OntoLex-FrAC: Standardizing the Corpus-Lexicon Interface

Christian Chiarcos
University of Augsburg
Germany
christian.chiarcos
@philhist.uni-augsburg.de

Anas Fahad Khan
ILC/CLR
Italy
Fahad.Khan
@ilc.cnr.it

Maxim Ionov
University of Cologne
Germany
mionov
@uni-koeln.de

Elena Simona Apostol
Ciprian-Octavian Truica
Uppsala University, Sweden
{elena-simona.apostol,
ciprian-octavian.truica}
@it.uu.se

Besim Kabashi
University of
Erlangen-Nuremberg
Germany
besim.kabashi
@fau.de

Katerina Gkirtzou
Athena Research Center
Athens, Greece
katerina.gkirtzou
@athenarc.gr

Relevant UniDive working groups: WG2+WG1

1 Background

Recent years have seen an increased adoption of RDF and Linked Data technologies for modelling, linking and sharing lexical resources over the web. In this context, Linked Data technology yields two key benefits over earlier formalisms: (1) unified mode of access and query to lexical data, (2) unified access and query over distributed data (federated search), (3) established, W3C-standardized wrapper technologies for manifold formats (CSV, relational databases, native RDF, XML, JSON), (4) semantically typed linking between dictionaries or between dictionaries and corpora, (5) standardized core vocabularies, and (6) support for schema-free database backends (freely extensible vocabularies).

For lexical data, this leads to novel applications: (1) transitive on-the-fly search across dictionaries and other lexical databases (Chiarcos and Sérasset, 2022), (2) linking of digital editions and annotated corpora with dictionaries (Tittel et al., 2018; Fantoli et al., 2022), (3) standardized web services and distributed, interoperable, and linked infrastructures for lexical data, e.g., for Latin and Greek (Mambrini et al., 2021), and (4), the conjoint development of knowledge graphs and dictionaries in digital lexicography (Bellandi et al., 2017)

2 OntoLex-Lemon

For lexical resources in RDF, OntoLex (McCrae et al., 2017) has become the dominant community standard, and along with the rising number of applications, novel requirements have arisen and led to the development of novel modules that complement the OntoLex core vocabulary (OntoLex-Lemon) for needs articulated by specific communities or use cases. At the moment, this includes a designated OntoLex module for lexicography (Lonke and Bosque-Gil, 2019), a model for morphology in OntoLex dictionaries (Chiarcos et al., 2022b), and the emerging OntoLex module for Frequency, Attestations, and Corpus-Based Information (OntoLex-FrAC) summarized in this poster (Chiarcos et al., 2020, 2021, 2022a,a).

Primary data structures of the OntoLex-Lemon core vocabulary (Fig. 1) are `ontolex:LexicalEntry` (lexeme), `ontolex:Form` (word form), `ontolex:LexicalSense` (word sense), `ontolex:LexicalConcept` (lexicalization-independent concept), and ontological concept (any URI), and these are the elements that observations can be made about in OntoLex-FrAC.

3 OntoLex-FrAC

OntoLex-FrAC, or, briefly, FrAC, is designed to complement OntoLex-Lemon with the vocabulary to represent major types of information found in or automatically derived from corpora, for applications in both language technology and the language sciences, so that these can be included in machine-readable dictionaries. For lexical forms (which can be counted), lexical entries (which can be illustrated with attestations or corpus examples), lexical senses or lexical concepts (which can be found as annotations in corpora), FrAC introduces a generalization over the OntoLex core elements,with the notion of frac:Observable, as a lexical unit that can be observed in natural language, i.e., in a corpus. About an observable, observations can be made, and a frac:Observation is any information found in, based on or created from a corpus, and the observations supported by the FrAC vocabulary are corpus frequency, attestation, collocation, similarity and embeddings. FrAC observations should have the following core properties: rdf:value (value of an observation, specific for each type of observation), dc:description (human-readable information about how the value was calculated), and frac:corpus (link from the observation to the data over which the observation took place).

