Linking Discourse Marker Inventories

Christian Chiarcos
Applied Computational Linguistics Lab, Goethe University Frankfurt, Germany

Maxim Ionov
Applied Computational Linguistics Lab, Goethe University Frankfurt, Germany

Abstract

The paper describes the first comprehensive edition of machine-readable discourse marker lexicons. Discourse markers such as *and*, *because*, *but*, *though* or *thereafter* are essential communicative signals in human conversation, as they indicate how an utterance relates to its communicative context. As much of this information is implicit or expressed differently in different languages, discourse parsing, context-adequate natural language generation and machine translation are considered particularly challenging aspects of Natural Language Processing. Providing this data in machine-readable, standard-compliant form will thus facilitate such technical tasks, and moreover, allow to explore techniques for translation inference to be applied to this particular group of lexical resources that was previously largely neglected in the context of Linguistic Linked (Open) Data.

2012 ACM Subject Classification Computing methodologies → Discourse, dialogue and pragmatics; Information systems → Graph-based database models

Keywords and phrases discourse processing, discourse markers, linked data, OntoLex, OLiA

Digital Object Identifier 10.4230/OASIcs.LDK.2021.40

Supplementary Material *Software (Source Code)*:
`https://github.com/acoli-repo/rdf4discourse`

Funding This work was funded by the project "Prêt-à-LLOD" within the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825182, as well as the project "Linked Open Dictionaries" (LiODi), funded within the eHumanities program of the German Ministry of Education and Science (BMBF, 2015-2021).

1 Motivation and Background

Natural language does not exist in isolation, but always fulfills a communicative purpose, be it to inform an addressee about a specific state of affairs, to motivate them to perform certain acts, to bond or to interact with them otherwise (e.g., convince the addressee of a certain belief). Much of this information, however, resides outside the scope of classical machine-learning based natural language processing: off-the-shelf NLP tools tended to focus on sentences, their components and the grammatical (and semantic) relations between them. With the rising maturity of solutions for more elementary NLP tasks, the automated processing of pragmatics and discourse information did, however, come back into the focus of the discipline and has been subject to a considerable number of shared tasks in recent years, e.g., the 2016 CoNLL Shared Task on Shallow Discourse Parsing[1], the 2019 Shared Task on Discourse Representation Structure Parsing [1] and others.

Discourse markers are a key to the analysis of discourse structure as they represent explicit (albeit not unambiguous) signals of semantic or pragmatic relations that link an utterance with its communicative context (discourse relations), and this has been explored to synthesize training data [39], and is generally considered to be a fast way to light-weight, practical discourse annotation [9].

[1] See `https://www.conll.org/previous-tasks`.

As the result of the COST Action IS1312 "Structuring Discourse in Multilingual Europe" (TextLink, 2014-2018),[2] a considerable number of multilingual discourse marker lexicons has been produced [40], largely following the model of the German DimLex collection [41], but mapped to the sense inventory of the Penn Discourse Treebank [36]. Building on this work and other discourse marker inventories, this paper describes the publication of an interlinked, multilingual discourse marker lexicons on the basis of machine- rather than human-readable form in accordance with web standards and best practices established in computational lexicography, namely as (Linguistic) Linked Open Data [19] and in conformance to OntoLex-Lemon [20].[3]

Motivation for doing so is two-fold: On the one hand, discourse markers inventory becomes more easily accessible for its potential use by off-the-shelf tools, both in individual data sets and as a multilingual graph. The format of the original data sets has considerable variation, even within the TextLink data set: Even though all TextLink discourse marker lexicons are available as XML using the original DimLex lexicon format as a template, they do not adhere to a consistent schema, many contain language-specific extensions, not all are XML-valid, and certain language editions even went so far to translate the original English element and attribute names into the object language. In the language-specific sense inventories, while all based on the Penn Discourse Treebank [36, PDTB], versions 2.0 or 3.0, we also see a certain degree of variation. As a result, this data, while being unquestionably valuable, cannot be directly applied for any NLP task. A better curated version of this data does exist as part of the "Connective-lex" database [40],[4] but the database provides human-readable information only and without explicit licensing information (i.e., restricted).

On the other hand, we also aimed to create further links between discourse marker lexicons and more general lexical resources already available in OntoLex-Lemon. This includes, for example, the Open Multilingual WordNet[5], linked across different languages by means of the Collaborative Interlingual Index[6] [6]. More relevant for the specific case of discourse marker lexicons may be, however, general bidictionaries as provided, for example, as part of the ACoLi Dictionary Graph [16], a large-scale collection of more than 3000 machine-readable bi-dictionaries in OntoLex-Lemon, covering more than 400 language varieties. By linking with this kind of data, it becomes possible to explore techniques to extrapolate discourse marker inventories for low-resource languages by means of techniques as similarly applied for translation inference [38, 28].

Related research on modelling discourse relations and discourse connectives in RDF and/or as Linked Data exists in the form of a suggested discourse extension [27] of the General Ontology of Linguistic Description [23] employed for discourse parsing [4]. At the time of writing, this resource is no longer publicly available but can be partially reconstructed from associated publications [31]. The FRED machine reading system [26] produces OWL output from an NLP stack that also incorporates an off-the-shelf discourse parser [7], but it does not seem[7] to provide a vocabulary for discourse relations.

