Inducing Discourse Marker Inventories from Lexical Knowledge Graphs

Prêt-à-LLOD

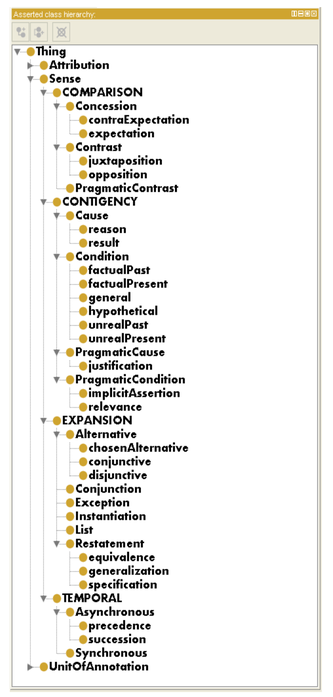
Christian Chiarcos
chiarcos@cs.uni-frankfurt.de
Applied Computational Linguistics (ACoLi), Goethe-Universität Frankfurt, Germany
Institut für Digital Humanities (IDH), Universität zu Köln, Germany

Discourse marker inventories: Discourse markers => discourse relations (e.g., for Penn Discourse Treebank)

```
# FORM          level 0 level 1     level 2                 level 3
"weil"@de       True    CONTINGENCY CONTINGENCY:Cause       CONTINGENCY:Cause:Reason
"weiterhin"@de  True    EXPANSION   EXPANSION:Conjunction   EXPANSION:Conjunction
```

Lexical induction:
Given discourse marker inventories in multiple source languages
Translate discourse markers into target languages, keep (propagate) relations
Aggregate over multiple sources => confidence scores [=> filters]
Iterate with indirect translations

Pre-requisites:
Discourse marker inventories in machine-readable (interoperable) formats
Normalization of discourse relations against a uniform relation taxonomy
Large collection of dictionaries in machine-readable (interoperable) formats

