Data model for citations in dictionaries

Katrien Depuydt, Jesse de Does

Abstract

In a dictionary article, lexicographers cite other resources to substantiate their analysis. These citations can be found everywhere in the dictionary, eg. in the etymological section of a dictionary article, in the definition of an entry etc.

A special form of citations are those that function as examples to give evidence, elucidate meaning or illustrate features (spelling variation, syntax, collocation, register (etc)). These citations are a representative selection of occurrences of the headword from either a corpus of texts or a corpus of snippets from texts. These citations, i.e. attestations, usually occur in the form of a quote with bibliographical information. In some cases, only a few context words are given, not the actual attested word form, in other cases, only the bibliographical reference is given.

Attestations are attributed examples in a dictionary. There are usage examples that are not attributed. They are either invented or based on corpus material, but adapted/simplified by the lexicographer. We will focus on attestations.

1 Introduction

Lexicographers use examples to support their analysis of the headword. The examples can either be authentic (exact quotations), adapted (modified versions of authentic examples) or invented examples ([**?**], p. 283-285). Authentic examples are attributed quotations (citations), which not only elucidate meaning and illustrate features of the headword (spelling, syntax, collocation, register etc.), but also function as attestations and are used provide evidence of the existence of a headword ([**?**], p. 453-458). We therefore call these examples "attestations". Adapted examples, (cf˙ a.o. [**?**], p. 251) or invented examples, which often occur in learner's dictionaries and bilingual dictionaries ([**?**]), will not be discussed here.

2 CELEX

The CELEX database has been published first in 1990. It contained extensive descriptions of word forms and lemma's of three languages: Dutch, English and German. It has been a major lexical source for many years. The data is distributed through a number of agencies, like ELRA1. Moreover, it was accessible through a web interface hosted at the Max Planck Institute for Psycholinguistics2. In 2001 the CELEX-Project was halted because of financial reasons. In 2015 the web interface was transferred to the *Instituut voor de Nederlandse Taal*. The data was cleansed to some extent. That version of the data has been the basis for the present work.

The main elements of the CELEX lexicon are lemma's and word forms. The word forms are linked to the corresponding lemma's.

```
:celex_lemma_47843 a ontolex:LexicalEntry .
:celex_lemma_47843 a ontolex:Word .
:celex_lemma_47843 ontolex:canonicalForm :celex_wform_147510; rdfs:label "kattekwaad"@nl .
:celex_lemma_54132 a ud:NOUN .
:celex_lemma_54132 ud:Gender ud:Neut_Gender .
:celex_lemma_47843 ontolex:lexicalForm :celex_wform_147510 .

:celex_wform_147510 a ontolex:Form .
:celex_wform_147510 ontolex:writtenRep "kattekwaad"@nl .
:celex_wform_147510 intskos:syllabified "kat-te-kwaad"@nl .
:celex_wform_147510 ontolex:phoneticRep "k-t-kwat"@nl-fonipa .
:celex_wform_147510 lexinfo:number lexinfo:singular .
```

```
:celex_lemma_108599 decomp:constituent :celex_comp_92551; rdf:_1 :celex_comp_92551; decomp:constituer
:celex_comp_92553 a decomp:Component, ud:NOUN; gold:stem [ontolex:writtenRep "riem"@nl] .
:celex_comp_92553 a decomp:Component; lexinfo:partOfSpeech lexinfo:Noun; gold:stem [ontolex:writtenR
:celex_comp_92554 a decomp:Component; decomp:correspondsTo :celex_lemma_124345 .
:celex_comp_92555 a decomp:Component; decomp:correspondsTo :celex_lemma_54132 .

:celex_lemma_124345 a ontolex:Affix .
:celex_lemma_124345 rdfs:label "e"@nl .

:celex_lemma_54132 a ontolex:LexicalEntry .
:celex_lemma_54132 a ontolex:Word .
:celex_lemma_54132 ontolex:canonicalForm :celex_wform_165612; rdfs:label "kwaad"@nl .
:celex_lemma_54132 a ud:NOUN .
:celex_lemma_54132 ud:Gender ud:Neut_Gender .
ud:NOUN a ud:PartOfSpeech .
:celex_lemma_54132 ontolex:lexicalForm :celex_wform_165612 .
:celex_lemma_54132 ontolex:lexicalForm :celex_wform_165714 .

:celex_wform_165612 a ontolex:Form .
:celex_wform_165612 ontolex:writtenRep "kwaad"@nl .
:celex_wform_165612 intskos:syllabified "kwaad"@nl .
:celex_wform_165612 ontolex:phoneticRep "kwat"@nl-fonipa .
:celex_wform_165612 lexinfo:number lexinfo:singular .

:celex_wform_165714 a ontolex:Form .
:celex_wform_165714 ontolex:writtenRep "kwaden"@nl .
:celex_wform_165714 intskos:syllabified "kwa-den"@nl .
:celex_wform_165714 ontolex:phoneticRep "kwa-d"@nl-fonipa .
:celex_wform_165714 lexinfo:number lexinfo:plural .
```

```
:aap a diamant:Beast .
:aap diamant:eats diamant:Nuts .
:Nuts diamant:kindof diamant:Food .
```

