Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC

Christian Chiarcos[1,2] and **Elena-Simona Apostol**[3,4] and **Besim Kabashi**[5] and **Ciprian-Octavian Truică**[3,4]
[1]Applied Computational Linguistics, Goethe University Frankfurt, Germany
[2]Institute for Digital Humanities, University of Cologne, Germany
[3]Department of Information Technology, Uppsala University, Sweden
[4]Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Romania
[5]Computational and Corpus Linguistics, University of Erlangen-Nuremberg, Germany
chiarcos@cs.uni-frankfurt.de, elena-simona.apostol@it.uu.se
besim.kabashi@fau.de, ciprian-octavian.truica@it.uu.se

Abstract

OntoLex-Lemon has become a de facto standard for lexical resources in the web of data. This paper provides the first overall description of the emerging OntoLex module for Frequency, Attestations, and Corpus-Based Information (OntoLex-FrAC) that is intended to complement OntoLex-Lemon with the necessary vocabulary to represent major types of information found in or automatically derived from corpora, for applications in both language technology and the language sciences.

1 Background

The OntoLex-Lemon vocabulary has become the dominant vocabulary for modelling machine-readable dictionaries on the web of data, i.e., by means of RDF. And indeed, publishing lexical resources in RDF has a number of advantages, including the ease of integration of dictionary information not only with ontologies and knowledge graphs (this was the original domain of application), but also with other lexical data.

Figure 1 illustrates the OntoLex-Lemon core vocabulary. Primary data structures are `ontolex:LexicalEntry` (lexeme), `ontolex:Form` (word form), `ontolex:LexicalSense` (word sense), and `ontolex:LexicalConcept` (lexicalization-independent concept), so that lexical entries can be described, but also fine-grained differences in meaning and surface form.

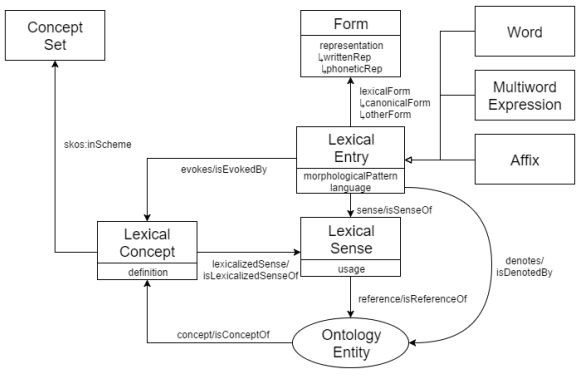


Figure 1: OntoLex-Lemon core module

While these aspects are advanced, stable and widely used, there is no complete module described in the current literature that enables interoperability and integration between lexical and textual resources and the distributional semantics of words, lexical senses and concepts, and collocation properties. By employing the usage of L(L)OD (Linked Linguistic Open Data) technologies, we describe the consolidation of OntoLex-FrAC (Frequency, Attestation, and Corpus Information), an OntoLex-Lemon model that *(1)* addresses the requirements of corpus-based lexicography (frequency and collocation information) and digital philology (linking lexical resources with corpus data), and *(2)* provides a standard for encoding, storing, and exchanging vector representations of words along with their lexical concepts, senses, and lemmas.

2 Core Concepts

So far, the development of FrAC has been conducted in a bottom-up fashion, where uses cases were analyzed and sub-vocabularies for different phenomena have been proposed. This includes frequency and attestations (Chiarcos et al., 2020), embeddings and similarity (Chiarcos et al., 2021) and collocations (Chiarcos et al., 2022). We complement these efforts with a top-down perspective, and we suggest three top-level classes to structure the model as a whole. In addition to that, we provide an OWL2/DL ontology to formalize the vocabulary. Restructuring the module entails a number of minor revisions regarding naming and scope of properties and classes, however, we aimed to stay faithful to the original definitions while integrating them into a more coherent overall picture.

Proceedings of the 29th International Conference on Computational Linguistics, pages 4018–4027
October 12–17, 2022.

The FrAC vocabulary is about information from or derived from corpora that can be included in machine-readable dictionaries and other forms of lexical or ontological resources, i.e., information about 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). For these, FrAC introduced a generalization over the OntoLex core elements (and any other entity FrAC-related information is to be expressed about), and introduced the notion of `frac:Observable`, i.e., a lexical unit that can be observed in natural language, e.g., in a corpus. The corpus class was a another vocabulary element introduced with FrAC, and it is understood here in the more general sense of structured (collections of) primary data.[1] In addition to representing the primary data itself, it can also provide the total number of tokens in the corpus `frac:total`.

