Building an Owl-ontology for representing, linking and querying SemAF discourse annotations

Christian Chiarcos, Purificação Silvano, Mariana Damova, Giedre Valunaite Oleškeviciene, Chaya Liebeskind, Dimitar Trajanov, Ciprian-Octavian Truică, Elena-Simona Apostol, Anna Baczkowska

Angaben zur Veröffentlichung / Publication details:

Chiarcos, Christian, Purificação Silvano, Mariana Damova, Giedre Valunaite Oleškeviciene, Chaya Liebeskind, Dimitar Trajanov, Ciprian-Octavian Truică, Elena-Simona Apostol, and Anna Baczkowska. 2023. "Building an Owl-ontology for representing, linking and querying SemAF discourse annotations." *Rasprave Instituta za hrvatski jezik i jezikoslovlje* 49 (1): 117–36. https://doi.org/10.31724/rihjj.49.1.6.

UDK: 81.322
004.8:81
Izvorni znanstveni rad
Rukopis primljen 10. XII. 2022.
Prihvaćen za tisak 6. VII. 2023.
https://doi.org/10.31724/rihjj.49.1.6

Christian Chiarcos[1], Purificação Silvano[2], Mariana Damova[3], Giedre Valunaite Oleškeviciene[4], Chaya Liebeskind[5], Dimitar Trajanov[6], Ciprian-Octavian Truică[7], Elena-Simona Apostol[8], Anna Bączkowska[9]

christian.chiarcos@uni-a.de, *msilvano@letras.up.pt*, *mariana.damova@mozajka.co*, *gvalunaite@mruni.eu*, *liebchaya@gmail.com*, *dimitar.trajanov@finki.ukim.mk*, *ciprian-octavian.truica@it.uu.se*, *elena.apostol@upb.ro*, *anna.baczkowska@ug.edu.pl*

BUILDING AN OWL-ONTOLOGY FOR REPRESENTING, LINKING AND QUERYING SEMAF DISCOURSE ANNOTATIONS

Linguistic Linked Open Data (LLOD) are technologies that provide a powerful instrument for representing and interpreting language phenomena on a web-scale. The main objective of this paper is to demonstrate how LLOD technologies can be applied to represent and annotate a corpus composed of multiword discourse markers, and what the effects of this are. In particular, it is our aim to apply semantic web standards such as RDF and OWL for publishing and integrating data. We present a novel scheme for discourse annotation that combines ISO standards describing discourse relations and dialogue acts – ISO DR-Core (ISO 24617-8) and ISO-Dialogue Acts (ISO 24617-2) in 9 languages (cf. Silvano and Damova 2022; Silvano et al. 2022). We develop an OWL ontology to formalize that scheme, provide a newly annotated dataset and link its RDF edition with the ontology. Consequently, we describe the conjoint querying of the ontology and the annotations by means of SPARQL, the standard query language for the web of data. The ultimate result is that we are able to perform queries over multiple, interlinked datasets with complex internal structure. This is a first, but essential step, in developing novel, powerful, and groundbreaking means for the corpus-based study of multilingual discourse, communication analysis, or attitudes

[1] Applied Computational Linguistics, University of Augsburg;[2] Faculty of Arts and Humanities of the University of Porto, Centre of Linguistics of the University of Porto; [3]Mozaika, Ltd.; [4]Mykolas Romeris University; [5] Jerusalem College of Technology; [6]Department for Information Systems and Network Technologies, Faculty of Computer Science and Engineering Ss. Cyril and Methodius University; [7]Department of Information Technology, Uppsala University, Sweden; [8]Faculty of Automatic Control and Computers, University Politehnica of Bucharest; [9]Institute of English and American Studies / Eviden, Big Data and Security
Corresponding author Mariana Damova, orcid.org/0000-0003-3684-4726

discovery.

1. Introduction

1.1. Multiword discourse markers and LLOD

The role of discourse markers is to create coherence linkages between clauses and sentences (Das 2014; Taboada 2006), indicating hesitation, turn-taking, theme changing, marking turn borders, hedging, revealing attitude, managing the connection with the interlocutor, seeking acceptance (Jucker and Ziv 1998), and denoting transitions in conversation and dialogue (Heeman and Allen 1999). They are single-word or multiword expressions (MWE) composed of conjunctions, adverbials, and prepositional phrases (Fraser 2009a). Expressions like *you know*, *you see*, and *I mean* are also considered discourse markers (Schiffrin 2005; Maschler and Schiffrin 2015).

Some academics differentiate between relational and non-relational discourse markers, based on their semantic or interactional function (Crible 2016), and taxonomies for both kinds have been suggested (e.g., Rhetorical Structure Theory (RST) by Mann and Thompson 1988; Cognitive approach to Coherence Relations (CCR) by Sanders et al. 1992; Penn Discourse Treebank (PDTB) by Prasad et al. 2008). Discourse markers have been studied for detecting and analysing discourse relations (e.g., Sanders et al. 1992; Knott and Dale 1993; Marcu 2000; Silvano 2010; Das 2014; Bunt and Prasad 2016; Das and Taboada 2019). As a consequence, a substantial number of annotated corpora with discourse relations signalled by discourse markers were generated, e.g. the RST-DT English corpus (Carlson et al. 2003); the Penn Discourse TreeBank (PDTB, Prasad et al. 2008); and the SDRT Annodis French corpus (Afantenos et al. 2012).