3 Citations and attestations in scholarly dictionaries

In "scholarly dictionaries" (OED, ANW, the major historical dictionaries of Dutch[1], TLFI, Lewis and Short, Liddell and Scott, Böhtlingk and Roth, Hanyu Da Cidian, ...), the main role of citations and quotations is

1. To illustrate usage
2. To provide evidence for the lexicographical interpretation of a word sense ("attestation")

As stated by Hawke, [?] p. 176, "The quotation evidence is the bedrock of any historical dictionary. The relationship between the definitions of each sense in a historical dictionary and the quotations that accompany it is particulary close. In the compilation of a historical dictionary (at least in modern times) the quotation evidence provides a sample of the empirical data on which the definitions have been based"

Thus, it is important to be explicit about the nature of the evidence provided. The dictionary should provide information on the reliability of attestations; this information should also be represented in the knowledge base.

3.1 Reliability of attestations

An attestation is as good as the corpus text or quotation the attestation is taken from. The analysis of the lexicographer as such can also be subject to discussion.

[1] http://gtb.ivdnt.org

Figure 1: Celex kattekwaad example

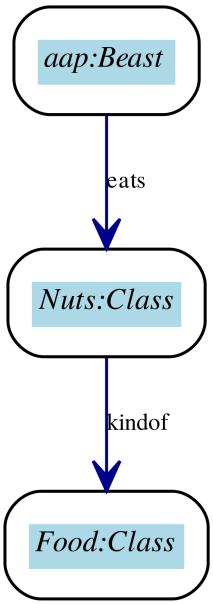

Figure 2: Apen eten noten

Uncertain reading or reconstruction

For historical dictionaries, the source material consists of text editions of historical documents. The further we go back in time, the more challenging it is to find enough material. Sometimes, sources have come down to us in one manuscript or a few manuscript fragments, with parts that are hard to read, or only partially legible (think of a hole in a manuscript). Or the word in question is misspelled.

The impact of this on the lexicographer's work is, that some of the "evidence" in a quotation is in fact a reconstruction[2].

Some examples In the Dictionary of Old Dutch, the entry with the Dutch adverb *fora* has related entries, one of which refers to a placename *foraskōta*. The editor suggests a correction of a spelling in the third quotation.

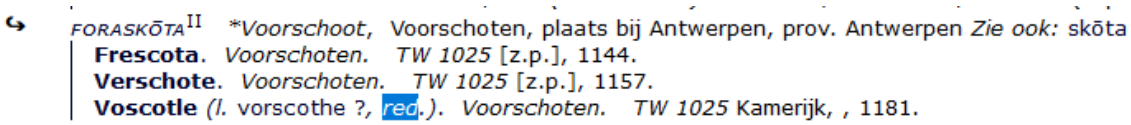
FORASKŌTA[II] *Voorschoot, Voorschoten, plaats bij Antwerpen, prov. Antwerpen *Zie ook:* skōta
Frescota. *Voorschoten.* *TW 1025* [z.p.], 1144.
Verschote. *Voorschoten.* *TW 1025* [z.p.], 1157.
Voscotle *(l.* vorscothe ?, red.*)*. *Voorschoten.* *TW 1025* Kamerijk, , 1181.

