Feb 12, 2021 OntoLex-FrAC

Frequency, Attestations, Corpus Information

Christian Chiarcos
Goethe University Frankfurt

Nexus Linguarum

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Motivation

a. Corpus-based lexicography
   i. Frequency, corpus links, “word sketches” (similarity and collocation analysis)
b. Lexical data for Natural Language Processing
   i. Frequency, dictionary links, similarity and collocation analysis, sense embeddings

Requirements

- representation of frequency information
  - corpus-based lexicography (Electronic Penn Sumerian Dictionary, ePSD)

a [WATER] (N) 4683 in

a [WATER] N (4683x) Early Dynastic IIIa, Early Dynastic IIIb, Ebla, Old Akkadian, Lagash II, Ur III, Old Babylonian, Middle Assyrian, Middle Babylonian, Neo-Assyrian, Neo-Babylonian, Hellenistic, Uncertain, unknown wr. a; a_2; e; e·n; g·a; u "progeny; semen; water; watercourse"

[1]		a
[2]		a_2
[3]		e
[4]		e·n
[5]		g·a
[6]		u

	PC	ED IIIa	ED IIIb	Ebla	OAkk	Lag II	Ur III	OB	Post-OB	(unknown)
[1]		20	254	3	74	96	2299	1523	316	71
[2]								1		
[3]								21	2	1
[4]								1		
[5]										
[6]									1	

35 distinct forms attested; click to view forms table.

Requirements

- representation of frequency information
  - corpus-based lexicography (ePSD)

lexical entry *a [WATER]*

a [WATER] (N)

a [WATER] N (4683x) Early Dynastic IIIa, Early Dynastic IIIb, Ebla, Old Akkadian, Lagash II, Ur III, Old Babylonian, Middle Assyrian, Middle Babylonian, Neo-Assyrian, Neo-Babylonian, Hellenistic, Uncertain, unknown wr. a; a_2; e; e·n; g·a; u "progeny; semen; water; watercourse"

sense and language definition

total lexeme frequency

Form/written representations (original and transcript)

corpus distribution over time and region

[1]		a
[2]		a_2
[3]		e
[4]		e·n
[5]		g·a
[6]		u

	PC	ED IIIa	ED IIIb	Ebla	OAkk	Lag II	Ur III	OB	Post-OB	(unknown)
[1]		20	254	3	74	96	2299	1523	316	71
[2]								1		
[3]								21	2	1
[4]								1		
[5]										
[6]									1	

35 distinct forms attested; click to view forms table.

Motivation

- representation of frequency information
  - corpus-based lexicography
  - frequency dictionaries (e.g., stop word lists for NLP)
- attestations
  - in dictionaries: real-world examples that illustrate a particular lexical entry, a lexical form, a sense or a concept

probatio , onis, f. [probo] .
I. A trying , proving; a trial , inspection , examination (class.): athletarum probatio, Cic. Off. 1, 40, 144 : futura, id. Verr. 2, 1, 54, § 142 ; Varr. R. R. 1, 20, 1: oesypi, Plin. 29, 2, 10, § 36 : croci sinceri, id. 21, 6, 17, § 32 : pumicis, id. 36, 21, 42, § 155 : gemmae recusan limae probationem, id. 37, 13, 76, § 200 : equitum, a review , Val. Max. 2, 2, 9 .--
II. In partic.

Latin, Lewis & Short,
http://www.inrebus.com/latindictionary.php

Motivation

- representation of frequency information
  - corpus-based lexicography
  - frequency dictionaries (e.g., stop word lists for NLP)
- attestations
  - in dictionaries: real-world examples that illustrate a particular lexical entry, a lexical form, a sense or a concept
  - more generally: any link from a lexical resource into a corpus

Motivation

- representation of frequency information
  - corpus-based lexicography
  - frequency dictionaries (e.g., stop word lists for NLP)
- attestations
- corpus data
  - similarity clusters (NLP: Brown clusters)
  - similarity metrics (in lexicography or NLP)
  - collocations (=> lexicographical similarity metrics)
  - embeddings (=> NLP similarity metrics)
  - multimodality (what about data other than text?)

Progress since October 2018

- representation of frequency information
  - corpus-based lexicography
  - frequency dictionaries (e.g., stop word lists for NLP)
- attestations
- corpus data
  - similarity clusters (NLP: Brown clusters)
  - similarity metrics (in lexicography or NLP)
  - collocations (=> lexicographical similarity metrics)
  - embeddings (=> NLP similarity metrics)
  - multimodality (what about data other than text?)

Progress since October 2018

- representation of frequency information
  - corpus-based lexicography
  - frequency dictionaries (e.g., stop word lists for NLP)
- attestations
- corpus data
  - similarity clusters (NLP: Brown clusters)
  - similarity metrics (in lexicography or NLP)
  - collocations (=> lexicographical similarity metrics)
  - embeddings (=> NLP similarity metrics)
  - multimodality (what about data other than text?)

Progress since October 2018

- representation of frequency information
  - corpus-based lexicography
  - frequency dictionaries (e.g., stop word lists for NLP)
- attestations
- corpus data
  - similarity clusters (NLP: Brown clusters)
  - similarity metrics (in lexicography or NLP)
  - collocations (=> lexicographical similarity metrics)
  - embeddings (=> NLP similarity metrics)
  - multimodality (what about data other than text?)

OntoLex-FrAC

- Draft and samples on GitHub
  - https://github.com/ontolex/frequency-attestation-corpus-information
- Core classes
- Frequency
- Attestation
- Embeddings

General structure

Observable and Observation

Nexus Linguarum

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Observable and Observation

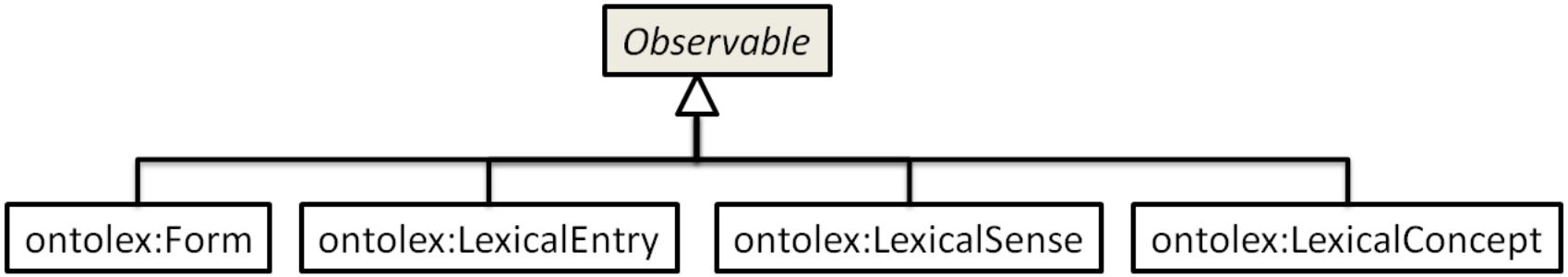
- FrAC aims at complementing lexical data with all relevant kinds of information drawn from a corpus
  - A corpus is any kind of collection of external data
  - It does not have to be a digital corpus or an annotated corpus
  - Any external data *sample* qualifies as a corpus

