Towards a Representation of Citations in Linked Data Lexical Resources

Anas Fahad Khan, Federico Boschetti
Istituto di Linguistica Computazionale "Zampolli", CNR, Pisa
E-mail: fahad.khan@ilc.cnr.it, federico.boschetti@ilc.cnr.it

Abstract

In this article we look at the modelling of citations in lexical resources in linked data. We start by discussing the treatment of citations in linked data and in TEI; we also look at the idea of different conceptual levels as posited by models such as TEI and FRBR. We argue that in representing citations in lexical resources it is important not to confuse different levels of information, and that at least in the case of attestations it is important to model the purpose of a citation, or the claim that is being made by that citation, separately. We develop this point with two separate examples before presenting *lemonBib*, our extension of the lemon model based around the idea of a lexical attestation. We also give a treatment of part of one of the examples described previously in the article.

Keywords: linked data, citations, bibliographic data, lexical resources

1 Introduction

Up until quite recently most lexical resources published as linked (open) data have tended to be born-digital (having been developed in many cases with specific NLP tasks in mind), lately, however, there has begun to be an increased interest in the provision of retrodigitized lexical resources on the Semantic Web, and especially of those legacy resources regarded as authoritative or which are thought to hold some particular historical interest. Publishing such works as linked open data has the obvious advantage of making the information contained in them much more accessible and available to a wider public than was previously possible. At the same time, that information is structured according to a common data framework, RDF, which makes individual resources more interoperable as well as more amenable to various kinds of automated or semi-automated processing, more so than if they were text files, say. In addition, thanks to the fact that the Semantic Web offers data modelers a simple, standardized way of creating links between individual datasets, it also makes it easier to enrich an original lexical resource with links to other datasets such as, say, biographical or geographical ones. What's more, the Semantic Web offers modelers the possibility of rendering the links between resources meaningful by giving them an explicit, formal 'semantics', and thus clarifying the ways in which individual datasets can help to augment the knowledge contained in other datasets. The process of converting or migrating retrodigitized lexicographic resources into RDF brings to the fore a number of different modeling challenges that concern aspects of lexicons that are usually less prominent in born-digital, NLP-oriented lexical resources. One such challenge is that of the correct modelling of lexical attestations: that is, of citations used to attest to different properties of individual lexical entries. For instance, in the case when a given lexical entry cites a particular text as exemplifying the use of a word with, say, a given sense or given orthography, it would be useful to be able to link to that text in the linked data version of the entry, and to information about the work and the author, and perhaps also to the secondary literature; the Semantic Web seems to be particularly apt in cases such as these. Clearly we would like to be able to represent as much of the information contained in the

original lexical entry as possible using the graph-based data framework of RDF, but it is also crucial (given the formal, 'semantic' nature of the Semantic Web) that we respect the conceptual differences between the kinds of information present in a citational act, and this calls for a more detailed and specific treatment. At the end of the day, the fact that a lexicographer or group of lexicographers decided, during the compilation of a lexicographic work, to attest to the existence of the property of a word by citing a relevant text is a salient piece of data, and one that it is worthwhile trying to model properly (even if this kind of information hasn't featured as strongly in previous lexical linked datasets).

In this article we take a detailed look at a number of issues which arise when it comes to modelling lexical citations as linked data; we will look at examples taken from retrodigitized lexicographic resources or that concern linked data versions of print resources as contexts in which certain of these issues become much more conspicuous (although of course they also apply to born-digital resources). We will work towards a provisional set of properties and classes that, together with already existing RDF vocabularies, will help to capture some pertinent aspects of citations and attestations in lexicons. On the way we will discuss some of the pre-existing vocabularies and models for representing this kind of information with a view to their adequacy to the case at hand.

2 Background

In this section we will discuss related work that deals with the representation of bibliographic records and citations, both in the specialised case of computational lexicons, as well as within the general framework of linked data resource. We will begin, in Section 2.1, by looking at a useful distinction that is made within the TEI model between different ways of viewing lexical datasets (and which will be relevant for the discussion which follows in the rest of the paper) before moving on to describe the TEI approach to representing lexical citations in detail. In Section 2.2 we give a brief overview the influential FRBR model which has had an important impact on the representation of citations in linked data, as well as in the field of library science more generally. A more detailed discussion of citations in linked data is given in Section 2.3.