The vocabulary distinguishes four main classes as subclasses of frac:Observation, i.e., frequency, attestation, collocations, embeddings, and similarity as summarized in Fig. 2.

4 frac:Observations

The concept frac:Attestation formalizes the linking of lexical resources with corpus evidence, i.e., a quotation or excerpt from a source document that exhibits a particular lexical entry, form, sense, lexeme or features such as spelling variation, morphology, syntax, collocation, register. An attestation should have a quotation or an attestation gloss (value) and must define a locus or corpus object to identify the source of this material. The frac:CorpusFrequency class gives the absolute number of attestations, i.e., rdf:value, of a single frac:Observable considering a specific corpus

A frac:Collocation is an expression containing two or more juxtaposition words with a quantification of their cooccurrence likelihood according to one or multiple metrics. Collocations are modeled as an aggregate (rdfs:Container, ordered or unordered) of frac: Observables, based on their co-occurrence within the same context window and characterized the head word of the collocation (frac:head) and the collocation strength (various sub-properties of rdf:value) in a particular corpus (frac:corpus). For asymmetric collocations scores the frac:head property is used to identify the elements' order.

In FrAC, a frac:Embedding is a structure-preserving projection (mapping) from a given domain into a numerical representation. The most popular example of embeddings in language technology is a more restricted form of embeddings in that sense, i.e., the topological space of the resulting embeddings is represented by frac:FixedSizeVector (resp., tensors as aggregates of such vectors). Other embedding subclasses are frac:BagOfWords (for unweighted or weighted bags of words), and frac:TimeSeries (for sequences of fixed-size vectors). Both representations are similar to embeddings in the NLP sense in that they represent a projection into a numerical feature space and that the primary function of this projection is to provide distance measurements. For bags of words, these are represented by confidence scores for weighted bag of words models (or booleans for unweighted bags of words) for *every word in the vocabulary* (at least, this would be a possible mathematical interpretation; in practice, such data is not represented as a vector, but as a hashtable – or, for unweighted bags of words, a set –, so that only words with positive scores are listed). Along with static embeddings for observables (forms, lexical entries, lexical senses or concepts), *contextualized* embeddings for a phrase, a lexical unit or another observable can be represented in FrAC as (a property of the) attestation of the observable in a corpus: frac:attestationEmbedding assigns an attestation an embedding.

In FrAC, contextual similarity is represented using the frac:Similarity class, an aggregate (set, or bag) of FrAC observables, that represents a relation between two or more embeddings (frac:Embeddings) along with the similarity value (rdf:value), a corpus and a dc:description of the method of comparison. As frac:Similarity is modelled as a concept (rather than a property), it can be used to represent either similarity between two observables or a similarity cluster comprising two or more observables.

5 Outlook: Modelling Queries and Annotations

At the time of submission, OntoLex-FrAC is close to finalization. For the aspects described above, the model is considered mature and stable, with only minor rewordings expected to occur. Accordingly, it is expected to be published as a W3C vocabulary

(W3C Community Report) later this year. One aspect that is still being explored is whether and how to include corpus queries into the model.

However, there are future tasks on the horizon. In particular, this includes the questions of how to anchor an attestation object in a corpus. At the moment, there are several conflicting standards being applied for the purpose, most notably the NLP Interchange Framework (Hellmann et al., 2013, NIF) and the Web Annotation standard of the W3C (Ciccarese et al., 2013), and these co-exist with a number of other pre-RDF standards based on XML, CSV, TSV (e.g., CoNLL-U, TEI, LAF/GrAF, KAF, NAF) on the one hand, and domain vocabularies on the other hand, e.g., special-purpose vocabularies for mediating RDF with common formats for interlinear glossed text (Ionov, 2021, Ligt) or tab-separated values (Chiarcos and Fäth, 2017, CoNLL-RDF).