Discourse annotations are highly heterogeneous and specific to distinct subcommunities, along with their specific tools, formats and vocabularies. The application of knowledge representation standards, most notably OWL and RDF, have long been suggested as a means to address such issues, especially for modelling language data and linguistic data categories, e.g., in GOLD (Farrar and Langendoen 2003). In particular, Schmidt et al. (2006), Chiarcos (2008) and

Dimitriadis et al. (2009) argued that overlapping hierarchies and multiple inheritance in the OWL formalization of description logics (OWL/DL) allow researchers to formalize and elucidate complex, overlapping definitions in a more transparent and reusable way. Using RDF for publishing data and publishing it under an open license under resolvable URIs on the web of data allows us to link concepts and data across different frameworks and theories of discourse. This is the very idea of Linguistic Linked Open Data (LLOD).

Although the potential of RDF, OWL and LLOD technology as a more sustainable way of publishing and sharing discourse annotations has been pointed out already by Chiarcos (2014), and RDF and OWL have been applied to discourse annotations even before that (Goecke et al. 2005; Lüngen et al. 2010), it seems that OWL and RDF have never been applied as native data formats for discourse-annotated corpora. Chiarcos et al. (2021) described the linking of such ontologies with discourse marker inventories. However, we are not aware of any discourse-annotated corpus data published in accordance with LLOD. The potential of this technology to overcome structural barriers to compare, integrate and query information from different sources is well-known, and interoperability is a notorious problem in discourse. The work of Burchardt et al. 2008 successfully applied this approach to the conjoint querying of semantic and syntactic annotations and still seems to be best solved with RDF technologies, e.g. Chiarcos and Fäth 2019. In the extension of this technology to the discourse-annotated corpora, we see one of the main contributions of this paper, a native OWL representation of the discourse and dialog specifications of ISO 24617, coupled with the preparation of a parallel, discourse-annotated corpus in accordance with LLOD specifications.

1.2. The multilingual parallel corpus

Our multilingual corpus comprises utterances extracted from TED talks in 9 languages: English, Lithuanian, Bulgarian, Portuguese, Macedonian, Polish, Romanian, Hebrew, Italian and German. The extraction was based on the occurrence of a discourse marker in the selected utterance. The list of discourse markers that guided the selection of utterances consists of multiword expressions, based on Schiffrin 2005 and the classification by Fraser 2009.

The single language corpora counted more than 10K utterances on average, each utterance being uniquely identified with a combination of three types of IDs. As the goal was to provide a multilingual parallel corpus, all 9 language corpora were automatically compared, the intersection of all 9 corpora was singled out, and a parallel corpus of 56 utterances for each language was created. For easier further processing by human experts, the parallel corpus was split into 9 files with aligned utterances in English and in one of the covered languages (cf. Table 1[2]).

Table 1. Structure of the parallel corpus files

Column Name	Cell content
0.1	6240
vid	724
Lid	143
DM EN	Of course
Text EN	And this, of course, is the basis of much of Eastern philosophy,
Context EN	from the other person, is your skin. Remove the skin, you experience that person's touch in your mind. You've dissolved the barrier between you and other human beings. And this, of course, is the basis of much of Eastern philosophy, and that is there is no real independent self, aloof from other human beings, inspecting the world,
DM Presence EN	1
targetLangtext	Разбира се, това е основата на голяма част от източната философия
targetLangContext	от другото лице, е вашата кожа. Отстранете кожата, и ще усетите докосването на онзи човек в ума си. Размита е бариерата между вас и другите човешки същества. Разбира се, това е основата на голяма част от източната философия - че няма истинска независима самоличност, настрана от другите същества, наблюдаваща света,
DM Presence targetLang	1
DM targetLang	разбира се

[2] The columns names and the cell contents are represented vertically due to space constraints of the formatting of the article and for better readability.

The first 3 columns are IDs, followed by 4 columns dedicated to the English utterance, e.g. the multiword expression describing the discourse marker, a short context where it is found in the text, a larger context window, and an annotation whether the multiword expression is a discourse marker in the text or not. For target languages other than English, the next 4 columns provide the same content for the target language. This structure has been enriched with further columns to allow annotation based on the annotation scheme described in Section 1.3. below.

1.3. The annotation scheme

The framework captures relational and non-relational uses (Crible 2016) of multiword discourse markers that compose our corpus. It harmonizes two parts of the ISO 24617 Semantic annotation framework (SemAF): Part 8 – Semantic relations in discourse, core annotation schema (DR-core)– ISO 24617-8 (ISO, 2016) and Part 2: Dialogue acts (ISO, 2020) (Silvano and Damova 2022; Silvano et al. 2022). Similar to Bunt (2019, 2020), this is composed of a host annotation scheme based on Part 8 and an annotation plug-in to Part 2. However, instead of plugging Part 2 into Part 8, we use the plug-in mechanism in the opposite direction.

ISO 24617-8 proposes a core set of discourse relations, meaning relations between discourse units, which are decisive in explaining how discourse is organized and accounting for different linguistic problems. Existing frameworks (Hobbs 1985; Mann and Thompson 1987; Sanders et al. 1992; Kehler 2002; Asher and Lascarides 2003) differ along several aspects, namely discourse relations' designations, definitions, nature, number, and their arguments' type, adjacency, directionality, and relevance. It is precisely the diversity of the different proposals that prompts ISO 24617-8, which aims to establish a set of interoperable discourse relations.

