Towards an Ontolex-Lemon module for Frequency, Attestations and Corpus Information (FrAC)

Christian Chiarcos & Maxim Ionov

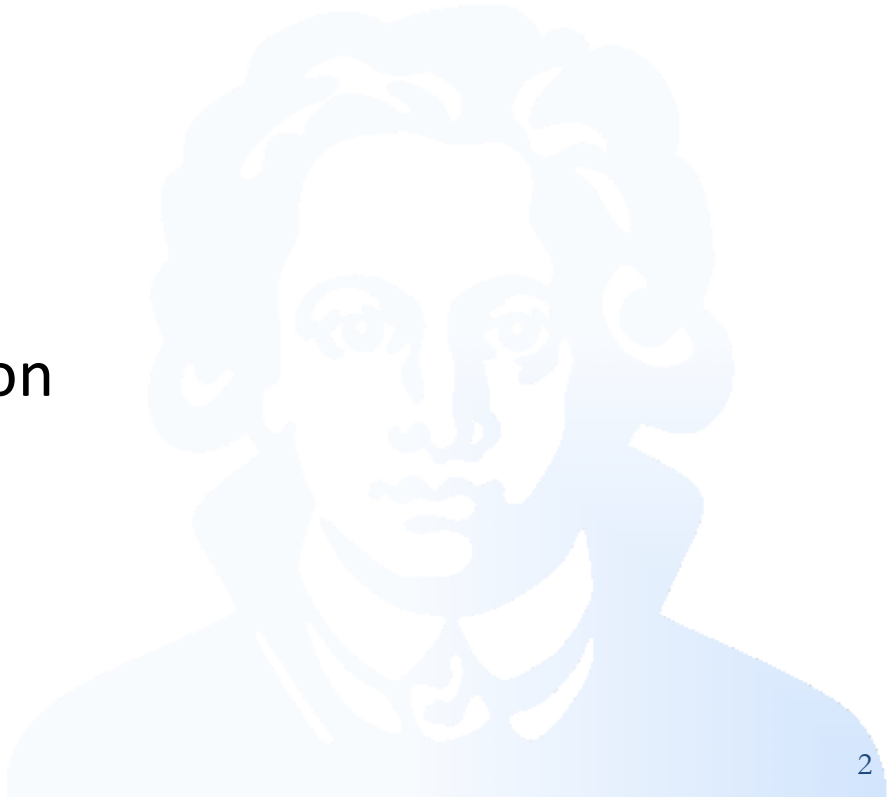
Applied Computational Linguistics / LiODi

Goethe-Universität Frankfurt, Germany

:acoli dc:isPartOf <http://uni-frankfurt.de>.

Towards an Ontolex-Lemon module for Frequency, Attestations and Corpus Information

- Motivation
- FrAC
  - Frequency
  - Attestations
  - Corpus(-derived) Information
    - Collocations
    - (distributional) Similarity
    - Embeddings



Towards an Ontolex-Lemon module for Frequency, Attestations and Corpus Information

- Motivation
- FrAC
  - Frequency
  - Attestations
  - Corpus(-derived) Information
    - Collocations
    - (distributional) Similarity
    - Embeddings

here and today

concept draft developed at
the ACoLi Lab, GU Frankfurt
not a proposal, but a basis for discussion
https://acoli-repo.github.io/ontolex-frac/ (page)
https://github.com/acoli-repo/ontolex-frac (dev)

todo: decide and specify scope,
next steps and relevant use cases

Next steps

- biweekly telcos, starting mid-July 2019
  - who would commit to participate?
    - we need 5 participants from different institutions
  - after/alternating with morphology telcos? different slot?
- until then
  - creating and populating a wiki with requirements, use cases and data sets
    - cf. morphology module

Corpus Frequency

Side note: „Corpus“ here is used in the original, broader sense as „structured data collection“. This can, but does not have to be a corpus in the language resource sense. It could also be, e.g., a(nother) dictionary

FrAC – Motivation: Frequency

frequency dictionaries? Freitag, 29. Juni 2018 16:25:48

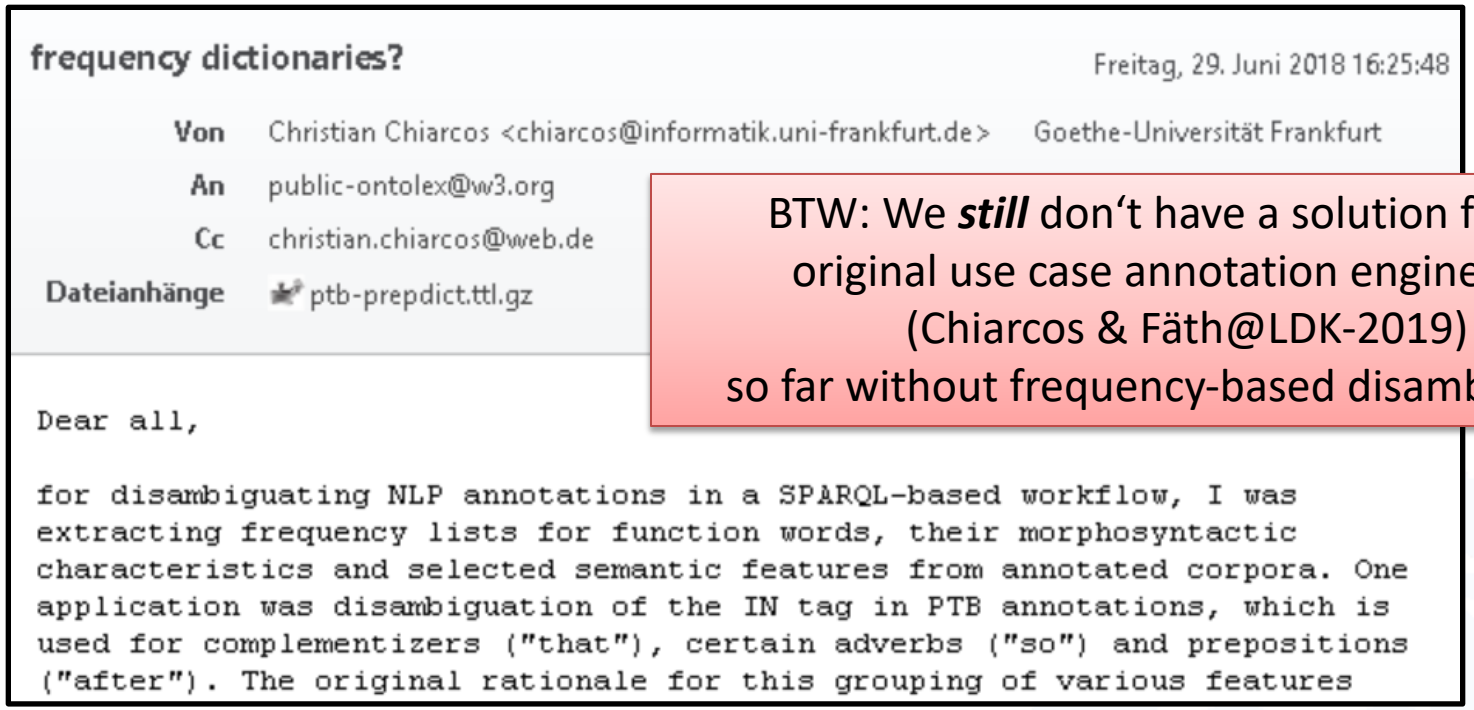
Von Christian Chiarcos <chiarcos@informatik.uni-frankfurt.de> Goethe-Universität Frankfurt
An public-ontolex@w3.org
Cc christian.chiarcos@web.de
Dateianhänge ptb-prepdict.ttl.gz

Dear all,

for disambiguating NLP annotations in a SPARQL-based workflow, I was extracting frequency lists for function words, their morphosyntactic characteristics and selected semantic features from annotated corpora. One application was disambiguation of the IN tag in PTB annotations, which is used for complementizers ("that"), certain adverbs ("so") and prepositions ("after"). The original rationale for this grouping of various features

- There's no easy way
  - not in(to) lexicography module
    - frequency *is* lexicographically relevant, **but** not only there
  - can only be defined relative to a corpus, i.e., reification
    - not a plain lexinfo property

FrAC – Motivation: Frequency



frequency dictionaries? Freitag, 29. Juni 2018 16:25:48

Von Christian Chiarcos <chiarcos@informatik.uni-frankfurt.de> Goethe-Universität Frankfurt

An public-ontolex@w3.org

Cc christian.chiarcos@web.de

Dateianhänge ptb-prepdict.ttl.gz

Dear all,

for disambiguating NLP annotations in a SPARQL-based workflow, I was extracting frequency lists for function words, their morphosyntactic characteristics and selected semantic features from annotated corpora. One application was disambiguation of the IN tag in PTB annotations, which is used for complementizers ("that"), certain adverbs ("so") and prepositions ("after"). The original rationale for this grouping of various features

BTW: We ***still*** don't have a solution for that original use case annotation engineering (Chiarcos & Fäth@LDK-2019) so far without frequency-based disambiguation

- There's no easy way
  - not in(to) lexicography module
    - frequency *is* lexicographically relevant, **but** not only there
  - can only be defined relative to a corpus, i.e., reification
    - not a plain lexinfo property

FrAC – Beyond Frequency

- frequency can only be defined relative to a corpus

⇒ if we need to provide novel vocabulary anyway, other, corpus-related information can be included here, too

  - Leiden f2f meeting
    - attestations (originally suggested for *lexicog* module)
  - corpus-derived information for digital lexicography
    - collocations (and NLP)
    - distributional similarity (clusters) [as produced by common tools]
  - SOTA NLP resources (beyond lexicography)
    - embeddings? (esp., sense embeddings, cf. Rothe & Schütze 2017)

Corpus Frequency in EPSD2

http://oracc.museum.upenn.edu/epsd2/sux

frequency

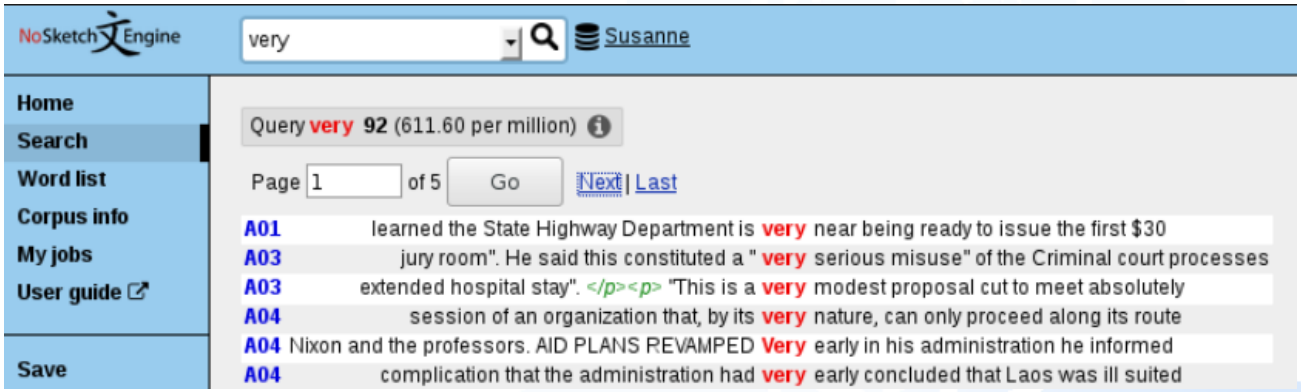
lugal [KING] N (39743x) Early Dynastic IIIa, Early Dynastic IIIb, Old Akkadian, Lagash II, Ur III, Old Babylonian, Middle Assyrian, Middle Babylonian, Neo-Assyrian, Neo-Babylonian, Hellenistic, Uncertain, unknown wr. lugal; lugal-la-a; lugal-lugal; lugal-x; lugal^{al}; lugal^{mušen}; lu_2-gal "king; lord; master; owner"

corpus-based lexicography
taking corpus frequencies into account
(at least, displaying them)

Corpus Frequency in NoSketchEngine

https://www.sketchengine.eu/nosketch-engine/

basically for everything
that you could find in
a corpus
⇒ informs lexicographers
about commonness, etc.