Figure 3: example: Spelling correction proposed by editor

Another example is from the Dictionary of Early Middle Dutch, where the fourth attestation has an occurrence of the verb *staen* (to stand), which is partly a reconstruction: s[tut].

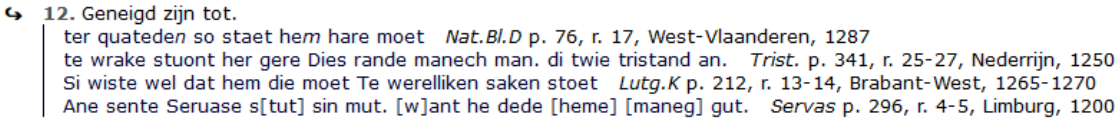
12. Geneigd zijn tot.
ter quateden so staet hem hare moet *Nat.Bl.D* p. 76, r. 17, West-Vlaanderen, 1287
te wrake stuont her gere Dies rande manech man. di twie tristand an. *Trist.* p. 341, r. 25-27, Nederrijn, 1250
Si wiste wel dat hem die moet Te werelliken saken stoet *Lutg.K* p. 212, r. 13-14, Brabant-West, 1265-1270
Ane sente Seruase s[tut] sin mut. [w]ant he dede [heme] [maneg] gut. *Servas* p. 296, r. 4-5, Limburg, 1200

Figure 4: example: reconstructed reading

[2]Another matter, which will briefly discussed in section 3.4, is to be explicit about what the attestation actually provides evidence for in terms of the temporal and regional distribution of the phenomenon attested. For this purpose, the metadata provided with the quotations should accurately represent the nature of the source text from which the quotation has been taken.

These corrections or reconstructions are recognizable. According to the lexicographer, they are valid examples of the meaning of the entry. However, as representative of the spelling, they are not so reliable.

Uncertain interpretation

Sometimes, there may not be any doubt about the reading of a quoted passage of a source text, but the lexicographer is uncertain about his/her interpretation. In this example, the lexicographer gives a potential "misschien ook (maybe also)" other interpretation of an attestation.

In onderstaande aanh. is behalve een acc.pl. ***misschien ook*** een interpretatie als znw.v. met als bet. 'bank(instelling)' mogelijk (vgl. MNW IV, 746-749).
Oec mach die greue van vyanen lombarde houden binnen sinen sonderlinghen lande, *Corp.I* p. 2470, r. 20-21, Grimbergen, Brabant-West, 1298

Figure 5: example: doubt about interpretation

3.2 Citations as evidence: different types

There are several ways in which a citation as evidence can occur (list might not be exhaustive).

1. A bibliographical reference is given
2. The attested wordform is quoted in context + the bibliographical reference is given
3. Some context words are given, accompanied by a bibliographical reference.

Example of a and b In the Dictionary of the Dutch language, the quotation section begins with a listing of the dictionaries that describe the entry in the same meaning, followed by a group of quotations + bibliographical information with attestations (bold) of the entry in this quotation

2. (Bedr. en onz.) De feiten, de juiste toedracht, de waarheid kennen omtrent wat in het object of de voorzetselbep. wordt genoemd of bevraagd.
   a. Ter aanduiding van het bezitten van feitelijke kennis.
   ↪ α. Met een objectszin, ingeleid door een vragend vnw. of bijw.
   V. BERLAINMONT E iij v° *a* [1536].
   LAMBRECHT, *Naembouck* [1546].
   PLANT. [1573].
   KIL. [1588].
   HEXHAM [1648].
   V. DALE [1872→].
   — Doe gingen de kinderen Juda tot Josua ... ende Chaleb ... sprac tot hem Ghi **weet** wat die Here tot Moyse den man Gods seyde van minen ende uwen wege in Kades Barnea, *Bijbel v. Liesveldt, Jozua* 14 B [1526].
   Noch langt my. en eewelijck sal om **weten** Wie sij waren wt goeder dueghden gront, J. V.D. DALE 173 [1528].
   Zouden zij **gheweten** hebben wat daer inne was ende bevindende dattet ghelt was, hebbent tot mijns heere huuse ghebrocht, V. VAERNEWIJCK, *Ber. T.* 1, 174 [1566].
   Abraham berispte Abimelech ter oorsake eenes water-puts, die Abimelechs knechten met gewelt genomen hadden. Doe seyde Abimelech; Ick en hebbe niet **geweten,** wie dit stuck gedaen heeft, *Statenb., Gen.* 21, 26 [ed. 1637].
   Alzoo (*t.w. door middel van een paspoort*) kan men **weeten** bijna, waar ieder zich ophoudt: alzoo is het wijduitgestrekt Rusland, voor de meerderheid zijner bewoners, eene zoort van gevangenisse, V. WOENSEL, *Rusl.* 193 [1804].
   En kijk hare dochteren maar eens aan, als gij **weten** wilt, hoe de deftige matrone er in den bloei van vrolijk- en jolijkheid uitzag, POTGIETER 1, 6 [1841].
   Ik **weet** nog niet waer de ongelukkige Geronimo verbleven is, CONSC., *S. Turchi* 2, 32 [1859].

