Proceedings of the Globalex Workshop on Linked Lexicography, pages 1–9
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Modelling Frequency and Attestations for OntoLex-Lemon

Christian Chiarcos[1], Maxim Ionov[1], Jesse de Does[2], Katrien Depuydt[2], Anas Fahad Khan[3], Sander Stolk[4], Thierry Declerck[5], John P. McCrae[6]
[1]Goethe-Universität Frankfurt am Main
[2]Instituut voor de Nederlandse Taal, Leiden, the Netherlands
[3]Istituto di Linguistica Computazionale "A. Zampolli" (ILC-CNR) Pisa, Italy
[4]Leiden University, Leiden, the Netherlands
[5]DFKI GmbH, Multilinguality and Language Technology
[6]Data Science Institute, National University of Ireland Galway
[1]{chiarcos,ionov}@informatik.uni-frankfurt.de, [2]{dedoes,depuydt}@ivdnt.org
[3]fahad.khan@ilc.cnr.it, [4]s.s.stolk@hum.leidenuniv.nl, [5]declerck@dfki.de, [6]john@mccr.ae

Abstract

The OntoLex vocabulary enjoys increasing popularity as a means of publishing lexical resources with RDF and as Linked Data. The recent publication of a new OntoLex module for lexicography, *lexicog*, reflects its increasing importance for digital lexicography. However, not all aspects of digital lexicography have been covered to the same extent. In particular, supplementary information drawn from corpora such as frequency information, links to attestations, and collocation data were considered to be beyond the scope of *lexicog*. Therefore, the OntoLex community has put forward the proposal for a novel module for frequency, attestation and corpus information (FrAC), that not only covers the requirements of digital lexicography, but also accommodates essential data structures for lexical information in natural language processing. This paper introduces the current state of the OntoLex-FrAC vocabulary, describes its structure, some selected use cases, elementary concepts and fundamental definitions, with a focus on frequency and attestations.

Keywords: lexical resources, community standards, linguistic linked (open) data, OntoLex

1. Background

The primary community standard for publishing lexical resources as linked data is the OntoLex-Lemon vocabulary, which is based on the *lemon* model (McCrae et al., 2012), that has been designed as a model for complementing ontologies with lexical information in the Monnet project.[1] With its further development in the context of the W3C OntoLex Community Group, its scope was broadened and it developed towards the primary RDF vocabulary for lexical information. In 2016, the OntoLex vocabulary was published as a W3C Report[2] (Cimiano et al., 2016).

The model's primary element is the lexical entry (see Fig. 1), which represents a single lexeme with a single part-of-speech (when appropriate) and a set of grammatical properties. This entry is composed of a number of forms as well as a number of senses which enumerate its various meanings. The meanings of these senses can be defined formally by reference to an ontology or informally by a lexical concept, which defines a concept in a cross-lingual manner. This paper describes the on-going development of a novel OntoLex module for frequency, attestation and corpus information (OntoLex-FrAC). FrAC extends OntoLex and its recently published *lexicog* vocabulary[3] with the capability to represent important supplementary information used in digital lexicography (collocations, distributional similarity, attestations, frequency information). As this information is equally relevant for both digital lexicography and for applications in fields such as natural language processing, the W3C OntoLex Community decided to treat such information within a separate module and to remove the corresponding concepts from the lexicography module.

Important motivations to extend OntoLex core and lexicography modules are the Elexis project (Krek et al., 2019),[4] where strategies, tools and standards for extracting, structuring and linking lexicographic resources are developed for their inclusion in Linked Open Data and the Semantic Web, as well as the Prêt-à-LLOD project (Declerck et al., 2020)[5] on making linguistic linked open data ready-to-use for knowledge services across sectors.

The goal of the module is to complement the OntoLex-Lemon core elements with a vocabulary layer to represent lexicographical and semantic information derived from or defined with reference to corpora and external resources in a way that (1) generalizes over use cases from digital lexicography, natural language processing, artificial intelli-

[1]A European Union Funded project in multilingual ontologies that ran from 2010-2013.

[2]`https://www.w3.org/2016/05/ontolex/`

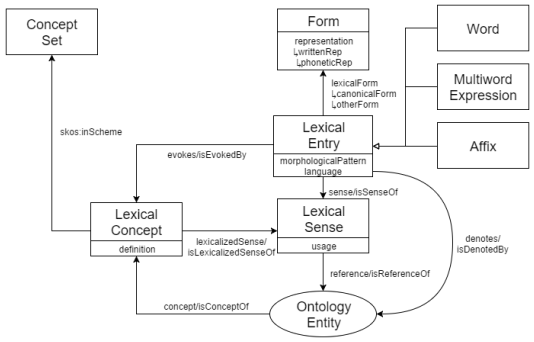


Figure 1: OntoLex-Lemon core model

[3]`https://www.w3.org/2019/09/lexicog/`

[4]See also `https://elex.is/`.

[5]See also `http://www.pret-a-llod.eu`.