Observable and Observation

- FrAC aims at complementing lexical data with all relevant kinds of information drawn from a corpus
- The information that is extracted must be about an entity that can be observed or detected in a corpus
  - It does not have to be in there, but if it is, we must be able to observe it
  - frac:Observable
    - includes all elements of OntoLex core (entry, sense, form, concept)
    - can be applied to elements of *any* ontology (*ontolex:reference*)
    - can be extended to other entities (morph?)

Observable and Observation

- FrAC aims at complementing lexical data with all relevant kinds of information drawn from a corpus
- The information that is extracted must be about an entity that can be observed or detected in a corpus
  - frac:Observable



Observable and Observation

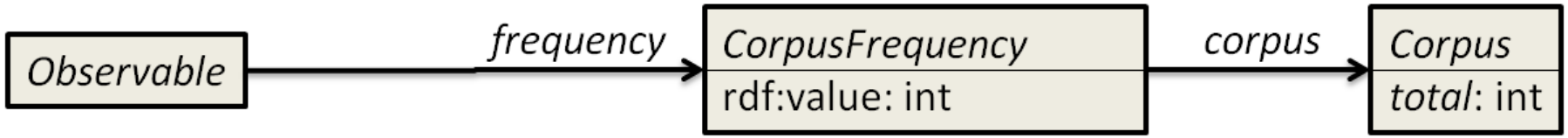
- FrAC aims at complementing lexical data with all relevant kinds of information drawn from a corpus
- The information that is extracted must be about an entity that can be observed or detected in a corpus
  - frac:Observable
- Every type of observation is a separate class, linked with a designated property

  frac:Observable => ontolex:frequency => ontolex:CorpusFrequency

  frac:Observable => ontolex:attestation => ontolex:Attestation

  frac:Observable => ontolex:embedding => ontolex:Embedding

frac:frequency



Nexus Linguarum

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

frac:frequency

a [WATER] (N) 4683 in

a [WATER] N (4683x) Early Dynastic IIIa, Early Dynastic IIIb, Ebla, Old Akkadian, Lagash II, Ur III, Old Babylonian, Middle Assyrian, Middle Babylonian, Neo-Assyrian, Neo-Babylonian, Hellenistic, Uncertain, unknown wr. a; a_2; e; e·n; g·a; u "progeny; semen; water; watercourse"

[1]		a
[2]		a_2
[3]		e
[4]		e·n
[5]		g·a
[6]		u

	PC	ED IIIa	ED IIIb	Ebla	OAkk	Lag II	Ur III	OB	Post-OB	(unknown)
[1]		20	254	3	74	96	2299	1523	316	71
[2]								1		
[3]								21	2	1
[4]								1		
[5]										
[6]									1	

35 distinct forms attested; click to view forms table.

frac:frequency

- representation of frequency information
  - corpus-based lexicography (ePSD)

lexical entry *a [WATER]*

a [WATER] (N)

a [WATER] N (4683x) Early Dynastic IIIa, Early Dynastic IIIb, Ebla, Old Akkadian, Lagash II, Ur III, Old Babylonian, Middle Assyrian, Middle Babylonian, Neo-Assyrian, Neo-Babylonian, Hellenistic, Uncertain, unknown wr. a; a_2; e; e·n; g·a; u "progeny; semen; water; watercourse"

sense and language definition

total lexeme frequency

Form/written representations (original and transcript)

corpus distribution over time and region

[1]		a
[2]		a_2
[3]		e
[4]		e·n
[5]		g·a
[6]		u

	PC	ED IIIa	ED IIIb	Ebla	OAkk	Lag II	Ur III	OB	Post-OB	(unknown)
[1]		20	254	3	74	96	2299	1523	316	71
[2]								1		
[3]								21	2	1
[4]								1		
[5]										
[6]									1	

35 distinct forms attested; click to view forms table.

frac:frequency

CorpusFrequency (Class) provides the absolute number of attestations (rdf:value) of a particular frac:Observable in a particular language resource (frac:corpus).

SubClassOf: rdf:value exactly 1 xsd:int, frac:corpus exactly 1

frequency (ObjectProperty) assigns a particular frac:Observable a frac:CorpusFrequency.

Domain frac:CorpusFrequency

Range frac:Observable

Corpus (Class) represents any type of linguistic data or collection thereof, in structured or unstructured format. At the lexical level, a corpus consists of individual elements (tokens, 'words'), and data providers should provide the total number of elements. It should also provide provenance information, e.g., the tokenization strategy, preprocessing steps, etc.

SubClassOf: frac:total exactly 1 xsd:int

corpus (Property) assigns a corpus to a particular frac:CorpusFrequency.

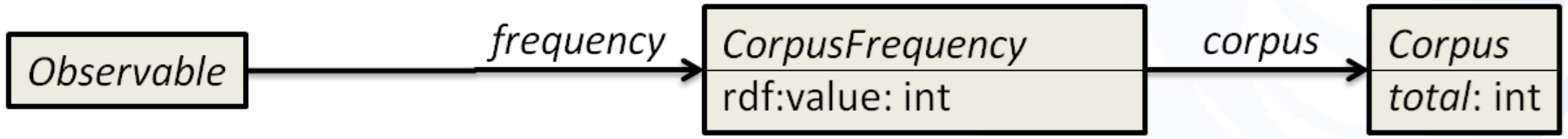
Domain: frac:CorpusFrequency

Range: frac:Corpus

total (Property) assigns a corpus the total number of elements that it contains. In the context of OntoLex, these are instantiations of lexemes, only, i.e., tokens ('words').