---

[2] `http://textlink.ii.metu.edu.tr/`

[3] We acknowledge that other authors established a level of machine-readability in earlier research by providing discourse marker inventories in XML rather than printed form [40]. While this establishes machine-readable syntax, we build and extend this work by establishing machine-readable *semantics*.

[4] `http://connective-lex.info/`

[5] `http://compling.hss.ntu.edu.sg/omw/`

[6] `https://github.com/globalwordnet/ili`

[7] The full vocabulary of the FRED system is not publicly documented. The observations above are insights obtained from example queries.

2 Discourse markers and discourse marker lexicons

Discourse markers, also referred to as discourse cues or discourse connectives, do not constitute a homogenous class of grammatical devices in most languages, but rather involve different aspects of grammar, in particular, if described from a cross-linguistic perspective. Accordingly, what constitutes a discourse marker may be defined differently in different theoretical frameworks and for different languages. In most European languages, prototypical discourse markers include conjunctions (such as English *and*, *but*, *if*, etc.), adverbials (such as *thereafter*, *so*), interjections (e.g., *indeed*), but can also be phrasal expressions (*in order to*). In addition to this, certain uses of punctuation (in written language) can be considered to serve as discourse cues, e.g., commas (as markers of lists or enumerations) or hyphens (as markers of contrast or elaboration). Morphological features may serve as discourse cue as well. In this paper, however, we focus on *lexical* discourse markers, i.e., expressions consisting of one or multiple lexemes.

We follow the Penn Discourse Treebank [36, PDTB] in assuming that discourse markers trigger (or indicate) the discourse relation that connects the (proposition expressed by the) local utterance (ARG2 in PDTB terminology, the utterance that contains the discourse marker) with an element in the context (ARG1 in PDTB terminology), so that the type of discourse relation is taken to be the sense of this relation. A discourse marker lexicon is then defined as a dictionary of discourse markers that minimally provides the form(s) of the discourse marker along with one or multiple discourse relations, as well as additional information, e.g., grammatical features, information about uses of the expression other than as discourse marker, frequency and usage information, provenance. It is to be noted that the discourse relations under consideration should be defined as a closed set with fixed identifiers, e.g., defined by an annotation manual. In particular, occasional, but often unsystematic remarks about uses of adverbs as found in traditional dictionaries (e.g., "adversative") are not sufficient to qualify for a discourse marker inventory.

In that sense, a minimal resource that qualifies as discourse marker lexicon is, for example, an aggregate excerpt from a discourse-annotated corpora that lists discourse markers along with the discourse relations these co-occur with, optionally with frequency information. Provided in machine-readable form, such information is an essential tool in technical challenges such as discourse parsing, natural language understanding and natural language generation, and this is where we see the primary application of the data addressed here.

Designated discourse marker lexicons in this sense have been produced since the 1990s, with early examples represented by Alistair Knott [30] and Stede and Umbach [41]. Knott's discourse marker lexicon is available as an appendix to his PhD thesis, and, effectively, has been represented as a plain list. DimLex, the model of Stede and Umbach, originally applied to data from German and English, has become particularly influential in the context of the TextLink initiative, which led to the creation of a relatively consistent set of multilingual discourse marker lexicons. By "relatively consistent", we mean that data is available in XML formats (inspired by the original DimLex XML format, but with language-specific adaptations), and that their sense information has been normalized against the discourse relation inventory of the Penn Discourse Treebank [36, PDTB]. However, the data is far from uniform, most have TextLink dictionaries have been updated to PDTB 3.0 specifications, but some remain at PDTB 2.0 (and the Czech and French datasets uses their own relation inventories, which we mapped as part of the conversion), and likewise, that there is variation in the XML format(s) being used, so that there is no DTD or schema that all these can be validated against. At the core of our data are the following discourse marker lexicons:

DimLex German [41], CC-BY-NC-SA 4.0; extended to Arabic and Bangla [21]; DimLex-XML, PDTB 3.0 relations.

PDTB English, excerpt from Penn Discourse Treebank guidelines [36]; DimLex-XML, PDTB 3.0 relations.

LICO Italian [24], CC-BY 4.0; modified DimLex-XML, PDTB 2.0/3.0 relations.

CzeDLex Czech, bootstrapped from Prague Discourse Treebank 2.0 [34], CC-BY-NC-SA 4.0; PML-XML, PDiT 2.0 relations [45]

LDM-PT Portuguese [33], CC-BY-NC-SA 4.0; DimLex-based XML, PDTB 3.0 relations.

LexConn French [37]; DimLex-inspired XML, SDRT relations [3].

DisCoDict Dutch [8], CC-BY-NC-SA 4.0; DimLex-XML, PDTB 3.0 relations.

A curated version of this data with extensions, consolidated formats and PDTB 3.0 sense linking is accessible from `http://connective-lex.info/`, but for browsing and search only, not for download. We did consult an older version of this data in a partially consolidated state as available from `https://github.com/discourse-lab/Connective-Lex.info`. Using this as a basis we performed format consolidation and linking to PDTB 2.0 for DimLex, DimLex-Arabic, DimLex-Bangla, and LDM-PT. Note that we went for PDTB 2.0 instead of PDTB 3.0 in order to facilitate interoperability with the OLiA Discourse Extensions. CzeDLex and LexConn were converted from the original sources.