In our annotation scheme, when a discourse marker conveys a semantic value, the discourse relations proposed by ISO 24617-8 are annotated choosing between symmetric or asymmetric discourse relations, and identifying the role of each argument. Table 2 gathers all the discourse relations that integrate the annotation scheme.

Table 2. Set of discourse relations proposed by ISO 24617-8 (Bunt and Prasad 2016)

Discourse Relations			
Asymmetric	**Semantic Role**		**Symmetric**
	Arg 1	Arg 2	
Cause	result	reason	Conjunction
Expansion	narrative	expander	Contrast
Asynchrony	before	after	Synchrony
Concession	expectation raiser	expectation-denier	Similarity
Elaboration	broad	specific	Disjunction
Exemplification	set	instance	Restatement
Manner	achievement	means	
Condition	consequent	antecedent	
Negative Condition	consequent	negated-antecedent	
Purpose	enablement	goal	
Exception	regular	exclusion	
Substitution	disfavoured-alternative	favoured-alternative	

When discourse markers bear a non-relational or pragmatic value, the plug-in into Part 2 of ISO 24617 is activated. This part of the standard puts forward a model for annotating dialogue acts establishing different dimensions, communicative functions, and qualifiers. The adopted annotation scheme incorporates the set of communicative functions and qualifiers, represented in Table 3.

Table 3. Set of communicative functions and qualifiers proposed by ISO 24617-2 (Bunt et al. 2020)

Communicative functions		Qualifiers
General	Dimension-specific	
checkQuestion	autoPositive	conditional/ unconditional
inform	autoNegative	certain/uncertain
agreement	alloPositive	positive/ negative
disagreement	alloNegative	
correction	feedbackElicitation	
answer	stalling	
confirm	pausing	
disconfirm	interactionStructuring	
offer	opening	
promise	topicShift	
addressRequest	selfError	
acceptRequest	retraction	
declineRequest	selfCorrection	
addressSuggest	initGreeting	
acceptSuggest	initSelfIntroduction	
declineSuggest	apology	
request	thanking	
instruct	initGoodbye	
suggest	compliment	
addressOffer	congratulation	
acceptOffer	sympathyExpression	
declineOffer	contactCheck	

Figure 1 summarizes the scheme we designed to annotate semantic and pragmatic values of multiword discourse markers.

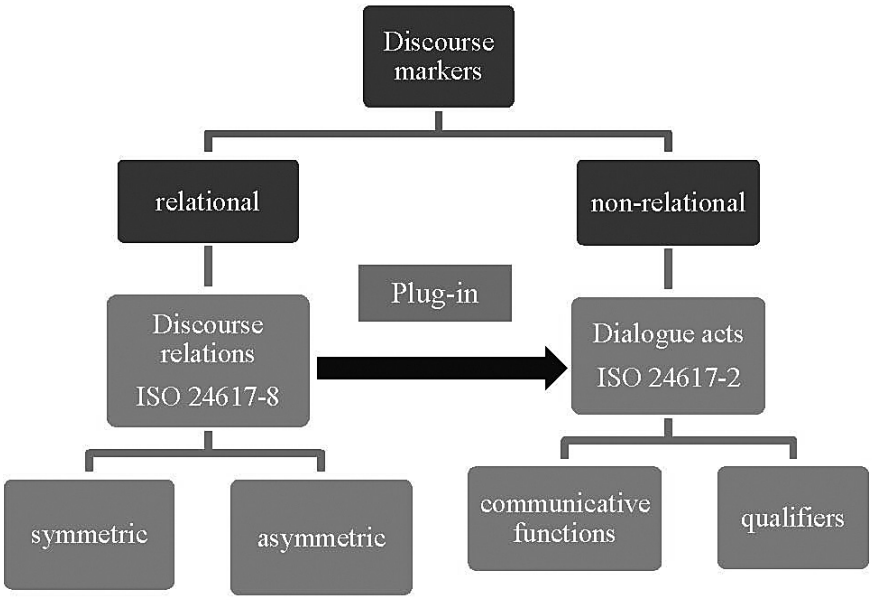

Figure 1. The annotation scheme for annotating discourse markers (**Silvano et al. 2022**)

2. Linguistic Linked Open Data approach

The Linguistic Linked Open Data approach uses an OWL ontology to represent meanings and functions of the discourse marker in compliance with the ISO SemAF model. We describe the creation of the ontology and its linking to an RDF edition of the parallel corpus. Finally, the resulting data is queried to demonstrate the viability of the LLOD approach.

2.1. OWL ontology

For modelling ISO 24617 (SemAF) discourse and dialogue annotations in a single OWL ontology, some restructuring of the original model was necessary. In comparison with other discourse schemes, ISO 24617 is special in that the subclassification of discourse relations (e.g., *result* as a specialization of a more ge-

neric *Cause* relation) is not expressed in the type of the relation, but in the labels of its arguments. The difference is illustrated in Figure 2. Most frameworks of discourse annotation (including PDTB and RST) define *result* as a relation between two spans or elementary discourse units, where one argument (either the less central one – as in RST –, or the one directly associated with the discourse marker that marks the relation – as in PDTB) points to the other argument. In ISO 24617, however, no such weighting or directionality is assumed. Instead, every *Cause* relation has a *result* and a *reason* argument.