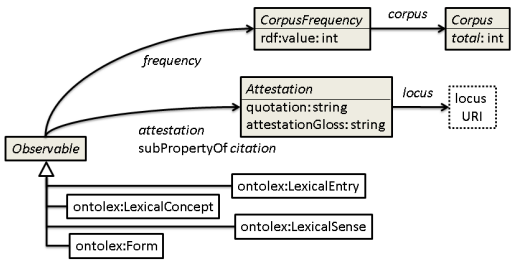


Figure 2: OntoLex-FrAC module structure

gence, computational philology and corpus linguistics, that (2) facilitates exchange, storage and re-usability of such data along with lexical information, and that (3) minimizes information loss.
The scope of the model is three-fold:

1. Extending the OntoLex-lexicog model with corpus information to support existing challenges in corpus-driven lexicography.
2. Modelling existing lexical and distributional-semantic resources (corpus-based dictionaries, collocation dictionaries, embeddings) as linked data, to allow their conjoint publication and inter-operation by Semantic Web standards.
3. Providing a conceptual / abstract model of relevant concepts in distributional semantics that facilitates building linked data-based applications that consume and combine both lexical and distributional information.

Based on this, the following parts of the module can be distinguished: (1) Frequency, (2) attestations, and (3) corpus-derived information.
This paper provides an account for frequency and attestations, for which a consensus model has already been reached. Corpus information beyond that includes various information about lexically relevant concepts that can be created on grounds of corpora. This includes, for example, distributional similarity scores, collocation vectors or embeddings.
The overall structure is presented in Figure 2, which reflects the current state of modelling. Extensions for embeddings, collocations and similarity are still under development.
For OntoLex, we assume that frequency, attestation and corpus information can be provided about *every* linguistic content element in the core model and the OntoLex modules. This includes `ontolex:Form` (token frequency, etc.), `ontolex:LexicalEntry` (frequency of disambiguated lemmas), `ontolex:LexicalSense` (sense frequency), `ontolex:LexicalConcept` (e.g., synset frequency), `lexicog:Entry` (if used for representing homonyms: frequency of non-disambiguated lemmas), etc. Formally, we define the domain of FrAC properties by the concept `frac:Observable` that we introduce as a generalization over these concepts:[6] Everything for which we provide frequency, attestation or corpus information must be observable in a corpus or another linguistic data source.

[6]It is to be expected that other, subsequent OntoLex modules may require a similar generalization, and then, it would be advisable to create a class `ontolex:LexicalElement` (or the like) in the core model and use that one, instead.

2. Frequency

Frequency information is a crucial component in human language technology. Corpus-based lexicography originates with the Brown corpus (Kučera and Francis, 1967) and, subsequently, the analysis of frequency distributions of word forms, lemmas and other linguistic elements has become a standard technique in lexicography and philology, and given rise to the field of corpus linguistics. Information on frequency is used in computational lexicography and is essential for NLP and corpus linguistics. The FrAC module includes terminology to capture such information, both absolute and relative frequency, in order to facilitate sharing and utilising this valued information.

2.1. Model

For modelling, we focus on absolute frequencies, as relative frequencies can be derived if absolute frequencies and totals are known.
In order to avoid confusion with `lexinfo:Frequency` (which provides lexicographic assessments such as *commonly used*, *infrequently used*, etc.), this is defined with reference to a particular dataset, a corpus.

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 frequency is always defined relative to a corpus. We do not provide a formal definition of what a corpus is (it can be any kind or collection of linguistic data at any scale, structured or unstructured), except that we expect it to define a total of elements contained (`frac:total`). In many practical applications, it is necessary to provide relative counts, and in this way, these can be easily derived from the absolute (element) frequency provided by the `CorpusFrequency` class and the total defined by the underlying corpus.

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)

Note that we expect a corpus to apply a specific tokenization strategy to define a total of elements. If different tokenization strategies of the same dataset occur, these result in different `frac:Corpus` elements.

2.2. Illustrative Example

The Electronic Penn Sumerian Dictionary (ePSD)[7] is an effort to provide an exhaustive dictionary of Sumerian, an isolate language of the ancient Near East, written between the 3rd and 1st millennium BCE being the oldest known written language. The Pennsylvania Sumerian Dictionary Project is carried out at the University of Pennsylvania Museum of Anthropology and Archaeology and funded by the National Endowment for the Humanities and private contributions. Its electronic edition has been developed in a corpus-based fashion, with information such as shown in Fig. 3: It provides frequency information per time period ("3000", "2500", "2000" etc.), orthographic variants ("[1]", "[2]", "[3]"), individual inflected forms (window "ePSD Forms"), and individual word senses ("1. (to be) strong" etc.), and it provides absolute and relative counts.