Domain: frac:Corpus

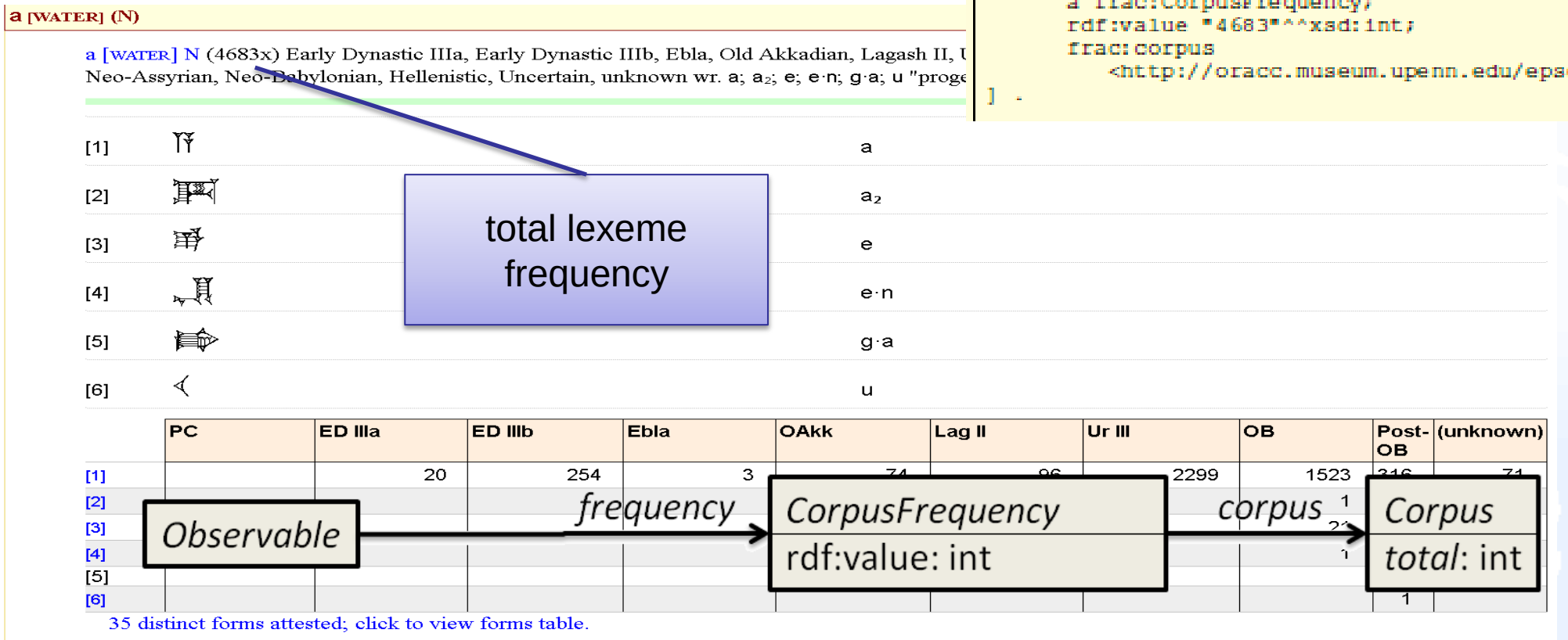
Range: integer (long)



frac:frequency

- representation of frequency information
  - corpus-based lexicography

```
# word frequency, over all form variants
epsd:a_water_n a ontolex:LexicalEntry;
   frac:frequency [
      a frac:CorpusFrequency;
      rdf:value "4683"^^xsd:int;
      frac:corpus
         <http://oracc.museum.upenn.edu/epsd2/pager>
] .
```



frac:frequency

- representation of frequency information
  - corpus-based lexicography

```
# word frequency, over all form variants
epsd:a_water_n a ontolex:LexicalEntry;
   frac:frequency [
      a frac:CorpusFrequency;
      rdf:value "4683"^^xsd:int;
      frac:corpus
         <http://oracc.museum.upenn.edu/epsd2/pager>
] .

# form frequency for individual orthographical variants
epsd:a_water_n a ontolex:canonicalForm [
   ontolex:writtenRep "a"@sux-Latn;
   frac:frequency [
      a frac:CorpusFrequency;
      rdf:value "4656"^^xsd:int;
      frac:corpus
         <http://oracc.museum.upenn.edu/epsd2/pager>
      ]
   ] .
```

a [WATER] (N)

a [WATER] N (4683x) Early Dynastic IIIa, Early Dynastic IIIb, Ebla, Old Akkadian, Lagash II, U
Neo-Assyrian, Neo-Babylonian, Hellenistic, Uncertain, unknown wr. a; a₂; e; e·n; g·a; u "proge

[1] a
[2] a₂
[3] e
[4] e·n
[5] g·a
[6] u

(row with) total form frequency

	PC	ED IIIa	ED IIIb	Ebla	OAkk	Lag II	Ur III	OB	Post-OB	(unknown)
[1]		20	254	3	74	96	2299	1523	316	71
[2]								1		
[3]								21		
[4]								1		
[5]										
[6]									1	

35 distinct forms attested; click to view forms table.

Observable —*frequency*→ *CorpusFrequency* (rdf:value: int) —*corpus*→ *Corpus* (*total*: int)

frac:frequency

- representation of frequency information
  - corpus-based lexicography

a [WATER] (N)

a [WATER] N (4683x) Early Dynastic IIIa, Early Dynastic IIIb, Ebla, Old Akkadian, Lagash II, U
Neo-Assyrian, Neo-Babylonian, Hellenistic, Uncertain, unknown wr. a; a_2; e; e·n; g·a; u "proge

[1]		a
[2]		a_2
[3]		e
[4]		e·n
[5]		g·a
[6]		u

(another row with) total form frequency

	PC	ED IIIa	ED IIIb	Ebla	OAkk	Lag
[1]		20	254	3	74	
[2]						
[3]						
[4]						
[5]						
[6]						

35 distinct forms attested; click to view forms table.

```
# word frequency, over all form variants
epsd:a_water_n a ontolex:LexicalEntry;
   frac:frequency [
      a frac:CorpusFrequency;
      rdf:value "4683"^^xsd:int;
      frac:corpus
         <http://oracc.museum.upenn.edu/epsd2/pager>
] .

# form frequency for individual orthographical variants
epsd:a_water_n a ontolex:canonicalForm [
   ontolex:writtenRep "a"@sux-Latn;
   frac:frequency [
      a frac:CorpusFrequency;
      rdf:value "4656"^^xsd:int;
      frac:corpus
         <http://oracc.museum.upenn.edu/epsd2/pager>
      ]
   ] .

epsd:a_water_n a ontolex:otherForm [
   ontolex:writtenRep "a2"@sux-Latn;
   frac:frequency [
      a frac:CorpusFrequency;
      rdf:value "1"^^xsd:int;
      frac:corpus
         <http://oracc.museum.upenn.edu/epsd2/pager>
      ]
   ] .
```

frac:frequency: metadata

- representation of frequency information
  - corpus-based lexicography (ePSD)

a [WATER] (N) 4683 ir

a [WATER] N (4683x) Early Dynastic IIIa, Early Dynastic IIIb, Ebla, Old Akkadian, Lagash II, Ur III, Old Babylonian, Middle Assyrian, Middle Babylonian, Neo-Assyrian, Neo-Babylonian, Hellenistic, Uncertain, unknown wr. a; a_2; e; e·n; g·a; u "progeny; semen; water; watercourse"