The different FrAC sub-vocabularies then defined different concepts that define the relation between observables and the corpus (or anyURI) object. A novel contribution of our paper is that we introduce a generalization over these FrAC-specific classes. In analogy with `frac:Observable`, we refer to this as `frac:Observation`. An observation in this understanding 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. We consider aggregate observations (frequency, collocations, embeddings, similarity clusters) to be observations in their own right, as long as their characteristics are solely defined by the underlying data. FrAC observations have a number of common properties:

(1) `rdf:value`: value of an observation, with characteristics depending on the specific observation class.

(2) `dc:description`: human-readable characterization of the methods involved in the observation. FrAC does not provide a vocabulary for provenance – if such information is to be provided, we recommended to use Prov-O (Lebo et al., 2013)

(3) `frac:corpus` link from the observation to the structured data from which the observation was created.

3 FrAC Observations

We propose four main classes as subclasses of `frac:Observation`, i.e., frequency, attestation, collocations, embeddings, and similarity as summarized in Fig. 2.

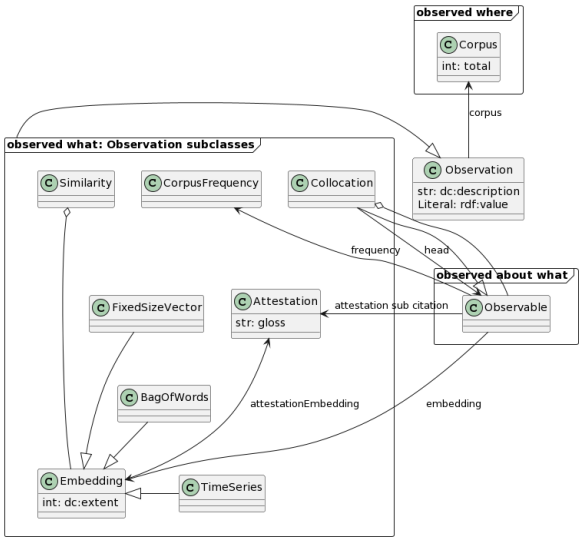


Figure 2: Revised FrAC vocabulary with the top-level classes `frac:Observable`, `frac:Observation` and `frac:Corpus`

3.1 Attestation

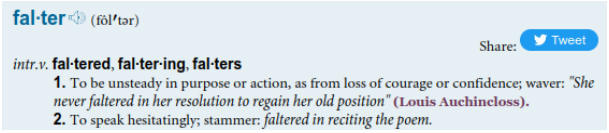
For attestation, the linking of lexical resources with corpus evidence, we distinguish three primary fields of application, i.e., lexicography (the use of references to corpora by a lexicographer to furnish evidence with reference to examples for the existence of a given lexical phenomena at a certain time period), language technology (linking a lexicon with the corpus from which information is derived), and corpus linguistics (linking a corpus [excerpt] with the lexical units or semantic annotations it provides). FrAC attestations are designed to support the different requirements in a unified way. FrAC defines `frac:Attestation` as an exact or normalized 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 and must define a locus object to identify the source of this material.

In the revised FrAC model, the attestation gloss – originally an independent property – is mod-

[1] In this more general sense, 'corpus' is also used in neighboring fields such as law (e.g., for Justinian's *Corpus Juris Civilis*) or archeology (e.g., for the Corpus Vasorum Antiquorum, a database of Greek vases). FrAC corpus thus comprises, but is not restricted to the sense of 'text corpus' (or speech corpus), i.e., a structured and/or electronically available and/or linguistically annotated collection of texts (or multimedia content).

elled as `rdf:value`. In its usage in lexicography, the attestation gloss differs from the quotation (`frac:quotation`) as it may include additional (human-readable) metadata about siglia, lines or versions that the actual primary data it refers to (the quotation) might not display. Similarly, the locus object (originally any URI as an object of `frac:locus`) is modelled as a corpus (`frac:Corpus` object of the `frac:corpus` property). This is in line with the fact that FrAC is underspecified as to the exact nature of `frac:corpus` objects, i.e., whether they represent the URI that resolves to the corpus that contains the lexical unit attested, or whether they represent the relevant excerpt of a corpus that contains the lexical unit, or whether they represent a metadata entry that stands in for a corpus which might not even exist in electronic form.