Aside from these discourse marker inventories that represent more or less direct results of TextLink, we also converted the DiscMar inventories for English, Spanish and Catalan by Lausa Alonso y Alemany [25], available from `https://cs.famaf.unc.edu.ar/~laura/shallowdisc4summ/discmar/` (HTML format, own relation set), as well as the discourse marker inventory of the TED-Multilingual Discourse Bank (TED-MDB) corpus [43], available under CC-BY from `https://github.com/MurathanKurfali/Ted-MDB-Annotations` (PDTB annotation format, PDTB 2.0/3.0 relations for English, German, Lithuanian, Polish, Portuguese, Russian and Turkish). For the latter, we provide a converter from PDTB annotation files to DimLex-XML, which can subsequently be applied to other PDTB-based corpora such as for Hindi [35] and Chinese [44] that are currently not covered.

For these data sources, we provide a conversion via DimLex-XML to OntoLex-Lemon, and further, a linking with the PDTB 2.0 ontology previously developed as part of the OLiA Discourse Extensions [13]. On this basis, at least two novel modes for querying the relations between discourse markers become possible:

- discourse marker ↦ PDTB concept ↦ discourse marker (from a given discourse marker, retrieve PDTB-equivalent discourse markers)
- discourse marker ↦ PDTB ontology ↦ discourse marker (use the PDTB ontology for imprecise matches, i.e., more general/more specific senses)

A third querying strategy allows to expand the sense information of a discourse marker lexicon, i.e., to apply it for annotation or disambiguation tasks for annotation schemes other than PDTB:

- discourse marker ↦ PDTB ontology ↦ OLiA Discourse Extensions ↦ discourse relations according to other schemes

Although the work described here is grounded on data sets that have been in existence before, with this paper, we describe the first application of Linked Data principles to this kind of data. As a result, improved means of querying local and web-accessible reference data become available only as a result of the conversion and linking activities described in

this paper. As we rely on the general accuracy of the original data, we do not evaluate qualitative performance; instead, subject of evaluation is the capability to formulate and execute these four types of cross-resource queries.

3 From DimLex-XML to OntoLex-Lemon

For the conversion of discourse marker lexicons, we focus on DimLex-XML. Most data sets required considerable pre-processing in order to either consolidate or to produce DimLex-XML, but as a first step, our processing aims to establish a DimLex-conformant level of representation to start with, either from the original XML discourse marker lexicon (DimLex-XML or otherwise), directly from PDTB-style annotations (for TED-MDB), or from a proprietary format (DiscMar).

Using an XSLT 2.0 script, the resulting DimLex file is then transformed to OntoLex-Lemon in Turtle. For reasons of space we do not provide an in-depth description of OntoLex-Lemon, but refer to the original specification [20]. The most relevant OntoLex elements in our context are:

ontolex:LexicalEntry unit of analysis of the lexicon, groups together one or more forms and one or more senses.
ontolex:Form string form of a lexical entry, e.g., written representation.
ontolex:LexicalSense word sense of a particular lexical entry.

Furthermore, a sense can be linked with an externally defined ontological entity by means of `ontolex:reference`. We will use this mechanism to link (lexical entries/senses of) discourse markers with discourse relations.

The OntoLex converter consists of two principal types of conversions, format-specific and generic. Format-specific transformations include:

- For every `entry` element,
  - create an instance of `ontolex:LexicalEnty`.
- For every `orth` element, attach to the entry an `ontolex:Form` by means of either
  - `ontolex:lexicalForm` (for DimLex dialects that do not define canonical forms), or
  - `ontolex:canonicalForm` (for every `orth` element with attribute `canonical="1"`), or
  - `ontolex:otherForm` (for every other `orth` element in a DimLex dialect that defines canonical forms), and
  - assign this form a `ontolex:writtenRep` that contains a language-typed string.
- For every `pdtb3_relation`,
  - create an `ontolex:Sense`, and
  - link it to the lexical entry by means of `ontolex:isSenseOf`.

All other components of the format are converted by generic transformations. For every element that contains (descendants with) attributes or CDATA content:

- identify the element created by the parent element as subject,
- create a property in the `dimlex:` namespace that takes the local name of the current element, and
- assign this property an object, this is either
  - the enclosed text as untyped literal (if the element carries neither attributes nor child elements), or
  - a blank node that serves as subject for properties generated from attributes or (recursively) from child elements.

  XML attributes are likewise preserved as datatype properties with untyped string values.

As namespace for the `dimlex:` elements, we resort to the DimLex DTD `https://github.com/discourse-lab/dimlex/blob/master/DimLex.dtd#`. However, note that several DimLex-style data sets do not validate against this DTD. In this way, we establish core data structures of OntoLex-Lemon, but perform a generic and lossless transformation of XML data structure. This converter is thus capable to support any DimLex dialect and (with minor modifications) related formats. In particular, all resource-specific extensions can be preserved.