An example of a and c In the Liddell Scott Jones lexicon, bibliographical references are given, or a description of the type of context the entry word was found in.

ἀνώμα^λ-ος , ον, (ἀ- priv., ὁμαλός)

A.*uneven, irregular,* "χώρα" **Pl.Lg.625d**; "φύσις" **Id.Ti.58a**; "τὸ 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,* "ὁμαλῶς 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*.

3.3 Linked data modeling of lexical citations

The concept of attestations is discussed in two recent publications. One is the publication on the model for the DiaMaNT lexicon ([?]). The most recent publication is the one by Khan and Boschetti [?]. This article on lexical attestations prompted us to re-evaluate the proposal for attestations in Depuydt en Does 2018. We will discuss briefly the model by [?] and will then come up with an alternative proposal.

Khan and Boschetti

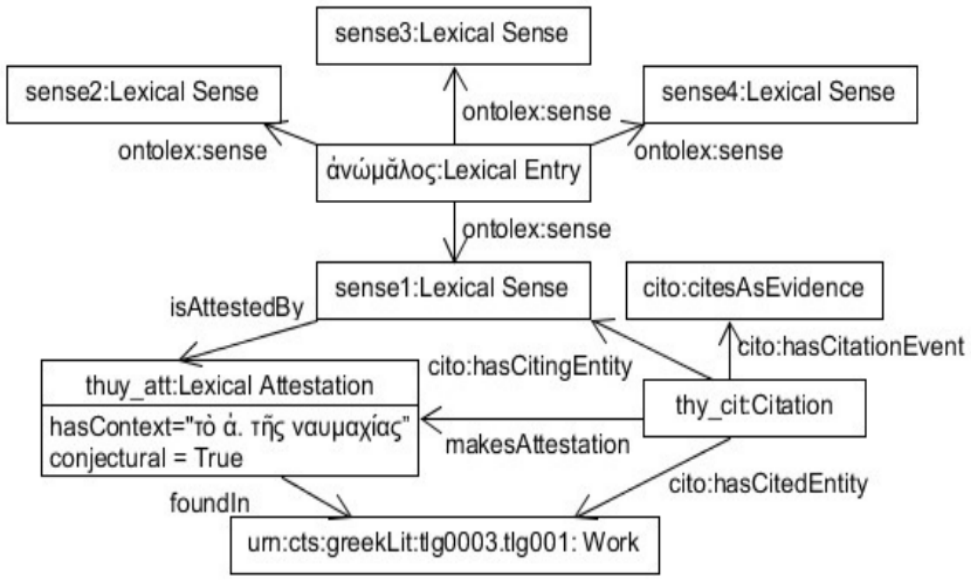

Figure 6: Example modeling from [?]

Khan and Boschetti's *lemonBib* model for lexiographical citations tackles some issues important for historical lexicography and proposes solutions based on the FRBR, Cito and Fabio ontologies[?]

- Distinction between citation in general and a citation which provides evidence for a certain lexical object (*Attestation*) (Word sense, word form, ...)
- Enabling the marking of text readings as conjectural
- The distinction between a "work" and its "Manifestation" in the form of a publication.

In our view, the resulting model[3] has some drawbacks.