Any observable can be linked with its attestation by means of `frac:attestation`, defined as a subproperty of a more general `frac:citation` property – which posits no constraints on its range and which has been introduced to accommodate the needs of lexicographers who want to include attestations from secondary sources (Khan and Boschetti, 2018). The object of `frac:citation` is thus any URI, but for objects other than attestations, FrAC users are encouraged to follow any of the existing vocabularies for bibliographic data in RDF (Saur, 1998; Peroni and Shotton, 2012).


fal·ter (fôl′tər)

Share: Tweet

intr.v. **fal·tered, fal·ter·ing, fal·ters**

1. To be unsteady in purpose or action, as from loss of courage or confidence; waver: *"She never faltered in her resolution to regain her old position"* **(Louis Auchincloss).**

2. To speak hesitatingly; stammer: *faltered in reciting the poem.*

Figure 3: Attestations in the American Heritage Dictionary (accessed 2022-05-17)

Figure 3 shows a sample entry from the (online) American Heritage Dictionary (AHDictionary) of the English Language, and the attestation for its second sense can be modelled as follows:

```
:le_falter_vi
  a ontolex:LexicalEntry ;
  ontolex:sense :ls_falter_vi_2 .
:ls_falter_vi_2
  a ontolex:LexicalSense ;
  rdfs:comment "To speak hesitatingly,
      ..." ;
  frac:attestation [
    a frac:Attestation ;
    rdf:value "faltered in reciting the
        poem" ] .
```

While this attestation does not point to a corpus, the original 1969 edition of the dictionary uses the Brown corpus as a basis for its attestations, and for an example that *would have* come from the Brown corpus, we could give the link to the relevant subsections of the corpus in its online edition provided by SketchEngine.

```
... a frac:Attestation ;
    frac:corpus <https://app.
        sketchengine.eu/#concordance?
        corpname=preloaded%2Fbrown_1&
        keyword=falter&showresults=1> ;
    rdf:value "faltered in reciting the
        poem"
```

3.2 Frequency

The frequency distribution of linguistic elements is one of the most fundamental corpus-based statistics. In general, frequency information is critical to corpus studies, linguistic analysis, and NLP. We can distinguish between absolute and relative frequencies. Relative frequencies are generally normalized and computed as frequencies per a pre-defined number of linguistic elements. The FrAC module considers both absolute and relative frequency in order to facilitate different necessities. However, in terms of modelling, the focus is on absolute frequencies that are defined in relation to a particular corpus. The `frac:CorpusFrequency` class gives the absolute number of attestations, i.e., `rdf:value`, of a single `frac:Observable` considering a specific language resource, i.e., `frac:corpus`. Auxiliary filter conditions can be added to extend the `frac:CorpusFrequency` class with views of different sub-corpora.

An example is to restrict the subcorpus to a particular period. By means of OWL restrictions, a corpus-specific subclass of `frac:CorpusFrequency` can be created, say, `my:XYZCorpusFrequency` for which values for `dc:description`, `frac:corpus` and other parameters are defined as fixed. If then, the object of `frac:frequency` is defined as a `my:XYZCorpusFrequency`, these values do not have to be repeated, but are, instead, inherited from the class definition. As an example for frequency, we again, resort to the Brown corpus as provided by SketchEngine:

```
:BrownCorpusFreq
 rdfs:subClassOf
  frac:CorpusFrequency ,
   [ a owl:Restriction ;
    owl:onProperty frac:corpus ;
    owl:hasValue <https://app.
        sketchengine.eu/#concordance?
```

```
        corpname=preloaded%252Fbrown_1>
        ] .
```

For *falter*, the number of hits returned by querying SketchEngine can be modelled as a `:BrownCorpusFreq` in FrAC, then:

```
:le_falter_vi
  frac:frequency
    [ a :BrownCorpusFreq ;
      rdf:value "6" ] .
```

When re-defining corpus frequency as a `frac:Observation`, no semantic changes are necessary, except that `rdf:value` and `frac:corpus` are inherited now rather than defined for corpus frequency.