The following listing shows the first entry of the German DimLex (with minor omissions):

```
<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>
```

For all XML elements and attributes below `entry`, the generated Turtle preserves this information faithfully (likewise simplified), in the same order and the same embedding depth:

```
  :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 ] ] .
```

In addition to the OntoLex properties, additional information from the XML format is provided by properties from the DimLex namespace that mirror the original structure of the original XML file. Note that in this way, all information of a DimLex entry can be captured in the RDF graph, but only as far as hierarchy and structure are concerned. The order of elements in the XML is lost in the graph, but also not deemed to be essential for subsequent processing.

A disadvantage of this modelling strategy is that (unless all discourse marker inventories validate against the same schema or DTD – which the publicly available data does not) the resulting DimLex vocabulary is open: Every DimLex-XML dialect can introduce novel datatype and object properties, so that it is not possible to provide an exhaustive class diagram of the DimLex vocabulary in RDF. But we capture the information about basic OntoLex data structures in an interoperable way.

4 Linking with the OLiA Discourse Extensions

The Ontologies of Linguistic Annotation [12, OLiA] have been developed to formalize annotation schemes and to link them with reference concepts, originally primarily for corpora with morphosyntactic and syntactic annotation, with regard to which OLiA covers more than 100 languages at the time of writing,[8] but also extended for pragmatic phenomena such as coreference, information structure, discourse structure and discourse relations. These OLiA Discourse Extensions [13] reside in a separate branch of the OLiA ontologies.[9] As they are still considered experimental, but with increasing maturity, they are about to be integrated with the OLiA.

In its conception, OLiA aimed to address what could be called the "standardization gap" of linguistic annotation. That means that a consistent and homogeneous standardization of linguistic annotation would either have to be reductionistic and neglect language specific characteristics (cf. Universal Dependencies tagset), or constantly grow in complexity with every new language added to it (cf. the evolution of morphosyntactic guidelines from EAGLES to MULTEXT-EAST) [14].

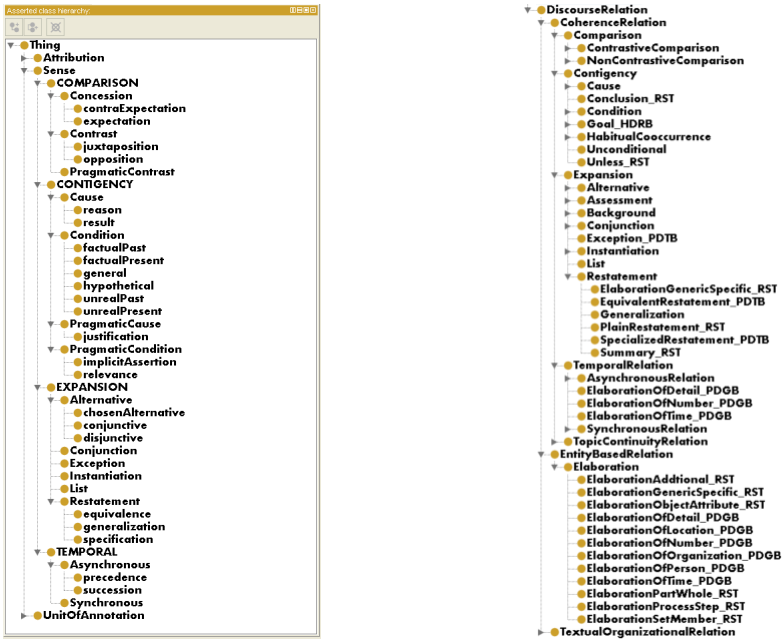
In order to avoid these problems, OLiA introduces an architecture of modular ontologies, formalized in OWL2/DL, to address and to distinguish the different aspects of

- defining concepts and tags relevant for the annotation of a language or a particular language resource (OLiA Annotation Models),
- identifying and defining conventionally used terms (OLiA Reference Model),
- interpreting annotation concepts against reference concepts (OLiA Linking Models, defining `rdfs:subClassOf` relationships between annotation model concepts and reference model concepts), and
- grounding conventionally used terms in reference concepts (external terminology repositories, linking defined as `rdfs:subClassOf` relationships between reference model concepts and externally defined concepts)

With multiple, distinct, but interlinked ontologies, published under CC-BY 3.0 and available under persistent, resolvable URIs, OLiA represents a prototypical application of Linked Data principles to leverage several distributed terminology repositories, and it became subsequently increasingly important as a terminology repository since the conception of the LLOD cloud in 2010, where it represents the central terminology hub for annotation terminology. As of 2020, OLiA is linked with a great number of external terminology repositories, including ISOcat, GOLD, the CLARIN Concept Registry, lexinfo, the Universal Dependency guidelines, the Unimorph guidelines, etc. [15], and it is developed as an open source project under `https://github.com/acoli-repo/olia`.

With the OLiA Discourse Extensions, the approach of modular ontologies and the application of Linguistic Linked Open Data principles to facilitate language resource interoperability can also been extended to the annotation of discourse relations and other aspects of pragmatics. As far as discourse relations are concerned, the OLiA Discourse Extensions cover five annotation schemes based on theoretical work on discourse relations [32, 30, 11, 42, 36] The discourse marker inventories described here are linked to the original PDTB ontology as part of the OLiA Discourse Extensions [13] and available under CC-BY 3.0 from `http://purl.org/olia/discourse/discourse.PDTB.owl`.