Corpus Frequency

- frequency relative to a corpus, i.e., reified

  ≠ lexinfo:frequency

  „The relative commonness with which a term occurs.”

  domain: lexinfo:Frequency = { lexinfo:commonlyUsed, lexinfo:infrequentlyUsed, lexinfo:rarelyUsed }

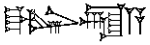
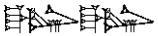
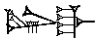
  no actual frequency counts

Corpus Frequency in EPSD2

http://oracc.museum.upenn.edu/epsd2/sux

frequency of lexical entry

lugal [KING] N (39743x) Early Dynastic IIIa, Early Dynastic IIIb, Old Akkadian, Lagash II, Ur III, Old Babylonian, Middle Assyrian, Middle Babylonian, Neo-Assyrian, Neo-Babylonian, Hellenistic, Uncertain, unknown wr. lugal; lugal-la-a; lugal-lugal; lugal-x; lugal^{al}; lugal^{mušen}; lu$_2$-gal "king; lord; master; owner"

[1]		lugal
[2]		lugal-la-a
[3]		lugal-lugal
[4]		lugal-x
[5]		lugal ^{al}
[6]		lugal ^{mušen}
[7]		lu$_2$-gal
+	--la=l+a (4x/0%); --la=l.a (122x/0%); --la$_2$=l.a (14x/0%); --le=l.e (4x/0%);	

frequency of lexical form (in different periods/ subcorpora)

	PC	ED IIIa	ED IIIb	Ebla	OAkk	Lag II	Ur III	OB	Post-OB	(unknown)
[1]		23	517		291	53	36588	1987	249	11
[2]										
[3]								1		

Senses:

1. king (39729x/100%)
2. lord
3. master (3x/0%)
4. owner (11x/0%)

Akk. *bēlu, šarru.*

frequency of lexical senses

Corpus Frequency beyond lexical entries

WordnetTools

Wordnet-tools is a collection of functions that operate on wordnets (Fellbaum 1998, Vossen et al. 1998) represented in Wordnet-LMF format (Vossen et al. 2013).

The following functions are provided:

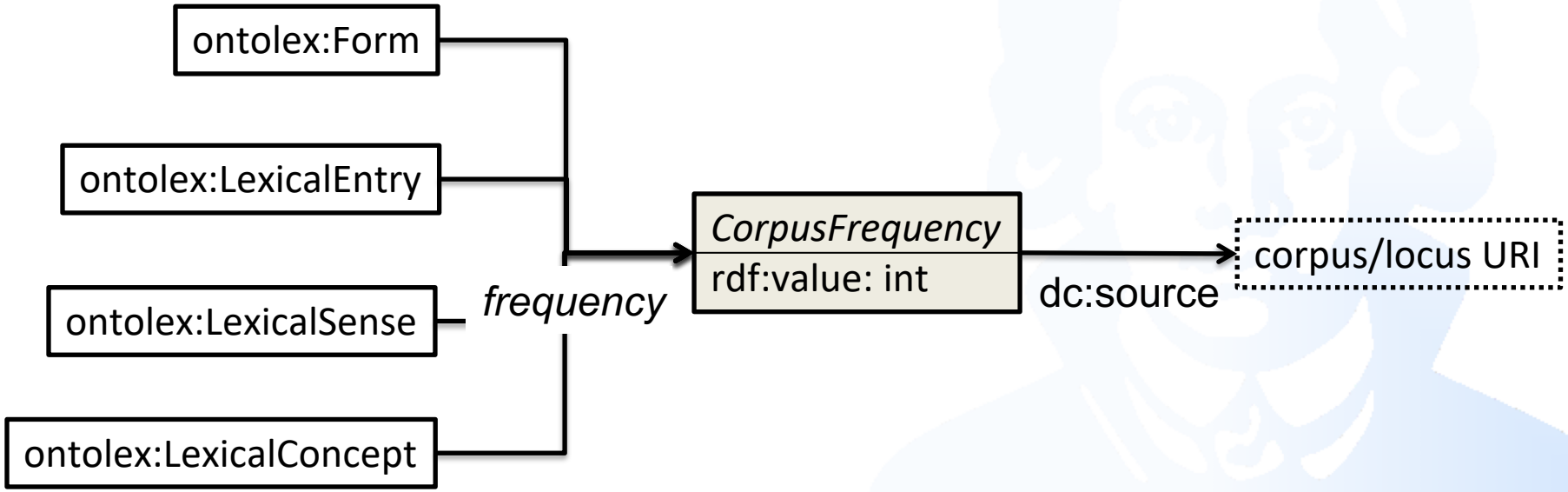
1. **Apply one of the following similarity measures to a pair of words or synsets**: path, Jiang Conrath, Leacock & Chodorow, Resnik, Lin, Wu & Palmer
2. **Create a cumulated frequencies for lowest common subsumbers (wordnet hypernyms) from a corpus frequency file**: takes a list of words with corpus frequencies and accumulates these frequencies to the full wordnet hierarchy, passing it to the hypernyms of each meaning of the word. The output is a file with hypernym synset ids and their cumulated frequencies. This file is used by various semantic similarity measures (Resnik 1995).

http://www.cltl.nl/results/software/wordnettools/

=> corpus frequency for lexical concepts !

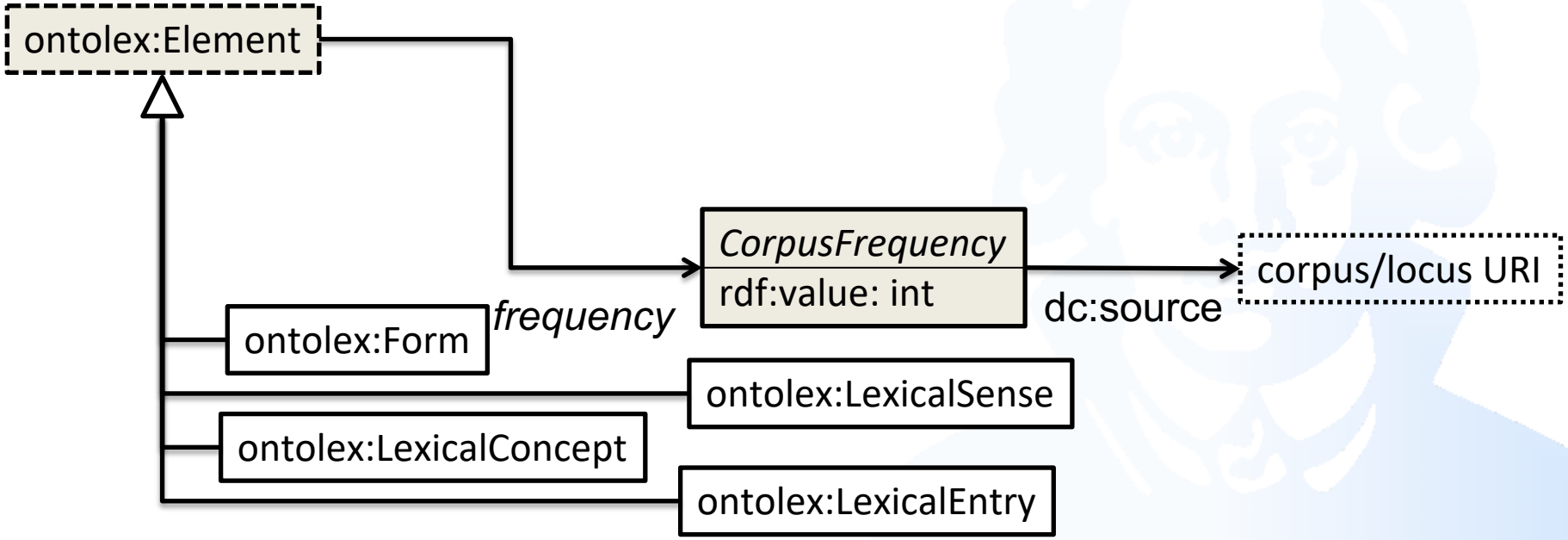
Corpus Frequency

- everything in ontolex can be counted against a corpus
  - minimum information: absolute frequency & provenance, i.e. reference to source corpus (any URI)



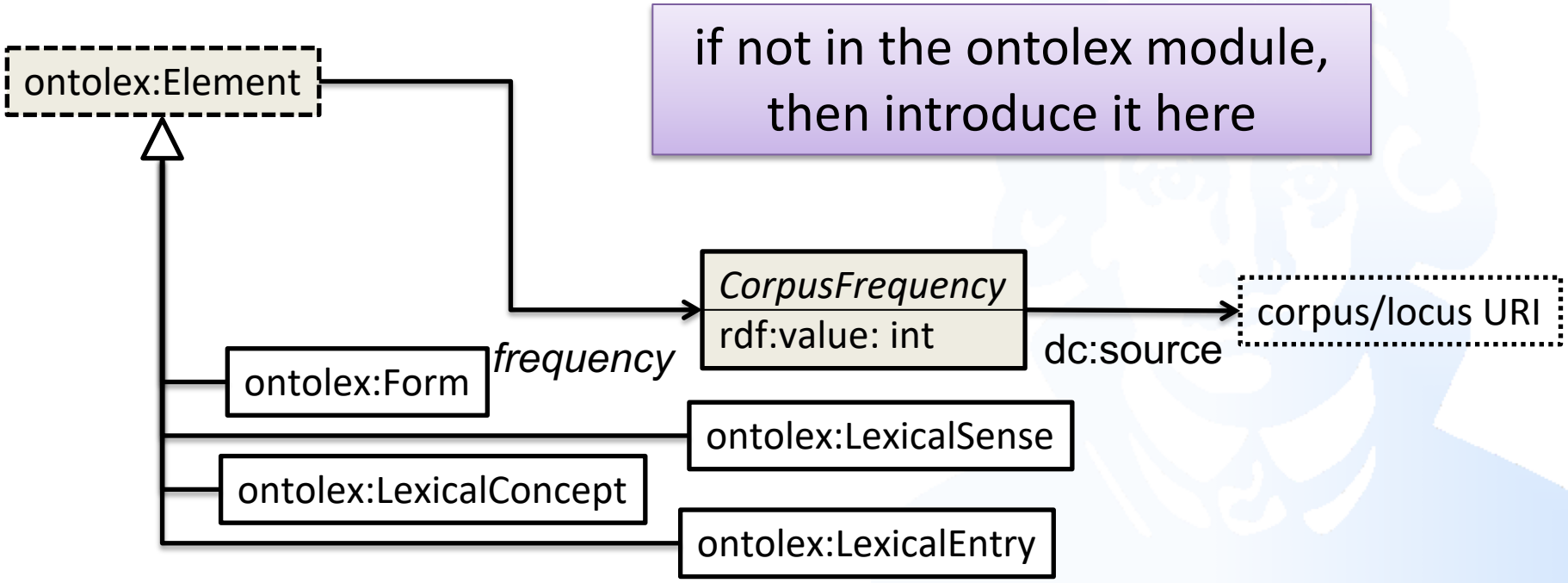
Corpus Frequency

- everything in ontolex can be counted against a corpus
  - similarly for every subsequent piece of corpus information => generalization