3.3 Collocation

A collocation is an expression containing two or more juxtaposition words that statistically appear together more frequently than by chance. The individual words of a collocation are characterized by the property of limited compositionality with each other, since they are predictable and, when they occur together, e.g. in the case of multi-word expressions, compound nouns, etc., they can have meanings that are different from their meanings when they occur alone or in other word combinations. Thus, some words can be freely combined with each other, others tend to combine only with certain words. They are word combinations that lie in a range between free and fixed. Collocation analysis is used in natural language processing, especially in automatic machine translation, in text generation, e.g. to make the output text as natural as possible, and to avoid untypical word combinations (Manning and Schütze, 1999; Evert, 2008).

In FrAC, collocations are modeled as an aggegate (`rdfs:Container`) of `frac:Observables`. Fixed word order collocations are defined using `rdf:Seq`) as a sequence, while variable word order collocations are defined using `rdf:Bag` as an ordered set. Collocations obtained by quantitative methods are characterized by their method of creation (`dc:description`), first word (`frac:head`), collocation strength (`rdf:value`), and the corpus used to create them (`frac:corpus`). Furthermore, collocations share these characteristics with other types of contextual relations. In previous FrAC proposals, these were thus inherited from the abstract class `frac:ContextualRelation` for the relation between two or more lexical elements. In our revised FrAC vocabulary,

LARGE CTy 82; CF 208; CTe 57	EF	IF	RF	TC	DI
LARGE AMOUNT OF DATA aha	2	2	*	2	4

Figure 4: Collocation analysis for the head word *large* and the collocation *(a) large amount of data* over the Brown corpus according to Kjellmer (1994) as given by Johansson (1998, p.339)

`frac:ContextualRelation` has been superseded by `frac:Observation`. A FrAC collocation is thus an aggregate (bag or sequence) of observables based on their co-occurrence within the same context window and characterized the head word of the collocation (`frac:head`) and the collocation score (`frac:cscore`) in a particular source corpus (`frac:corpus`).

Collocations are `frac:Observables`, they can also be given an `frac:attestation`, `frac:embedding` or `frac:frequency`. Using the embeddings, we can determine nested collocation by computing a similarity metric (e.g., cosine similarity). The collocation score (`frac:cscore`) is a subproperty of `rdf:value` that provides a specific corpus dependent collocation score. In FrAC we define multiple symmetric and asymmetric collocation metrics as sub-properties of `frac:cscore`, e.g., `frac:rel_freq` for the relative frequency (asymmetric), `frac:pmi` for the pointwise mutual information (symmetric), `frac:chi2` for Person's Chi-square test (asymmetric), etc. For asymmetric collocations scores the `frac:head` property is used to identify the elements' order.

As an example for collocation analysis over the Brown corpus, we refer to the classical work by Kjellmer (1994), who provides (candidate) collocations along with different scores in a tabular format (Fig. 4). Kjellmer's work differs from more recent works in collocation analysis in that he focuses on absolute frequencies rather than designated collocation scores. For this kind of data, it is sufficient to resort to `frac:CorpusFrequency` (resp., designated subclasses such as `:Inclusive-BrownFrequency`, with fixed values for `frac:corpus` and `dc:description`), and as collocations are both observations and observables, this is possible in FrAC:

```
:coll_laod
  a frac:Collocation, rdf:Seq ;
  rdf:_1 :le_large ;     # lexical
  rdf:_2 :le_amount ;    # entries
  rdf:_3 :le_of ;
  rdf:_4 :le_data ;
  frac:head :le_large ; # head
  frac:frequency  # frequencies
```

```
[ a :InclusiveBrownFrequency ; rdf:
    value "2" ] .
```

More conventional scores can be expressed with the designated subproperties of `frac:cscore` (or, if this is unambiguous, `rdf:value`).

3.4 Embeddings Subclasses

In the context of FrAC, the notion of embedding has been understood in a sense established in mathematics. An 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 fixed-size vectors (resp., tensors as aggregates of such vectors), but as there are other forms of numerical representations that serve similar or identical functions, FrAC introduces a more general class of `frac:Embedding` along with a specific sub-class of embeddings `frac: FixedSizeVector` to for embeddings as typically found in NLP. 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). Mathematically, bags of words could also be described as infinite-size embeddings (if the vocabulary is not completely known in advance), and indeed, earlier methods for dimensionality reduction motivated embeddings as a compact form of bags of words (Schütze, 1992, with slightly different wording).