[8] `http://purl.org/olia`
[9] `http://purl.org/olia/discourse`



(a) PDTB ontology as provided by the OLiA Discourse Extensions, visualized using Protégé.

(b) OLiA discourse model: Reference concepts for the OLiA Discourse Extensions.

Figure 1 PDTB ontology and OLiA Discourse Extensions.

The PDTB ontology is summarized in Fig. 1a. We focus on the `pdtb:Sense` branch alone, where PDTB asserts the existence of four major types of discourse relations, contrastive relations (COMPARISON), causal relations (CONTIGENCY), temporal relations (TEMPORAL) and additive relations (EXPANSION), with two levels of further refinement. In real-world annotation, an annotator may decide to assign a discourse marker the most specific relation they find in that taxonomy (e.g., *reason*), but likewise, a more abstract relation if none of the subclasses match precisely or seem to be equally applicable (e.g., *Cause*, or even *CONTIGENCY*). We take this to be an implicative hierarchy, i.e., that any more specific discourse relation automatically entails the applicability of a more generic one – albeit this kind of reasoning seems to be rarely applied in current PDTB practice. Instead, the hierarchy has been exploited for evaluating the performance of discourse parsing, where accuracy can be evaluated against different levels of granularity, ranging from top-level (4 discourse relations plus entity relations and no relation) over second-level relations (15 discourse relations) to the full inventory. The discourse marker inventories are linked with the PDTB ontology by means of a simple SPARQL update: If the label of a discourse relation of the PDTB ontology matches the literal value of `dimlex:sense`, insert an `ontolex:reference`, i.e., link the PDTB ontology as an external ontology:

```
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))       };
```

For the linked version of the data set, we perform an additional pruning step and omit all properties from the `dimlex:` namespace, i.e., aspects of the original XML content that have, so far, not been interpreted into machine-readable information. (Remember that the

Table 1 Statistics and accessibility information for discourse marker inventories, PDTB-linked, OntoLex-Lemon edition.

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

`dimlex` namespace is merely a placeholder for generic XML information that has not found an interpretation against OntoLex or another RDF vocabulary.) However, the original RDF data is preserved and can be consulted for future extensions.

Table 1 gives an overview over the linking statistics and also provides the persistent URIs for the respective linked data sets. Note that these URIs resolve, and that machine-readable license information is provided, so that the result of conversion and linking represents fully qualified Linguistic Linked (Open) Data.

All resulting data is available under the respective original license from our GitHub repository (`https://github.com/acoli-repo/rdf4discourse/tree/master/discourse-markers/linked`). After conversion and linking, the resulting data has been enriched with machine-readable metadata about the respective license (`dct:license`), and the location of the original data (`dcr:source`). Human-readable details on attribution are provided as `rdf:comment` of the `lime:Lexicon` element that represents the respective discourse marker inventory. Note that not all data sets have an explicit license statement. This includes LexConn, DiscMar and Arabic. As for the first three, the information contained in them corresponds *exactly* to a respective appendix of the accompanying documentation [36, 37, 2]. We consider this as unproblematic in terms of copyright, as the discourse marker inventories created on this basis represent collections of (fully attributed) non-literal quotations. In order to preserve the intended band-width of scientific citations, we assert a CC-BY-NC 3.0 license for these, using the original literature references as attributions.[10] The copyright situation of the Arabic discourse marker lexicon is still unresolved, full attribution is provided, but in case of complaints, it will be withdrawn from the public release.

[10] We adopt CC-BY-NC 3.0 instead of CC-BY-NC 4.0 as the 3.0 BY clause allows authors to enforce the use of a specific title, i.e., a particular form of citation, rather than alternative means of attribution (e.g., by publication URI).

5 Querying

As mentioned before, our evaluation consists of demonstrating the capability to query discourse marker inventories in combination with discourse relation inventories, both the PTDB ontology and the OLiA Discourse Extensions.

With discourse marker inventories, sense definitions and annotation schemes interconnected by means of Linked Data technology, it now becomes possible to traverse the paths in a graph, e.g., in order to retrieve translations or alternative lexicalizations of discourse markers. Note that this functionality is currently not provided by the Connective-Lex database [40], so that this is a novel functionality. The following query retrieves an English word from the DiscMar inventory and its PDTB sense:[11]

```
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>
PREFIX pdtb: <http://purl.org/olia/discourse/discourse.PDTB.owl#>
PREFIX rst: <http://purl.org/olia/discourse/discourse.RST.owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

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
```

Simplifying the query using property paths and adding a second path with a different language filter, we can now apply this query to derive translations, e.g., between English and German connectives:

```
SELECT distinct ?en ?pdtb ?de  # prefixes are omitted for space reasons
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
```

As a general rule, we would expect that DiscMar results are more coarse-grained than DimLex results. In the SPARQL query, this can be captured by extending the search to retrieve DimLex lexicalization for indirect *subclasses of* DiscMar entries (assuming that the PDTB ontology is loaded in the default graph):

```
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
```

[11] This query as well as all subsequent queries can be directly executed with the online SPARQL service provided under `http://www.sparql.org/sparql.html` and have been tested for this purpose. No additional configuration is necessary, as the `FROM` statements indicates the RDF graphs to read from. Alternatively, they can run on *any* local SPARQL end point once the ontologies are loaded.