2.1 TEI: Zero, One, Two Dimensional Views and Citations

The Text Encoding Initiative (TEI) refers both to a widely used standard for encoding digital texts in XML, as well as to the consortium that maintains and develops the standard[1]. TEI, the standard, is available as a set of guidelines[2] (Burnard & Bauman 2008) which are used to define an XML schema. These guidelines are divided up into several parts and include a number of specialist modules, each of which deals with a different type of text. Dictionaries, in particular, have their own dedicated module TEI-DICT[3]. One interesting aspect of these dictionary guidelines, which will turn out to be extremely pertinent in what follows, is the explicit distinction that they make between three different ways of viewing dictionary data, these are: (a) **the typographic view**; (b) **the editorial view**; and (c) **the lexical view**. It will be useful to go into some detail on this threefold distinction since, at the very least, it will motivate our own separation of lexical citations into different conceptual levels below. The first view, the typographic view, essentially concerns the layout of a page – so, for instance, where the line breaks are in a text, or how the entries are arranged visually on any single page; for obvious reasons the authors of the TEI guidelines refer to this view as the 'two-dimensional' view. The second view, the editorial one, deals with the properties of a text modeled as a sequence of tokens. Accordingly, if

1 http://www.tei-c.org/index.xml [accessed 29/03/18]

2 These guidelines can be found on the TEI website: http://www.tei-c.org/Guidelines/ [accessed 29/03/18]

3 http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html [accessed 29/03/18]

we take this view with respect to a print dictionary, then for any specific entry we will be interested in exactly which words are used in the entry and in which order, along with the exact placement of punctuation in the text. The authors of the guidelines identify the editorial view as 'one-dimensional', since it is effectively concerned with a *linear* sequence of tokens. The third and final view mentioned in the guidelines, the lexical, relates to the conceptual or linguistic content of a lexicon or dictionary as well as each of its individual entries: to the fact, for instance, that a particular lexicon focuses on the medical domain or that the grammatical category of a given entry is "verb"; we might tentatively refer to this view as the 'zero-dimensional' view, although this term is not used in the document itself. TEI-DICT is, avowedly, a model for encoding all three views, something which has resulted in a relatively complex set of modeling guidelines, in which there exist several different ways of modelling the same information. In effect however we can lump the first two views together as dealing with the mode of presentation in a lexical resource, and isolate the last view as describing the content or meaning of the information itself. In the next subsection we will discuss the FRBR model which makes a similar classification of the different kinds of information that can be potentially referred to in a bibliography. As we shall see, this classification is not entirely orthogonal to the tripartite TEI classification discussed above. Before we move on, however, we should look at what provision the TEI-DICT guidelines offer for the encoding of citational information. In the case of citations that include quotations, the TEI guidelines recommend the use of the <cit> element; bibliographic references to other works can then be added using the <bib> element. As an illustration of the recommendations made by the TEI-DICT guidelines we will take an example from the Perseus TEI-XML encoding of the Liddell Scott Jones Ancient Greek-English lexicon (Liddell et. al. 1925), a hugely influential and authoritative Ancient Greek-English dictionary which was first published in 1843 and which is currently still in print in its 9th edition. The example in question is the following (presented in its original formatting):

> **Ἄβρων ,** ωνος, ὁ,
> **A.***Abron,* an Argive, proverbial for luxurious living, "Ἄβρωνος βίος" Suid., Zen.1.4.

According to the TEI guidelines we can serialize the example as follows in XML:

```
<entryFree id="n210" key="*)/abrwn" type="main" opt="n">
        <orth extent="full" lang="greek" opt="n">Ἄβρων</orth>
        ,
        <itype lang="greek" opt="n">ωνος</itype>
        ,
        <gen lang="greek" opt="n">ὁ</gen>
        ,
        <sense id="n210.0" n="A" level="1" opt="n">
                <tr opt="n">Abron,</tr>
                an Argive, proverbial for luxurious living,
                <cit>
                        <quote lang="greek">Ἄβρωνος βίος</quote>
                        <bibl default="NO"><author>Suid.</author></bibl>
                </cit>
                ,
                <bibl n="Perseus:abo:tlg,0596,001:1:4"       default="NO">
                        <author>Zen.</author>
                        <biblScope>1.4</biblScope>
                </bibl>
        .
        </sense>
</entryFree>
```

Here the two citations given in the original dictionary text as '"Ἄβρωνος βίος" Suid.' and 'Zen.1.4' respectively are encoded in the first case with a <**cit**> element containing a <**quote**> and <**bibl**> element, and in the second case with a <**bibl**> element.