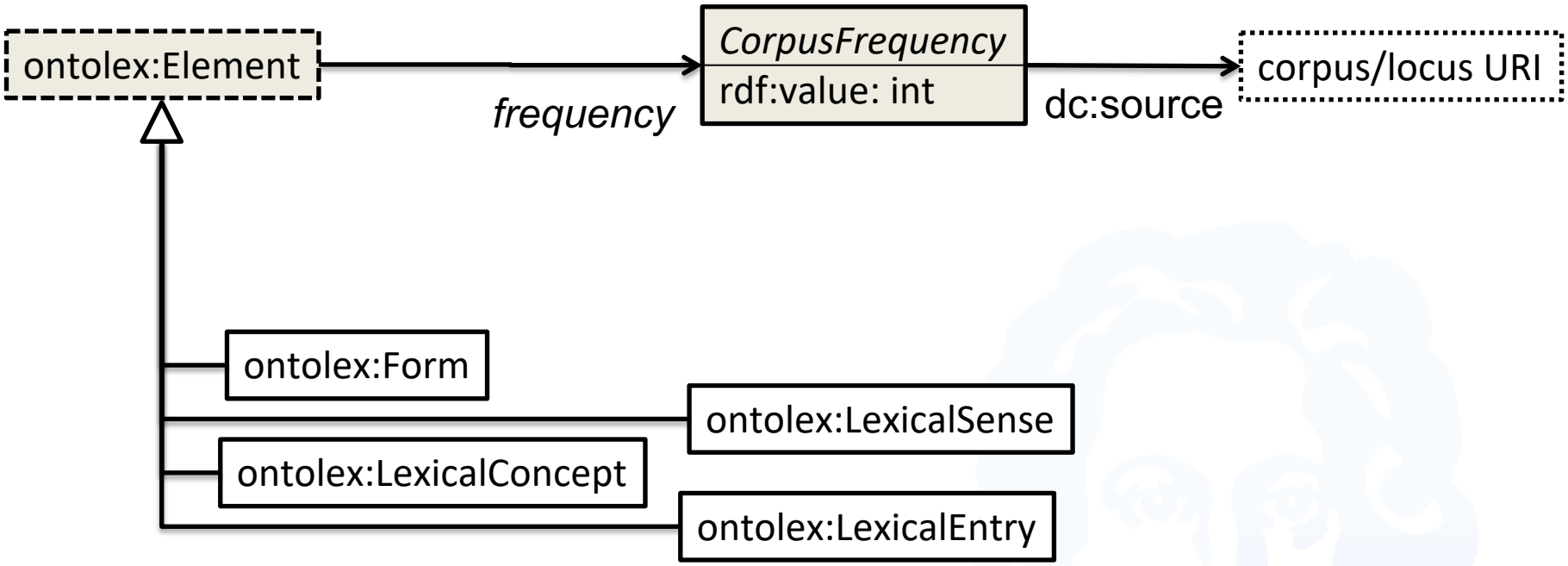


Corpus Frequency

- everything in ontolex can be counted against a corpus
  - similarly for every subsequent piece of corpus information => generalization



Frequency: Discussion

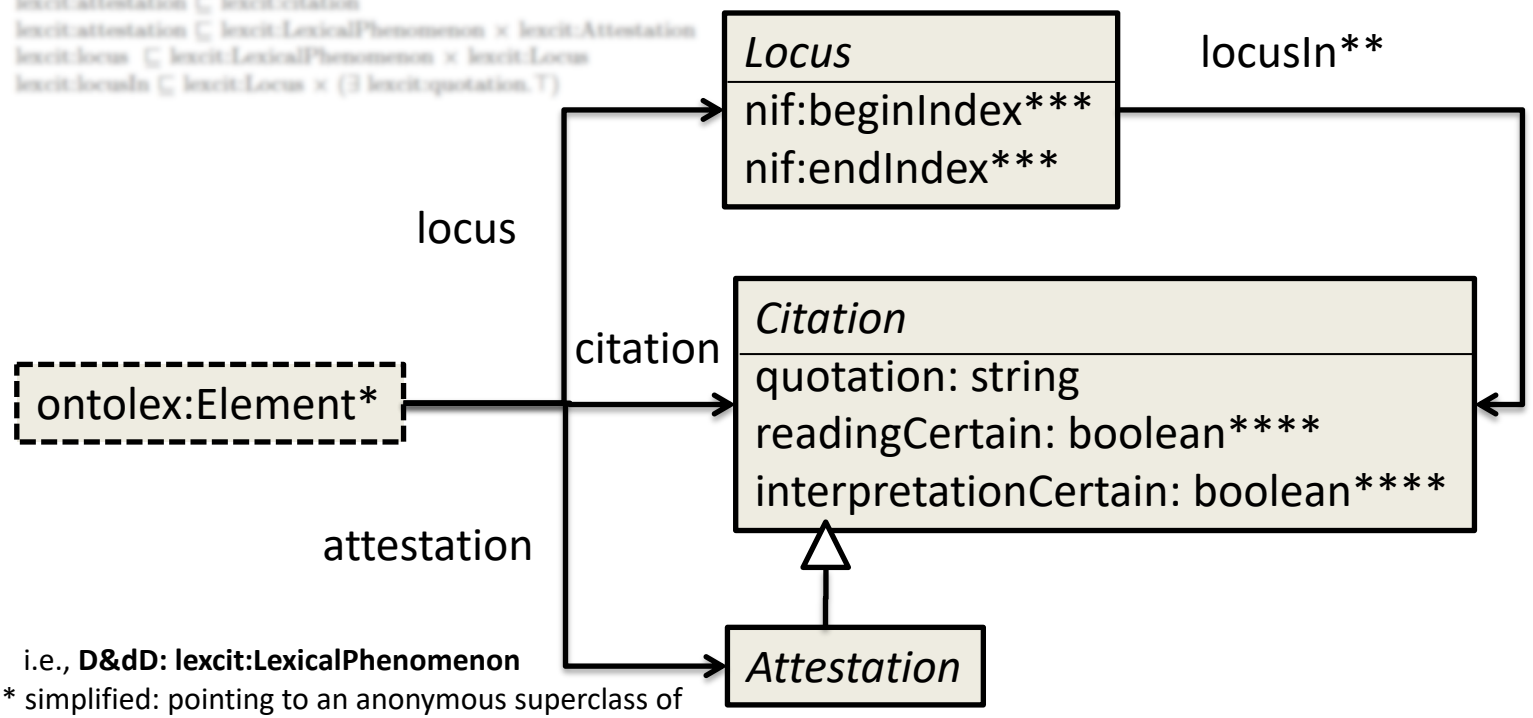


Attestations

Attestations

- long and vivid discussion in the *lexicog* model
  - Leiden f2f meeting: merge with module on corpus-related issues
- based on two existing proposals
  - Khan and Boschetti (2018) and Depuydt and de Does (2018)
- approach:
  - focus on linking with an edition/corpus
  - move “simple” lexicographic properties to lexinfo
  - support citations, but stay agnostic wrt. vocabulary for bibliography

attestations after Depuydt & de Does 2018



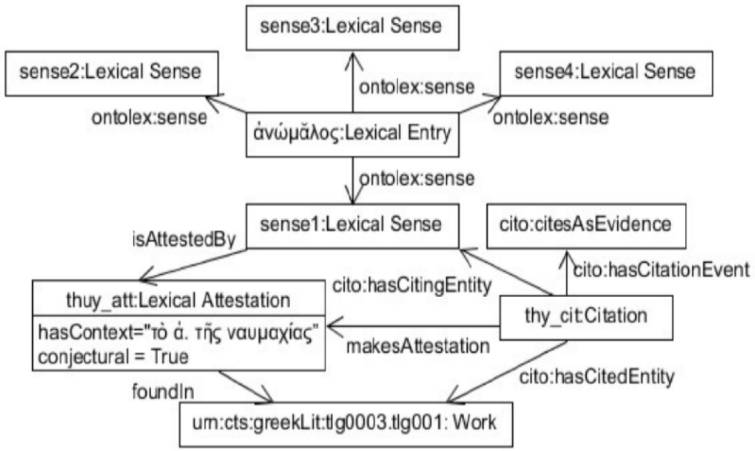
* i.e., **D&dD: lexcit:LexicalPhenomenon**

** simplified: pointing to an anonymous superclass of something that contains a quotation

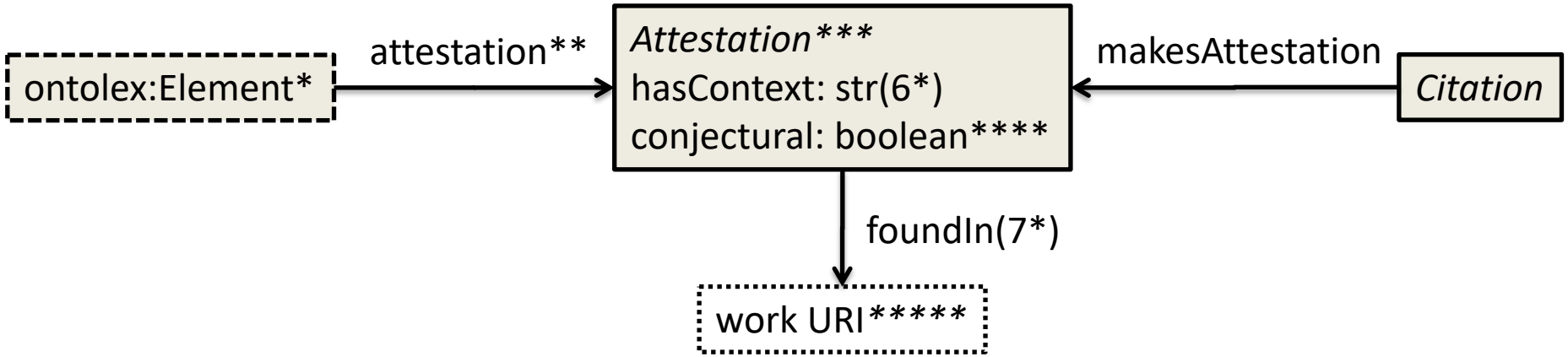
*** I would prefer not to prescribe nif properties in order to permit other means of cross-referencing, e.g., nif URIs, WebAnnotation selectors, etc.