This query now returns 7,040 different translation pairs for German and English, whereas the former query retrieved only 828 translation pairs. However, note that many of these translations are imprecise because of differences in granularity. As an example, DimLex differentiates between subclasses of causal relations (PDTB *reason* and *result*), whereas DiscMar only identifies clausal relations as PDTB *Cause*. Transitive queries, e.g., along the `rdfs:subClassOf` axis as expressed by the Kleene star in the example, are an efficient way to deal with differences in granularity. An alternative is to enable the RDFS entailment regime in the SPARQL end point. Then, the original query (without the Kleene star) does return a comparable result.[12]

But the linked graph can also be used in other ways. If the PDTB linking model (`http://purl.org/olia/discourse/discourse.PDTB-link.rdf`) is imported into the default graph, we arrive at reference model concepts from the OLiA Discourse Extensions. with other linking and annotation models connected, it becomes possible, then, for example, to "translate" the PDTB relations to RST relations [32], illustrated for the marker *because* according to the PDTB inventory:

```
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.

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

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

This query returns 11 possible RST relations and also gives information about the path that connects them with the original definition:

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

[12] The difference is in the binding for the variable `?pdtb`: In the query with the Kleene star, we retrieve the more specific sense, as expressed in DimLex. In the query without the Kleene star but RDFS entailment enabled, we retrieve the more general sense, as the system can infer the superclass `pdtb:Causal` from the DimLex-provided senses `pdtb:reason` and `pdtb:result`. The translation pairs of English and German expressions, however, are identical.

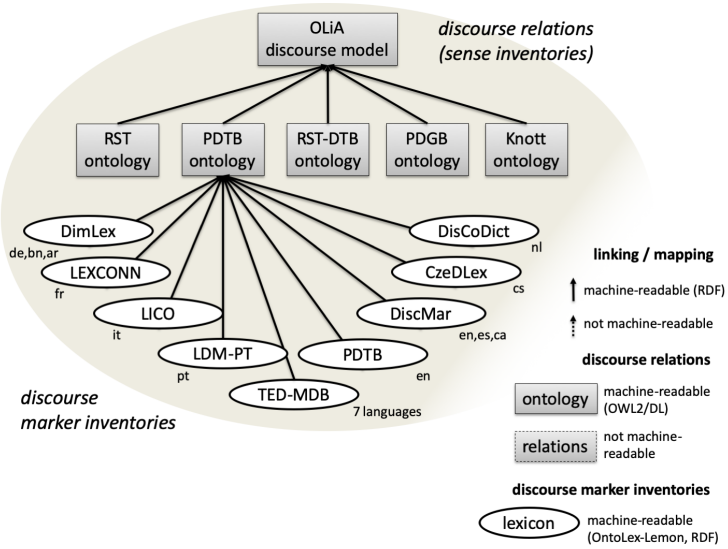


Figure 2 Discourse marker and discourse relation inventories as Linked Data.

Such queries can be further refined if confidence scores or relative sense frequencies are taken into consideration. From the current set of discourse marker lexicons, however, only PDTB and the German DimLex provide such information. For encoding frequency information at a later stage of development, we plan to apply the OntoLex module for Frequency, Attestation and Corpus Information that is currently being developed [17, OntoLex-FrAC].

Overall, we have been able to show that the linked data edition of the discourse marker lexicons and its linking with the OLiA Discourse Extensions provide improved means of querying this data. The example queries have addressed three types of queries:

- discourse marker ↦ PDTB concept ↦ discourse marker (from a given discourse marker, retrieve PDTB-equivalent discourse markers)
- discourse marker ↦ PDTB ontology ↦ discourse marker (use the PDTB ontology for imprecise matches, i.e., more general/more specific senses)
- discourse marker ↦ PDTB ontology ↦ OLiA discourse model ↦ RST ("translate" PDTB relations into another theoretical framework)

To the best of our knowledge, none of these functionalities have been possible before.

A specific benefit of publishing this data as Linked Open Data and under resolvable and persistent URLs is that such queries can be executed independently from any local data base installation. Instead, generic web tools such as the "general purpose SPARQL processor" from `http://sparql.org` can be employed to execute such queries.

6 Summary and Outlook

In this paper we described the conversion of existing discourse marker lexicons into RDF, their linking with the PDTB ontology of the OLiA Discourse Extensions and their publication as Linked Data. This contribution is an important step in formation of a small group of discourse-related resources within the Linguistic Linked Open Data cloud. The general structure and the relation between the resources introduced or described in this paper is illustrated in Fig. 2.

The respective discourse marker lexicons are provided as plain RDF dumps (preserving all information from the original XML file in the `dimlex:` namespace, but lacking PDTB linking) and linked OntoLex-Lemon data sets (preserving only statements that involve OntoLex properties or classes, extended with `ontolex:reference` links to the PDTB ontology. As part of the conversion, we introduced BCP47 language tags to identify the participating languages. It is thus possible to load all discourse marker lexicons into a single RDF graph and query, for example, for correspondences between languages. Moreover, machine-readable language identification and adherence to web standards allows us now to explore synergies with other OntoLex-Lemon datasets, e.g., the ACoLi Dictionary Graph [16], e.g., to enrich conventional bilingual dictionaries with machine-readable sense information for discourse markers (in this regard, the PDTB ontology, and the OLiA discourse model can serve a similar function as WordNet for lexical semantics). Likewise, it becomes possible now to explore conventional dictionaries to bootstrap PDTB-linked discourse marker inventories for other languages.