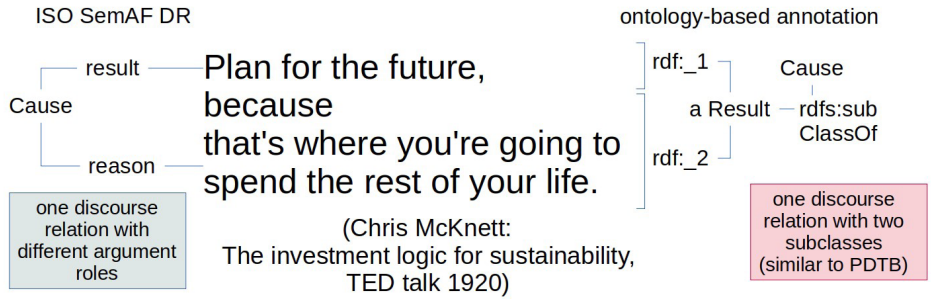

Figure 2. Comparing ISO 24617 categories with ontological modelling

In the case of discourse marker annotation, a natural imbalance exists between the argument span to which a discourse marker is syntactically attached (internal argument) and the other, external argument to which it is not attached. When annotating the discourse function of discourse markers, the annotation of the internal argument naturally takes priority over the annotation of the external argument, simply because the former is usually unambiguous whereas the external argument can be uncertain (or might not even be annotated). For this particular constellation, we can thus re-conceptualize the argument roles of ISO 24617 as subclasses. For the internal argument (or the discourse marker), the corresponding class is thus *result* (with superclass *Cause*), with the implication that the external argument is a *reason*. ISO would require explicit, and separate annotations of both arguments and the overall relation as *result, Cause* and *reason*, respectively. It is to be noted that this modelling is consistent with PDTB (although they use a different and more fine-grained relation inventory), and that the compatibility between both ways of modelling has previously been demonstrated (Prasad and Bunt 2015).

Technically, the discourse relation (an instance of *Result*) is an ordered list (*rdfs:Seq*) whose first element (*rdf:_1*, corresponding to the ARG1 in PDTB) is the external argument, and whose second element (*rdf:_2*, corresponding to the PDTB ARG2) is the internal argument. The class of the discourse relation indicates the type of the internal argument (the ISO SemAF DR argument type), and (by subsumption inference), also the ISO SemAF DR relation type (as its superclass), as well as the type of the external argument (implicitly, by not annotating the other subclass of *Cause*).

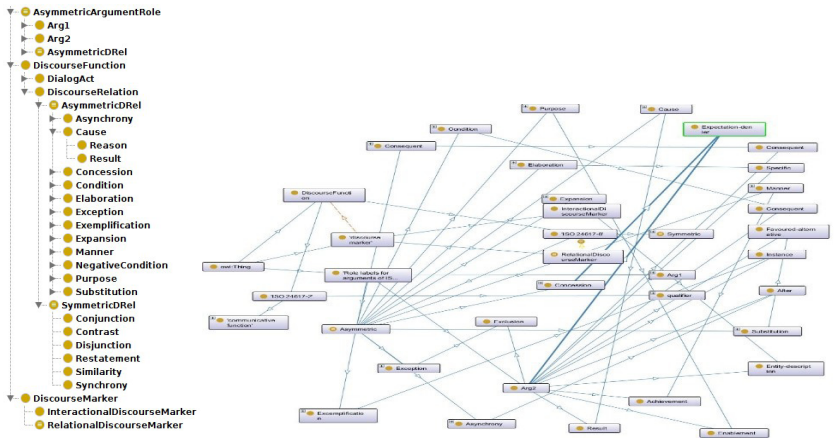

Figure 3. Excerpts of the discourse ontology: Concept hierarchy and selected cases of multiple inheritance

Aside from the relation taxonomy, we also model (and mark) asymmetric discourse relations. Every ISO argument role is thus not only a subclass of an ISO relation but also a subclass of either the class *Arg1* or the class *Arg2*, depending on which ISO DR argument they pertain to. Our ontological modelling is thus semantically lossless – and this is possible only by multiple inheritance.

In a similar way, ISO 24617-2 concepts are modelled in the corresponding subclasses of *DialogAct*. Again, multiple inheritance can be exploited to model the overlap between communicative functions and qualifiers. Figure 3 shows the OWL ontology structure.

2.2. Parallel corpus as LLOD

Querying discourse corpora has been notoriously problematic due to the multitude and complexity of tool- and theory-specific formats involved. For ISO SemAF annotations, we are not aware of any software to produce, convert and query this data. An RDF conversion is a natural way to make sure that off-the-shelf technology can be used to query, consume, exchange and store discourse annotations, once they are provided as a graph. This will be illustrated with our corpus. For data modelling, we follow Cimiano et al. 2020, chapters 5 and 6, to combine three main vocabularies:

- the NLP Interchange Format (NIF, Hellmann et al. 2013) that defines words (*nif:Word*), sentences (*nif:Sentence*) and the relations between them (*nif:nextWord*, *nif:nextSentence*),
- the CoNLL-RDF vocabulary (Chiarcos et al. 2021) that defines word-level annotations (*conll:WORD, conll:POS*, *conll:HEAD*) and the subset of NIF concepts and properties to be used,
- the POWLA vocabulary (Chiarcos 2012) that implements labelled relations and hierarchical structures in accordance with ISO 24612.