**** I would prefer to leave certainly etc. to lexinfo

Khan and Boschetti



attestations after Khan & Boschetti (2018)



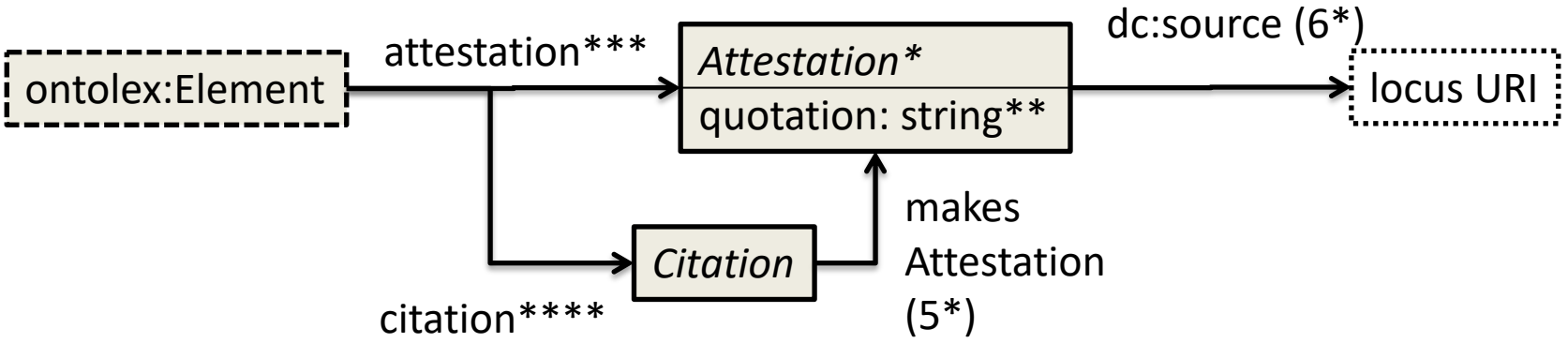
* originally restricted to LexicalSense
** originally „isAttestedBy“
*** originally „LexicalAttestation“
**** I would prefer to leave certainty etc. to lexinfo
***** originally „Work“
(6*) maybe rename such that that also context-free examples fit

(7*) I had dc:source at other occasions

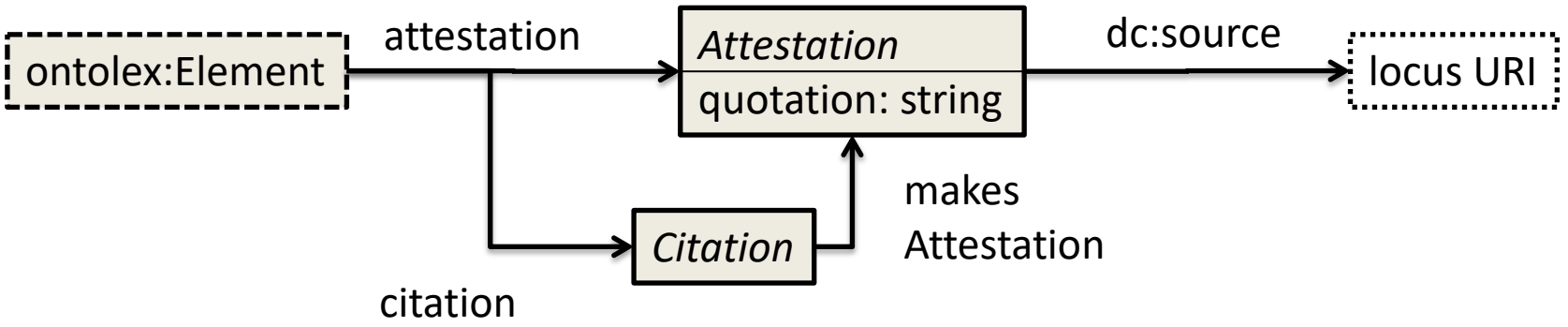
as in the other proposal, cito properties and concepts are skipped, these are beyond the module

attestations: proposal for a minimal consensus (Nov 2018)

- we do not cover:
  - scientific citations (should refine Citation, using external vocabularies)
  - corpus pointers (can be NIF objects, NIF URIs, WebAnnotation selectors, CTS URNs, URLs, etc.)



* D&dD: Citation
** K&B: hasContext
*** D&dD: attestation
**** K&B: inv of hasCitingEntity
5* D&dD: implicit
6* D&dD: inv of locusIn, K&B: foundIn



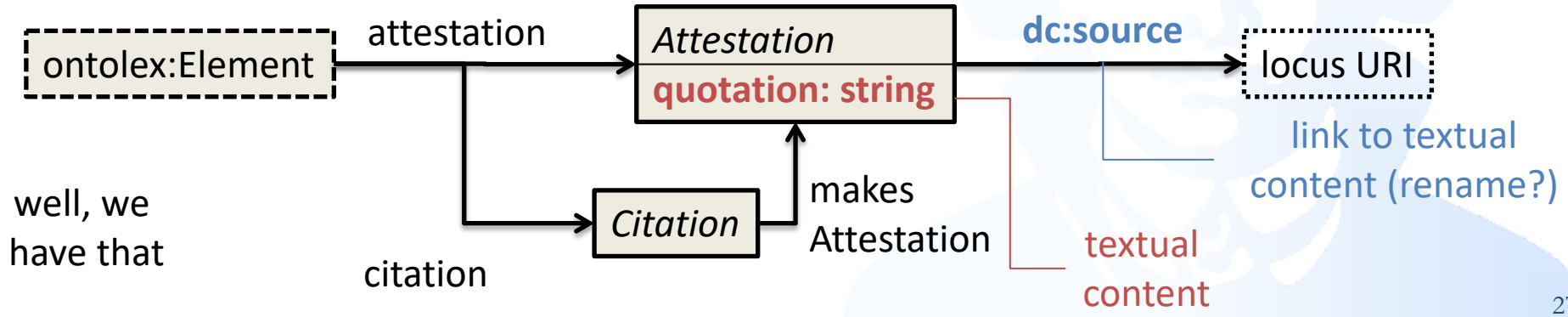
proposed minimal consensus	D&dD	K&B
ontolex:Element	LexicalPhenomenon	(LexicalSense)
attestation	attestation	isAttestedBy
Attestation	Attestation	Attestation
Citation	~ Citation	Citation
citation	~ citation	(indirectly via makesAttestation)
makesAttestation	[identity/subClassOf]	makesAttestation
dc:source	^locusIn	foundIn
quotation	quotation	hasContext
(locus URI)	Locus	Work
(indirectly via attestations)	locus	(hasCitingEntity/hasCitedEntity)

Attestations: initial discussion

- Fahad (May 15, 2019)
  - I see an attestation as evidence of the use of some linguistic element or convention and therefore attestations relate together entries, senses, etc, with any source that might manifest evidence of that use, e.g., a book, an inscription, a corpus; it is not simply a quotation as in Christian's model, although there should be *datatype* properties which link an attestation with its textual context;
- currently covered by *dc:source* and *quotation*
  - the source is the *dc:source* element, the textual content is *quotation*
    - do we need a frac-specific source identifier?

Attestations: initial discussion

- Fahad (May 15, 2019)
  - I see an attestation as evidence of the use of some linguistic element or convention and therefore attestations relate together entries, senses, etc, with any source that might manifest evidence of that use, e.g., a book, an inscription, a corpus; it is not simply a quotation as in Christian's model, although there should be *datatype* properties which link an attestation with its textual context;



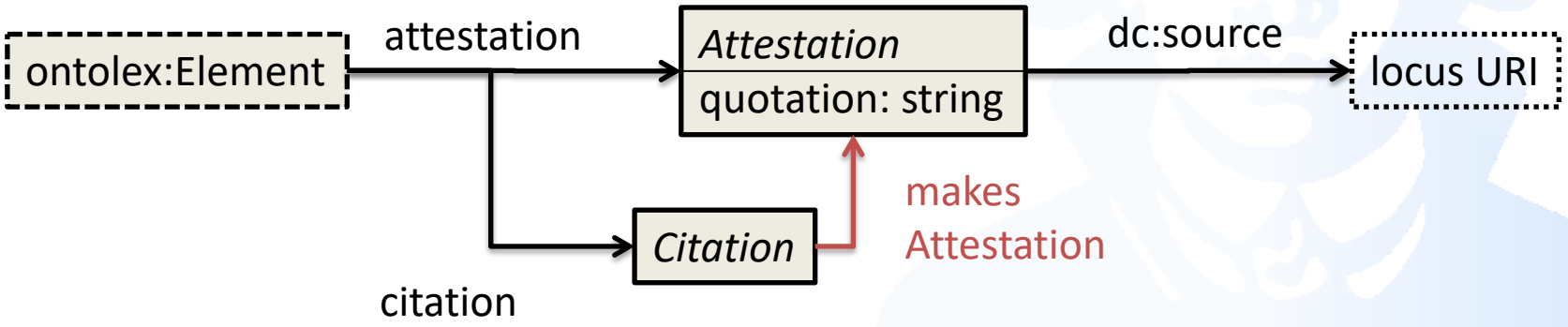
Attestations: initial discussion

- Fahad (May 15, 2019)
  - uncertainty levels to attestations and in certain cases might actually be referring to a conjectural, reconstructed text; in other cases a legacy resource might actually be incorrect in regarding a certain source as attest and it would be useful to be able to annotate; we should make provision for this;
- not specific to attestations => lexinfo (& provo)
  1. **eluku** ‘offered (?)‘ (uncertain sense)
  2. **lemais** ‘Lemai‘, THEO (?) (uncertain POS)
  3. **ti[– – –]** ‘?‘ (defective form)
  4. got. *raíhts*, aisl. *réttr*, ags. *riht*, as. ahd. *reht* ... (= kelt. **rektu-* ...) (reconstructed lexeme)

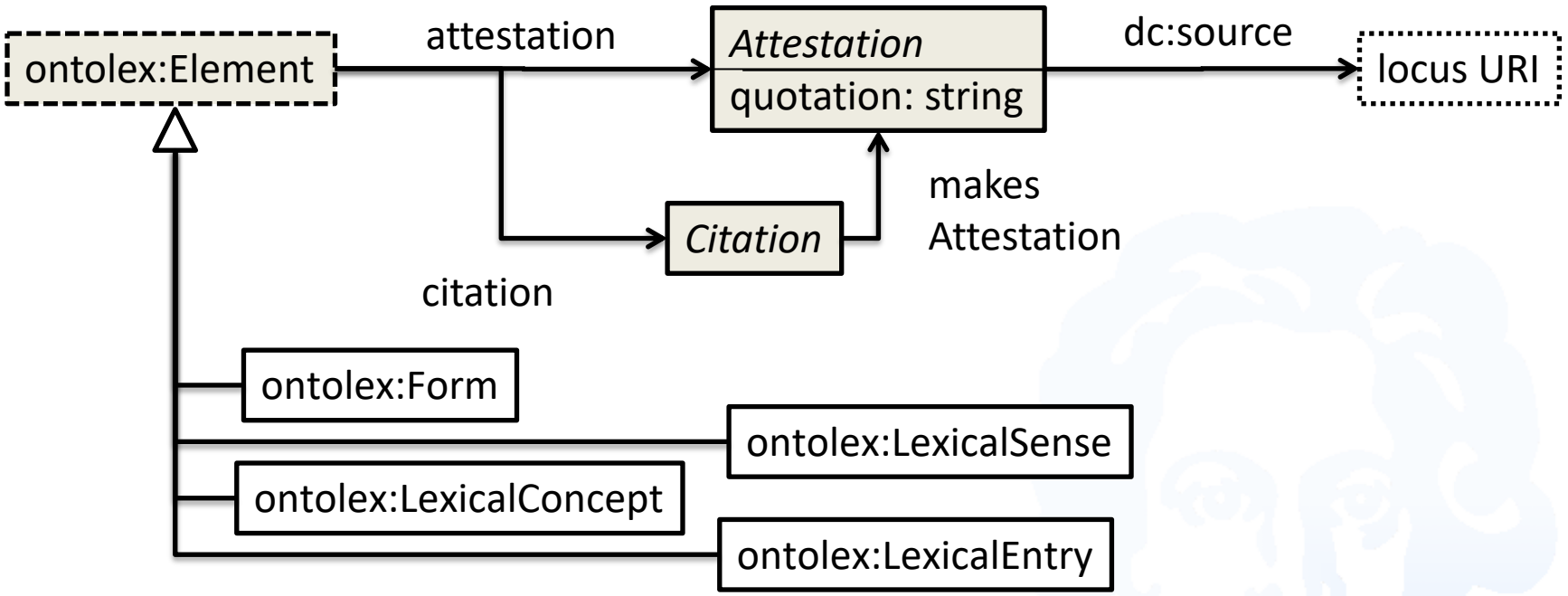
(examples 1-3 from Wallace 2008. *Zikh Rasna. A Manual of the Etruscan Language and Inscriptions*, Beech Stave Press, example 4 from Pokorny, J. (1969). Indogermanisches etymologisches Wörterbuch (Vol. 2). Francke, entry 1589)