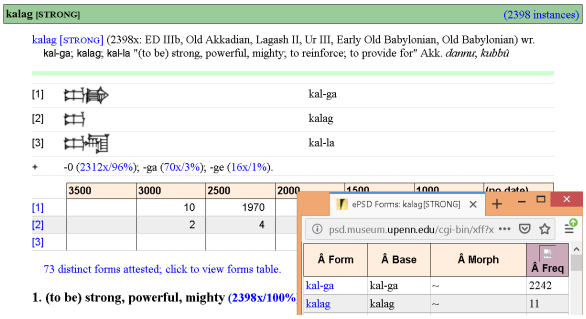


Figure 3: Electronic Penn Sumerian Dictionary (ePSD), sample entry *kalag*

Within the ePSD, frequency information is assigned to *any* element in the dictionary (at least forms, entries, senses), and separately for a large number of subcorpora (defined by time periods and regions/cultures).
An example in Listing 1 illustrates word and form frequencies for the Sumerian word *kalag* (n.) "(to be) strong" and the frequencies of the underlying corpus.

2.3. Shorthands for Data Modelling

The model sketched above is relatively verbose: It requires full provenance information to be provided with every frequency count. It is necessary to provide the link to the underlying corpus *for every frequency assessment* because the same element may receive different counts over different corpora. For querying and retrieval, having this information explicitly given is a very good means to ease access and processing. From the perspective of data modelling, however, it is highly redundant and should be avoided.

As corpus-derived information requires provenance and other metadata, the FrAC module uses reification (class-based modelling) for concepts such as frequency or embeddings. In a dataset, this information will be recurring, and for redundancy reduction, we recommend to provide resource-specific subclasses of concepts that provide metadata by means of `owl:Restrictions` that provide the value for the respective properties. This is illustrated in Listing 2 for the relevant FrAC classes.

For data modelling and sharing, we thus define a corpus- or collection-specific subclass of `frac:CorpusFrequency` with an invariant link to the underlying corpus (and additional provenance information, if required). For specifying absolute frequencies, we thus refer to this *constrained* frequency type.

This leads to more compact data and is more robust against information loss (i.e., if an RDF dump is incomplete, we either lose frequency metadata completely or we maintain its provenance, but it will not be incomplete).

Listing 1: Word and form frequencies in ePSD

```
# word frequency, over all form variants
epsd:kalag_strong_v a ontolex:LexicalEntry;
   frac:frequency [
      a frac:CorpusFrequency;
      rdf:value "2398"^^xsd:int;
      frac:corpus
         <http://oracc.museum.upenn.edu/epsd2/pager>
] .

# form frequency for individual orthographical variants
epsd:kalag_strong_v a ontolex:canonicalForm [
   ontolex:writtenRep "kal-ga"@sux-Latn;
   frac:frequency [
      a frac:CorpusFrequency;
      rdf:value "2312"^^xsd:int;
      frac:corpus
         <http://oracc.museum.upenn.edu/epsd2/pager>
      ]
   ] .

epsd:kalag_strong_v a ontolex:otherForm [
   ontolex:writtenRep "kalag"@sux-Latn;
   frac:frequency [
      a frac:CorpusFrequency;
      rdf:value "70"^^xsd:int;
      frac:corpus
         <http://oracc.museum.upenn.edu/epsd2/pager>
      ]
   ] .
```

Listing 2: Reifying provenance information for ePSD

```
: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:kalag_strong_v frac:frequency [
   a :EPSDFrequency;
   rdf:value "2398"^^xsd:int
].
```

[7] http://psd.museum.upenn.edu/

`frac:CorpusFrequency` can be extended with additional filter conditions to define sub-corpora. For example, we can restrict the subcorpus to a particular time period, e.g., the Neo-Sumerian Ur III period:

```
# 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:kalag_strong_v frac:frequency [
   a :EPSDFrequency_UrIII;
   rdf:value "1916"^^xsd:int
].
```

3. Attestations

According to Kilgarriff (1997):

> "the scientific study of language should not include word senses as objects in its ontology. Where 'word senses' have a role to play in a scientific vocabulary, they are to be construed as abstractions over clusters of word usages ... the basic units are occurrences of the word in context (operationalised as corpus citations)."

While dispensing with word senses is not an option for modelling dictionaries and lexica, one should take into account that it is by analysing corpus material that lexicographers, using their expert knowledge, can provide a careful description of the meanings of each word in the dictionary, together with the corpus evidence in the form of dictionary citations. Both the current OntoLex core model and the lexicography module lack a way to include this evidence. The main objective of modelling attestations for OntoLex is to do justice to the character of 'scholarly' lexicographical work by allowing us to "put the corpus into the dictionary".

3.1. Model

There are at least two different ways of linking lexical information with corpus evidence, each of which arises from a different tradition, in the first case that of scholarly lexicography and in the second case that of computational linguistics. These are:

- 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;
- Linking a computational lexicon with the corpora from which the lexical information is derived.

