An OWL Ontology for ISO-Based Discourse Marker Annotation

Christian Chiarcos,
Applied Computational Linguistics, Goethe-Universität, Robert Mayer Straße 10 60325 Frankfurt am Main, Germany, chiarcos@cs.uni-frankfurt.de
Purificação Silvano,
Faculty of Arts and Humanities of the University of Porto, Centre of Linguistics of the University of Porto, Via Panorâmica, 4150-564 Porto, Portugal, msilvano@letras.up.pt
Mariana Damova,
Mozaika, Ltd., Solunska 52, Sofia 1000, mariana.damova@mozajka.co
Giedrė Valūnaitė-Oleškevicienė,
Mykolas Romeris University, Ateities 20, Vilnius, Lithuania, gvalunaite@mruni.eu
Chaya Liebeskind,
Jerusalem College of Technology, Havaad Haleumi 21, Givat Mordechai 91160, liebchaya@gmail.com
Dimitar Trajanov,
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, dimitar.trajanov@finki.ukim.mk
Ciprian-Octavian Truica,
Department of Information Technology, Uppsala University, Sweden, Lägerhyddsvägen 1, 75105 Uppsala, Sweden
Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Splaiul Independen¸tei nr. 313, sector 6, Bucuresti, Romania, ciprian-octavian.truica@it.uu.se
Elena-Simona Apostol,
Department of Information Technology, Uppsala University, Sweden, Lägerhyddsvägen 1, 75105 Uppsala, Sweden
Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Splaiul Independen¸tei nr. 313, sector 6, Bucuresti, Romania,elena.apostol@upb.ro
Anna Baczkowska,
Institute of English and American Studies, University of Gdansk, anna.k.baczkowska@gmail.com

Purpose: Discourse markers are linguistic cues that indicate how an utterance relates to the discourse context and what role it plays in conversation. The authors are preparing an annotated corpus in nine languages, and specifically aim to explore the role of Linguistic Linked Open Data (/LLOD) technologies in the process, i.e., the application of web standards such as RDF and the Web Ontology Language (OWL) for publishing and integrating data. We demonstrate the advantages of this approach.

Design/methodology/approach: (1) Survey of existing (community) standards for discourse (marker) annotation; (2) decision to follow ISO SemAF Discourse Relations (Core), using ISO SemAF Dialog Annotation standards [1,2] as a plugin; (3) provide annotation in a simplistic tabular format [3,4]; (4) formalize the annotation schema in an ontology [main contribution here]; and (5) convert annotations to RDF, link with ontology, and perform conjoint queries.

Findings: ISO SemAF discourse and dialog annotations can be formalized conjointly in a single OWL ontology. Some aspects needed to be systematically restructured. Most importantly, this concerns how ISO SemAF implements the subclassification of asymmetric discourse relations (as labels on arguments, not as subclasses of discourse relations). In discourse marker annotation, this needs to be converted into a flat hierarchy of relation types (more in line with previous research, e.g., Penn Discourse Treebank). A key advantage of OWL (in comparison to hierarchical annotation schemes) is that it supports fully fledged description logics, so we can annotate communicative functions and their qualifiers using logical (set/class) operators, e.g., *AcceptOffer Conditional*.

Research limitations/implications: Pilot annotations have been conducted for nine languages. The ontology is published under https://purl.org/olia/discourse/discourse.Nexus.owl. The LLOD transformation and linking of the *annotations* have not yet been performed. This paper provides the necessary prerequisites for doing so and discusses modelling challenges.

Practical implications: To the best of our knowledge, this is the first application of ISO SemAF for cross-lingual discourse marker annotation. Using LLOD technologies allows for the seamless integration of *any data* (here: annotations and ontologies), e.g., for querying.

Originality/Value: To the best of our knowledge, this is the first application of ISO SemAF for cross-lingual discourse marker annotation. Furthermore, this is one of the first concrete applications of LLOD technologies for implementing aspects of discourse annotation.

References:

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

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

[3] 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". In Proceedings of the 13th Edition Language Resources and Evaluation Conference (LREC 2022), pages 2739-2749. European Language Resources Association (ELRA). ACL anthology.

(http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.293.pdf

[4] Silvano, Purificação & Damova, Mariana (2022). ISO-DR-core plugs into ISO-dialogue acts for a crosslinguistic taxonomy of discourse markers. DiSLiDaS 2022 workshop, NexusLinguarum, Jerusalem, Israel, May 2022.Web Ontology Language (OWL) https://www.w3.org/OWL/