[3]http://lari-datasets.ilc.cnr.it/lemonBib

- In the K&B model, the presence of context entails attestation (the *hasContext* data property has domain *Attestation*). The model does not take into account citations with context snippets that are *not* attestations, eg. citations in a section on etymology, where an authority is quoted to back up or contradict an analysis.
- There appears to be an unneccessary amount of relation-reifying objects in the model.

In the simplest application of the Cito model, "cites" is essentially a relation between two resources representing publications, without any intermediate reifying objects. Ported to the lexical domain, this would consist of a relation between a lexical sense and a work containing an occurrence of this sense. Since we want to attach some characterizations to the citation, some degree of reification is necessary, for which purpose Cito proposes the Citation class. K& B proceed by modeling Citation and Attestation as different resources with a relation *attestationCitation/makesAttestation* between them. We do not see why one could not declare Attestation as a subclass of Citation and characterize them, for instance, by (e.g.)

- Attestation is a subclass of Citation
- Attestation is a subclass of (hasCitationCharacterization cito:citesAsEvidence)

Alternative proposal

In Depuydt & De Does 2018, we proposed a model for attestations in which all dictionary quotations with context that illustrate word senses are also attestations (which reflects the reality of our dictionaries). The "Attestation" object proposed there included a pointer to the location of the headword in the quotation. We considered this useful because

- It allows attestations of word forms (so users can related specific word forms to document metadata, e.g. period or location)
- It allows the immediate use of dictionary quotations for computational applications like WSD

After reading Khan and Boschetti's paper we realized it was important to take into account that the presence of context and the fact that the quotation provides evidence for the word sense are distinct dimensions.

In fact, a model should be able to deal with the following situations:

1. attestation, no doubt about reading or interpretation
   (a) only mentioning the source
   (b) source with a context [with or without keyword in it]
   (c) acknowledgment of the source with context and pointer
2. attestation, uncertain reading
   (a) only source indication, is attestation, uncertain reading
   (b) source with context [whether or not with keyword in it], is attestation, uncertain reading
   (c) source with context and pointer, is attestation, uncertain reading
3. attestation, uncertain interpretation
   (a) only source indication, is attestation, uncertain interpretation
   (b) source with context [with or without keyword in it], is attestation, uncertain interpretation
   (c) source indication with context and pointer, attestation,uncertain interpretation
4. not an attestation, different type of quotation
   (a) only source indication, is not an attestation
   (b) source mention with context [with or without keyword in it], is not an attestation
   (c) [theoretically] source with context and pointer, is not an attestation

Classes
lexcit:Citation $\sqsubseteq$ cito:Citation *(a lexcit Citation is also a cito:Citation)*
lexcit:Attestation $\sqsubseteq$ lexcit:Citation
lexcit:Attestation $\sqsubseteq \exists$ cito:hasCitationCharacterization . cito:citesAsEvidence
(an Attestation has citation characterization citesAsEvidence)
lexcit:Locus $\sqsubseteq (\exists$ nif.beginIndex.$\top) \sqcap (\exists$ nif.endIndex.$\top) \sqcap (\exists$ lexcit.locusIn.$(\exists$ lexcit:quotation.$\top)))$
(a Locus has a begin and end index and points to something which has a quotation)
(ontolex:Form $\sqcup$ ontolex:LexicalSense) $\sqsubseteq$ lexcit:LexicalPhenomenon
Data properties
lexcit:quotation $\sqsubseteq$ lexcit:Citation $\times$ xs:String *(domain is Citation, range is string)*
lexcit:readingCertain $\sqsubseteq$ lexcit:Citation $\times$ xs:Boolean
lexcit:interpretationCertain $\sqsubseteq$ lexcit:Citation $\times$ xs:Boolean
nif:beginIndex $\sqsubseteq$ lexcit:Locus $\times$ xs:Integer
nif:endIndex $\sqsubseteq$ lexcit:Locus $\times$ xs:Integer
Object properties
lexcit:citation $\sqsubseteq$ lexcit:LexicalPhenomenon $\times$ lexcit:Citation
lexcit:citation $\sqsubseteq$ cito:hasCitingEntity$^-$ *(citation is subset of the converse of cito hasCitingEntity)*
lexcit:attestation $\sqsubseteq$ lexcit:citation
lexcit:attestation $\sqsubseteq$ lexcit:LexicalPhenomenon $\times$ lexcit:Attestation
lexcit:locus $\sqsubseteq$ lexcit:LexicalPhenomenon $\times$ lexcit:Locus
lexcit:locusIn $\sqsubseteq$ lexcit:Locus $\times (\exists$ lexcit:quotation.$\top)$