The attestation part of the FrAC module is intended to model both of these approaches in a unified way. It is important to have a flexible vocabulary to characterize the properties of attestations in dictionaries, allowing us to take account of, for instance, the presence of a context snippet and aspects of a cited attestation which relate to its being a scholarly hypothesis. Khan and Boschetti's *lemonBib* model for lexicographical citations (Khan and Boschetti, 2018) tackles some important issues relevant to the characterization of evidence in lexicography and proposes solutions based on the FRBR[8](Saur, 1998), CiTO and FaBIO ontologies (Peroni and Shotton, 2012). In particular (Khan and Boschetti, 2018) mention:

- The distinction between citations in general and citations which provide evidence (attestation)
- Enabling the marking of text readings as conjectural

In fact we can identify at least five axes of classification:

1. Attestation (Citation provides evidence for the word sense) versus other types of citation in a lexical entry.
2. Degree of certainty with regard to the source text (e.g., given a reconstructed text how sure can we be that the word was present in the original?)
3. Degree of certainty of the interpretation (e.g., is this really an instance of the relevant word sense?)
4. Is any textual context for the cited usage of the word given in the form of a quotation?
5. Is the occurrence (or multiple occurrences) of the headword in the context/snippet explicitly marked?

The attestation part of the module tries to provide the necessary vocabulary for the representation of this data.

- There always is an instance of an object for any type of citation. It is always linked to the `frac:Observable` with the `citation` object property. Several vocabularies for modelling citation information have been introduced, FrAC is thus underspecified with respect to the exact definition but relies on using such vocabularies. One candidate vocabulary is the previously mentioned CITO ontology which provides fine-grained information, e.g., the type of citation (cites as evidence, agrees with, etc.) can be reflected in the value of `cito:hasCitationCharacterization` property and by subclasses of `Citation`.
- (Un)certainty of source text reading and/or lexicographic interpretation can be modeled by two distinct boolean data properties associated with the `Citation` object.
- Presence of context is simply reflected by a non-empty value for the `quotation` data property.
- The `locus` object property can optionally be used to mark the place in the snippet in which the headword occurs (this is useful for computational applications use of dictionary quotations in e.g.). For expressing the locus, external vocabularies such as NIF or WebAnnotation can be used.

3.1.1. Classes and Concepts

Attestations constitute a special form of citation that provide evidence for the existence of a certain lexical phenomena; they can elucidate meaning or illustrate various linguistic features.
In scholarly dictionaries, attestations are a representative selection from the occurrences of a headword in a textual

[8] `http://purl.org/vocab/frbr/core#`

corpus. These citations often consist of quotation accompanied by a reference to the source. The quoted text usually contains the occurrence of the headword.

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.

A **Citation** is "a conceptual directional link from a citing entity to a cited entity, created by a human performative act of making a citation, typically instantiated by the inclusion of a bibliographic reference in the reference list of the citing entity, or by the inclusion within the citing entity of a link, in the form of an HTTP Uniform Resource Locator (URL), to a resource on the World Wide Web".

This definition is taken from CITO (Peroni and Shotton, 2012). The FrAC module does not prescribe a specific vocabulary for the citation object. If the CITO vocabulary is used, FrAC Citations can be defined as the subclass of CITO citations having `frac:Observable` as citing entity and attestations would correspond to citations with the `cito:hasCitationCharacterization` value `citesAsEvidence`.

In many applications, it is desirable to specify the location of the occurrence of a headword in the quoted text of an attestation, for example, by means of character offsets. Different conventions for referencing strings by character offsets do exist, representative solutions are string URIs as provided by RCF5147 (for plain text) and NIF (all mimetypes),[9] and the selector mechanism of WebAnnotation.[10] As different vocabularies can be used to establish locus objects, the FrAC vocabulary is underspecified with respect to the exact nature of the locus object. Accordingly, the `locus` property that links an attestation with its source takes any URI as object.

3.1.2. Properties

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: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.

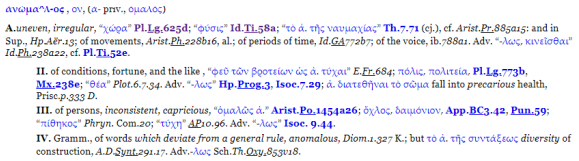
frac:locus (domain: frac:Attestation) points to the location at which the relevant word(s) can be found.

3.1.3. Relation with other Vocabularies

When the dictionary citations refer to an accessible corpus, we could consider the link between corpus and lexicon as a (e.g. word sense) annotation of the corpus. Different vocabularies for this purpose exist.

[9] `https://tools.ietf.org/html/rfc5147`, `http://persistence.uni-leipzig.org/nlp2rdf/`

[10] `https://www.w3.org/TR/annotation-model/`


ανωμα^λ-ος , ον, (α- prv., ομαλος)

A. *uneven, irregular,* "χωρα" **Pl.Lg.625d**; "φυσις" **Id.Ti.58a**; "το ἀ. τῆς ναυμαχίας" **Th.7.71** (cj.), cf. *Arist.Pr.885a15*: and in Sup., *Hp.Aër.13*; of movements, *Arist.Ph.228b16*, al.; of periods of time, *Id.GA772b7*; of the voice, ib.*788a1*. Adv. "-λως, κινεῖσθαι" *Id.Ph.238a22*, cf. **Pl.Ti.52e**.