With the FINTAN converter platform (Fäth et al. 2020), it is possible to convert data into RDF, to manipulate or create annotations with SPARQL update scripts and to provide the result in RDF, and we used this tool to convert the original data. The initial annotation of the corpus was done in a spreadsheet software, using a tabular representation, and after exporting to a tab-separated format, these data could be directly processed with FINTAN. FINTAN already provides built-in support for CoNLL-RDF and NIF, and with task-specific SPARQL update scripts, we could convert the corpus annotations to an RDF representation and link it with both other language editions of the same text and the discourse ontology. A snippet of resulting data is given in Figures 4 and 5.

```
:s1_0 a nif:Sentence .
:s1_1 a nif:Word; conll:WORD "Plan"; nif:nextWord :s1_2 .
:s1_2 a nif:Word; conll:WORD "for"; nif:nextWord :s1_3 .
:s1_3 a nif:Word; conll:WORD "the"; nif:nextWord :s1_4 .
:s1_4 a nif:Word; conll:WORD "future"; nif:nextWord :s1_5 .
:s1_5 a nif:Word; conll:WORD "because"; nif:nextWord :s1_6 .
:s1_6 a nif:Word; conll:WORD "that"; nif:nextWord :s1_7 .
:s1_7 a nif:Word; conll:WORD "'s"; nif:nextWord :s1_8 .
...

:s1_5 a semaf:RelationalDiscourseMarker .
:s1_5 semaf:function :s1_5_relation .
:s1_5_relation a semaf:Result .
:s1_5_relation rdf:_1 :s1_5_relation_arg1 .
:s1_5_relation rdf:_2 :s1_5_relation_arg2 .
:s1_5_relation_arg1 a powla:Node; powla:hasChild :s1_1, :s1_2, :s1_3, :s1_4 .
:s1_5_relation_arg2 a powla:Node; powla:hasChild :s1_5, :s1_6, :s1_7, :s1_8, :s1_9,
```

Figure 4. Excerpt of converted corpus data

In Figure 4, we see selected aspects of the sentence under the URI*:s1_0*. It is defined as a *nif:Sentence* and contains several *nif:Word*s, each with its unique URI. In terms of annotations, we only provide the surface string here, but additional annotations can be added. One such addition is the annotation of a *semaf:RelationalDiscourseMarker*. Note that the URI *:s1_5* has been defined above for a *nif:Word*, so that the word and the discourse marker are actually treated as a single entity. Relational discourse marker is the root class for the annotation of discourse relations, and thus, one such relation is assigned as its *semaf:function*. The URI of this function is derived from the URI of the discourse marker, and it is defined as an instance of *semaf:Result*. The argument spans *...arg1* and *... arg2* are modelled as *powla:Node*s that take their respective *nif:Word*s as children. This snippet already shows the linking between the SemAF ontology and the corpus data. The resulting knowledge graph from the conversion of the corpus in RDF format can be queried with standard RDF technology (SPARQL).

2.3. Querying example

Querying with SPARQL differs from conventional corpus query languages in that it is largely unconstrained in its expressivity (and complexity). A special feature of RDF technology is that SPARQL allows to query over both local and remote data (federation) and to integrate information from heterogeneous data from the web – as long as this is exposed as RDF.

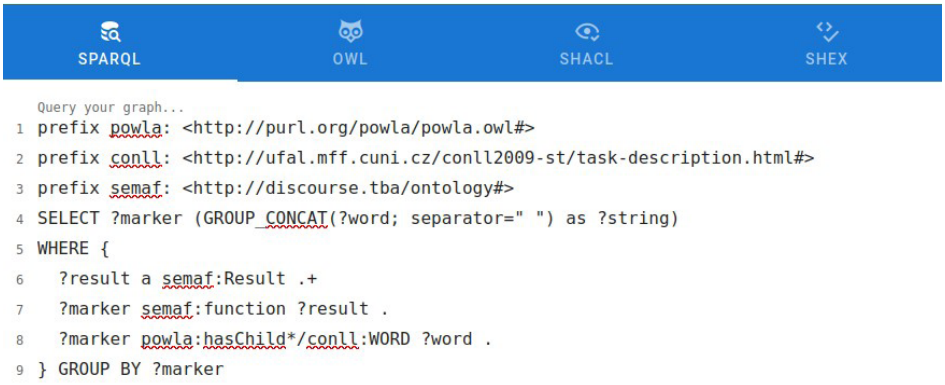

Figure 5. Sample query to retrieve discourse markers along with their function.

A simple example query is shown in Figure 6, and it illustrates how to retrieve a *Result* relation, the associated discourse marker (linked via *semaf:function*) along with all the words contained in that marker, resp., their surface strings (*conll:WORD*). The result of that query, if applied to the data from Figure 4, includes the marker with the URI *:s1_5* and its surface representation "*because*". Analoguously, we can query for the arguments of the discourse marker by the triples *?result rdf:_1 ?arg1* and *?result rdf:_2 ?arg2*, respectively, and then retrieve their surface strings and URIs. By means of these URIs, it is then possible to inspect the wider textual context of discourse markers and arguments. Figure 7 shows the results of the SPARQL query from Figure 6. The strings describing the two arguments and the discourse markers are returned in the first three columns and the type of the discourse marker and the role of the discourse marker are also displayed.