Figure 7: Simple data model for quotations

So there are at least five dimensions:

1. Attestation (Citation provides evidence for the word sense) $\leftrightarrow$ other type of citation
2. Certainty of the reading of the source text (is the word really there?)
3. Certainty of the interpretation (is this really an instance of the relevant word sense?)
4. Is a context snippet given?
5. Is the occurrence (or multiple occurrences) of the headword in the context/snippet explicitly marked?

The simple model of figure 7 (for readability, written in a loose description logic augmented with cartesian product) tries to provide the necessary degrees of freedom.

- There always is a "Citation" object for any type of citation. it is always linked to the lexical sense (or other "lexical phenomena") with the "citation" object property. The type of citation (cites as evidence, agrees with, etc, cf the CITO ontology[4]) can be reflected in the value of cito:hasCitationCharacterization property and by subclassing 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 snippet data property.
- The "locus" object 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.).

The examples (figure 8), illustrate an attestation of a word sense, and an example of an attestation of both a word form and a sense.

3.4 Document metadata

When dealing with historical text, it is important to distinguish:

- The "text" as conceived by the author(s)

[4] types of citation from https://sparontologies.github.io/cito/current/cito.html: agrees withop, citationop, cites as authorityop, cites as data sourceop, cites as evidenceop, cites as metadata documentop, cites as potential solutionop, cites as recommended readingop, cites as relatedop, cites as source documentop, cites for informationop, compilesop, confirmsop, contains assertion fromop, correctsop, creditsop, critiquesop, derideop, describesop, disagrees withop, discussesop, disputesop, documentsop, extendsop, includes excerpt fromop, includes quotation fromop, links toop, obtains background fromop, obtains support fromop, parodiesop, plagiarizesop, qualifiesop, refutesop, replies toop, retractsop, reviewsop, ridiculesop, speculates onop, supportsop, updatesop, uses conclusions fromop, uses data fromop, uses method inop