II. of conditions, fortune, and the like , "φεῦ τῶν βροτείων ὡς ἀ. τύχαι" *E.Fr.684*; πόλις, πολιτεία, **Pl.Lg.773b**, **Mx.238e**; "θέα" *Plot.6.7.34*. Adv. "-λως" **Hp.Prog.3**, **Isoc.7.29**; ἀ. διατεθῆναι τὸ σῶμα fall into *precarious* health, Prisc.*p.333 D.*

III. of persons, *inconsistent, capricious,* "ὁμαλὰς ἀ." **Arist.Po.1454a26**; ὄχλος, δαιμόνιον, **App.BC3.42**, **Pun.59**; "πίθηκος" *Phryn. Com.20*; "τύχη" *AP10.96*. Adv. "-λως" **Isoc. 9.44**.

IV. Gramm., of words *which deviate from a general rule, anomalous, Diom.1.327* K.; but τὸ ἀ. τῆς συντάξεως *diversity* of construction, *A.D.Synt.291.17*. Adv.-λως Sch.*Th.Oxy.853v18*.

Figure 4: The entry for ἀνώμαλος

The NLP Interchange Format NIF, for example, provides vocabulary to point to a more precise location of the relevant word(s) within the quotation:

nif:beginIndex (range: xs:Int) Initial character offset of the word to which the lexicographical interpretation is attached

nif:endIndex (range: xs:Int) Final character offset of the word to which the lexicographical interpretation is attached

Similarly, the Web Annotation Framework can be used for modelling loci (listing 6). In particular, Web Annotation provides a vocabulary to formalize loci by means of offsets as in NIF, but also by other means, e.g., XPath.

4. Use Cases

4.1. The Liddell-Scott-Jones Ancient Greek Lexicon

Our first use-case shows the application of the FrAC module to the modelling and publication of legacy lexical resources as linked data. In our particular case we will be working with the Liddell-Scott-Jones Ancient Greek Lexicon (LSJ) a scholarly dictionary in Ancient Greek-English originally published in the 19th century by Henry George Liddell and Robert Scott and then revised in 1940 by Henry Stuart Jones. The LSJ is still regarded as an authoritative lexicographic resource in Ancient Greek scholarship and is currently in print in its ninth edition (Liddell et al., 1996). In 2007 the Perseus project published a digital edition of the work which was made available on their website both in HTML and as a TEI source[11], which we take as a starting point of our work[12]. As may be imagined, the LSJ is an extremely rich resource and one that is particularly valuable with respect to its sense based attestations which it takes from the surviving corpus of Ancient Greek literature. We will look at one entry from that work and then show how the attestations may be modelled using the classes and properties which have been provisionally developed as part of the FrAC module. The entry in question is that for the word ἀνώμαλος 'uneven, irregular', from which the English word *anomalous* derives, see Fig. 4.

We will focus on the first sense of the word (the sense preceded by a bold capital letter 'A') which has 9 attestations, for some of which links are given. We will look at the TEI-XML source for the first three of these, see Fig. 5. The `<cit>` element is described as containing "a quotation

[11] Text Encoding Initiative, `https://tei-c.org/`

[12] `http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:1999.04.0057`

Figure 5: The TEI encoding for ἀνώμαλος

from some other document, together with a bibliographic reference to its source". Additionally, in dictionaries it "may contain an example text with at least one occurrence of the word form, used in the sense being described, or a translation of the headword, or an example"; this obviously fits the citation in the third attestation. Note that in each case the quotation itself is contained within a`<quote>` element and the bibliographic reference in the `<bibl>` element. In cases where there isn't a quotation, as in, for example, the fourth, fifth and sixth attestations in the entry, the `<bibl>` element has been used by itself. In fact there is no single mechanism for representing attestations in TEI since, depending on the particular feature content in a dictionary, and the practice of the project regarding bibliographic information, a number of different mechanisms can be used including: `<cit>`, `<bibl>`, `<ref>` as well as pointer attributes like `@source`.[13]

In the FrAC module, however, our proposal is to define a generic mechanism to model the fact that a given lexical phenomenon, i.e., a given word sense, form, subcategorisation and valency information, etc., described in a lexical resource is attested to by a text, and to distinguish this from other kinds of citations. Returning to the example given above, looking at the first sense we see the following:

- Instances of attestations for words both with and without associated quotations;
- Instances of attestations where the quotation contains the headword and others where it does not;
- An instance of an attestation where the text referred to is conjectural (it has been reconstructed and may or may not be accurate), marked by the Latin *cj.*;
- A citation (marked as 'cf.', an abbreviation for the Latin *confere* 'compare') which may not be an attestation of the sense in question.