Attestations: initial discussion

- Fahad (May 15, 2019)
  - citations are (speech) acts which can't be correct or incorrect, they're either successful or unsuccessful, they should be related to attestations, because a lexical entry will often *cite* a book in order to specify an attestation (but this is far from the only way citations can be used in a lexical entry)



Attestation: Discussion



Embeddings

and related distributional resources

Embeddings

- in a broader sense: incl. bags of words, etc.
  - if representable as a vector
- why?
  - most important data structure in distributional semantics, often laborsome to replicate
    - ⇒ reuse encouraged
  - applicable to form, lexeme, sense,* concept*
    - but detached from their definition
  - often distributed as CSV files => vocabulary for CSV2RDF rendering
    - nobody expects embeddings to be distributed in Turtle ;)

* cf. Rothe & Schütze (2017), AutoExtend, for embedding induction over WordNet senses and synsets

Embeddings: Co-occurrence

- simple, uncompressed form of embedding
  - based on a reference list of vocabulary items, where every reference word is associated with a fixed position, e.g., *ship* with position 1, *ocean* with 2, *sky* with 3, etc.
- Sample corpus for *frak* (from Wikiquote)

It's in the frakking ship!
Have you lost your frakkin' mind?
Oh, for frak's sake, let me see if I can make heads or tails of it.
It's a frakking Cylon.
Our job isn't to be careful, it's to shoot Cylons out of the frakking sky!

list of reference words: `(ship, ocean, lose, find, brain, mind, head, sky, Cylon, ...)`

vector `(1,0,1,0,0,1,1,1,2,...)`

Word Embeddings

- Word vectors are always defined relative to a base corpus*
  - for co-occurrence statistics
    - source needs to be explicated
- Word vectors can be compressed
  - using *different* statistical or neural methods
    - needs to be explicated
  - dimensions are usually no longer interpretable

* “corpus“ in a broader sense, could be knowledge base, etc.

Sense Embeddings

- Normally extrapolated from word embeddings plus a lexical resource
  - e.g., WordNet word senses (= ontolex:Sense) and synsets (= ontolex:LexicalConcept)

(Rothe & Schütze 2017)

⇒ Sense definition/underlying KB version should be explicated, e.g., by an URI

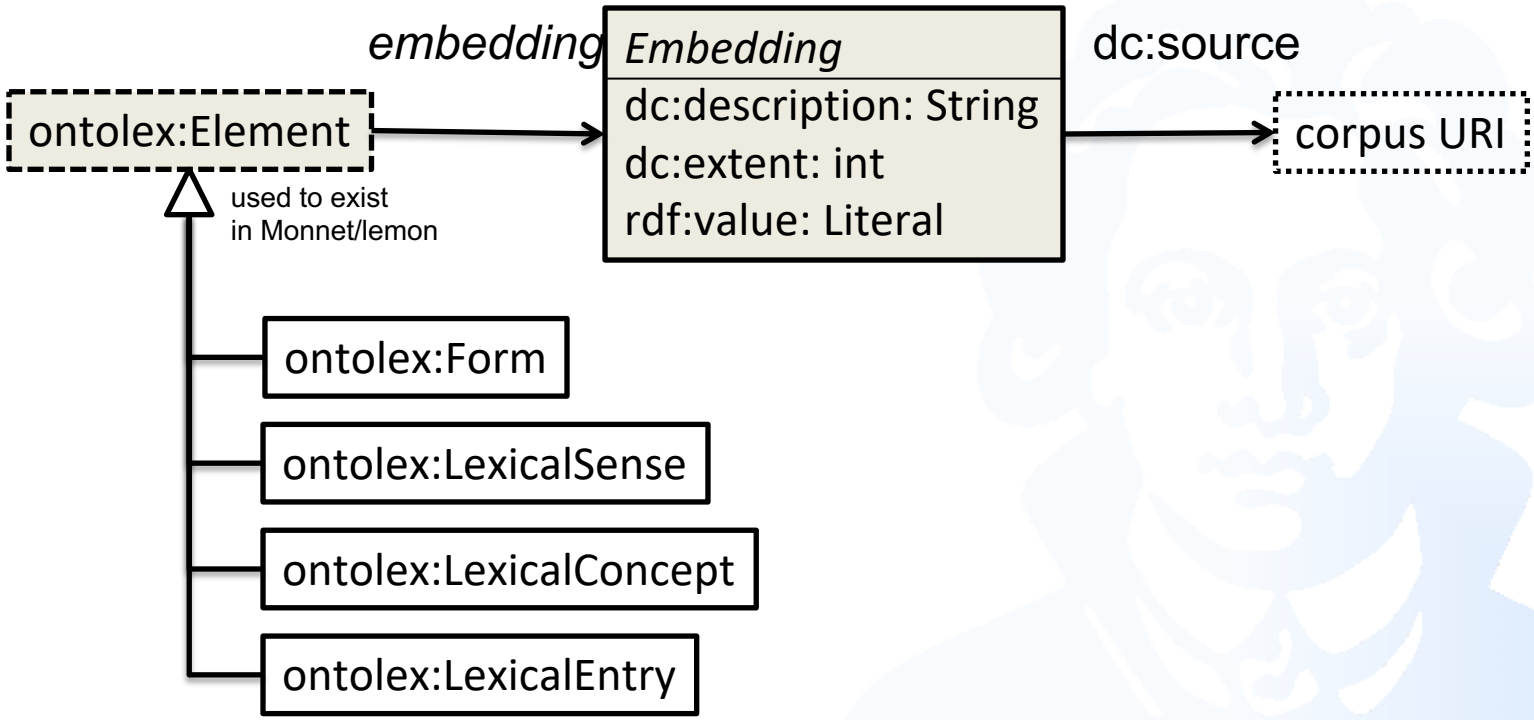
Embeddings in OntoLex

- typical CSV format
  - first column contains string/lemma/sense id
  - followed by ,-separated float numbers
- suggestion: shallow model / „BLOB“
  - first column => a *ontolex:Element* (LexEnt/Form/Sense/Concept)
  - following columns => *rdf:value* (single string) + *dct:extent* (dimensionality)
  - *dct:description*: method, e.g., „CBOW“, „co-occurrence counts“
  - *dc:source*: underlying corpus (URI)

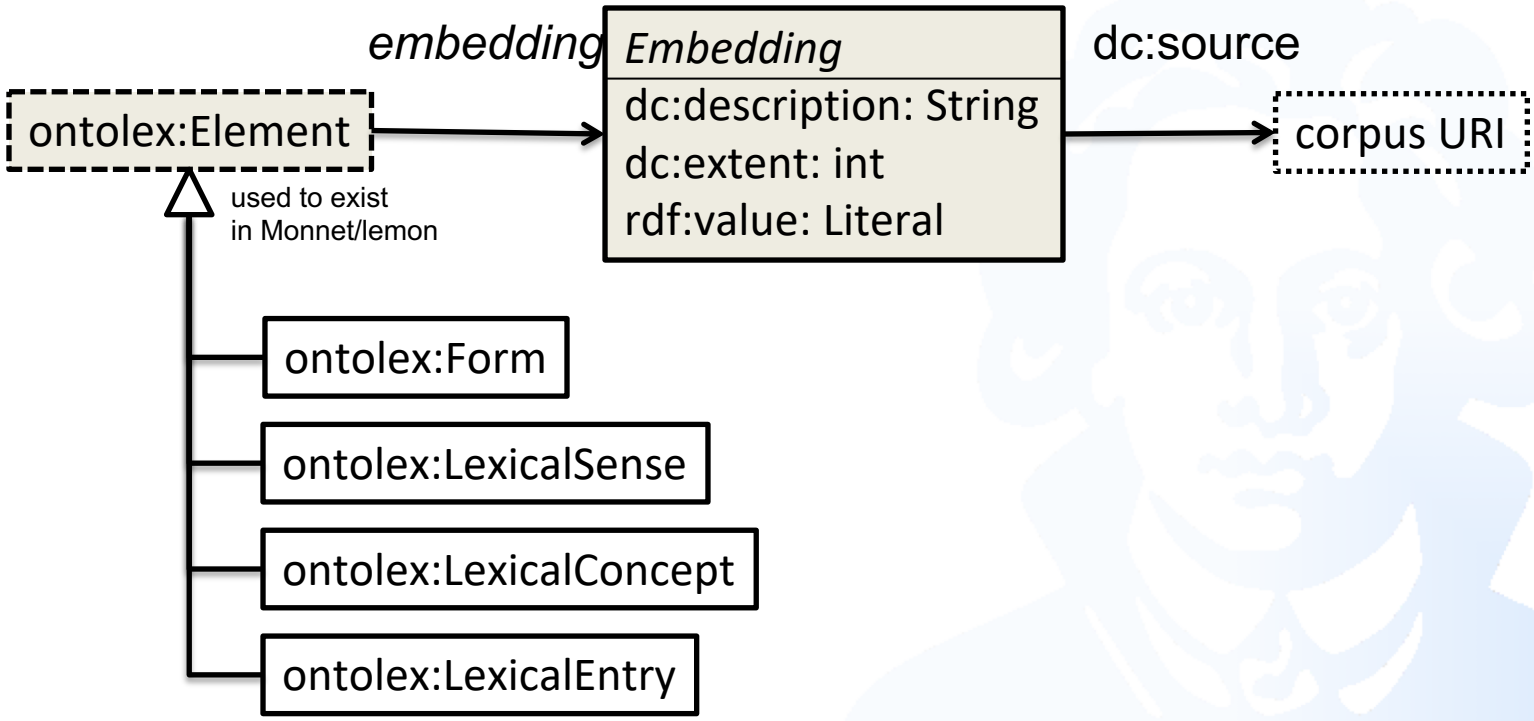
Embedding

```
frak 0.015246 -0.30472 0.68107 -0.59727 -0.95368 -1.0931 0.58783 -0.19128 0.49108 (
-0.86604 -0.91168 0.26087 -0.42067 0.60649 0.80644 -1.0477 0.67461 0.34154 -0.07251
-1.187 -0.11523 -0.078265 0.29849 0.22993 -0.12354 0.2829 1.0697 0.015366
```