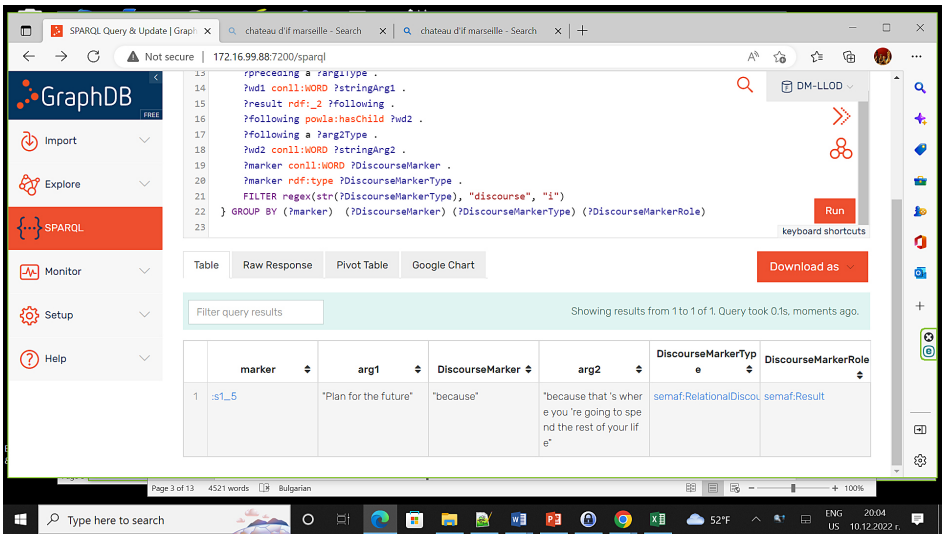

Figure 6. SPARQL query result

Writing SPARQL queries using aggregate functions and operating over multiple named graphs requires some expertise, and might not necessarily be an innate skill of every linguist. Initial steps are laid to make this easier with research projects that integrate LLOD technologies in the development of infrastructure portals (cf. Khan et al. 2022), and for discourse studies, our paper represents yet another such building block.

3. Conclusion and outlook

In this paper, we described the creation of a novel, multilingual resource with discourse annotations, specifically focusing on multiword discourse markers and their annotation with discourse functions (discourse relations, dialogue acts). To this end, we provide annotated data, but we also operationalized the ISO SemAF specifications for discourse and dialogue annotation into an annotation schema, consequently represented as an OWL2/DL ontology. We argued that this would be advantageous as it allows us to define overlaps between different categories instead of enforcing a single, and strict hierarchical schema. But also, it allows the application of LLOD technologies for accessing and querying the data.

Hence, we described the data modelling of our data in RDF and the conversion from its original table representation to a collection of RDF graphs.

Ultimately, this allows for the conjoint querying of the ontology and the annotations by means of SPARQL, the standard query language for the web of data. We are thus able to perform queries over multiple, interlinked datasets with complex internal structure without the need of any specialised software. This is a first, but essential step in developing novel and powerful means for the corpus-based study of multilingual discourse, communication analysis, or attitudes discovery. In particular, the study on discourse markers enables the discovery of lexical elements that may be included in lexicons and the definition of new distinguishing characteristics. We consider the creation and publication of discourse data and annotation scheme in accordance with LLOD principles a best practice as it facilitates the subsequent use and re-use of data published in that way. This technology does not create any dependencies from specific software packages, libraries or any particular programming language, but *every programming language* and *every database* that implements web standards such as RDF and/or SPARQL will be capable of accessing, querying and transforming our data. For the field of discourse studies, which is plagued by complex, and rather idiosyncratic formats, often specific to legacy tools, this represents a great improvement already as is.

References

Afantenos, Stergos; Asher, Nicholas; Benamara, Farah; Bras, Myriam; Fabre, Cécile; Ho-Dac, Mai; Le Draoulec, Anne; Muller, Philippe; Péry-Woodley, Marie-Paule; Prévot, Laure; Rebeyrolle Josette; Tanguy Ludovic; Vergez-Couret Marianne; Vieu Laure. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*. Eds. Calzolari, Nicoletta et al. European Language Resources Association. Istanbul. 2727–2734.

Asher, Nicholas; Lascarides, Alex. 2003. *Logics of Conversation*. Cambridge University Press.

Bunt, Harry; Prasad, Rashmi. 2016. ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. *Proceedings of the 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)*. Ed. Bunt, Harry. TiCC, Tilburg center for Cognition and Communication. Portorož. 45–54.

Bunt, Harry. 2019. Plug-ins for content annotation of dialogue acts. *Proceedings of the 15th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-15)*. Ed. Bunt, Harry. Gothenburg. 34–45.

Bunt, Harry; Petukhova; Gilmartin, Emer; Pelachaud, Catherine; Fang, Alex; Keizer, Simon; Prévot, Laurent. 2020. The ISO Standard for Dialogue Act Annotation, Second Edition. *Proceedings of the 12th Language Resources and Evaluation Conference (LREC'20)*. Eds. Calzolari, Nicoletta et al. European Language Resources Association. Marseille. 549–558.

Burchardt, Aljoscha; Padó, Sebastian; Spohr, Dennis; Frank, Anette; Heid, Ulrich. 2008. Formalising Multi-layer Corpora in OWL DL - Lexicon Modelling, Querying and Consistency Control. *Proceedings of the 3rd International Joint Conference on Natural Language Processing*. Volume-I. Asian Federation of Natural Language Processing. Hyderabad. 389–396.

Carlson, Lynn; Marcu, Daniel; Okurowski, Mary Ellen. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. Current and New Directions in Discourse and Dialogue. CCurrent and New Directions in Discourse and Dialogue. *Text, Speech and Language Technology* 22. 85–112. https://doi.org/10.1007/978-94-010-0019-2_5.