In the context of UniDive, we would like to discuss the place of both OntoLex and FrAC in the context of standardizing the lexicon-corpus interface (WG2), and OntoLex and its interplay with the aforementioned corpus standards in the context of standardizing corpora and linguistic annotations on the web (WG1+WG2).

References

Andrea Bellandi, Emiliano Giovannetti, Silvia Piccini, and Anja Weingart. 2017. Developing lexo: a collaborative editor of multilingual lexica and termino-ontological resources in the humanities. In *Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017)*.

Christian Chiarcos, Thierry Declerck, and Maxim Ionov. 2021. Embeddings for the lexicon: Modelling and representation. In *Proceedings of the 6th Workshop on Semantic Deep Learning (SemDeep-6)*, pages 13–19.

Christian Chiarcos and Christian Fäth. 2017. Conll-rdf: Linked corpora done in an nlp-friendly way. In *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings 1*, pages 74–88. Springer.

Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Besim Kabashi, Fahad Khan, and Ciprian-Octavian Truică. 2022a. Modelling collocations in ontolex-frac. In *Proceedings of Globalex Workshop on Linked Lexicography within the 13th Language Resources and Evaluation Conference*, pages 10–18.

Christian Chiarcos, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. 2022b. Computational morphology with ontolex-morph. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 78–86.

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

Christian Chiarcos and Gilles Sérasset. 2022. A cheap and dirty cross-lingual linking service in the cloud. In *8th Workshop on Linked Data in Linguistics (LDL-2022)*.

Paolo Ciccarese, Stian Soiland-Reyes, and Tim Clark. 2013. Web annotation as a first-class object. *IEEE Internet Computing*, 17(6):71–75.

Margherita Fantoli, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. Linking the lasla corpus in the lila knowledge base of interoperable linguistic resources for latin. In *Proceedings of the Linked Data in Linguistics Workshop@ LREC2022*, pages 26–34.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating nlp using linked data. In *The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II 12*, pages 98–113. Springer.

Maxim Ionov. 2021. Apics-ligt: Towards semantic enrichment of interlinear glossed text. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Dorielle Lonke and Julia Bosque-Gil. 2019. Applying the ontolex-lemon lexicography module to k dictionaries' multilingual data. *K Lexical News (KLN)*.

Francesco Mambrini, Eleonora Litta, Marco Passarotti, and Paolo Ruffolo. 2021. Linking the lewis & short dictionary to the lila knowledge base of interoperable linguistic resources for latin. In *CLiC-it*.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Sabine Tittel, Helena Bermúdez-Sabel, and Christian Chiarcos. 2018. Using rdfa to link text and dictionary data for medieval french. In *Proceedings of the 6th Workshop on Linked Data in Linguistics (LDL-2016): Towards Linguistic Data Science. European Language Resources Association (ELRA), Paris, France, Miyazaki, Japan.*