[1]		a
[2]		a_2
[3]		e
[4]		e·n
[5]		g·a
[6]		u

	PC	ED IIIa	ED IIIb	Ebla	OAkk	Lag II	Ur III	OB	Post-OB	(unknown)
[1]		20	254	3	74	96	2299	1523	316	71
[2]								1		
[3]								21	2	1
[4]								1		
[5]										
[6]									1	

35 distinct forms attested; click to view forms table.

corpus distribution over time and region

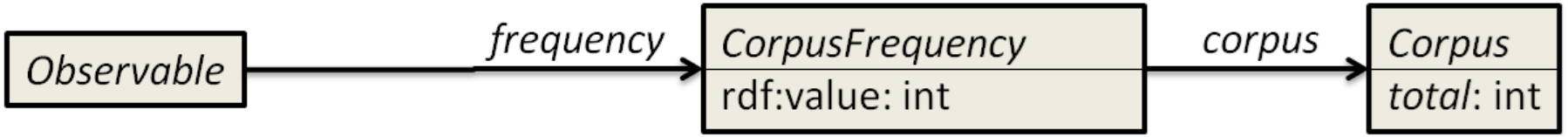
frac:frequency: metadata

- adding provenance
  - can be expressed on every individual frequency object
  - but can also be inherited from a corpus-specific subclass of frequency
  - we recommend the latter

```
:EPSDFrequency rdfs:subClassOf frac:CorpusFrequency .

:EPSDFrequency rdfs:subClassOf [
   a owl:Restriction ;
   owl:onProperty frac:corpus ;
   owl:hasValue
      <http://oracc.museum.upenn.edu/epsd2/pager>
] .

# frequency assessment
epsd:a_water_n frac:frequency [
   a :EPSDFrequency;
   rdf:value "4683"^^xsd:int
].
```

```
# EPSD frequency for the Ur-III period (aat:300019910)
:EPSDFrequency_UrIII
   rdfs:subClassOf :EPSDFrequency;
   rdfs:subClassOf [
      a owl:Restriction ;
      owl:onProperty dct:temporal ;
      owl:hasValue aat:300019910
   ] .

# frequency assessment for sub-corpus
epsd:a_water_n frac:frequency [
   a :EPSDFrequency_UrIII;
   rdf:value "2299"^^xsd:int
].
```

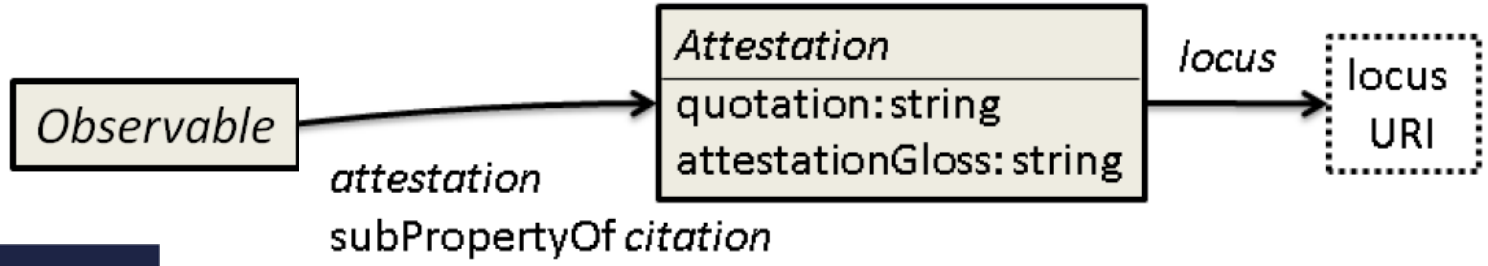
frac:frequency



Limitations:

- At the moment, we represent absolute frequencies only
- For relative frequencies, use the *frac:total* property of the *frac:Corpus* class
  - except for being able to have a *frac:total*, *frac:Corpus* is undefined

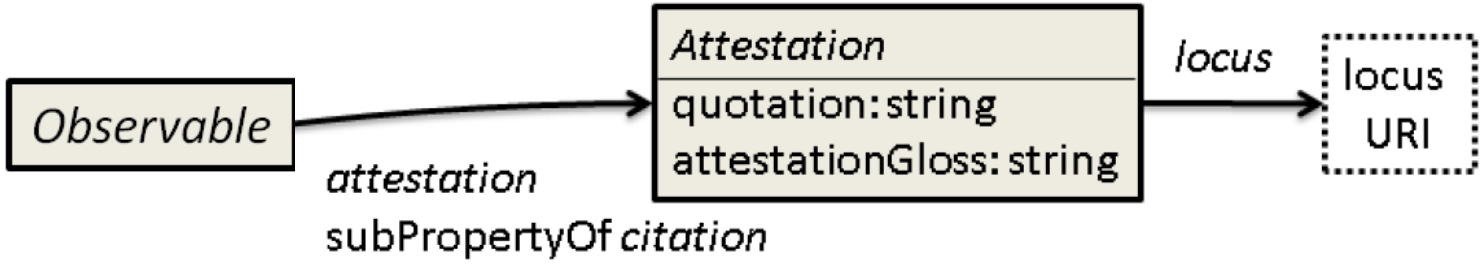
frac:attestation



Nexus Linguarum

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

frac:attestation



frac:Attestation class represents an exact or normalized quotation or excerpt from a source document that illustrates a particular form, sense, lexeme or features such as spelling variation, morphology, syntax, collocation, register.

frac:citation (domain: frac:Observable) Associates a citation to the frac:Observable citing it.

frac:attestation (domain: frac:Observable, range: frac:Attestation) Associates an attestation to the frac:Observable. This is a subproperty of frac:citation using it as evidence.