In the following we will make some remarks on the OntoLex-FrAC encoding of the example in RDF; the whole example is available on the Github repository. Listing below presents the entry with frequency information which lists its frequency in a corpus, which in this case is composed of Strabo's *Geography*:

[13]Personal Communication, Jack Bowers.

```
:lsjEntry_ent_n10947 a ontolex:LexicalEntry;
   frac:frequency [
      a frac:CorpusFrequency;
      rdf:value "18"^^xsd:int
      frac:corpus
      <http://www.perseus.tufts.edu/hopper/text?
      doc=Perseus:text:1999.01.0197>] .
```

The first sense here is associated with 9 frac:Attestation resources:

```
:sense_n10947_0 a ontolex:LexicalSense ;
   frac:attestation :att_n10947_0_bib0,
      :att_n10947_0_bib1,
      :att_n10947_0_bib2,
      :att_n10947_0_bib3,
      :att_n10947_0_bib5,
      :att_n10947_0_bib6,
      :att_n10947_0_bib7,
      :att_n10947_0_bib8,
      :att_n10947_0_bib9 ;
ontolex:isSenseOf :lsjEntry_ent_n10947 .
```

The first attestation is encoded in RDF as follows:

```
:att_n10947_0_bib0 a frac:Attestation ;
   cito:hasCitedEntity :n10947_0_bib0 ;
   att:hasBiblScope "625d" ;
   att:attestationGloss
      "uneven, irregular "χωρα" Pl.Lg.625d".
```

Here we can see the use of a new datatype properties which complement the newly proposed FrAC properties. The first `hasBiblScope` is directly inspired by the corresponding TEI element `<biblScope>` which is defined as giving the "scope of a bibliographic reference". The listing also demonstrates the use of the FrAC property `attestationGloss` which gives the exact written text accompanying an attestation (this is important in the case of legacy and retrodigitized resources). We also use the property `hasCitedEntity` from the CITO vocabulary[14] (Peroni and Shotton, 2012) to link the attestation to a bibliographic record `:n10947_0_bib0` (the latter is described using the FRBR vocabulary). The second attestation is represented as follows in RDF with FrAC:

```
:att_n10947_0_bib2 a frac:Attestation ;
   cito:hasCitedEntity :n10947_0_bib2 ;
   att:hasBiblScope "7.71" ;
   frac:quotation "το α. τηςναυμαχιας" ;
   att:attestationGloss "'το α. τηςναυμαχιας'..." ;
   rdfs:seeAlso :cit_n10947_0_1, :cit_n10947_0_2;
   :conjectural 'True' .
```

Note the use of the `quotation` property here (since the quotation in the attestation gloss includes the word itself), as well as the use of `conjectural` here. We also use the rdfs property `seeAlso` to encode the two citations `cit_n10947_0_1` and `cit_n10947_0_2`.

4.2. Attestations in DiaMaNT

DiaMaNT (*Diachroon seMAntisch lexicon van de Nederlandse Taal*), is a diachronic semantic computational lexicon of Dutch, currently under development at the *Instituut voor de Nederlandse Taal* (Dutch Language Institute). This lexicon is the third component of the lexicographical infrastructure for historical Dutch, which is being developed at the Institute. The core of the infrastructure is formed by the four scholarly historical dictionaries of Dutch: the *Woordenboek der Nederlandsche Taal (WNT)* (Dictionary

[14]`https://sparontologies.github.io/cito/current/cito.html`

of the Dutch Language), the *Middelnederlandsch Woordenboek (MNW)* (Dictionary of Middle Dutch), the *Vroegmiddelnederlands Woordenboek (VMNW)* (Early Middle Dutch Dictionary) and the *Oudnederlands Woordenboek (ONW)* (Dictionary of Old Dutch). The four dictionaries cover a language period from ca. 500 – 1976.

The first component of this infrastructure is the historical dictionary portal. The portal gives online access to the dictionaries so that a user can look up the meaning of a word. The second component is the morphosyntactic lexicon GiGaNT, containing information on possible variation in spelling and form of historical Dutch language, by means of which searching in historical texts was made easier. The third component is the DiaMaNT lexicon. It forms a layer on top of GiGaNT. It aims to resolve the issue of historical semantic variation. The main purpose of this lexicon is to enhance text accessibility and to foster research in the development of concepts, by interrelating attested word forms and semantic units (concepts), and tracing semantic variation through time. Core of the DiaMaNT lexicon is on the one hand the senses of the dictionaries and on the other hand the attestations. The latter give information as to the time period a certain sense occurred in. A first Linked Open Data version (Depuydt and de Does, 2018) has been elaborated and published in the Dutch CLARIAH infrastructure. The DiaMaNT lexicon is also available at `http://diamant.ivdnt.org/diamant-ui/`. For an example of the use of the attestations, see Fig. 6.

An excerpt of the lexicon using the FrAC module to model attestations is presented in Listing 3.

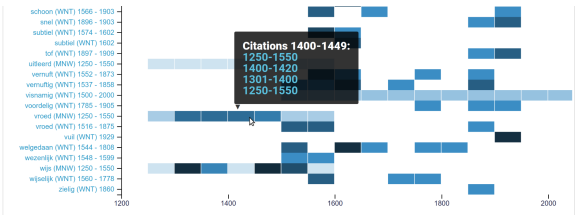


Figure 6: Application: chronology of synonyms; the DiaMaNT lexicon.