Cimiano, Philipp; Chiarcos, Christian; MacCrae, John; Gracia, Jorge. 2020. *Linguistic Linked Data*. Springer International Publishing. https://doi.org/10.1007/978-3-030-30225-2.

Chiarcos, Christian. 2008. An ontology of linguistic annotations. *GLDV-Journal for Language Technology and Computational Linguistics* 23/1. 1–16. https://doi.org/10.21248/jlcl.23.2008.98.

Chiarcos, Christian. 2012. A generic formalism to represent linguistic corpora in RDF and OWL/DL. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Eds. Calzolari, Nicoletta et al. European Language Resources Association. Istanbul. 3205–3212.

Chiarcos, Christian. 2014. Towards interoperable discourse annotation. Discourse features in the Ontologies of Linguistic Annotation. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Eds. Calzolari, Nicoletta et al. European Language Resources Association. Reykjavik. 4569–4577.

Chiarcos, Christian; Fäth, Christian. 2019. Graph-Based Annotation Engineering: Towards a Gold Corpus for Role and Reference Grammar. *2nd Conference on Language, Data, and Knowledge (LDK 2019)*. Eds. Eskevich, Maria et al. Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. https://doi.org/10.4230/OASIcs.LDK.2019.9.

Chiarcos, Christian; Declerck,Thierry; Ionov, Maxim. 2021. Embeddings for the

Lexicon: Modelling and Representation. *Proceedings of the 6th Workshop on Semantic Deep Learning (SemDeep-6)*. Espinosa-Anke, Luis et al. Association for Computational Linguistics. Online. 13–19.

Cribble, Ludivine. 2016. Discourse Markers and Disfluencies: Integrating Functional and Formal Annotations. *Proceedings of the 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)*. Ed. Bunt, Harry. TiCC, Tilburg center for Cognition and Communication. Portorož. 38–44.

Das, Debopam; Taboada, Maite. 2019. Multiple signals of coherence relations. *Discours* 24. https://doi.org/10.4000/discours.10032.

Das, Debopam. 2014. *Signalling of coherence relations in discourse*. Doctoral thesis, Arts & Social Sciences. Department of Linguistics, Simon Fraser University, British Columbia, Canada. 211 pp.

Dimitriadis, Alexis; Windhouwer, Menzo; Saulwick, Adam; Goedemans, Rob; Bíró, Tamás. 2009. How to integrate databases without starting a typology war: The Typological Database System. *The Use of Databases in Cross-Linguistic Studies*. Eds. Everaert, Martin; Musgrave, Simon; Dimitraidis, Alexis. 155–207. https://doi.org/10.1515/9783110198744.155.

Farrar, Scott; Langendoen, Donald Terrence. 2003. A linguistic ontology for the semantic web. *GLOT international* 7/3. 97–100.

Fraser, Bruce. 2009. An account of discourse markers. *International review of Pragmatics* 1/2. 293–320. https://doi.org/10.1163/187730909X12538045489818.

Goecke, Daniela; Lüngen, Harald; Sasaki, Felix; Witt, Andreas; Farrari, Scott. 2005. GOLD and discourse: Domain - and community - specific extensions. *Proceedings of the E-MELD Workshop on Morphosyntactic Annotation and Terminology: Linguistic Ontologies and Data Categories for Language Resources*. Boston.

Heeman, Peter A.; Allen, James F. 1999. Speech repains, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue. *Computational Linguistics* 25/4. 527–572.

Hellmann, Sebastian; Lehmann, Jens; Auer, Sören; Brümmer, Martin. 2013. Integrating NLP using linked data. *Proceedings of the 12th International semantic web conference (ISCW 2013). Lecture Notes in Computer Science 8219*. Eds. Alani, Harith et al. Springer. Berlin – Heidelberg. 98–113. https://doi.org/10.1007/978-3-642-41338-4_7.

Hobbs, Jerry R. 1985. *On the coherence and structure of discourse*. Technical report, CSLI-85-37, Center for the Study of Language and Information – USC Information Science Institute. Marina del Rey, CA, USA.

ISO. 2016. *Language resource management- Semantic annotation framework (SemAF) - Part 8 - Semantic relations in discourse, core annotation schema (DR-core)*. Standard. Geneva.

ISO. 2020. Language resource management- Semantic annotation framework (SemAF) - Part 2 - Dialogue acts. Standard. Geneva.

Jucker, Anreas. H.; Ziv, Yael. 1998. Discourse Markers: introduction. *Discourse Markers*. 1–12.

Kehler, Andrew. 2002. *Coherence, Reference, and the Theory of Grammar*. CSLI Publications.

Khan, Anas Fahad; Chiarcos, Christian; Declerck, Thierry; Gifu, Daniela; González-Blanco García, Elena; Gracia, Jorge; Ionov; Maxim; Labropoulou, Penny; Mambrini, Francesco; McCrae, John P.; Pagé-Perron, Émilie; Passarotti, Marco; Ros Muñoz, Salvador; Truica, Ciprian-Octavian. 2022. When linguistics meets web technologies. Recent advances in modelling linguistic linked data. *Semantic Web* 13/6. 987–1050. https://doi.org/10.3233/SW-222859.

Knott, Alistair; Dale, Robert. 1993. Using linguistic phenomena to motivate a set of rhetorical relations. *Discourse Processes* 18/1. 35–62. https://doi.org/10.1080/01638539409544883.

Lüngen, Harald; Bärenfänger, Maja; Hilbert, Mirco; Lobin, Henning; Csilla Puskás. 2010. Discourse relations and document structure. *Linguistic modeling of information and markup languages*. 97–123. https://doi.org/10.1007/978-90-481-3331-4_6.