50-dimensional GloVe 6B (Wikipedia 2014+Gigaword 5)



```
:frak a ontolex:LexicalEntry;
  ontolex:canonicalForm/ontolex:writtenRep "frak"@en;
  frac:embedding [
    a frac:Embedding;
       rdf:value "0.015246 -0.30472 0.68107 ...";
       dct:source
         <http://dumps.wikimedia.org/enwiki/20140102/>,
         <https://catalog.ldc.upenn.edu/LDC2011T07>;
       dct:extent 50^^^xsd:int;
       dct:description "GloVe v.1.1, documented in Jeffre
```

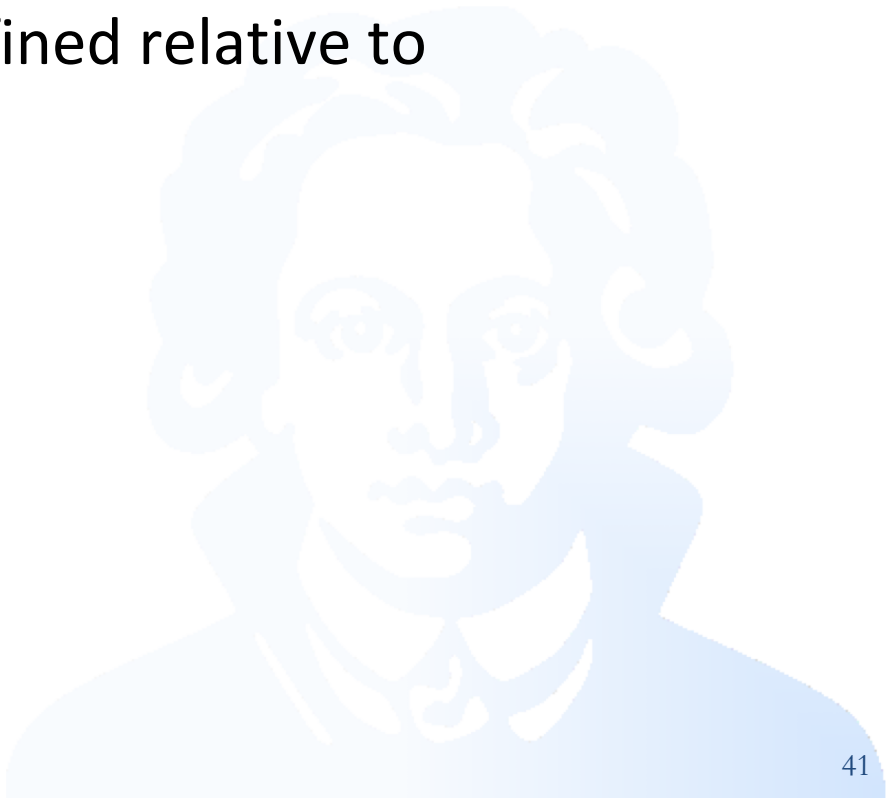


Afterthoughts on embeddings

- wrt. embeddings and RDF, we most not confuse the following two aspects:
  1. vector processing / machine learning
     requires encoding structured data in embeddings
  2. embedding publication and sharing
     requires interoperable representation of embeddings
- current research on RDF embeddings solely focuses on processing (1)
  - Here, we are interested in resource sharing and access (2), *not* in processing. Of course, subsequent processing of embeddings requires a machine learning component, but here, we are ignorant on how this is handled.

Discussing embeddings

- Not yet ...
  - let‘s look into collocations and synonyms, first
  - with a generalized notion of embeddings, corpus-based synonyms can be defined relative to embeddings



Corpus-based Associations

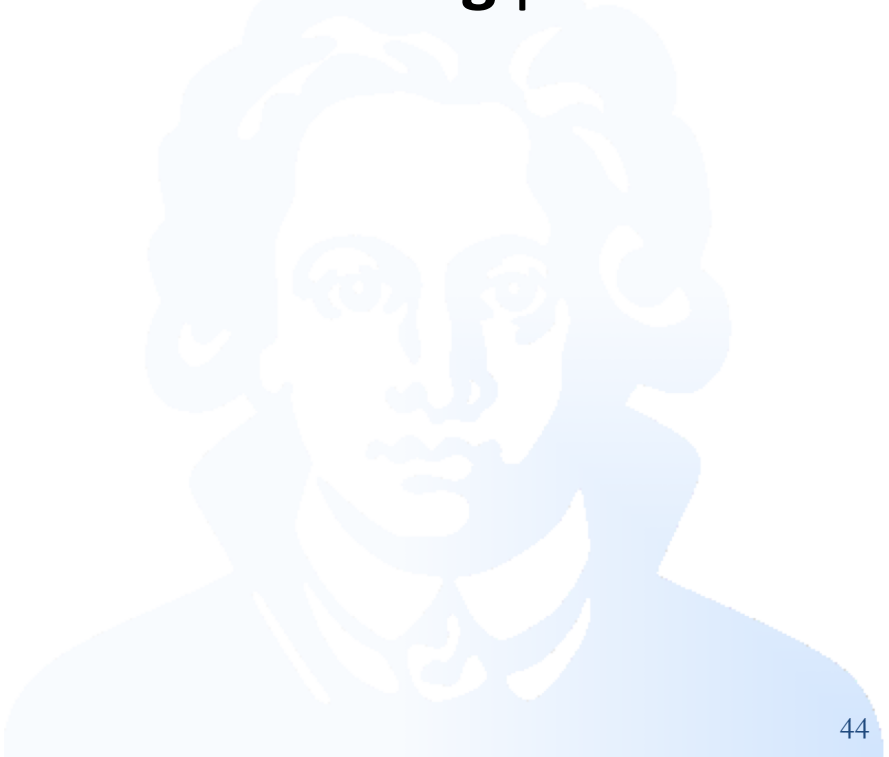
Collocation & Synonymity

Corpus-based associations

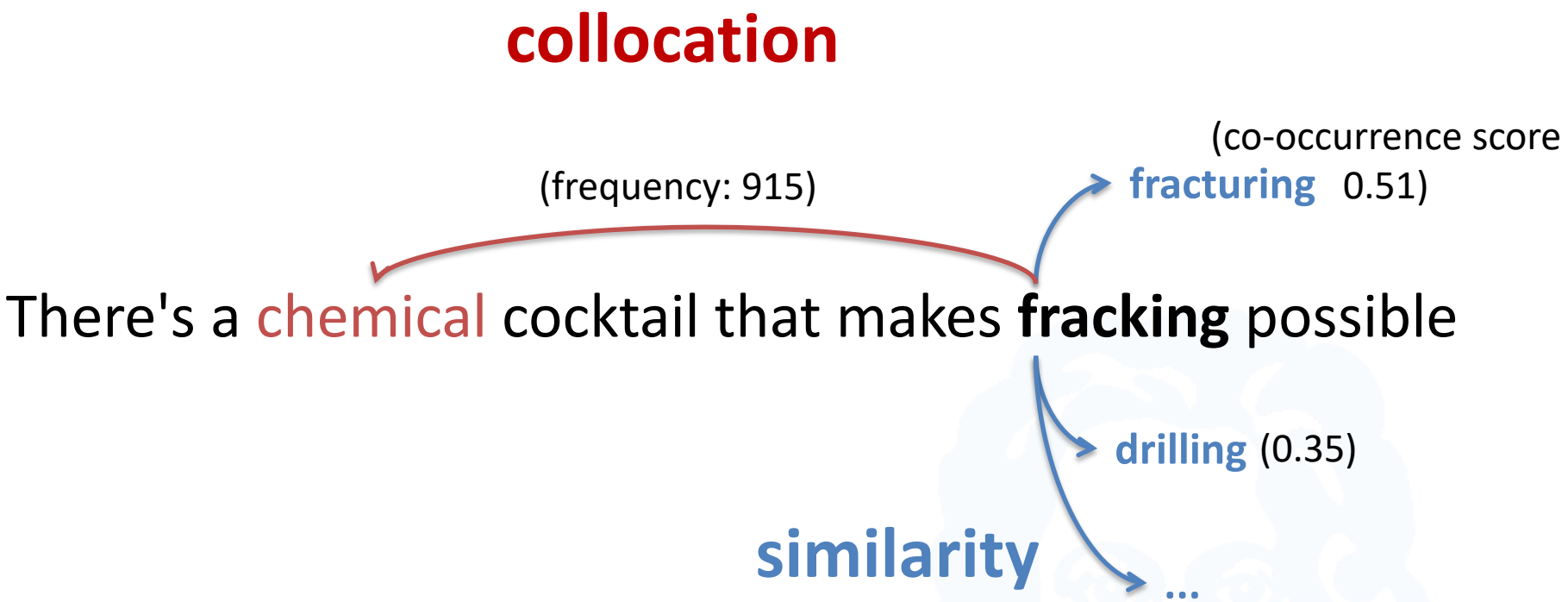
- collocation
  - two (or more) words that regularly co-occur with each other, can be lexicalized multi-word expressions
    - corpus-based approximation of idiomatic expressions
    - syntagmatic relation between words: likelihood to occur simultaneously in the same context
- similarity
  - two (or more) words that are characterized by their occurrence in similar contexts
    - corpus-based approximation of synonymity
    - paradigmatic relation between words: likelihood to stand in for each other in the same context

Corpus-based associations

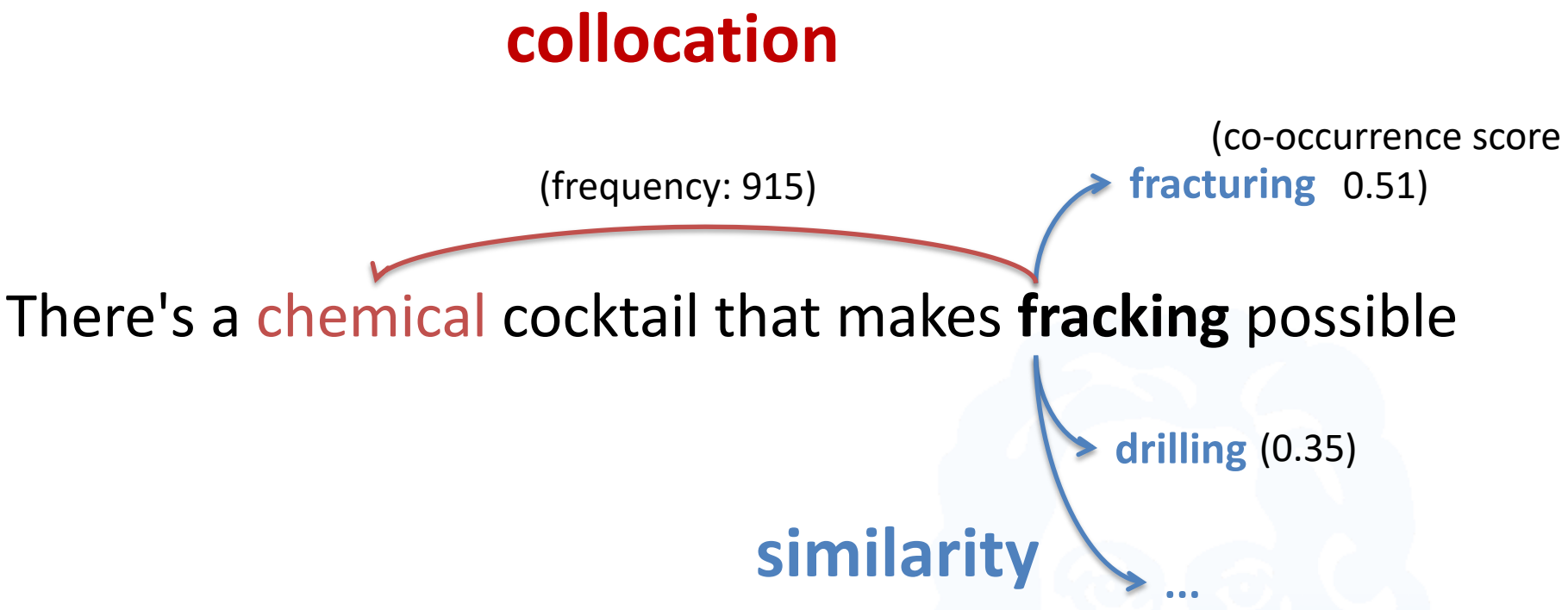
There's a chemical cocktail that makes **fracking** possible