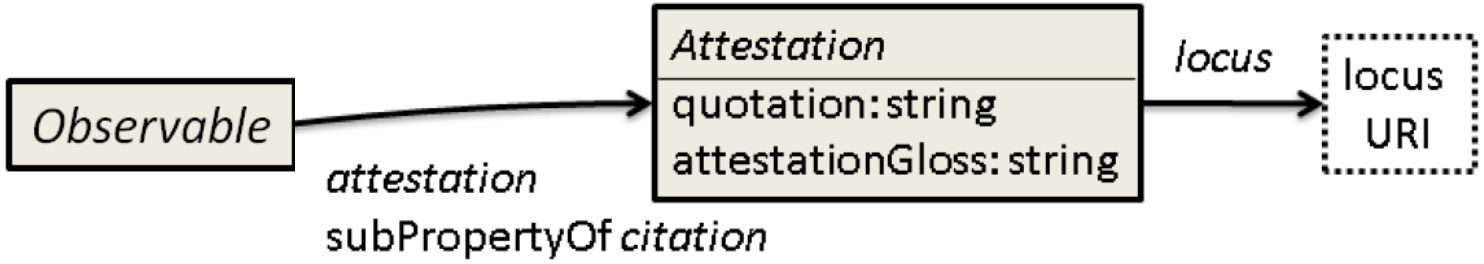
frac:quotation (range: xs:String) This contains the text content of the dictionary quotation.

frac:attestationGloss (domain: frac:Attestation, range: xs:String) This contains the text content of an attestation as represented within a dictionary. This may be different from a direct quotation because the target expression may be omitted or normalized.

frac:locus (domain: frac:Attestation) points to the location at which the relevant word(s) can be found.

Note: object of locus *can* be a corpus, but does not have to be

frac:attestation



probatio , onis, f. [probo] .
I. A trying , proving; a trial , inspection , examination (class.):
athletarum probatio, Cic. Off. 1, 40, 144 : futura, id. Verr. 2, 1, 54, § 142 ; Varr. R. R. 1, 20, 1: oesypi, Plin. 29, 2, 10, § 36 : croci sinceri, id. 21, 6, 17, § 32 : pumicis, id. 36, 21, 42, § 155 : gemmae recusan limae probationem, id. 37, 13, 76, § 200 : equitum, a review , Val. Max. 2, 2, 9 .--
II. In partic.

Observable: lexical sense (here)
Attestation:

- quotation: *athletarum probatio*
- here: a citation
  - bibliographical metadata can be attached to Attestation
- locus: http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A2007.01.0047%3Abook%3D1%3Asection%3D40

Beyond FrAC

- how to identify an element in a corpus?
  - fragment URIs for different media types, e.g., plain text URIs for strings (etc.)
  - NLP Interchange Format (NIF) String URIs plus contexts
  - Web Annotation (WA) selectors for various formats; can be extended
  - NIF and Web Annotation are partially compatible with each other

    => working on synergies (https://ld4lt.github.io/linguistic-annotation/)
- how to model bibliographical data
  - as part of a citation
  - multiple ontologies are being applied
  - scope of that question is much more general than adequate for an OntoLex module

frac:attestation

DiaMaNT (Diachroon seMAntisch lexicon van de Nederlandse Taal)

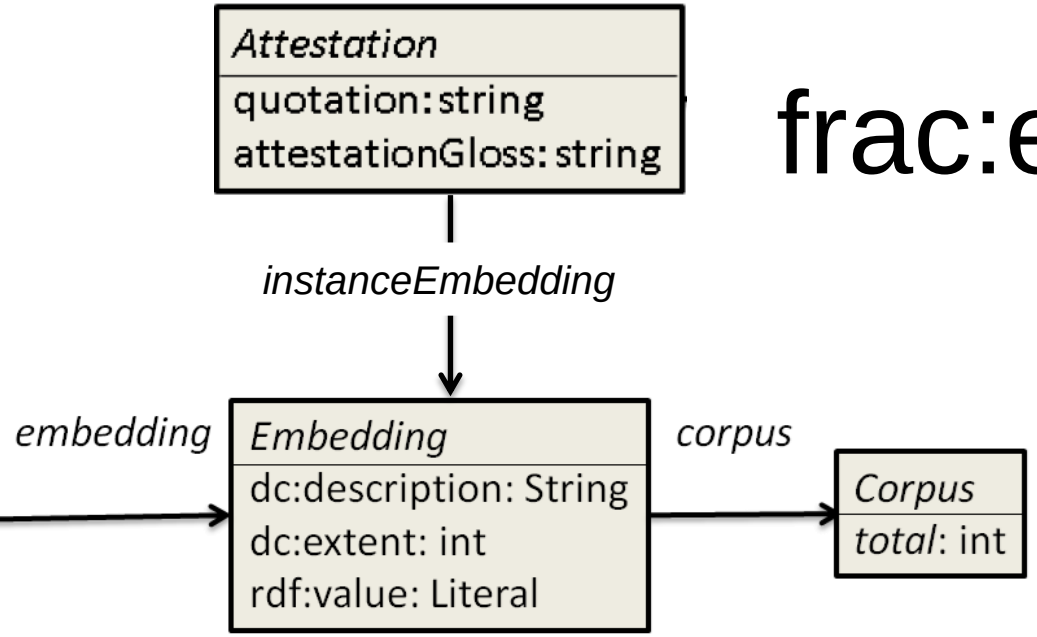
- diachronic semantic computational lexicon of Dutch
- under development at the Instituut voor de Nederlandse Taal (Dutch Language Institute)
- lexicon modelled using OntoLex
  - attestations => FrAC
  - corpus links => NIF
  - citations => CITO/FRBR
    - *Functional Requirements for Bibliographic Records* https://vocab.org/frbr/core
    - *Citation Typing Ontology* http://purl.org/spar/cito

```
diamant:entry_WNT_M030758 a ontolex:LexicalEntry ;
  ontolex:sense diamant:sense_WNT_M030758_bet_207 .

diamant:sense_WNT_M030758_bet_207 a ontolex:LexicalSense;
  rdfs:label "V.-" ;
  frac:attestation diamant:attestation_2108540 ;
  skos:definition "Iemand een kat (of de kat)
                   aan het been jagen .... iemand
                   in moeilijkheden brengen." .

diamant:attestation_2108540 a frac:Attestation ;
  cito:hasCitedEntity diamant:cited_document_WNT_332819 ;
  cito:hasCitingEntity diamant:sense_WNT_M030758_bet_207;
  frac:locus diamant:locus_2108540 ;
  frac:quotation "... dat men licht yemant de cat
                  aen het been kan werpen," .

diamant:locus_2108540 a diamant:Occurrence ;
  nif:beginIndex 107 ;
  nif:endIndex 110 .

diamant:cited_document_WNT_332819
  frbr:Manifestation ;
  frbr:embodimentOf diamant:expression_WNT_332819 ;
  diamant:witnessYearFrom 1621 ;
  diamant:witnessYearTo 1621 .

diamant:expression_WNT_332819 a frbr:Expression ;
  dcterms:creator "N. V. REIGERSB." ;
  dcterms:title "Brieven van Nicolaes
                 van Reigersberch aan Hugo de Groot" ;
  frbr:embodiment diamant:quotation_WNT_332819 .
```