2.2 Functional Requirements for Bibliographic Records

The Functional Requirements for Bibliographic Records (FRBR) entity relationship model is perhaps the single most influential conceptual model so far devised for the representation of bibliographic data in computational resources (linked data resources being no exception to the trend). It was developed by the International Federation of Library Associations and Institutions in the early 1990s (Tillett 2007), and then in a subsequent development was harmonized with the well-known CIDOC-CRM conceptual model and published as a formal ontology called FRBR-oo (the 'oo' standing for object oriented) (Boeuf 2012). An expression of the core concepts of FRBR has been made available in RDF, and there also exists an RDF version of FRBR-oo, and so the FRBR model is, in effect, ready to use in the construction of RDF datasets. With respect to the contents of the model, FRBR makes a fourfold distinction in describing bibliographic entities on the basis of the particular ontological status which each entity holds. We present the classification as it pertains to texts, although these categories can just as well be applied to other kinds of bibliographically referable entity. The categories are (in ascending levels of concreteness):

- Work: those aspects of a text that can be abstracted away from any particular linguistic representation: so that for instance all the translations of a text, e.g., Hamlet, Amleto, हैमलेट, 哈姆雷特, etc., refer to the same Work under this view; the 0-dimensional TEI Lexical View seems to largely overlap with this category;
- *Expression*: the specific linguistic form which a Work takes, this view includes all of that which gets lost in translating a Work from one language to another;
- *Manifestation*: a physical embodiment of an Expression, e.g., the 2015 Penguin Classics edition of Hamlet;
- *Item*: a specific instance of a Manifestation, so for instance, I could use this category to refer to the copy of the 2015 Penguin Classics edition of Hamlet that is currently held in my local library.

It is clear from this description that the TEI lexical view corresponds to the FRBR concept of Work, the editorial view to Expression, and the typographical to Manifestation (and perhaps also to Item). Of course, it is important that we make the disclaimer here that the conceptual distinctions made by TEI and FRBR should not be considered as watertight, in fact they turn out to be very difficult to apply in certain kinds of concrete instance (something which we discuss in more detail in Section 4.2). In many other cases, however, they have proven to be very useful approximations.

2.3 Bibliographies and Citations in Linked Data

There exist a number of vocabularies that allow for, or assist in, the representation of bibliographic information as linked data; we will mention only a few of the most popular ones here and do not aim at comprehensiveness. The most well-known of these vocabularies is undoubtedly the **Dublin Core (DC)** [4], which provides data modelers with a number of fundamental classes and properties allowing for the description of relations between bibliographic entities in linked data resources. However, as Peroni and Shotton (2012) point out, the generic nature of the DC vocabulary means that we are seriously restricted in the kinds of bibliographic information which we can use it to express, unless we make use of other vocabularies. Another important linked data bibliographic vocabulary is **FRBR** for which, as we mentioned in the previous section, there exist a number of versions in RDF. The

4 http://dublincore.org/ [accessed 29/03/18]

Bibliographic Framework (BIBFRAME), on the other hand, was developed as a replacement for the **MARC** standards which had been previously used in the library sector; BIBFRAME was specifically designed with linked data datasets in mind (Casalini 2017). It is interoperable with FRBR, although it uses a slightly different classification hierarchy to FRBR (the BIBFRAME concept *Work* encompasses both FRBR categories Work and Expression).