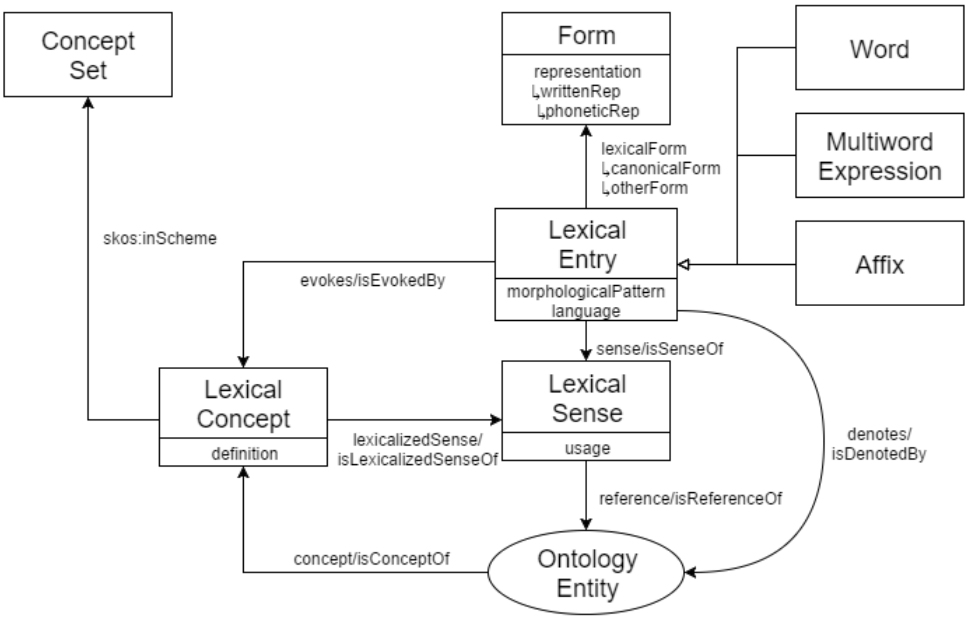

Figure 1: OntoLex-Lemon core module

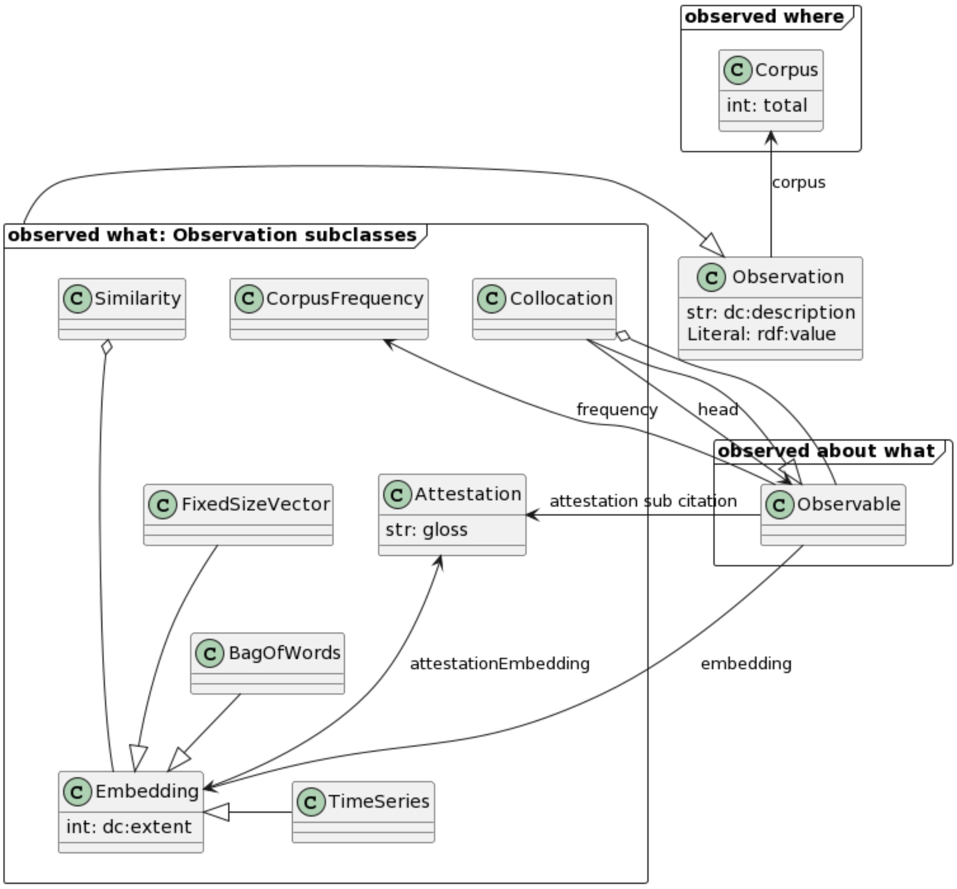

Figure 2: OntoLex-FrAC