FrAC

NIF

frac:embedding



Nexus Linguarum

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

frac:embedding

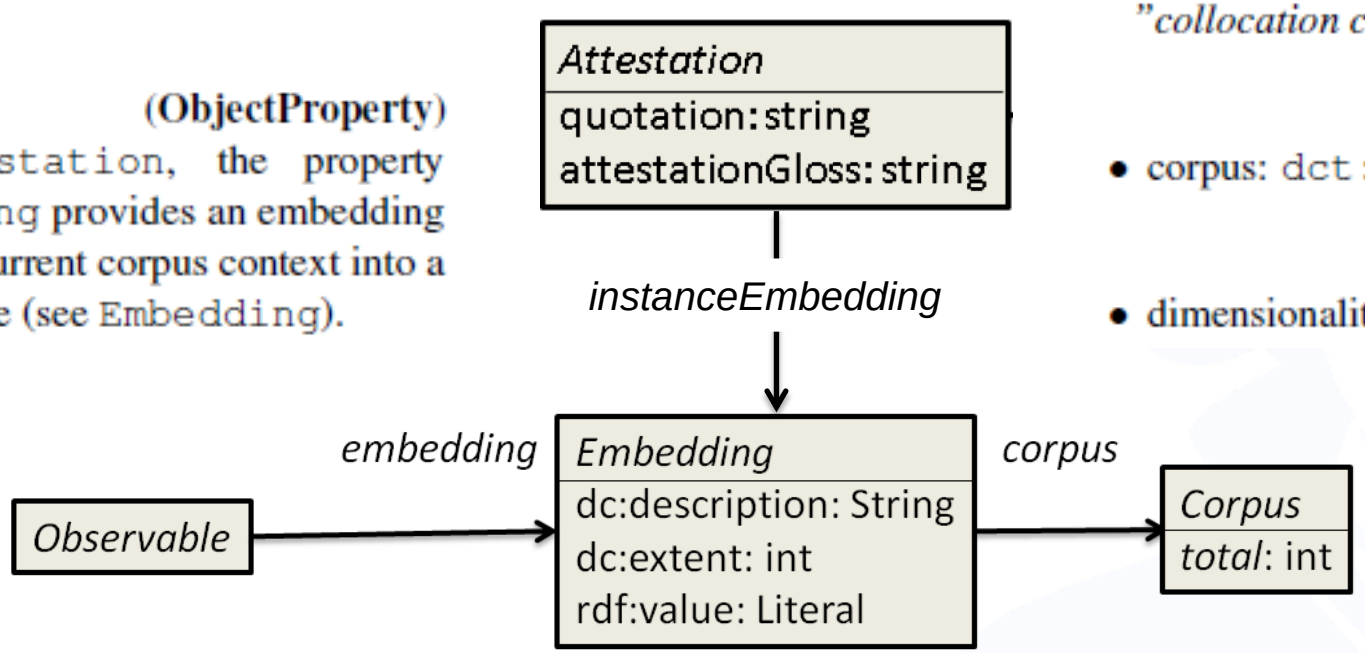
embedding (ObjectProperty) is a relation that maps an `Observable` into a numerical feature space. An embedding is a structure-preserving mapping in the sense that it encodes and preserves contextual features of a particular `Observable` (or, an aggregation over all its attestations) in a particular corpus.

instanceEmbedding (ObjectProperty) For a given `attestation`, the property `instanceEmbedding` provides an embedding of the example in its current corpus context into a numerical feature space (see `Embedding`).

Embedding (Class) is a representation of a given `Observable` in a numerical feature space. It is defined by the methodology used for creating it `dct:description`), the URI of the corpus or language resource from which it was created (`corpus`). The literal value of an Embedding is provided by `rdf:value`).

`Embedding` ⊑ `rdf:value exactly 1` ⊓ `corpus exactly 1` ⊓ `dct:description min 1`

- procedure/method: `dct:description` with free text, e.g., *"CBOW"*, *"SKIP-GRAM"*, *"collocation counts"*
- corpus: `dct:source`
- dimensionality: `dct:extent`



frac:embedding

- originally, we thought about embeddings in the NLP sense
  - Word Embeddings (GloVe, Word2Vec) => *Form*
  - Lemma Embeddings (-“-) => *LexicalEntry*
  - Sense Embeddings (AutoExtend) => *LexicalSense*
  - Concept Embeddings (-“-) => *LexicalConcept*
    - for word embeddings, RDF modelling is just unneccessary
    - for sense and concept embeddings, it can be very helpful to bundle embeddings together with the lexical graph that define them

anecdotal evidence: the original AutoExtend embeddings (Rothe & Schütze 2015) came without metadata, and for 5 years, nobody realized that they were pointing to the wrong WordNet version (... and nobody published any experiments over them)

frac:embedding

- note that NLP embeddings (word/sense/etc. vectors) belong to a larger group of data structures with similar uses
  - uses: similarity metrics (e.g., cosine distance) & clustering
  1. NLP embeddings: fixed size vector
     - fixed-size vectors, mapping positions to numerical scores
     - needed for NLP and computational lexicography