The **SPAR** suite of formal ontologies offers users a collection of vocabularies that permit them to model a wide number of different aspects of the semantic publishing and referencing domains in RDF (Peroni 2014). These ontologies have had a wide uptake in both scholarly and industrial domains, having been used by, among others, *Nature*, Europeana, and the Open University. We will single out two of these ontologies in what follows: **FRBR-aligned Bibliographic Ontology (FABiO)** and the **Citation Typing Ontology (CiTO)** (Peroni & Shotton 2012). FABiO, which carries the fact of its FRBR-aligned status in its very name, deals with RDF versions of bibliographic records; it also encompasses a number of other vocabularies, such as DC Terms and SKOS, in addition to a series of newly defined properties intended to facilitate the production of semantically rich bibliographic metadata. CiTO on the other hand allows for the elaboration of different kinds of rhetorical and factual relationship between two or more bibliographic objects in a network of citations. Here it is important to note that the CiTO model defines a citation as "a conceptual directional link from a citing entity to a cited entity, created by a human performative act of making a citation". This definition will have important consequences in the development of our own vocabulary for lexical attestations below. In addition, the SPAR **Document Components Ontology (DoCO)** groups together a number of vocabulary terms for describing both the structural and rhetorical makeup of a text; it will also be pertinent in what follows (Constantin et al. 2011).

So much then for the bibliographic and citational side of things, for the time being at least; when it comes to the representation of lexical information in linked data on the other hand, our options are a little bit more restricted. Indeed, the **lemon** model for representing lexical data in RDF (McCrae et. al. 2011), recently published in a updated version as ontolex-lemon (McCrae et. al. 2017), has come to take on the status of a de facto standard for representing lexical resources in RDF, and so, in view of its popularity, its dominance of the field as it were, we have chosen to use it as the basis of the work presented in this article. However lemon, unlike TEI-DICT, focuses on capturing the conceptual content of a lexicon; that is, it takes a primarily lexical view of lexical resources, treating them as Works according to the basic FRBR conceptual scheme. Hence there is no conflict here between the demands of fidelity to the text in its lexical view and the text in its editorial and typographical view as there is in TEI; lemon simply prioritizes the former.

Neither lemon nor its successor ontolex-lemon make any specific provision for lexical citations, which brings us onto one of the main arguments of our article, namely that there is a necessity for a specific vocabulary (in our case based on lemon) to do just this in the important case (and also likely the majority case when it comes to citations in lexicons) in which a citation is being used to attest a lexical entry or one of its properties. Why not, then, use the 'citation' class provided by CiTO to do this? The reason is that there are (at least) two ways of viewing such a citation, both of which we may want to capture separately when modelling a lexicon or a dictionary. One of these views pertains to the lexical/Work view and regards the purpose of the citation, that is, to attest to the existence, in language use, of an association of a given lexical entry with a given linguistic property; the other seems to pertain more to the Expression level or to the editorial view: to a lexicon viewed as a bibliographic entity enmeshed in a web of bibliographically-salient relations with other bibliographic resources. The object properties furnished by the CiTO vocabulary refer to this latter view. Our proposal is to create a RDF-based vocabulary that deals with the level of the former view. We elaborate on this point in the next section through the provision of two detailed examples.

3 Two Illustrative Examples

In this section we try and support one of the central claims of this article, namely, that a proper encoding of citations attesting to lexical properties must take into consideration at least two different kinds of conceptual entity: citations and attestations. In the following subsections, 3.1 and 3.2, we present two different examples of lexicographic encoding in which the difference between the two kinds of entity comes out as particularly transparent.

3.1 If at First You Don't Succeed...

Our first example has a strong Dantesque flavor to it, and serves to illustrate how two authoritative lexical resources can completely disagree on the meaning of a citation, even one as famous as the quotation which we discuss in the example[5]. The example centers around the Italian word *riprovare*, which means both 'to try something again' (deriving in this instance from the word *provare* 'to try' and the prefix *ri-* which adds the sense of repetition), as well as 'to scold, rebuke' (in this sense it is cognate with the English verb *reprove*): we are in this case dealing with a pair of homonyms. The popular Italian dictionary *il vocabolario Treccani* (Simone et. al. 2010)[6] lists these as two separate entries: *riprovare*^{1} ('to try again')[7] and *riprovare*^{2} ('to scold')[8]; we will refer to the two homonyms in the same way in what follows. The entry for *riprovare*^{1} makes an etymological reference to the entry for *provare* in the same dictionary and cites both the motto of the short lived 16th scientific society *L'Accademia del Cimento,* i.e., *provando e riprovando* ('trying and trying again'), and the *terzina* of the Divine Comedy from which the motto was adapted ('*Quel sol che pria d'amor mi scaldò 'l petto,/di bella verità m'avea scoverto,/provando e riprovando, il dolce aspetto*.' Par. III, 1-3)[9] – where however, as the entry itself points out, it means *riprovare*^{2}: that is although Dante's use of *riprovare* is cited in the entry for *riprovare*^{1} the entry does not make the claim that this use attests to *riprovare*^{1}. The Treccani entry for *riprovare*^{2} also cites the same use of *riprovando* in Dante, but in this case the claim is that it does *attest* to the entry in question. On the other hand *Il Grande Dizionario della Lingua Italiana* (GDLL)[10] (Battaglia 1961) cites both the motto of L'Accademia del Cimento and the *terzina* from the Divine Comedy mentioned above under its entry for *riprovare*^{1} (which recall has the meaning 'to try again') – just as in Treccani – but with the contradictory claim, this time round, that both cited texts do attest to the entry in question, namely *riprovare*^{1}.