With this kind of data, machine-readable inventories of discourse markers, discourse relations and corpora (resp., their annotation schemes, as formalized in the OLiA Discourse Extensions), it now becomes possible to integrate them into local applications, general web tools, or perform queries against them, as well as enrich them with further information other Linguistic Linked Open Data sets, e.g., general purpose dictionaries provided in OntoLex-Lemon. As the same time, we would like to emphasize that we see prospective users of this technology not so much among specialists in discourse and semantics, but more among developers of technical solutions for studying discourse as well as NLP specialists and knowledge engineers interested in more advanced levels of linguistic analysis and semantic relations beyond individual sentences. As far as the field of discourse studies is concerned, we consider this implementation to provide a practical benefit, but we also assume that general web technologies, e.g., the RDF data model, the Turtle format, and the SPARQL query language, require an additional layer of abstraction in order to be effective tools in the hands of linguist. Such tools are becoming increasingly available for different aspects of linguistic inquiry (e.g. [5, 22, 29]). For discourse studies, such an infrastructure currently does not exist, nor is the use of RDF technologies particularly established in the field, but it is to be noted that the potential for such an application is enormous, as Linked Data provides natural support for standoff and multi-layer annotations [10], all of these are notorious problems for discourse studies [18], as well as for information integration across heterogeneous and distributed data in general, as demonstrated here for discourse marker inventories. By publishing essential data for this field in accordance with Linked Data principles, our work represents an initial step towards the development of advanced tools and improved information aggregation for applications in discourse parsing and discourse analysis.

The OLiA discourse extensions, including the PDTB ontology are published under `http://purl.org/olia/discourse` and available as a code bundle under CC-BY 3.0 from `https://github.com/acoli-repo/olia/tree/master/owl/experimental/discourse`. The discourse marker inventories and the scripts to produce them are currently available under a Apache 2.0 license from `https://github.com/acoli-repo/rdf4discourse`. The data itself remains under the same license as the original data as described above. Code and data is publicly available from our GitHub repository[13] as Open Source under an Apache 2.0 license.

[13] `https://github.com/acoli-repo/rdf4discourse`

References

1 Lasha Abzianidze, Rik van Noord, Hessel Haagsma, and Johan Bos. The first shared task on discourse representation structure parsing. In *Proc. of the IWCS Shared Task on Semantic Parsing*, 2019.

2 Laura Alonso. *Representing discourse for automatic text summarization via shallow NLP techniques*. PhD thesis, Tesis doctoral. Barcelona: Universitat de Barcelona, 2005.

3 Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. *Logics of conversation*. Cambridge University Press, 2003.

4 Maja Bärenfänger, Mirco Hilbert, Henning Lobin, and Harald Lüngen. Using owl ontologies in discourse parsing. *OTT'06*, 1:87, 2007.

5 Andrea Bellandi, Emiliano Giovannetti, Silvia Piccini, and Anja Weingart. Developing lexo: a collaborative editor of multilingual lexica and termino-ontological resources in the humanities. In *LOTKS-2017*, 2017.

6 Francis Bond and Kyonghee Paik. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71, Matsue, 2012.

7 Johan Bos. Open-domain semantic parsing with boxer. In *Proceedings of the 20th nordic conference of computational linguistics (NODALIDA 2015)*, pages 301–304, 2015.

8 Peter Bourgonje, Jet Hoek, Jacqueline Evers-Vermeul, Gisela Redeker, Ted Sanders, and Manfred Stede. Constructing a lexicon of dutch discourse connectives. *Computational Linguistics in the Netherlands Journal*, 8:163–175, 2018.

9 Peter Bourgonje and Manfred Stede. Exploiting a lexical resource for discourse connective disambiguation in german. In *Proc. of the 28th International Conference on Computational Linguistics*, pages 5737–5748, 2020.

10 Aljoscha Burchardt, Sebastian Padó, Dennis Spohr, Anette Frank, and Ulrich Heid. Formalising Multi-layer Corpora in OWL/DL – Lexicon Modelling, Querying and Consistency Control. In *Proc. of the 3rd International Joint Conf on NLP (IJCNLP)*, pages 389–396, Hyderabad, India, 2008.

11 Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie W. Smith, editors, *Current and New Directions in Discourse and Dialogue*, Text, Speech, and Language Technology; 22, chapter 5. Kluwer, Dordrecht, 2003.