frac:embedding

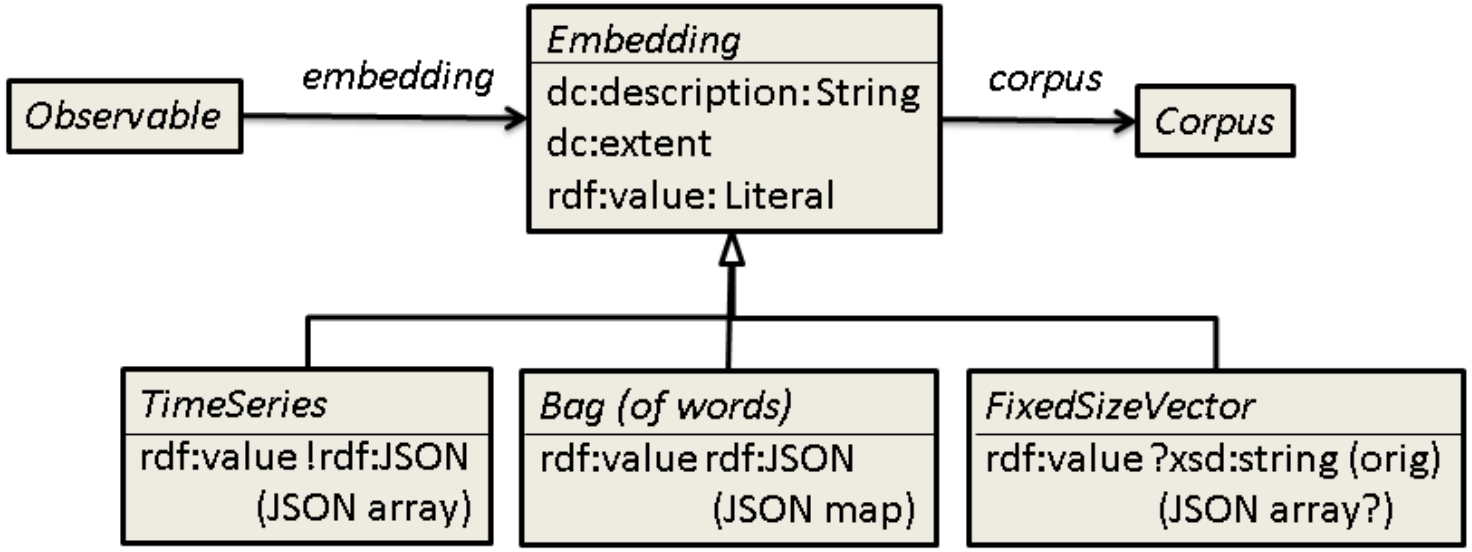
- note that NLP embeddings (word/sense/etc. vectors) belong to a larger group of data structures with similar uses
  - uses: similarity metrics (e.g., cosine distance) & clustering
  1. NLP embeddings: fixed size vector
     - fixed-size vectors, mapping positions to numerical scores
  2. collocation lists (bag of words): weighted multiset
     - infinite-size hashtable, mapping collocates to numerical scores
     - needed for computational lexicography and corpus linguistics

frac:embedding

- note that NLP embeddings (word/sense/etc. vectors) belong to a larger group of data structures with similar uses
  - uses: similarity metrics (e.g., cosine distance) & clustering
  1. NLP embeddings: fixed size vector
     - fixed-size vectors, mapping positions to numerical scores
  2. collocation lists (bag of words): weighted multiset
     - infinite-size hashtable, mapping collocates to numerical scores
  3. time series: sequences of a fixed number of observations
     - infinite-size sequence of fixed size vectors like (1)
     - needed for sign languages; useful for sequence models in NLP

frac:embedding

- three subclasses of frac:Embedding



e.g., recording of movements

e.g., collocation vector, weighted by frequency or otherwise

e.g., word (or sense) embeddings

frac:FixedSizeVector

- GloVe embeddings:
  - original data CSV file, first column is the word, followed by floats

```
frak 0.015246 -0.30472 0.68107 ...
```

FixedSizeVector (Class) is an `Embedding` that represents a particular `Observable` as list of numerical values in a k-dimensional feature space. The property `dc:extent` defines provides the value k.

```
:frak a ontolex:LexicalEntry;
  ontolex:canonicalForm/
    ontolex:writtenRep "frak"@en;
  frac:embedding [
    a frac:FixedSizeVector;
        rdf:value "0.015246 ...";
        dct:source
          <https://catalog.ldc.... >;
        dct:extent 50^^^xsd:int;
        dct:description "GloVe v.1.1,
          ..."@en. ].
```

other embeddings

BagOfWords (Class) is a `frac:Embedding` that represents a particular `Observable` by a set of collocate terms or a mapping from collocates to numerical scores. The value of `dc:extent` can be used to specify either the maximum size of bags of words, or, the actual size of a particular bag of words. The `rdf:value` should be a JSON literal, e.g., a dictionary.

TimeSeries (Class) is a `frac:Embedding` that represents a particular `Observable` or its `Attestation` as a sequence of a fixed number of data points recorded over a certain period of time. The value of `dc:extent` must be used to specify the number of data points per observation. The `rdf:value` should be a structured JSON literal.

frac:instanceEmbedding

- for modelling contextual embeddings as relevant for more recent architectures on neural NLP
- domain is not a *frac:Observable*, but a *frac:Attestation*
  - may be left underspecified
  - embeddings represents the embedding of the target expression *in this particular context*
  - encoding of embedding itself otherwise identical to *frac:embedding*

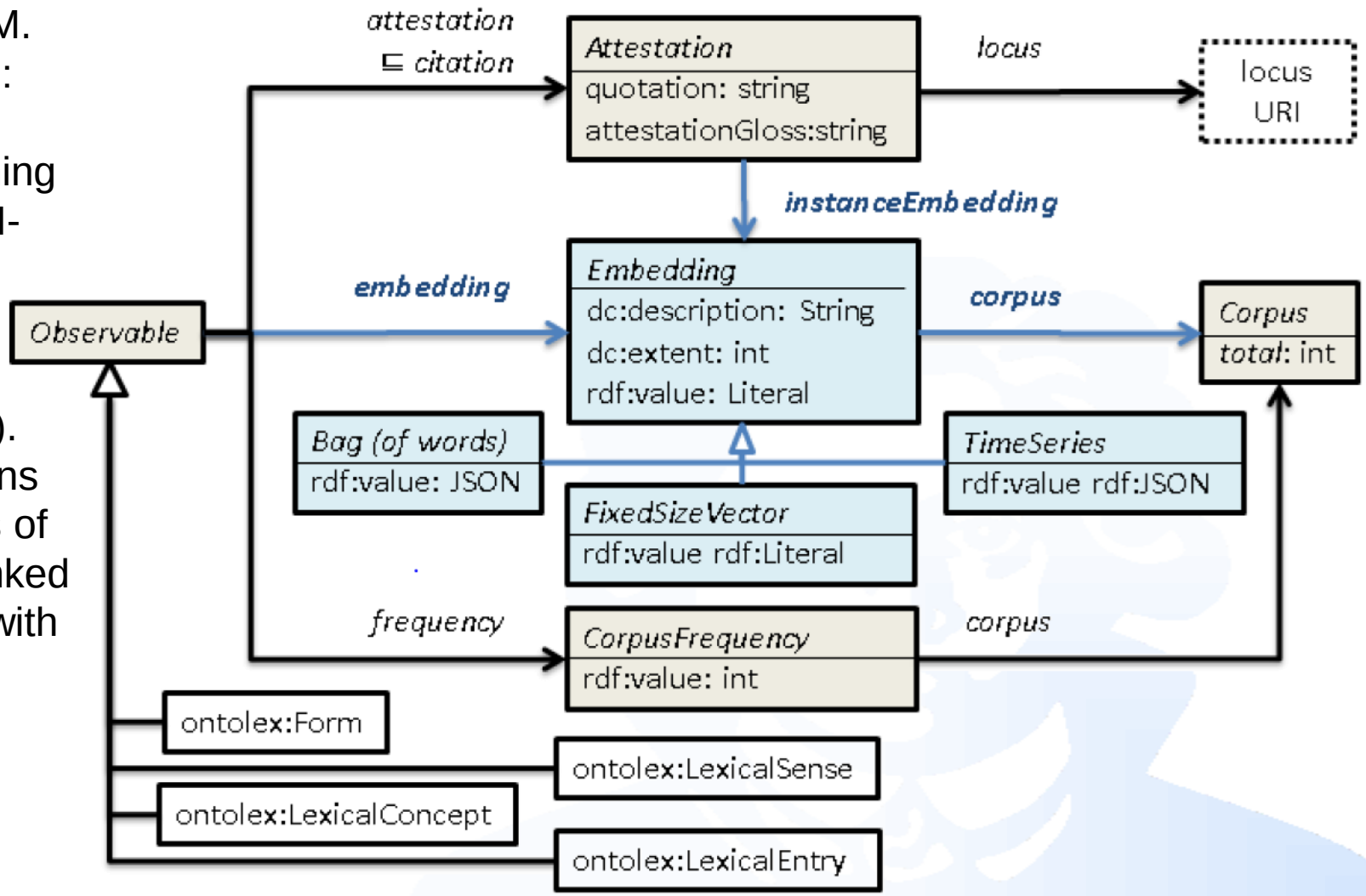
```
wn31:play_n a ontolex:LexicalEntry;
  ontolex:sense wn31:07032045-n,
    wn31:play_n_4, ...
wn31:07032045-n
    a ontolex:LexicalSense;
    frac:attestation [
        frac:quotation "the play
        lasted two hours";
        frac:locus wn31:07032045-n;
        frac:instanceEmbedding
            wn31-bert:07032045-n-1
        ].
wn31-bert:07032045-n a
    frac:FixedSizeVector;
    dc:extent "300"^^xsd:int;
    rdf:value "0.327246 0.48170 ...";
    dc:description "...";
    frac:corpus <http://wordnet-rdf.
       princeton.edu/static/wordnet.
       nt.gz> .
```