Time series data is another form of infinite-size embeddings, but here, it is an infinite-size series of finite-size vectors. In language technnology, a stream of text, mapped to word embeddings, is such a structure – but normally not stored. A lexicographically more relevant use case is in sensor data, e.g., for the recording of gestures for sign languages. Such recordings can then be compared with each other using techniques such as dynamic time warping (Gold and Sharir, 2018), and then be the basis for automated clustering, etc.

3.5 Word and Concept Embeddings

In FrAC, any observable can be assigned an embedding. This includes lexical form, lemmas (lexical entries), word senses (lexical senses), lexical concepts and other entities, multi-word expressions (as lexical entries) and groups of observables (FrAC collocations). This also partially answers the question on why embeddings (esp., fixed-size vectors for NLP embeddings) are a necessary data structure for FrAC. In many use cases, word embeddings are created on the fly and not shared across different applications – but as their creation involves a non-deterministic element, they cannot be easily compared across languages or corpora. For this reason (and because, historically, the creation of embeddings from large-scale corpora was a matter of weeks or months of processing), applications often use precompiled embeddings for either subsequent fine-tuning or directly. As far as word embeddings are concerned, it does – again – not seem to be necessary to represent these in RDF. The typical structure of an embedding file is a table, with the first column representing the token, the following columns representing the embedding with one value per cell. As long as applications refer to the same embedding file with the same parameters (same length, same tokenization, same normalization for strings [e.g., lowercasing], same vector normalization function – e.g., to the spans of either $[0, ..., 1]$ or $[-1, ..., 1]$ –, same underlying corpus data), they will operate in the same embedding space, and with libraries such as TextTorch (TorchText, 2022) or repositories like HuggingFace (Wolf et al., 2020), there is an established infrastructure to retrieve identical word embeddings using standard identifiers. However, this can nevertheless be problematic, especially if different applications retrieve their embeddings from different sources. While the retrieval of, say, GloVe embeddings (Pennington et al., 2014) via TextTorch or via the original provider *should* lead to the same result, this cannot be automatically validated as neither source provides machine-readable metadata – nor is there any transparent relation between both methods of access unless TextTorch code is manually inspected.

RDF metadata can help here to resolve ambiguities. And ambiguities exist, and will intensify with the adoption of current techniques and data in new programming languages and future ecosystems. A classical example are the infamous 'Collobert & Weston embeddings' which is a term applied to two different and unrelated sets of embeddings also known as 'SENNA embeddings' on the one hand and 'Turian embeddings' on the other (Collobert, 2011).

Another objective to provide embeddings in RDF is that the basic table format in which embeddings are shared is insufficient if these embeddings are detached from the definition of the elements they are assigned to. An important category here are embeddings of lexical concepts and lexical senses as derived, for example, from underlying word embeddings, and thus residing in the same feature space, e.g., the classical AutoExtend embeddings (Rothe and Schütze, 2017) whose synset identifiers are ambiguous as to which WordNet version they refer to. While this can be solved with using persistent URIs as synset identifiers, for lexical resources for which no publicly accessible, resolvable or persistent URIs can be provided, an alternative solution is to *bundle* embeddings and the underlying knowledge graph into a data structure from which both the graph and the embeddings can be accessed, and FrAC provides the vocabulary to provide that from the RDF perspective.

3.6 Contextualized Embeddings

Another aspect in which static word embeddings have been superseded by more recent developments is the rise of transformer architectures operating with subsymbolic embeddings and the processing of text spans rather than static lexemes. *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. For encoding multiple contextual embeddings for a particular lexical entry, say, *tree*, it is necessary to create one attestation with one attestation embedding property. If possible, the attestation object should be linked to the respective passage in the corpus, but in the spirit of the open world assumption in RDF semantics, this information is optional, so that a minimal encoding of contextual embeddings in a lexical resource can use the following template:

```
:le_tree a ontolex:LexicalEntry ,
           frac:Observable ;
  frac:embedding [
    a frac:FizedSizeVector;
    dc:extent "50";
    rdf:value "[ 0.0001, ... ]" ];
  frac:attestation [
    frac:attestationEmbedding [
      a frac:FixedSizeVector;
      dc:extent "50";
      rdf:value "[ 0.5352, ... ]" ] ].
```

From the perspective of a corpus, both contextualized and context-free embeddings can be encoded correspondingly. If we use the CoNLL-RDF vocabulary for identifying tokens in a corpus (FrAC can be used with any vocabulary for this purpose, e.g., Web Annotation or NIF), and the token `doc:s1_5` has already been defined, then, it can just be linked with the attestation:

```
... frac:attestation [
      frac:corpus doc:s1_5;
      frac:attestationEmbedding
        [ ... ] ] .
```