4.3. Attestations in the DOE Web Corpus

The Dictionary of Old English Web Corpus (DOEC) has been compiled for the Dictionary of Old English at the University of Toronto and consists of "at least one copy of every surviving Old English text" (diPaolo Healey et al., 2009), amounting to over 3 million written words.

Originally available as a set of TEI-XML files, the DOEC is currently accessible online as a Web corpus.

In this paper, we will illustrate modelling an attestation in DOEC of a lexical sense described in the Thesaurus of Old English (Roberts et al., 2000). The thesaurus provides an onomasiological ordering of the lexis that was available to speakers of Old English. This ordering allows users to traverse a hierarchy of meanings, described in present-day English, to Old English lexical items that express that meaning. This information has recently been transformed to Linguistic Linked Data (Stolk, 2019). This new form of the

Listing 3: Representation of attestations in the DiaMaNT lexicon

```
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 .
```

lexicographic work offers identifiers (or IRIs) for its concepts of meaning, its lexical entries, and its lexical senses. Thus, the single recorded sense of the entry *gēardagum* in TOE has its own IRI and is categorized under the concept named "Formerly, long ago".[15] Listing 5 shows an RDF sample of the entry, its sense, and the concept that expresses its meaning.

The lexical sense of *gēardagum* in TOE is attested in a number of Old English texts, including the poem *Beowulf*. In fact, its first occurrence is in the second line of the single surviving copy of the poem. Listing 4 shows that very occurrence, in bold, as it is presented in the DOEC.

The URL that provides access to the information above, is the following: `https://tapor.library.utoronto.ca/doecorpus/cgi-bin/oec-idx?type=bigger&byte=982592&q1=geardagum` . This Web address includes information on the type of visualization (i.e., 'type=bigger'), the location of the current corpus reference (i.e., 'byte=982592') and the query string to highlight using a bold font (i.e., 'geardagum'). The type of visualization, as can be seen in the snippet, includes a small context surrounding the currently selected token in the corpus. The three lines are preceded by sentence numbering in *Beowulf* (i.e., 0001, 0002, and 0003 respectively) and the line number on which the given sentence starts in the manuscript (i.e., 1, 1, and 4 respectively).

Rather than duplicating all the information from DOEC on

[15] Information on this lexical sense in the linguistic linked data form of A Thesaurus of Old English has been made available on the digital platform Evoke: `http://evoke.ullet.net/app/#/view?source=toe&iri=http://oldenglishthesaurus.arts.gla.ac.uk/sense/%23id%3D21808`.

Listing 4: Snippet from DOEC on *geardagum* in the first lines of the Old English poem *Beowulf*

```
[0001 (1)] Hwæt.
[0002 (1)] We Gardena in geardagum, þeodcyninga, þrym gefrunon, hu ða æþelingas ellen fremedon.
[0003 (4)] Oft Scyld Scefing <sceaþena> þreatum, monegum mægþum, meodosetla ofteah, egsode eorlas.
```

Listing 5: RDF sample of TOE as linguistic linked data

```
@base <http://oldenglishthesaurus.arts.gla.ac.uk/> .

<entry/#id=21808> a ontolex:LexicalEntry ;
   rdfs:label "gēardagum"@ang ;
   ontolex:canonicalForm [
      ontolex:writtenRep "gēardagum"@ang
   ] ;
   ontolex:sense <sense/#id=21808> .

<sense/#id=21808> a ontolex:LexicalSense ;
   ontolex:isLexicalizedSenseOf <category/#id=9880> .

<category/#id=9880> a ontolex:LexicalConcept ;
   skos:prefLabel "Formerly, long ago"@eng .
```

Listing 6: RDF representing the attestation in DOEC of the lexical sense of *gēardagum* from TOE

```
@base <http://oldenglishthesaurus.arts.gla.ac.uk/> .

ex:attestation412 a oa:Annotation ;
   oa:motivation oa:identifying ;
   oa:hasBody <sense/#id=21808> ;
   oa:hasTarget [
      # the source corpus is DOEC
      oa:hasSource
         <https://tapor.library.utoronto.ca/doecorpus/> ;
      # for selecting the entire sentence in DOEC
      oa:hasSelector [
         a oa:TextPositionSelector ;
         oa:start 982592 ;
         oa:end 982708 ;
      ] ;
      # for selecting the exact token in DOEC
      oa:hasSelector [
         a oa:TextPositionSelector ;
         oa:start 982607 ;
         oa:end 982616 ; ] ; ] .
```

the context of the particular attestation of *gēardagum* in the thesaurus, it would be more valuable to link that information to the relevant lexical sense in the thesaurus instead. Doing so will enable users from either resource to benefit from the complementary information provided by the other resource. Moreover, an additional advantage is that no licensing rights are violated in this manner: links between the two sources would simply refer to them without redistributing their content. Those who have the right to access the material can simply follow these links (from one resource to another) or query them integrally if they also have the means to do so. In this specific case, links such as the one proposed will allow for further examinations of both the accuracy of the definitions in the lexicographic resource and the aspects of, for instance, the distribution and frequency of specific senses as found in a body of texts. Thus, lexicographers and corpus linguists can benefit from these connections.