OntoLex-FrAC as of January 2021

Chiarcos, C., Declerck, T. and Ionov, M. (2021), Embeddings for the Lexicon: Modelling and Representation. 6th Workshop on Semantic Deep Learning (SemDeep-6), co-located with IJCAI-PRICAI 2020. Japan, January 2021

Chiarcos, C., Ionov, M., de Does, J., Depuydt, K., Khan, F., Stolk, S., Declerck, T. and McCrae, JP (2020). Modelling Frequency and Attestations for OntoLex-Lemon. In Proceedings of the 2020 Globalex Workshop on Linked Lexicography (pp. 1-9), co-located with LREC 2020, France, May 2020

https://github.com/ontolex/frequency-attestation-corpus-information



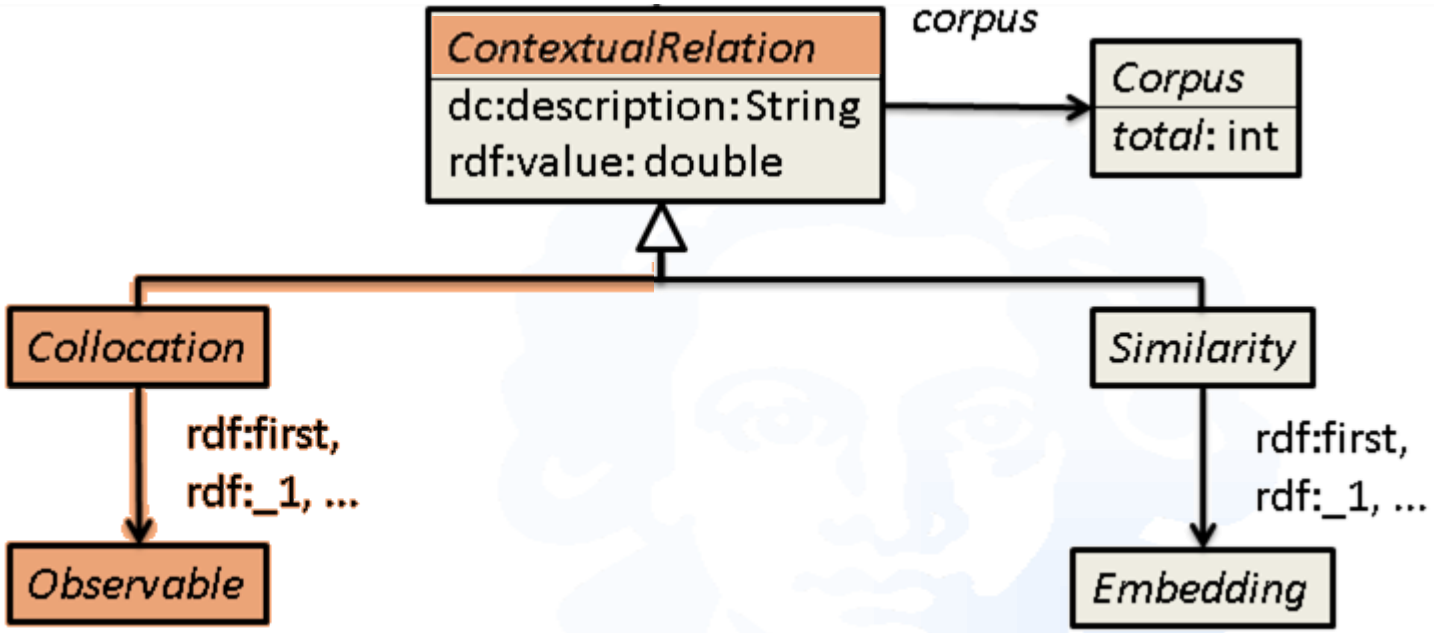
On-going discussions

Nexus Linguarum

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Collocation and Similarity

- We have begun to discuss *frac:Similarity*
  - as an RDF collection of embeddings
  - can represent similarity clusters (Brown clusters) and similarity relations (pair-wise similarity)
- We have not discussed collocations
  - nor their relation with *frac:Similarity*



Internal consolidation

- We anticipate that the discussion of similarity and collocations can have a profound impact on other components of the vocabulary
  - BagOfWords also represents collocates
- After these have been addressed, a thorough review of all vocabulary components is required
  - reduce properties, improve readability, avoid mis-interpretations
- We expect to deliver a result by late 2021
  - depending on use cases, so, please join our calls ;)
  - bi-weekly, see OntoLex mailing list and Nexus Linguarum calendar