For the token `doc:s1_5`, we can now easily retrieve various kinds of embeddings:

```
doc:s1_5 ^frac:corpus ?att.
?att frac:attestationEmbedding
       ?contextualEmbedding .
?att ^frac:attestation
   [ a ontolex:LexicalEntry;
     frac:embedding ?entryEmbedding ].
doc:s1_5
  ^frac:corpus/^frac:attestation
   [ a ontolex:LexicalForm;
     frac:embedding ?formEmbedding ].
doc:s1_5
  ^frac:corpus/^frac:attestation
   [ a ontolex:LexicalSense;
     frac:embedding ?senseEmbedding ].
doc:s1_5
  ^frac:corpus/^frac:attestation
   [ a ontolex:LexicalConcept;
     frac:embedding ?conceptEmbedding ].
```

These partial queries operate on separate attestations for every kind of observables. However, the model is generic enough to also follow indirect links: Using OntoLex core data structures, a sense attestation can server as an anchor to retrieve embeddings for lexical sense, but also lexical concept or lexical entry: For the Brown corpus, a concrete application can be seen in the SemCor corpus (Fellbaum et al., 1997), a layer of semantic annotation (WordNet senses), and with the following query we can retrieve AutoExtend synset embeddings and contextualized sense embeddings:

```
SELECT ?contextualEmbedding
       ?synsetEmbedding
WHERE {
  ?att frac:attestationEmbedding [
```

```
  rdf:value ?contextualEmbedding ] .
?att ^frac:attestation ?sense.
?sense
  a ontolex:LexicalSense;
    ontolex:isLexicalizedSenseOf
      ?synset.
?synset a ontolex:LexicalConcept;
  frac:embedding/rdf:value
      ?synsetEmbedding .
```

Such data can then be used, for example, to train a mapping from contextual embeddings to synset embeddings.

In their earlier formulation of the FrAC vocabulary, `frac:Embedding` had the following attributes: *(1)* `dc:extent` dimensionality of embeddings (for fixed-size vectors), or the number of data points per observation (in time series data) *(2)* `rdf:value` value of the embeddings, according to the examples, this should be a JSON literal, e.g., an array (of floats) or a hashtable of keys (e.g., context words) and numerical weights. *(3)* `dc:description` human-readable description of embedding type and parameters *(4)* `frac:corpus` URI of the underlying corpus data

With the revised upper model, embedding inherits `rdf:value`, `dc:description` and `frac:corpus` from `frac:Observation`. As part of the generalization, the restriction of `rdf:value` to JSON literals is abandoned – and this may, indeed, conincide with external requirements to the FrAC vocabulary, as it seems easier at times to encode embeddings (fixed size vectors or bags of words) just as plain strings, as such data can be more easily created from existing resources.

3.7 Similarity

Similarity relates to computing the strength of the semantic relationships between different elements, e.g., forms, lexemes, and phrases. There are various similarity metrics, but in the FrAC context, similarity is obtained through a numerical description of the contexts for each of the analysed elements, i.e., their embeddings (Sect. 3.4).

In FrAC, 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`). In the revised FrAC model, the earlier `frac:ContextualRelation` superclass of `frac:Similarity` from which its properties were inherited has been replaced by `frac:Observation`, no further renaming necessary. FrAC similarity can be applied to both similarity relations (sets of two observables) and similarity clusters (sets of two or more observables) characterized by a single value.

The value of a similarity is an `rdf:value` calculated according to the employed Embedding model, e.g., the number of shared dimensions – in a bag-of-word model. This value is published together with its corresponding metadata, *(2)* one or more source corpora, i.e., `frac:corpus`, *(2)* the description of the comparison method, i.e., `dc:description`.

We can use `frac:Similarity` for different scenarios, two being exemplified in FrAC, i.e., the similarity between two words, and similarity clusters. Similarity clusters are useful in computational linguistics for tasks that rely on cognate recognition and language similarity. When applied to a particular corpus, similarity cluster offers a generalization score over all the pairs of similes. This generalization method can use different approaches, e.g., the minimal similarity between all members in the cluster, or a score given by the clustering algorithm. The used approach must be explained in `dc:description`.