Mann, William C.; Thompson, Sandra A. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8/3. 243–281. https://doi.org/10.1515/text.1.1988.8.3.243.

Marcu, Daniel. 2000. *The theory and practice of discourse parsing and summarization*. MIT press.

Maschler, Yael; Schiffrin, Deborah. 2015. Discourse markers: Language, meaning, and context. *The handbook of discourse analysis* 2. Eds. Tannen, Deborah; Hamilton, Heidi E.; Schiffrin, Deborah. John Wiley & Sons, Inc. 189–221. https://doi.org/10.1002/9781118584194.ch9.

Prasad, Rashmi; Dinesh, Nikhil; Lee, Alan; Miltsakaki, Eleni; Robaldo, Livio; Joshi, Aravind; Webber, Bonnie. 2008. The Penn Discourse TreeBank 2.0. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Eds. Calzolari, Nicoletta et al. European Language Resources Association. Marrakech. 2961–2968.

Prasad, Rashmi; Bunt, Harry. 2015. Semantic relations in discourse: The current state of ISO 24617-8. *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*. Ed. Bunt, Harry. Association for Computational Linguistics. London.

Sanders, Ted J.M.; Spooren, Wilbert P.M.; Noordman, Leo G.M. 1992. Toward a Taxonomy of Coherence Relations. *Discourse Processes* 15/1. 1–35. https://doi.org/10.1080/01638539209544800.

Schiffrin, Deborah. 2005. Discourse markers: Language, meaning, and context. *The handbook of discourse analysis*. Eds. Tannen, Deborah; Hamilton, Heidi E.; Schiffrin, Deborah. Blackwell Publishers Ltd. 54–75. https://doi.org/10.1002/9780470753460.ch4.

Schmidt, Thomas; Chiarcos, Christian; Lehmberg, Timm; Rehm, Georg; Witt, Andreas; Hinrichs, Erhard. 2006. Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources. *E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art.* Institut für Deutsche Sprache, Bibliothek. 1–25.

Silvano, Purificação. 2010. *Temporal and rhetorical relations: the semantics of sentences with adverbial subordination in European Portuguese*. Doctoral thesis. University of Porto. 430 pp.

Silvano, Purificação; Damova, Mariana. 2022. ISO-DR-core plugs into ISO-dialogue acts for a crosslinguistic taxonomy of discourse markers. *DiSLiDaS 2022 workshop*. Jerusalem.

Silvano, Purificação; Damova, Mariana; Oleškevičienė, Giedrė Valūnaitė; Liebeskind, Chaya; Chiarcos, Christian; Trajanov, Dimitar; Truică, Ciprian-Octavian; Apostol, Elena-Simona; Baczkowska, Anna. 2022. ISO-Based Annotated Multilingual Corpus For Discourse Markers. *Proceedings of the 13th Edition Language Resources and Evaluation Conference (LREC 2022)*. Eds. Calzolari, Nicoletta et al. European Language Resources Association. Marseille. 2739–2749.

Izrada OWL ontologije za prikaz, povezivanje i pretraživanje SemAF diskursnih oznaka

Sažetak

Diskursni markeri jezični su znakovi koji pokazuju kako se iskaz odnosi na kontekst diskursa i koju ulogu ima u razgovoru. Lingvistički povezani otvoreni podatci (LLOD) tehnologije su u nastajanju koje omogućuju snažan instrument za prikaz i tumačenje jezičnih fenomena na razini *weba*. Glavni je cilj ovoga rada pokazati kako se tehnologije lingvistički povezanih otvorenih podataka (LLOD) mogu primijeniti za prikaz i označavanje korpusa višerječnih diskursnih markera te koji su učinci toga. Konkretno, naš je cilj primijeniti standarde semantičkoga *weba* kao što su RDF i Web Ontology Language (OWL) za objavljivanje i integraciju podataka. Autori predstavljaju novu shemu za označavanje diskursa koja kombinira ISO standarde za opis diskursnih odnosa i dijaloških činova – ISO DR-Core (ISO 24617-8) i ISO-Dialogue Acts (ISO 24617-2) na devet jezika (usp. Silvano et al. 2022a; Silvano et al. 2022b). Razvijamo OWL ontologiju kako bismo formalizirali tu shemu, pružili nov označeni skup podataka i povezali njegovu RDF inačicu s ontologijom. U skladu s tim opisujemo zajedničko postavljanje

upita ontologiji i oznakama s pomoću SPARQL-a, standardnoga jezika upita za *web* podataka. Konačni je rezultat taj da možemo izvršiti upite nad višestrukim, međusobno povezanim skupovima podataka sa složenom unutarnjom strukturom bez potrebe za ikakvim specijaliziranim softverom. Umjesto toga upotrebljavaju se gotove tehnologije utemeljene na *web* standardima koje se bez napora mogu prenijeti na različite operativne sustave, baze podataka i programske jezike. Ovo je prvi, ali prijeloman korak u razvoju novih, snažnih i (u određenom trenutku) pristupačnih sredstava za korpusno utemeljena istraživanja višejezičnoga diskursa te za analizu komunikacije i otkrivanje stavova.

Keywords: LLOD, OWL ontology, RDF, discourse markers, ISO standard, parallel corpus
Ključne riječi: LLOD, OWL ontologija, RDF, diskursni markeri, ISO standard, paralelni korpus