Corpus-based associations (Wortschatz)



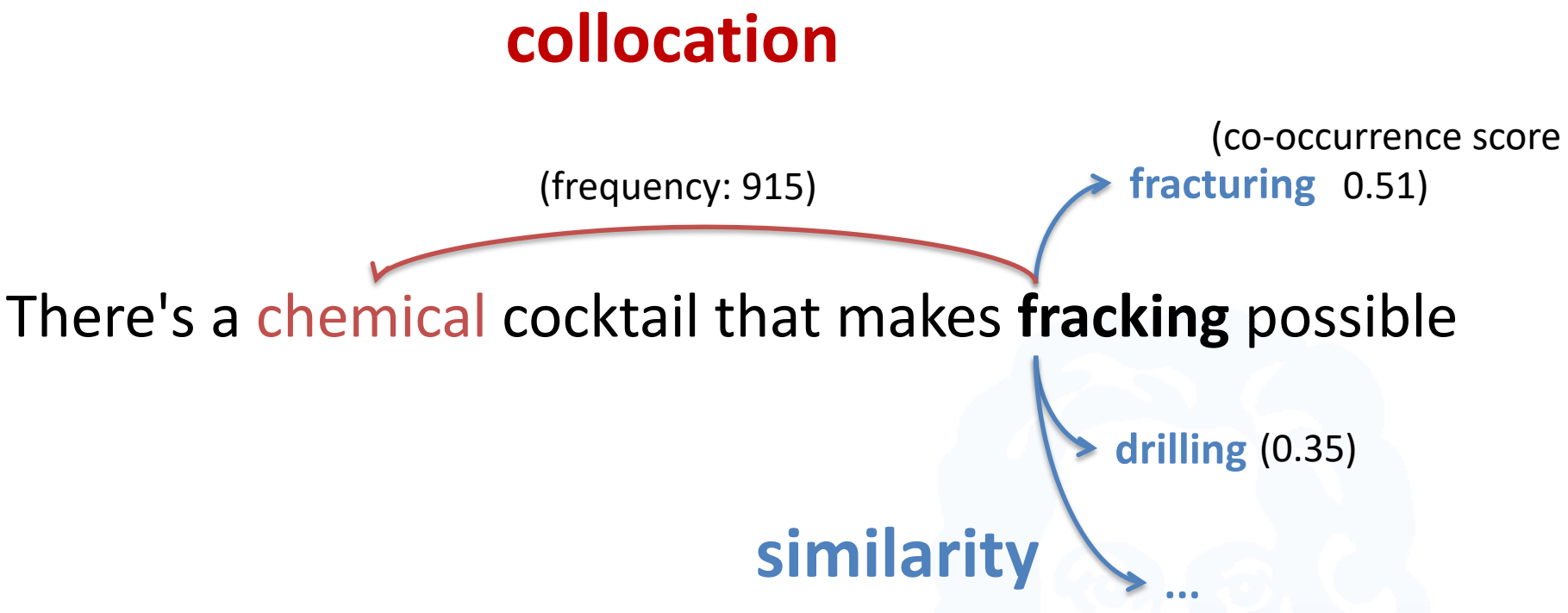
Corpus-based associations (Wortschatz)



- lexinfo:collocation ⊑ lemon:senseRelation
- lexinfo:relatedTerm ⊑ lemon:senseRelation

~ ontolex:senseRel : ontolex:LexicalSense x ontolex:LexicalSense

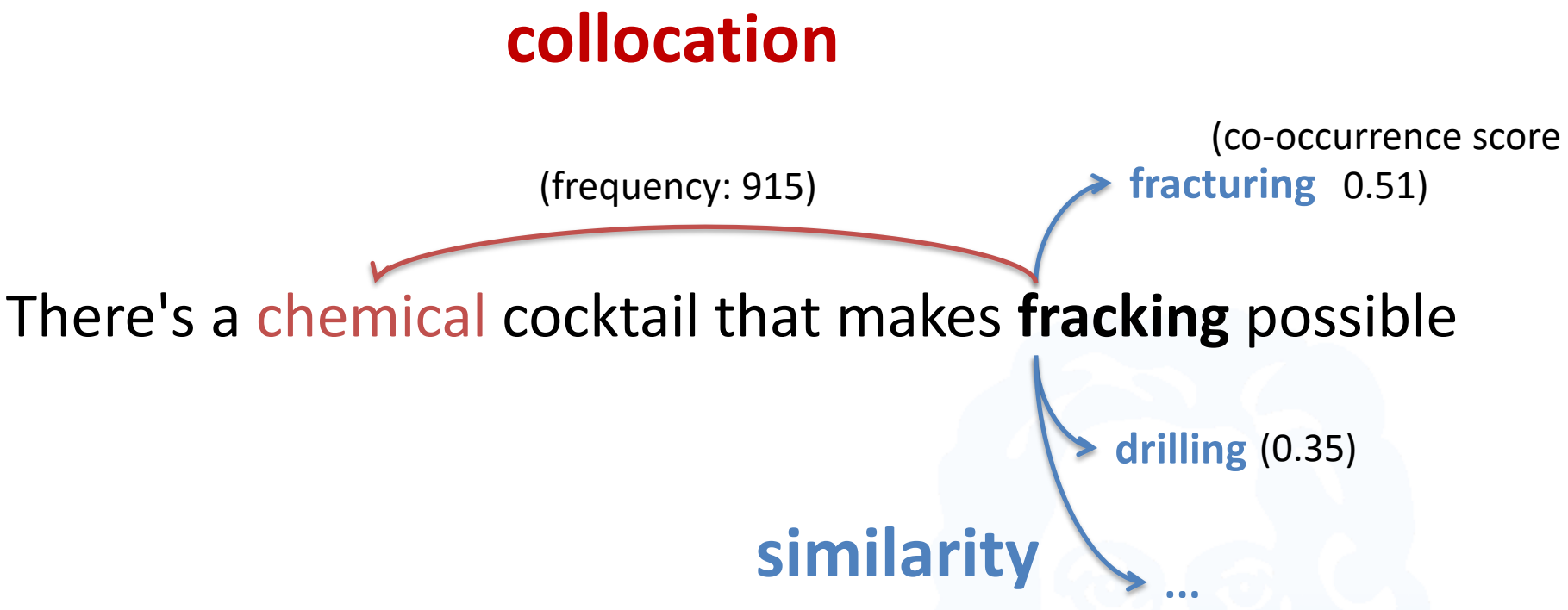
Corpus-based associations (Wortschatz)



- lexinfo:collocation ⊑ lemon:senseRelation
- lexinfo:relatedTerm ⊑ lemon:senseRelation

but we‘re normally **not** dealing with senses, here, more often with forms or lemmas => new vocabulary elements

Corpus-based associations (Wortschatz)



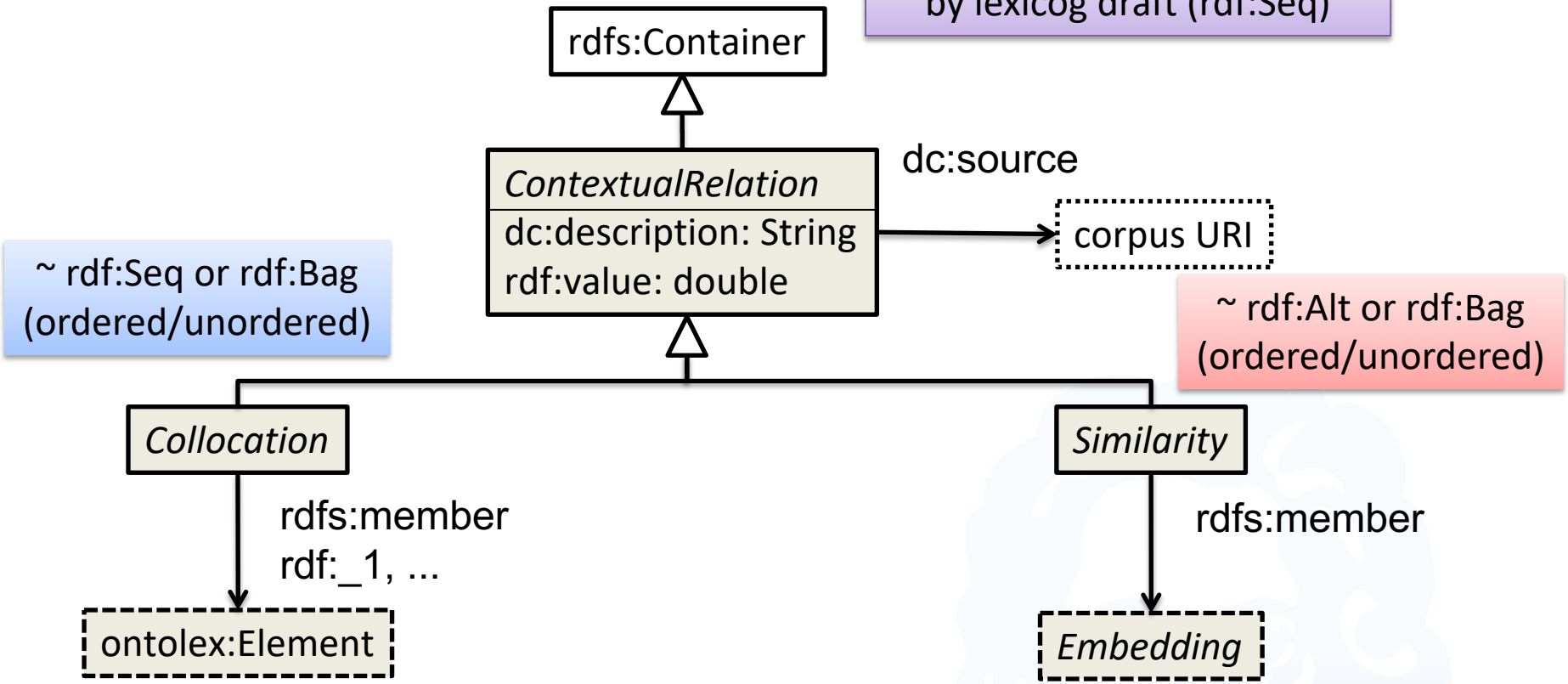
- lexinfo:collocation ⊑ lemon:senseRelation
- lexinfo:relatedTerm ⊑ lemon:senseRelation