A very simple example for Similarity is cosine similarity, as can be calculated between fixed size vectors. Using the AutoExtend embeddings, the cosine similarity between any two lexical concepts can be modelled as follows:

```
:ls_abc a ontolex:LexicalSense ;
  frac:embedding :ls_abc_embedding .
:ls_xyz a ontolex:LexicalSense ;
  frac:embedding :ls_xyz_embedding .
[ dc:description
  "cosine similarity" ]
  a frac:Similarity, rdfs:Bag ;
  rdfs:member :ls_abc_embedding ,
              :ls_xyz_embedding ;
  rdf:value "0.0036" .
```

4 Consolidation and Outlook

The revised FrAC vocabulary proposed in this paper introduces the novel class `frac:Observation` as a generalization over different kinds of phenomena that can be observed from or derived from corpus data and that are relevant for lexical resources. With small changes to previously proposed vocabulary elements, this revised top-level structure can be seamlessly applied to use cases and sample data featured in previous publications on FrAC.

The following changes have been proposed: *(1)* merge the `frac:locus` property of `frac:Attestation` into `frac:corpus`; *(2)* extend the understanding of `frac:Corpus` / range of `frac:corpus` to cover any piece of structured (collections of) primary data, including parts thereof; *(3)* merge `frac:quotation` and `rdf:value`; *(4)* abandon previous restrictions on the range of `rdf:value`; and *(5)* merge `frac:ContextualRelation` with the newly created class `frac:Observation`

Aside from inheriting common characteristics from a newly created generalization, we claim that this model is equivalent in expressivity to the current formulation of FrAC as available from the public draft of the vocabulary. but that it features a more systematic structure in that the common pattern exhibited for modelling the different types of observations and corpus-derived information is now explicitly encoded in the model, making the overall model both easier to describe, more compact and easier to formalize in RDFS semantics.

We illustrate the applicability of this model to a number of examples, mostly with reference to the Brown corpus. We argue that with this in formation, it becomes possible for the first time to encode both the majority of lexical information derived from the Brown corpus in a unified way, and to thus integrate those resources on a technical level. With FrAC, all these different aspects can be encoded in RDF and this representation can be the basis to define interoperable APIs for different web services, APIs or applications to produce, consume or integrate such data. At the moment, the state of the art in this area is probably best represented by the *proprietary* SketchEngine APIs, whose responses are, however, do not come with any guarantees for long-term stability or reproducibility. Furthermore, a number of aspects are not well-supported by SketchEngine: This includes, for example, the online reference to individual attestations (SketchEngine only provides resolvable URIs for query responses, but not for the individual matches), or the retrieval of embeddings (provided by SketchEngine but only as data dumps, not integrated in the API). Another possible application is to provide dumps of corpus-derived information (of attestations, embeddings, collocations, similarity clusters or frequency lists) along with the associated lexical graph.

As it provides uniform data structures on the basis of web standards, FrAC represents the fundament to develop consistent access protocols for the unified access, public exposure, exchange and integration of heterogeneous data as currently provided, for example, via the Linguistic Linked Open Data cloud (Chiarcos et al., 2011; Declerck et al., 2020), libraries such as NLTK (Bird, 2006) or via portals such as HuggingFace (Lhoest et al., 2021). At the same time, FrAC accomodates the needs of digital lexicography and the language sciences, and has partially been motivated by applicability to philological data (Chiarcos et al., 2020) and multimedia content (Chiarcos et al., 2011). As a result of applying OntoLex and FrAC, resources developed in lexicography become accessible and re-usable in the context of language technology, and resources and solutions developed in language technology become applicable to lexicographic and linguistic data and research challenges.

The second novel contribution of this paper is that we provide an OWL2/DL ontology as formalization of the revised FrAC vocabulary. This allows to automatically validate FrAC data, to detect inconsistencies and to perform reasoning (inferences) over FrAC data. In particular, types (classes) of observations and observables can be automatically (RDFS-)inferred from domain and range constraints of properties such as `frac:frequency`, `frac:attestation`, `frac:embedding`, etc., so that this information is in fact optional in data exchange. Likewise, formal OWL2/DL axioms allow users to define application-specific subclasses of `frac:CorpusFrequency`, etc., so that these can be used as a short hand for specific bundles of observations with `frac:corpus`, `rdf:value`, `dc:description` and other properties that constrain the corpus under consideration or that define hyperparameters used in the extraction process.