12 C. Chiarcos and M. Sukhareva. OLiA - Ontologies of Linguistic Annotation. *Semantic Web Journal*, 518:379–386, 2015.

13 Christian Chiarcos. Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In *LREC*, pages 4569–4577. Citeseer, 2014.

14 Christian Chiarcos and Tomaz Erjavec. OWL/DL formalization of the multext-east morphosyntactic specifications. In *LAW-2011*, pages 11–20, Portland, Oregon, USA, June 2011. ACL.

15 Christian Chiarcos, Christian Fäth, and Frank Abromeit. Annotation interoperability for the Post-ISOCat era. In *LREC-2020*, pages 5668–5677, 2020.

16 Christian Chiarcos, Christian Fäth, and Maxim Ionov. The ACoLi dictionary graph. In *LREC-2020*, pages 3281–3290, 2020.

17 Christian Chiarcos, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck, and John Philip McCrae. Modelling frequency and attestations for ontolex-lemon. In *Globalex-2020*, pages 1–9, 2020.

18 Christian Chiarcos, Julia Ritz, and Manfred Stede. Querying and visualizing coreference annotation in multi-layer corpora. In *DAARC-2011*, pages 80–92, 2011.

19 Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. *Linguistic Linked Data*. Springer, 2020.

20 Philipp Cimiano, John P. McCrae, and Paul Buitelaar. Lexicon Model for Ontologies. Technical report, W3C Community Report, 10 May 2016, 2016.

21 Debopam Das, Manfred Stede, Soumya Sankar Ghosh, and Lahari Chatterjee. DiMLex-Bangla: A lexicon of Bangla discourse connectives. In *LREC*, pages 1097–1102, Marseille, France, 2020. ELRA.

22 Gimena del Rio Riande and Valeria Vitale. Recogito-in-a-box: From annotation to digital edition. *Modern Languages Open*, 2020.

23 S. Farrar and D.T. Langendoen. A linguistic ontology for the semantic web. *Glot International*, 7(3):97–100, 2003.

24 Anna Feltracco, Elisabetta Jezek, Bernardo Magnini, and Manfred Stede. Lico: A lexicon of italian connectives. *CLiC it*, page 141, 2016.

25 Maria Fuentes Fort. *A flexible multitask summarizer for documents from different media, domain and language*. Universitat Politècnica de Catalunya, 2008.

26 Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. Semantic web machine reading with fred. *Semantic Web*, 8(6):873–893, 2017.

27 D. Goecke, H. Lüngen, F. Sasaki, A. Witt, and S. Farrar. GOLD and discourse: Domain-and community-specific extensions. In *E-MELD Workshop*, Cambridge, Massachusetts, July 2005.

28 Jorge Gracia, Besim Kabashi, Ilan Kernerman, Marta Lanau-Coronas, and Dorielle Lonke. Results of the translation inference across dictionaries 2019 shared task. In *TIAD*, pages 1–12, 2019.

29 Maxim Ionov, Florian Stein, Sagar Sehgal, and Christian Chiarcos. cqp4rdf: Towards a suite for rdf-based corpus linguistics. In *ESWC-2020*, pages 115–121. Springer, 2020.

30 Alistair Knott and Robert Dale. Using linguistic phenomena to motivate a set of coherence relations. *Discourse processes*, 18(1):35–62, 1994.

31 Harald Lüngen, Maja Bärenfänger, Mirco Hilbert, Henning Lobin, and Csilla Puskás. Discourse relations and document structure. In *Linguistic modeling of information and markup languages*, pages 97–123. Springer, 2010.

32 William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

33 Amália Mendes, Iria del Rio, Manfred Stede, and Felix Dombek. A lexicon of discourse markers for portuguese–ldm-pt. In *LREC-2018*, pages 4379–4384, 2018.

34 Jiří Mírovský, Pavlína Synková, Magdaléna Rysová, and Lucie Poláková. CzeDLex 0.5, 2017.

35 Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. The Hindi discourse relation bank. In *LAW III*, pages 158–161, 2009.

36 Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *LREC-2008*, pages 2961–2968, Marrakech, Morocco, 2008.

37 Charlotte Roze, Laurence Danlos, and Philippe Muller. Lexconn: a french lexicon of discourse connectives. *Discours*, 10, 2012.

38 Stephen Soderland, Oren Etzioni, Daniel S Weld, Kobi Reiter, Michael Skinner, Marcus Sammer, Jeff Bilmes, et al. Panlingual lexical translation via probabilistic inference. *Artificial Intelligence*, 174(9-10):619–637, 2010.

39 Caroline Sporleder and Alex Lascarides. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(3):369, 2008.

40 Manfred Stede, Tatjana Scheffler, and Amália Mendes. Connective-lex: A web-based multilingual lexical resource for connectives. *Discours*, 24, 2019.

41 Manfred Stede and Carla Umbach. Dimlex: A lexicon of discourse markers for text generation and understanding. In *COLING-ACL-1998*, pages 1238–1242, 1998.

42 Florian Wolf and Edward Gibson. Representing Discourse Coherence: A Corpus-Based Study. *Computational Linguistics*, 31(2):249–287, 2005.

43 Deniz Zeyrek, Amalia Mendes, Yulia Grishina, Murathan Kurfali, Samuel Gibbon, and Maciej Ogrodniczuk. Ted multilingual discourse bank (ted-mdb): a parallel corpus annotated in the PDTB style. *LREC-2019*, pages 1–38, 2019.

44 Yuping Zhou and Nianwen Xue. PDTB-style discourse annotation of chinese text. In *ACL-2012*, pages 69–77, 2012.

45 Šárka Zikánová, Jiří Mírovskỳ, and Pavlína Synková. Explicit and implicit discourse relations in the prague discourse treebank. In *TSD-2019*, pages 236–248. Springer, 2019.