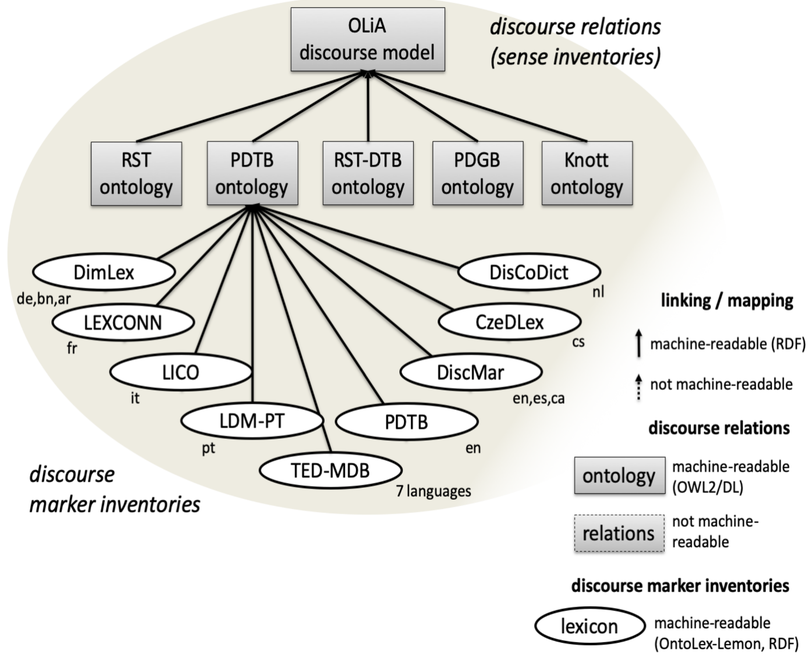


Machine-readable discourse marker inventories

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

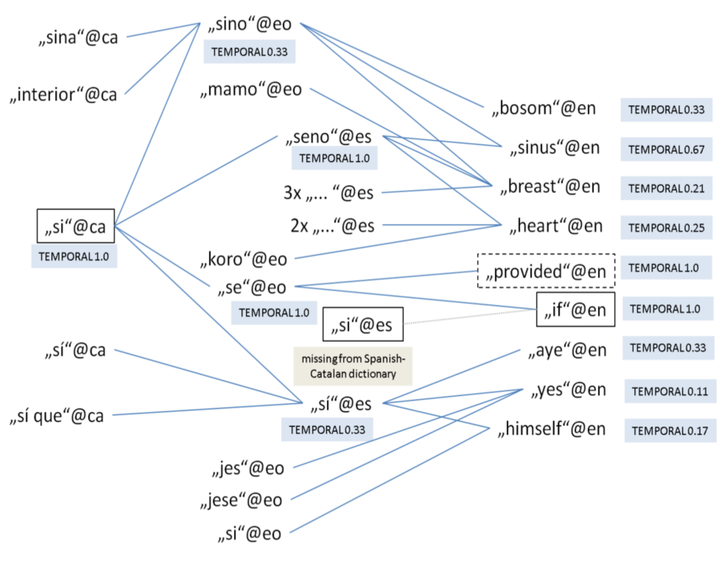
- 16 languages, 19 inventories (Chiarcos & Ionov 2021)
- Partially building on TextLink/Connective-lex.org
- **Data model**: OntoLex
- **Formats**: RDF (=> TSV, with SPARQL)
- **Discourse relation taxonomies**: **PDTB**, RST, CCR

  multiple taxonomies linked with an overarching ontology
  => "translate" between frameworks, using the shortest path



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)



Machine-readable dictionaries

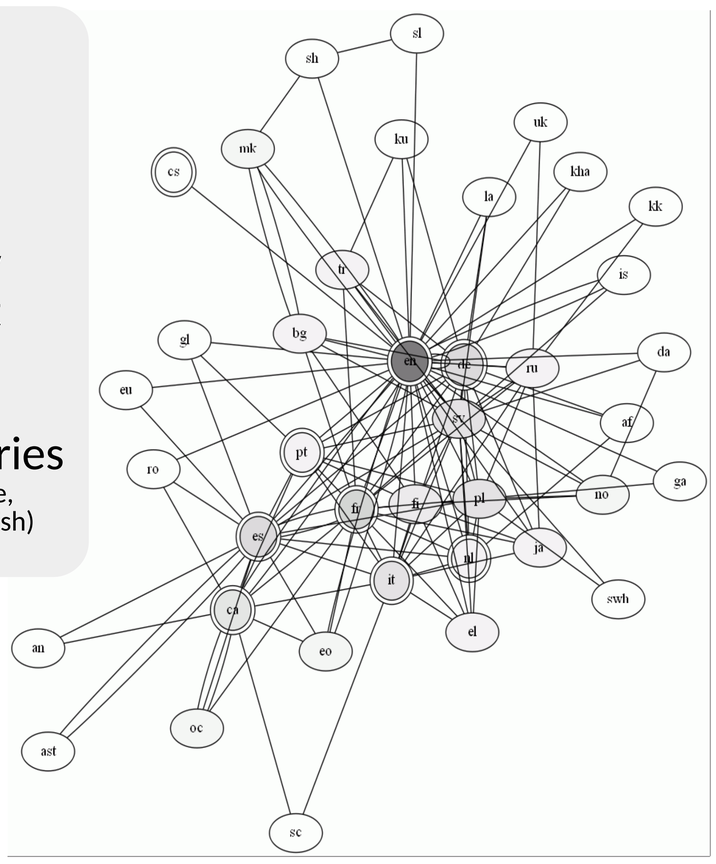
http://github.com/acoli-repo/acoli-dicts

- 430+ languages, 3000+ bi-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

Experimental Setup

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

- 11 inventories, 9 languages
- PDTB, each level separately
- evaluate prec, rec, f against target inventories
- Publish 10 induced inventories
  (Bulgarian, Greek, Esperanto, Finnish, Japanese, Norwegian, Polish, Russian, Swedish and Turkish)



Results (also see paper)

- Direct induction (e.g., from/to English) yields best results, but depends on dictionary quality
  - Apertium: "gold standard", but only 2 inventory languages with links to English
  - FreeDict/MUSE: insufficient coverage
- 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 generally small, < 1000 entries)