To summarizes, then: we have presented an example in which the same text is cited by two different sources and used to attest to two different homonyms of a word. The following statements describe the current example:

Treccani's entry for riprovare1 cites Par. III, 1-3.

1. Treccani's entry for *riprovare*^{2} cites Par. III, 1-3.
2. GDLL's entry for *riprovare*^{1} cites Par. III, 1-3.
3. *riprovare*^{1} is attested by Par. III, 1-3.
4. *riprovare*^{2} is attested by Par. III, 1-3.

5 The example is dealt with in more detail, and an attempt at an encoding in RDF given in Bellandi et al. (2017).

6 See also the online version: http://www.treccani.it/.

7 http://www.treccani.it/vocabolario/riprovare1/ [accessed 29/03/18]

8 http://www.treccani.it/vocabolario/riprovare2/[accessed 29/03/18]

9 Translated by Longfellow (Alighieri & Longfellow 1867) as 'That Sun, which erst with love my bosom warmed/ Of beauteous truth had unto me discovered/By proving and reproving, the sweet aspect.'

10 The *GDLL* holds something like the same status and authority in the Italian language as the *Oxford English Dictionary* does in English.

The first three items in the list are true statements about the lexicons which they refer to; they describe the existence of three successful citational speech ('performative') acts: speech acts which can be directly represented in RDF using the cites object property from CiTO or one of its subproperties. These statements do not deal *directly* with words or their usages, but rather they are concerned with documents or works and the rhetorical/organizational structure pertaining to them. The other two statements, those which I have numbered 4 and 5, instead describe the direct relationship between an item in a lexicon and a text which evidences, or better, attests to its past use. These latter statements are at the level of linguistic facts about words and other lexical entries, that is, at the TEI lexical level[11]. The fourth statement is false but the fifth one is true; however in neither case does this follow from the truth (or falsity) of the first three statements. Both 4 and 5 are only indirectly described by CiTO's cites object property; one of the core aims of the work described in this paper is to describe statements such as 4 and 5 directly in RDF. Note that this example is by no means an atypical one, as this kind of divergence between different lexical resources, for instance, is especially common when it comes to the treatment of word etymologies.

3.2 An Anomalous Example

The second example in this section is also our second example taken from the Liddell Scott Jones lexicon (LSJ). This time around the example entry is for the word *ἀνώμαλος*, *(anṓmalos)* from which the English word *anomalous* derives[12].

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.**

The LSJ lists one basic sense for the entry; this single sense is then divided into a number of subsenses. Each subsense is associated with a number of citations and these (in most cases) serve to elaborate further shades of meaning with respect to their corresponding subsenses. In what follows we will concentrate on the third and fourth citations, both of which belong to the first sense, that of 'uneven, irregular'. The third citation is interesting because of the appearance, in parentheses, of the abbreviation *cj*, which is usually found in critical apparatuses and which stands for the Latin *conicit*, 'conjectures'. The abbreviation signifies that the citation refers to a critical reconstruction of a work and that there is, therefore, a good chance that the text referred to might not actually attest to the word or sense in question at all: all we can be sure of is that a later scholar in attempting to reconstruct the text from the fragments that were available to him or her made the decision to include the word in his or her conjectural emendation; other lexicographers may decide not to include the citation due to its conjectural status. And this is indeed the case with the third citation, which has not been included by the *Diccionario Griego–Español* (Adrados et. al. 2008), a contemporary ancient Greek-Spanish lexical

11 It is not entirely clear whether all five statements belong to the TEI lexical view or not – or whether the first three regard the editorial view. Regardless we believe such examples make a strong case for defining a separate attestation relation.