One of the approaches explored with the FrAC module for modelling attestations in corpora (most notably online corpora) is to use the standardized Web Annotation vocabulary. This vocabulary, published in 2017, was developed by W3C. The vocabulary offers terminology to indicate a selection that one wishes to annotate. For the current case, we use a TextPositionSelector to indicate the start and end of our selection within the entire corpus of DOEC. For the sentence in which *gēardagum* occurs, this selection would start at 982592 (i.e., the value embedded as 'byte' in the URL for the DOEC snippet above) and end at 982708. If we were to select solely the token, however, the selection should start 15 characters (or bytes) later and be 9 characters (or bytes) long.

Thus, the selection would start at 982607 and end at 982616. Listing 6 shows the resulting RDF for both options. The body of the annotation is the lexical sense from TOE; its target is the selection of the token (or its sentence) in DOEC. The motivation for the annotation is one of 'identifying', indicating that the lexical sense offers details on the identity of the selection. Selecting the token in DOEC only, rather than its entire sentence, is preferable since it allows for fine-grained analyses. Additionally, feeding this more accurate starting position to the DOEC interface (i.e., embedding it as 'byte' in the URL) does not pose any issues: The website of the online corpus still presents the user with an appropriate context for this more accurate selection. In conclusion, the use-case of DOEC shows that the Web Annotation vocabulary provides enough expressivity to capture attestations in corpora.

5. Conclusion

In this paper, we introduced the OntoLex-FrAC vocabulary, an OntoLex extension for representation of frequency, attestation and corpus information for the needs of digital lexicography, natural language processing and corpus linguistics. We described its structure, some selected use cases, elementary concepts and fundamental definitions, with a specific focus on frequency and attestations.

The main goal of the paper is to document the progress achieved so far, and even more importantly, to elicit feedback from the language resource community.

The next step is to reach a consensus for representing additional corpus information such as collocations and similarity scores. Another important direction is to apply the model on a larger scale to further test its applicability.

Acknowledgements

This paper is supported by the Horizon 2020 research and innovation programme with the projects Prêt-à-LLOD (grant agreement no. 825182) and ELEXIS (grant agreement no. 731015). It is also partially based upon work from COST Action CA18209 - NexusLinguarum "European network for Web-centred linguistic data science". We also thank the anonymous reviewers for their helpful comments.

6. References

Cimiano, P., McCrae, J. P., and Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report.

Declerck, T., McCrae, J., Hartung, M., Gracia, J., Chiarcos, C., Montiel, E., Cimiano, P., Revenko, A., Sauri, R., Lee, D., Racioppa, S., Nasir, J., Orlikowski, M., Lanau-Coronas, M., Fäth, C., Rico, M., Elahi, M. F., Khvalchik, M., Gonzalez, M., and Cooney, K. (2020). Recent developments for the linguistic linked open data infrastructure. In Nicoletta Calzolari, et al., editors, *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*. ELRA, ELRA, 5.

Depuydt, K. and de Does, J. (2018). The Diachronic Semantic Lexicon of Dutch as Linked Open Data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May. European Language Resources Association (ELRA).

diPaolo Healey, A., Wilkin, J. P., and Xiang, X. (2009). Dictionary of Old English Web Corpus.

Khan, A. F. and Boschetti, F. (2018). Towards a Representation of Citations in Linked Data Lexical Resources. In Jaka Čibej, et al., editors, *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 137–147, Ljubljana, Slovenia, July. Ljubljana University Press, Faculty of Arts.

Kilgarriff, A. (1997). "I Don't Believe in Word Senses". *Computers and the Humanities*, 31(2):91–113.

Krek, S., McCrae, J., Kosem, I., Wissek, T., Tiberius, C., Navigli, R., and Sandford Pedersen, B. (2019). European lexicographic infrastructure (elexis). In *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts*, pages 881–892.

Kučera, H. and Francis, W. N. (1967). *Computational Analysis of Present-day American English*. Brown University Press.

Liddell, H., Scott, R., and Jones, H. S. (1996). *A Greek-English lexicon*. Oxford University Press, 9th edition.

McCrae, J., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6):701–709.

Peroni, S. and Shotton, D. (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.

Roberts, J., Kay, C., and Grundy, L. (2000). *A Thesaurus of Old English: In Two Volumes*. Rodopi.

Saur, K. (1998). Ifla study group on the functional requirements for bibliographic records. functional requirements for bibliographic records: final report.

Stolk, S. (2019). A Thesaurus of Old English as linguistic linked data: Using OntoLex, SKOS and lemon-tree to bring topical thesauri to the Semantic Web. In *Proceedings of the eLex 2019 conference*, pages 223–247, oct.