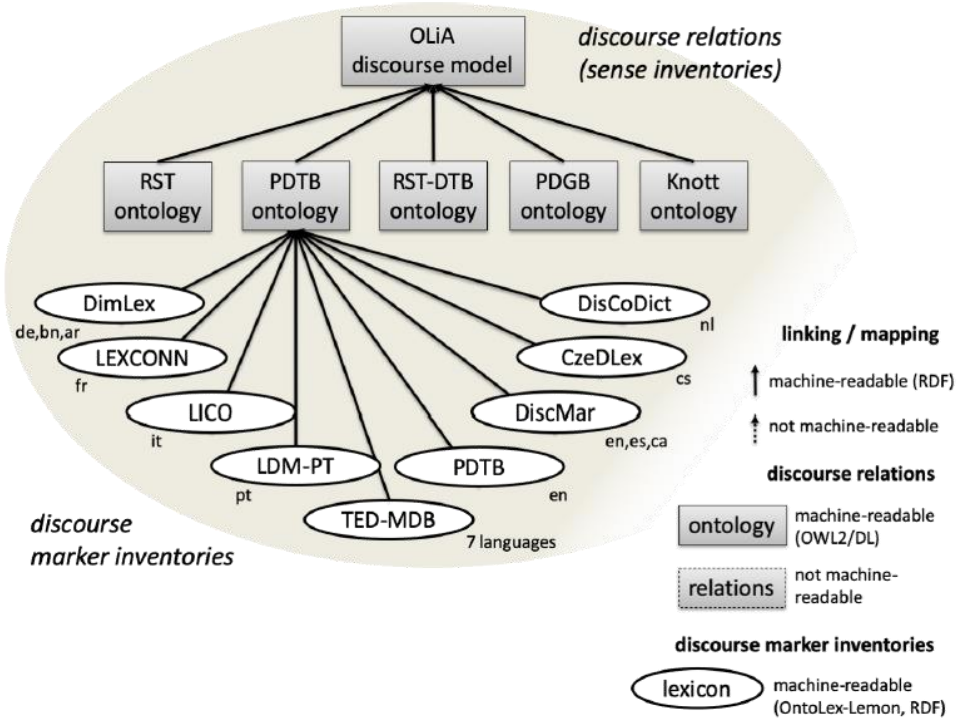
Verknüpfung maschinenlesbarer Diskursmarkerinventorien



- erweitert und formalisiert TextLink-Inventorien
  - Verknüpfung mit PDTB-Ontologie
  - Sprachübergreifende Suche
- Verknüpfung mit OLiA-Ontologie
  - Subsumption / Reasoning
  - Verknüpfung mit Modellen für andere Theorien
- Nutzung z.B. zur Induktion von Diskursannotationen
  - hier am Beispiel des Bayrischen

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

Verknüpfung maschinenlesbarer Diskursmarkerinventorien



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

Vielen Dank

für Ihre

Aufmerksamkeit!

Quellen

- Christian Chiarcos & Maxim Ionov (2021), *Linking Discourse Marker Inventories*.
  In: Proc. 3rd Conference on Language, Data and Knowledge (LDK 2021), Sep 2021, Saragossa, Spanien
  - siehe dort für weitere Referenzen
- Christian Chiarcos (2014), *Towards Interoperable Discourse Annotation*
  In: Proc. Conference on Language Resources and Evaluation (LREC-2014), Mai 2014, Reykjavik, Island.
  - OLiA Discourse Extensions
- Christian Chiarcos et al. (2020), *Translation Inference by Concept Propagation.*
  In: 2020 Globalex Workshop on Linked Lexicography (GLOBALEX-2020), May 2020, Marseille, Frankreich
  - Induktion von Konzepten über maschinenlesbare Wörterbücher hinweg, hier für WordNet
- Christian Chiarcos (ms.), *Cross-Lingual Discourse Marker Induction.*
  unpubliziertes Manuskript
  - hier angewendet auf das Bayrische

Ergänzungen

Daten:	Muss es denn unbedingt die(se) Bibel sein?
	Geht das auch ohne Paralleltext?
	Geht das auch für Sprache XYZ?
Nutzen und Nutzung:	Müssen wir jetzt SPARQL lernen?
Anfragen:	Subsumptionsinferenz
Indirekte Nutzung:	Evaluation von Diskurstheorien (bzw. deren Annotationen)
Erweiterung:	Geht das auch mit Semantik?

Daten: Andere Texte

- Die(se) Bibel ist in vielerlei Hinsicht speziell
  - Aber es gibt weitere Datengrundlagen

Daten: Induktion aus Wörterbüchern

vgl. Bayerisches Wörterbuch: *aft* „dann, nachher“

https://lexhelfer.bwb.badw.de/index.php?limit=&Bogen=087&Frage=8&onlySnippets=1

DimLex

dann	TEMPORAL:Asynchronous:Precedence
nachher	EXPANSION:Conjunction TEMPORAL:Asynchronous:Precedence CONTINGENCY:Condition:Arg2-as-cond TEMPORAL:Synchronous

Daten: Induktion aus Wörterbüchern

vgl. Bayerisches Wörterbuch: *aft* „dann, nachher“

⇒ einfache Mehrheit, häufigste Relation(en)

TEMPORAL:Asynchronous:Precedence

DimLex

dann	TEMPORAL:Asynchronous:Precedence
nachher	EXPANSION:Conjunction TEMPORAL:Asynchronous:Precedence CONTINGENCY:Condition:Arg2-as-cond TEMPORAL:Synchronous