12 The entry can be found online here in the Perseus published version of the lexicon; http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0057%3Aentry%3Da)nw%2Fmalos [accessed 29/03/2018].

resource based on the LSJ (along with a number of other Greek lexicons), in its entry for *ἀνώμαλος*[13]. Once again the example demonstrates the clear conceptual distinction that exists between the performative act of citing a piece of text as evidence – and there can be no reasonable doubt that the 1947 edition of the LSJ did indeed cite Thucydides 7.71 in its entry for *ἀνώμαλος* – and an instance of a word in a text attesting to a given sense – it is doubtful whether Thucydides did use the word in that sense in the passage in question: that is, the distinction between *citations* and *attestations*. Looking at the fourth citation on the other hand we see that it is prefaced by another abbreviation, *cf*, which stands, this time round, for the Latin term *confer* meaning 'compare' and is an instruction to readers to compare the use of the word the cited text (in this case, Aristotle's *Prior Analytics*) with its use in the text(s) previously cited. It is underspecified whether this kind of citation attests to the same lexical sense/sense/other lexical property, or whether it only provides some interesting contrast or comparison. We cannot therefore always be sure that we are dealing with an *attestation* for, in this case, a sense; we can on the other hand be certain that we are dealing with a *citation*.

In summary then we have an example, one that is, again, by no means an exceptional or a marginal one when it comes to scholarly dictionaries like LSJ, where the idea of two different levels of description, a citational/rhetorical one and a lexical one, arises very naturally. A citation can successfully reference a text with a view to attesting a lexical property even if in reality the text does not attest that property at all: citations can also have different rhetorical purposes other than that of attestation. In the following section we try and model this notion of a lexical attestation in an RDF-based model that extends lemon.

4 A Proposal for a Vocabulary for Lexical Attestations

We made a number of observations in the preceding two sections with a view to motivating our definition of a specialized linked data vocabulary for representing lexical attestations. Such a vocabulary, as we hope to have shown, is useful for modeling certain kinds of linguistic claims made via the use of citations: claims which are especially common in scholarly print-born lexicographic resources. We will detail our (minimal) proposal of such a vocabulary, called lemonBib, in Section 4.2. Before that however we turn to the discussion of a modeling issue, which turns out to be very relevant to the modelling of legacy lexicographic resources in RDF, and which also relates to the TEI/FRBR classifications that we mentioned above.

4.1 How to Model Different Textual Views on Computational Lexica

As we mentioned above, despite the fact that the distinctions between Work and Expression and between the 0D and 1D/2D views seem to be extremely useful at first sight (and indeed they turn out to be useful in the long run, too), in practice they are often difficult to apply to numerous ambiguous or fuzzy cases. How much sense, for example, does it make to separate out the conceptual Work part of a novel like *Finnegans Wake* from its realization in any specific language[14]? When it comes to lexicographic resources we have to deal with an additional problem that arises from the fact that a lexical entry, as well as being a conceptual component of a lexicographic work, also happens to be a document component of a text in the same way as a chapter, a table of contents, or a bibliography are – and arguably the same

13 The DGE entry for *ἀνώμαλος* can be consulted online at http://dge.cchs.csic.es/xdge/%E1%BC%80%CE%BD%E1%BD%BD%CE%BC%CE%B1%CE%BB%CE%BF%CF%82 [accessed 29/03/2018].