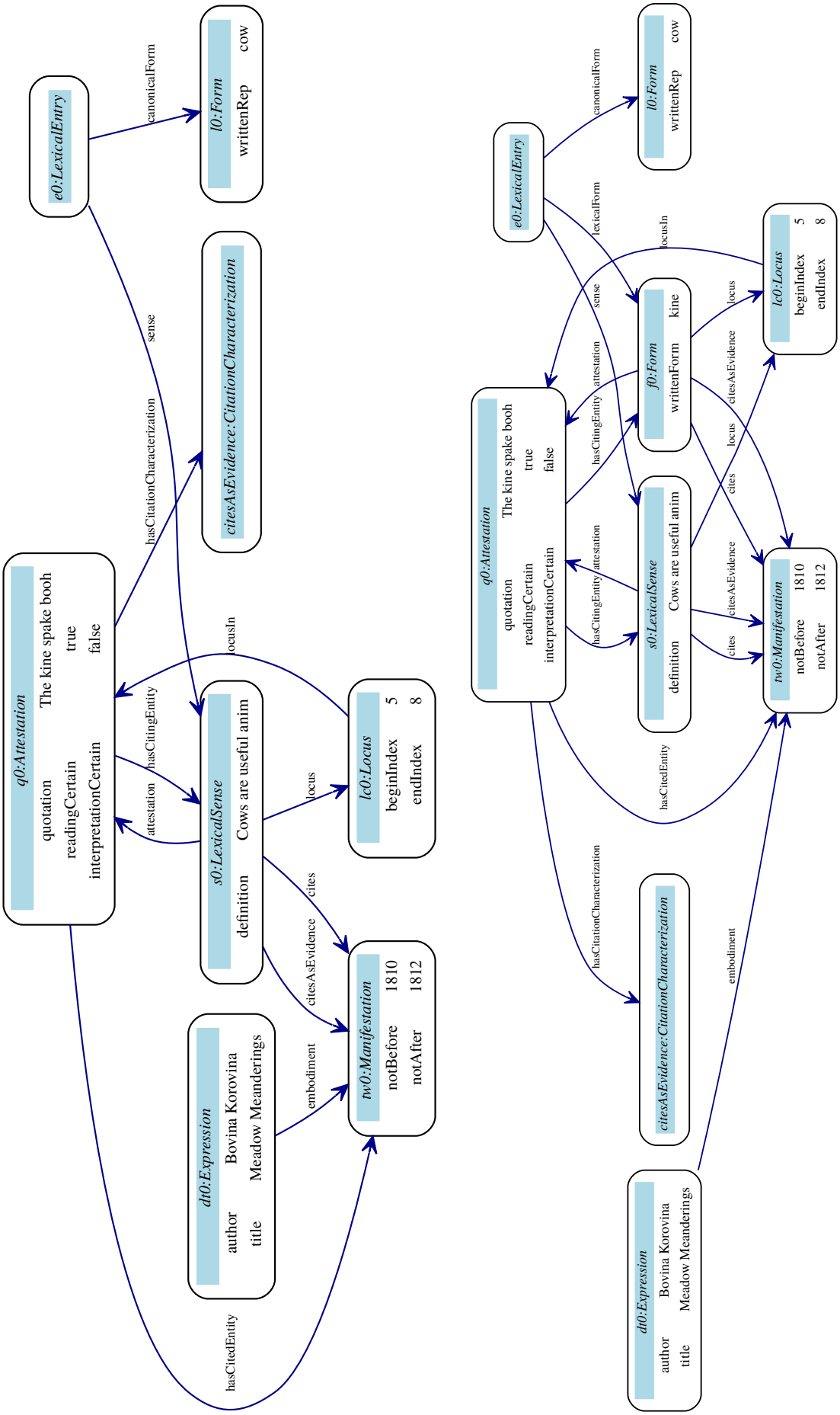

Figure 8: Attestation of a sense (left); attestation of a sense *and* a form (right)

– The "text witness" (the manuscript in which the text has come down to us)
– The edition which gives a representation of the text witness.

Each of these has its own metadata. For scholarly dictionaries, the text witness is important, because it determines the "date" and the regional characteristics of the language. It is not uncommon for e.g. medieval texts that the manuscript that contains the text is dated much later than the time the text was written, and may be representative of the dialect of the copyist rather than the native dialect of the author. The language in these cases is never exactly the same, therefore the date witness is the most important. As manuscripts are usually quoted from text editions, bibliographical information of the edition the text is quoted from, is important provenance information.

To model these aspects, (as pointed out by K&B), the FRBR/Fabio distinction between Works, Expressions, Manifestations and Items indeed provides us with the necessary vocabulary.

We do not need the nonlinguistic "frbr:Work" , since there are no non-linguistic data to attest in dictionaries. So the (possibly unavailable) text as conceived by the author (or translator) corresponds to "frbr:Expression" and each text witness can be considered a Manifestation.

However, both ontologies provide no means to say that the expression of the text is manifested in a manuscript which is in its turn manifested in a text edition.

A further typical characteristic of dictionaries is, that the bibliographical information accompanying a quotation, is a short reference (author, title and data witness). The full bibliographical description of the resource is given separately.

Minimal metadata information

Arriving at a proposal for a common standard for attestation metadata which includes all aspects touched upon in the previous subsection is not feasible in the time frame for this document. We propose the following simple minimal model, which includes author/title/date metadata but no proposal for localization yet.

Classes
lexcit:Document $\sqsubseteq$ frbr:Expression
lexcit:Witness $\sqsubseteq$ frbr:Manifestation
Data properties
dc:title $\sqsubseteq$ lexcit:Document $\times$ xs:String
dc:creator $\sqsubseteq$ lexcit:Document $\times$ xs:String
lexcit:notAfter $\sqsubseteq$ lexcit:Witness $\times$ xs:dateTime
lexcit:notAfter $\sqsubseteq$ lexcit:Witness $\times$ xs:dateTime
Object properties
frbr:embodiment $\sqsubseteq$ frbr:Expression $\times$ frbr:Manifestation