It is important to note that – to the best of our knowledge – no RDF vocabulary of similar scope was in existence prior to FrAC, and that the integration with OntoLex facilitates a relatively wide application across different disciplines and research networks. It is less certain whether there are comparable pre-RDF vocabularies in existence. We mentioned the SketchEngine API and exchange formats, but as far as *open (community) standards* are concerned, we are not aware of any related work of similar scope. Nevertheless, aspects of corpus-driven lexicography have been addressed in the Lexical Markup Framework (Romary et al., 2019,

for attestation) and the TEI guidelines (Burnard, 2013, for collocation), but we are not aware of any formal standard for embeddings or machine-readable similarity scores. By extending OntoLex with a designated module for frequency, attestation and corpus-based information in lexical resources, we are thus breaking novel ground.

Acknowledgments

The research described in this paper was conducted in the context of the Cost Action CA18209 *Nexus Linguarum. European network for Web-centred linguistic data science*. Ciprian-Octavian Truică was partially funded through the OPTIM Research project (POCU grant no. 62461/03.06.2022, SMIS code 153735). Moreover, the authors would like to thank Maxim Ionov, Anas Fahad Khan, and Julia Bosque-Gil for contributing to the development of OntoLex-FrAc as well as to all OntoLex-FrAC contributors.

References

Steven Bird. 2006. Nltk: The natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Lou Burnard. 2013. The evolution of the text encoding initiative: from research project to research infrastructure. *Journal of the Text Encoding Initiative*, (5).

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, Katerina Gkirtzou, Maxim Ionov, Besim Kabashi, Anas Fahad Khan, and Ciprian-Octavian Truică. 2022. Modelling collocations in OntoLex-FrAC. In *Proceedings of GlobaLex-2022*, Marseille, France.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2011. Towards a linguistic linked open data cloud: The open linguisticsworking group. *TAL Traitement Automatique des Langues*, 52(3):245–275.

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.

Ronan Collobert. 2011. SENNA. `https://ronan.collobert.com/senna/`, accesased 2022-05-16.

Thierry Declerck, John Philip McCrae, Matthias Hartung, Jorge Gracia, Christian Chiarcos, Elena Montiel-Ponsoda, Philipp Cimiano, Artem Revenko, Roser Sauri, Deirdre Lee, et al. 2020. Recent developments for the linguistic linked open data infrastructure. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5660–5667.

Stefan Evert. 2008. Corpora and collocations. In Kytö M Lüdeling A, editor, *Corpus Linguistics. An International Handbook*, pages 1212–1248. Mouton de Gruyter, Berlin, New York.

Christiane Fellbaum, Joachim Grabowski, and Shari Landes. 1997. Analysis of a hand-tagging task. In *In Proceedings of ANLP-97 Workshop on Tagging Text with Lexical Semantics, Why, What, and How?*, Washington, D.C.

Omer Gold and Micha Sharir. 2018. Dynamic time warping and geometric edit distance: Breaking the quadratic barrier. *ACM Trans. Algorithms*, 14(4).

Stig Johansson. 1998. Göran kjellmer. a dictionary of english collocations, based on the brown corpus. *International Journal of Corpus Linguistics*, 3(2):338–348.

Anas Fahad Khan and Federico Boschetti. 2018. Towards a Representation of Citations in Linked Data Lexical Resources. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 137–147, Ljubljana, Slovenia. Ljubljana University Press.

Göran Kjellmer. 1994. *A Dictionary of English Collocations. Based on the Brown Corpus*. Clarendon Press. 3 vols.

Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. 2013. Prov-o: The prov ontology. *W3C Recommendation*.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184. Association for Computational Linguistics.

Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press Ltd.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Silvio Peroni and David Shotton. 2012. Fabio and cito: ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17:33–43.

Laurent Romary, Mohamed Khemakhem, Fahad Khan, Jack Bowers, Nicoletta Calzolari, Monte George, Mandy Pet, and Piotr Bański. 2019. Lmf reloaded. *arXiv preprint arXiv:1906.02136*.

Sascha Rothe and Hinrich Schütze. 2017. Autoextend: Combining word embeddings with semantic resources. *Computational Linguistics*, 43(3):593–617.

KG Saur. 1998. IFLA study group on the functional requirements for bibliographic records.

Hinrich Schütze. 1992. Dimensions of meaning. In *Supercomputing'92: Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, pages 787–796. IEEE.

TorchText. 2022. TorchText 0.4.0 documentation. Technical report.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.