Daten: Induktion aus Wörterbüchern

vgl. Bayerisches Wörterbuch: *aft* „dann, nachher“

⇒ einfache Mehrheit, häufigste Relation(en)

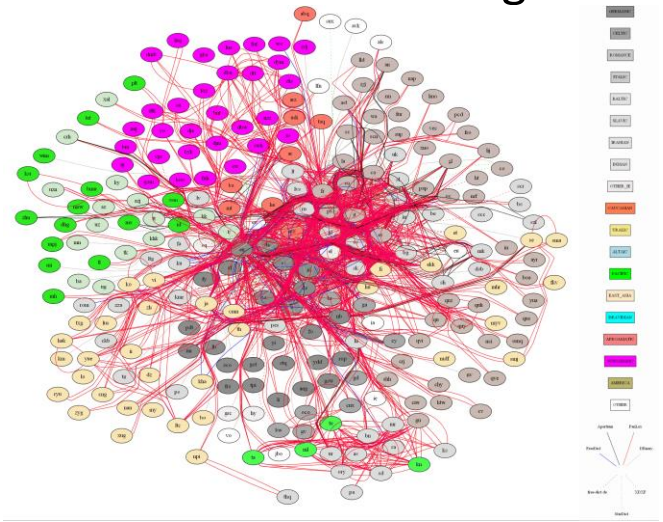
TEMPORAL:Asynchronous:Precedence

vgl. korpusbasiert induziertes Diskursmarkerinventorium


```
"Aft": {"confidence": 0.5084337349397591,
        "freq": 415,
        "rels": {
                 "EXPANSION:Conjunction": 0.5985385878489325,
                 "TEMPORAL:Asynchronous:Precedence": 0.5687689969604861,
                 "CONTINGENCY:Cause:Result": 0.4875,
```

Daten: Andere Sprachen

- Die beschriebene Induktion von Diskursinformation kann entweder auf Paralleltexten oder bilingualen Wörterbüchern aufbauen, z.B. mit



ACoLi Dictionary Graph: 430 Sprachen
(Chiarcos et al., LREC-2020)
https://github.com/acoli-repo/acoli-dicts

German, 16th c. (deu)	Germanic, external	Early New High German corpus	phrase structure syntax (mirror of an external resource)
German (deu)	Germanic, Biblical	Bible	verse-aligned, CES/XML
Gothic (got)	Germanic, Biblical	Bible	verse-aligned, CES/XML
Icelandic (is)	Germanic, Biblical	Bible	verse-aligned, CES/XML
Middle Low German (gml)	Germanic, Biblical	Bugenhagen's Passion of Christ	text edition
Norwegian (no)	Germanic, Biblical	Bible	verse-aligned, CES/XML
Swedish (sv)	Germanic, Biblical	Bible	verse-aligned, CES/XML
Achiar-Shiwiar (acu)	Biblical	Bible	verse-aligned, CES/XML
Aguaruna (agr)	Biblical	Bible	
Akewaio (ake)	Biblical	Bible	
Albanian (alb)	Biblical	Bible	
Amharic (amh)	Biblical	Bible	
Amuzgo (azm)	Biblical	Bible	
Arabic (ar)	Biblical	Bible	
Armenian (hye)	Biblical	Bible	
Aukan (djk)	Biblical	Bible	
Barasana (bsn)	Biblical	Bible	
Basque (eus)	Biblical	Bible	
Bulgarian (bul)	Biblical	Bible	

ACoLi Bible Corpus:
125 Sprachen Open Source
>700 Sprachen Build Scripts
(Chiarcos et al., LaTeCH-2014)
https://github.com/acoli-repo/acoli-corpora

Nutzen und Nutzung

- Nutzen: Verknüpfung von Informationen
  - Z.B. in Korpus- oder Analyseworkflows
  - Muss gemeinsam mit Nutzern für die Lösung von konkreten Forschungsfragen evaluiert werden
- Nutzung: Backend und Datenvorverarbeitung
  - RDF-Technologie und die Anfragesprache SPARQL richten sich nicht an Geisteswissenschaftler
  - ⇒ graphische Nutzerschnittstellen (z.B. https://github.com/acoli-repo/cqp4rdf)

Anfrage: Subsumptionsinferenz

```
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
```

Diskursmarker → PDTB-Ontologie → OLiA → RST-Ontologie

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

Indirekte Nutzung: Evaluation von Diskurstheorien

- Welche Diskurstheorie liefert die beste sprachübergreifende Generalisierung?

  Welche Annotation liefert die besten Evaluationsergebnisse?

  - Auswertung läuft noch (über 125 Sprachen)
    - Zwischenergebnisse
      - PDTB L.1 – CCR – PDTB L.2, PDTB L.3, ISO SemAF – RST
    - spiegelt aber vor allem wider, wie viele Klassen unterschieden werden

Erweiterung: Geht das auch mit Semantik?

- Inferenz von Diskursannotationen ist ein sehr spezifischer und kleiner Bereich
  - Die Technologie ist aber nicht darauf beschränkt
- Alle Arten lexikalischer Ressourcen, die in OntoLex-Lemon verfügbar sind

  Open Multilingual WordNet (http://compling.hss.ntu.edu.sg/omw/, 34 WordNets)

  ⇒ Automatische Induktion und Disambiguierung von Konzeptannotationen (Synsets)
    - auf Basis von bilingualen Wörterbüchern (Chiarcos et al., TIAD-2020)