14 This is not to say that there haven't been numerous attempts, as in the case of many other supposedly 'untranslatable' works of literature, at translating *Finnegans Wake* in other languages, or that these attempts were entirely fruitless. However the French translation is said to have taken its translator over 30 years to complete, and in the case of the Japanese translation the intellectual toll was so great that the first translator of the work simply disappeared and the second ended up going mad (see https://www.mhpbooks.com/the-challenge-of-translating-finnegans-wake/ [accessed 29/3/2018]).

is also true of senses in dictionaries. This raises the issue of where we should locate lexical entries and senses in our overall classification, on the grounds of the distinctions that we've already made between Work and Expression, 0D and1D, and attestations from citations. One option – and this is the strictly purist one – would be to extend the DoCO vocabulary with the classes Lexical Entry and Lexical Sense. Then, for instance, the order of entries in a dictionary – an ordering which, in most cases, has no systematic linguistic significance but is only there to help readers locate the word or entry that they're looking for in a physical copy of the dictionary – would be an ordering of Expression/DoCO Lexical Entries, but not of lemon/Work Lexical Entries. Of course we would want to associate corresponding Work/Expression Lexical Entries, and Lexical Senses with each other in each case. Citations would then belong to the Expression level and attestations to the Work level. Unfortunately, however, this would also lead to a doubling of entries and senses in the RDF version of a lexicon – the kind of prolixity that, conceptual purity notwithstanding, would probably make this quite an unpopular approach. We have therefore decided not to make an explicit distinction between the two views of lexical entries/senses in our example, but to merge the two conceptual levels together in the same entity.

4.2 LemonBib

And so it is that we finally come in the present section to the definition of our proposed extension of the ontolex-lemon model for modeling lexical attestations the ontolex-lemon model, *lemonBib*[15]. From our discussion above it is clear that our vocabulary should allow us to do the following:

- Relate attestations to their corresponding citations;
- Relate an attestation to the text which it refers to;
- Relate attestations to other, relevant citations.

We have decided to create a fairly minimal set of properties and classes that meet these requirements in order to make the vocabulary as re-usable as possible. Our proposed modular extension of is based around the definition of the new class Attestation. The idea is that Attestation reifies the relationship between a given lexical element in lemon – whether this is a Lexical Entry, a Lexical Sense, a Lexical Form, or something else – and a bibliographic item that contains a text exemplifying the use of the element in question; we will also be able to relate an Attestation with any citation that is associated with it. We define an object property, isAttestedBy relating a lexical element *e* with a member of the class Attestation *a*, with an inverse property attests going in the other direction. We also define the object property involvedinAttestation between an instance of the CiTO class Citation and Attestation with the inverse property attestationCitation. This allows us to relate together the entities which as we have argued in this article belong to two different conceptual levels. The object property foundIn relates an attestation with the bibliographic entity in which the attestation can be found. We also define two new data properties. The first, hasContext, relates an attestation together with the textual context in which the word is found; the second is the Boolean property conjectural, which is true when an attestation is based on a conjectural witness.

We now present a partial encoding of the ἀνώμᾰλος example in diagrammatic form in order to illustrate the features of lemonBib listed above[16]. Our lexical entry has three senses[17], we will focus on the first sense sense1 (sense **A** in the original entry above), and on the third attestation of that sense thuy_att (with the citation **Th.7.71**) along with its corresponding citation thuy_cit. Note that sense1 is

15 The lemonBib vocabulary is available at http://lari-datasets.ilc.cnr.it/lemonBib .

16 The example is available in an RDF version at http://lari-datasets.ilc.cnr.it/lsj_anomalos .

17 Note that we have not attempted to describe the hierarchical structure of the senses in our encoding so as not to make the example overly complicated; however a lexicographic extension of ontolex-lemon that will deal with such hierarchies of senses has been proposed and is currently being discussed in the W3C ontolex mailing list (https://www.w3.org/community/ontolex/).

linked to thuy_att via the isAttestedBy that we have defined and that thuy_att is linked in turn to the relevant text (here represented using a CTS URN) by the foundIn relation. In addition the attestation thuy_att is associated with a textual context via the hasContext data property and is specified as referring to a conjectured text by the property conjectural which is set to true. The attestation is then linked to the citation with which it is associated, namely thuy_cit, using the attestationinCit property and vice versa using the involvedinAttestation property. The citation thuy_cit is further associated with the sense sense1 as its citing entity as well as the cited text using object properties defined in CiTO; the type of the citation is also specified using the punned object property citesAsEvidence. Note that although we have not included it in the diagram, we can use the rdfs:seeAlso property to model the use of the cf abbreviation in the text.