but we‘re normally **not** dealing with senses, here, more often

similarity assessments are *the* primary purpose of (word) embeddings

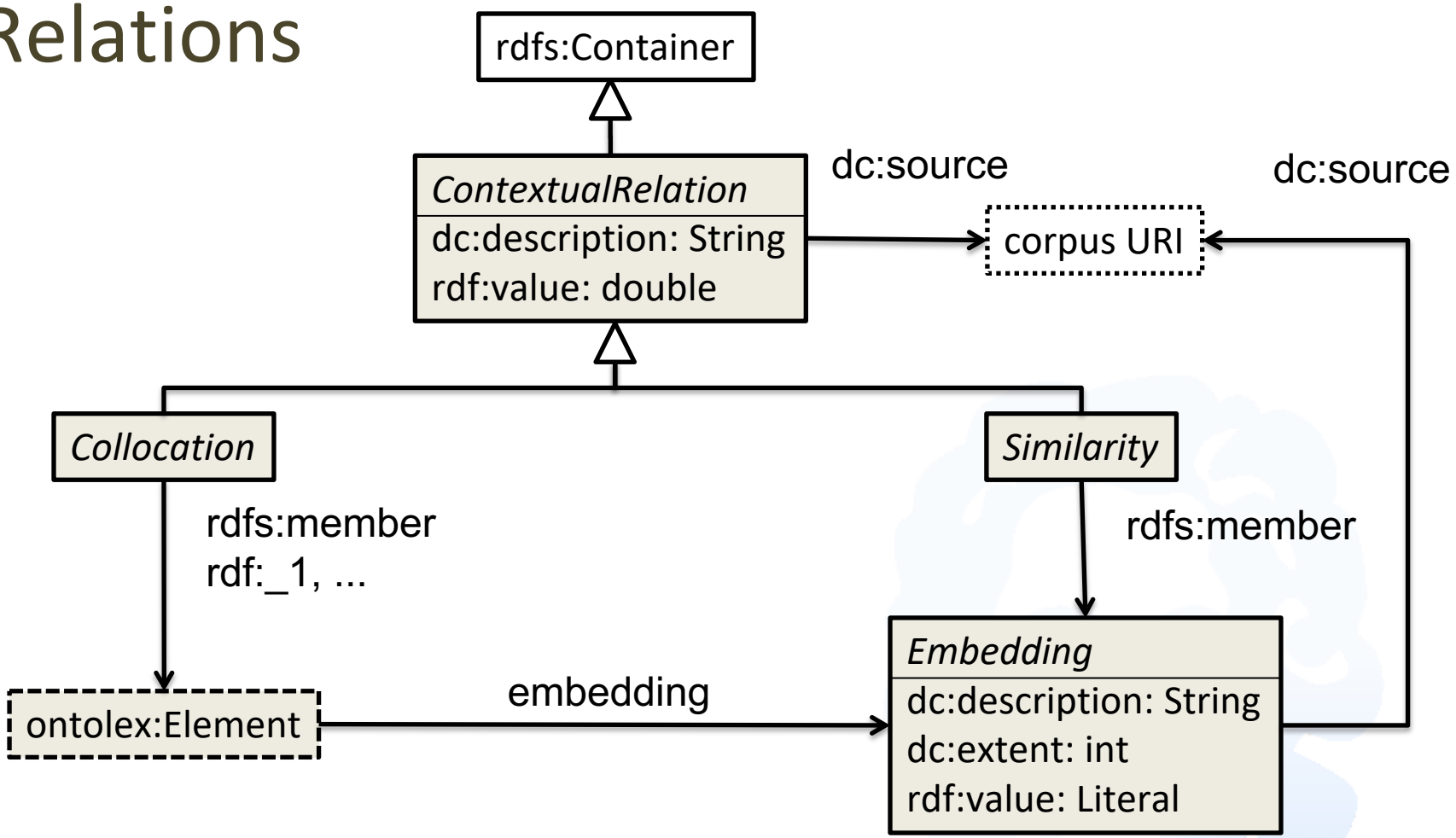
ContextualRelation

Container modelling inspired by lexicog draft (rdf:Seq)



- Collocation between ontolex:Elements
- (distributional) similarity between embeddings
  - where none are given, these can be blanks attached to ontolex:Elements via frac:embedding
  - classical bags of words are infinite-dimensional vectors ;)

Summary: Embeddings & Contextual Relations



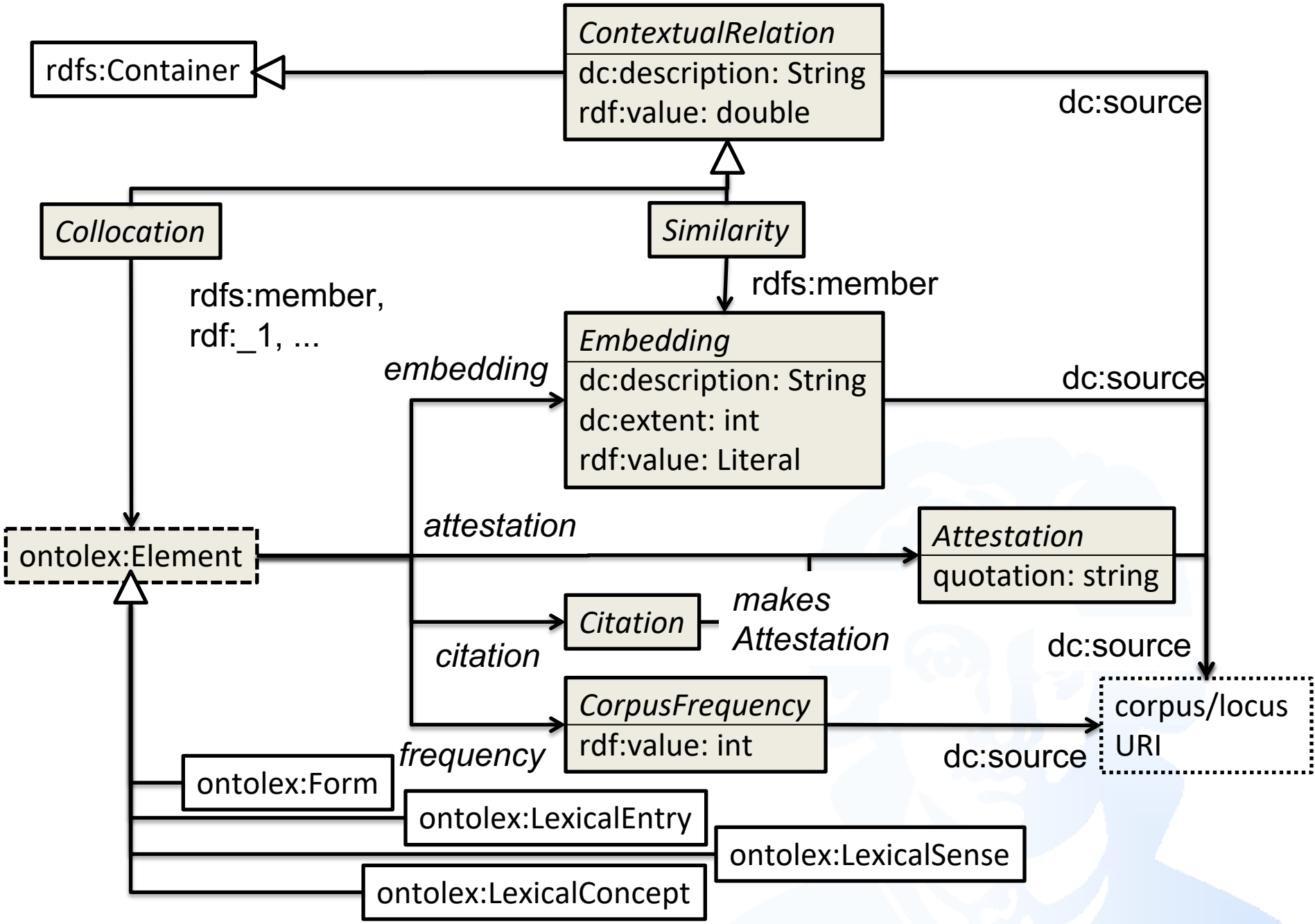
get all collocates for a given lexical entry

```
SELECT ?x (MAX(?strength) as ?s)
WHERE {
        ?le canonicalForm/writtenRep „bucket“.
        ?le ^rdfs:member ?coll.
        ?coll a Collocation; rdf:value ?strength;
        rdf:member ?sim.
        FILTER(?sim!=?le)
        ?sim canonicalForm/writtenRep ?x.
} GROUP BY ?x
```

note that we have no way of telling whether how many elements the collocation contains

Discussion

- frac:Embedding
  - more generalized than plain word embeddings, any better term?
- relation to lexinfo properties ?
- alternatives to the container modelling of contextual associations?
- suggested use cases:
  - contextual relation: wrapper for corpus-based word associations (word sketches)
  - embedding: wrapper for any kind of distributional representation, bridge to SOTA NLP resources


ContextualRelation
dc:description: String
rdf:value: double
rdfs:Container
dc:source
Collocation
Similarity
rdfs:member
rdfs:member,
rdf:_1, ...
Embedding
dc:description: String
dc:extent: int
rdf:value: Literal
embedding
dc:source
attestation
ontolex:Element
Attestation
quotation: string
Citation
makes
Attestation
citation
dc:source
CorpusFrequency
rdf:value: int
corpus/locus
URI
frequency
dc:source
ontolex:Form
ontolex:LexicalEntry
ontolex:LexicalSense
ontolex:LexicalConcept

How to proceed ...

- biweekly telcos, starting mid-July 2019
  - who would commit to participate?
    - we need 5 participants from different institutions
  - after/alternating with morphology telcos? different slot?
- until then
  - create and populate a wiki with requirements and use cases
    - cf. morphology module

Some more details

Corpus Frequency

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

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

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

Simplifications

introduce corpus-specific frequency classes

```
# Corpus Frequency in the EPSD corpus
:EPSDFrequency rdfs:subClassOf frac:CorpusFrequency.
:EPSDFrequency rdfs:subClassOf
 [ a owl:Restriction ;
   owl:onProperty dct:source ;
   owl:hasValue <http://oracc.museum.upenn.edu/epsd2/pager> ] .

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

Simplifications

additional restrictions => frequency classes for specific sub-corpora

```
# 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 ].
```

Simplifications: Resource-specific embedding class

```
# resource-specific embedding class
:GloVe6BEmbedding_50d rdfs:subClassOf frac:Embedding;
  rdfs:subClassOf
    [ a owl:Restriction;
          owl:onProperty dct:source;
          owl:hasValue
                  <http://dumps.wikimedia.org/enwiki/20140102/>,
                  <https://catalog.ldc.upenn.edu/LDC2011T07> ],
        [ a owl:Restriction;
          owl:onProperty dct:extent;
          owl:hasValue 50^^^xsd:int ],
        [ a owl:Restriction;
          owl:onProperty dct:description;
          owl:hasValue "GloVe v.1.1, documented in Jeffrey Pennino

# embedding assignment
:frak a ontolex:LexicalEntry;
  ontolex:canonicalForm/ontolex:writtenRep "frak"@en;
  frac:embedding [
    a :GloVe6BEmbedding_50d;
        rdf:value "0.015246 -0.30472 0.68107 ..." ].
```