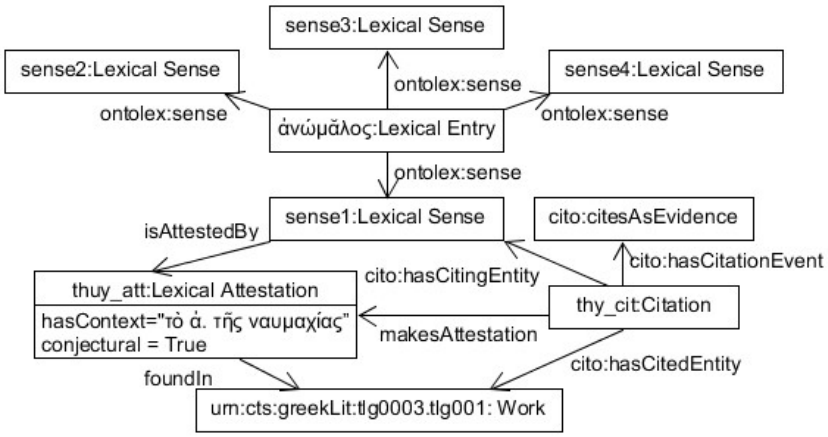

Figure 1: A (partial) encoding of the ἀνώμᾰλος example

5 Conclusion

In this article we have discussed the modeling of citations in lexical linked data resources and proposed an extension of ontolex-lemon for dealing with lexical attestations since, as we have argued, this case is not sufficiently covered by pre-existing vocabularies. We have concentrated on examples from traditional, print-based dictionaries because of the wealth of interesting cases that such resources offer. However, we are confident that our vocabulary will be useful for other kinds of resources, at least as a basis for the addition of further classes and properties. In further work we are planning to test the usefulness and the sufficiency of our vocabulary by using it to encode entire lexical resources.

References

Adrados, F. R., Elícegui, E. G., & Berenguer, J. A. (2008). Diccionario griego-español. Consejo Superior de Investigaciones Científicas.

Battaglia, S. (1961). Il Grande dizionario della lingua italiana di Salvatore Battaglia. UTET, Torino

Bellandi, A., Boschetti, F., Del Grosso, A. M., Khan, A. F., & Monachini, M. (2017). Provando e riprovando modelli di dizionario storico digitale: collegare voci, citazioni, interpretazioni. Proceedings of the AIUCD.

Boeuf, P. L. (2012, 06). A Strange Model Named FRBROO. Cataloging & Classification Quarterly, 50(5-7), 422-438. doi:10.1080/01639374.2012.679222

Casalini, M. (2017). Implications of BIBFRAME and Linked Data for Libraries and Publishers. “Roll With the Times, or the Times Roll Over You”. doi:10.5703/1288284316449

Constantin, A., Peroni, S., Pettifer, S., Shotton, D., Vitali, F. (201). The Document Components Ontology (DoCO). In Semantic Web – Interoperability, Usability, Applicability, 7 (2): 167-181. Amsterdam, The Netherlands: IOS Press. https://doi.org/10.3233/SW-150177

Functional requirements for bibliographic records: Final report. (2013). De Gryuter.

Il vocabolario Treccani. (1997). Istituto della Enciclopedia italiana, Fondata da Giovanni Trecanni.

Liddell, H. G., Scott, R., & Jones, H. S. (1925). A Greek-English lexicon. Clarendon Press.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In Proceedings of eLex 2017 conference, September (pp. 19-21).

McCrae, J., Spohr, D., & Cimiano, P. (2011, May). Linking lexical resources and ontologies on the semantic web with lemon. In Extended Semantic Web Conference (pp. 245-259). Springer, Berlin, Heidelberg.

Peroni, S. (2014). The Semantic Publishing and Referencing Ontologies. In Semantic Web Technologies and Legal Scholarly Publishing: 121-193. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-04777-5_5

Peroni, S., Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. In Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 17 (December 2012): 33-43. Amsterdam, The Netherlands: Elsevier. DOI: 10.1016/j.websem.2012.08.001

Simone, R., Berruto, G., & D'Achille, P. (2010). Enciclopedia dell'italiano: Il vocabolario Treccani. Istituto della Enciclopedia italiana.

Tillett, B. B. (2007). What is FRBR?: A conceptual model for the bibliographic universe. Library of Congress, Cataloging Distribution Service.