Situating Word Senses in their Historical Context with Linked Data

Fahad Khan
CNR - Istituto di Linguistica Computazionale "A. Zampolli", Pisa, Italy
fahad.khan@ilc.cnr.it

Jack Bowers
Austrian Academy of Sciences (ACDH), Vienna, Austria
Jack.Bowers@oeaw.ac.at

Francesca Frontini
Praxiling UMR 5267 CNRS, Université Paul-Valéry, Montpellier, France
francesca.frontini@univ-montp3.fr

Abstract

In this article we present a Semantic Web-based model for creating lexical resources in which the diachronic and, more broadly, contextual dimensions of word meaning can be explicitly represented as part of a graph-based data structure. We start by discussing why Linked Data is the right publishing approach for such diachronic datasets. We then describe our model, lemonEty, which utilizes the ontology engineering technique of perdurants in order to model lexical entries as dynamic processes. Next we go onto explain how to represent etymologies using our model, and in particular how to associate temporal information with word senses, taking examples from two different lexicographic resources. In addition, we will show how our model deals with cognates and attestations.

1 Introduction

Ontologies can be used to enrich computational lexical resources in numerous different ways, the most obvious of which is by using ontological entities to describe the semantics of individual lexical entries. In addition to this however ontologies can also be used to help describe and to reason about how languages change and evolve; and in particular, they allow us to model word meaning change. In this article our focus will be on lexical and ontological datasets for the **Semantic Web**, that is **linked data** datasets. This choice was made for various reasons, not the least of which is the fact that the Linked Data (LD) publishing paradigm has made it easier than ever before to link together different, individual datasets. This has, in turn, served to highlight the many benefits of augmenting single resources with (semantically specified) links to other datasets. Working with linked data also offers up access to a host of different Semantic Web tools and languages, amongst which: a dedicated Knowledge Representation language (OWL), an expressive query language (SPARQL) along with a semantic web rule language (SWRL). There can, however, be certain drawbacks to publishing datasets as Linked Data. One of the most restrictive of these relates to the fact that Linked Data best practices require that datasets be published using the standard RDF model, meaning that they must in effect be modeled as sets of subject-predicate-object (S-P-O) triples. This restriction to unary and binary relations prevents us from directly adding extra arguments to binary predicates in order to model *n*-ary relations for any *n* greater than 2. And so we cannot just specify the temporal validity of RDF statement representing an S-P-O triple by adding a time argument to P. This can be a particular problem as temporal information is a core part of many lexical datasets, especially those dealing with etymological data.

A number of solutions to the problem of representing *n*-ary relations in RDF have already been proposed each of which carries its own particular advantages and disadvantages. One popular solution, and the one which we adopt in this paper, is to model entities as *perdurants*, that is, as entities with an inherent temporal extent: effectively treating them as processes that unfold through time.

Unlike static entities, so called *endurants* – which retain their essential (identifying) properties at any of the different points in time in which they exist – looking at a 'snapshot' of a perdurant at any given point in time only gives us a part of the entity. So for example take a lexical entry, *l*, and two lexical senses *s1*, *s2*, such that *l* has the sense *s1* during interval *i1*, and *s2* during interval *i2*. We would like represent this as follows, *sense(l, s1, i1)* and *sense(l,s2,i2)*. But since the sense relation is binary we can't express the temporal duration of *s1* and *s2* in this way. Instead if we view *s1* and *s2* as perdurants then we can associate them with the respective temporal intervals *i1* and *i2* so that they are classified among the properties of *s1* and *s2*. The particular approach to perdurants which we take up in this article is the *fluent* approach proposed by Welty and Fikes (Welty and Fikes (2006)) and subsequently modified by Krieger (Krieger (2014)). In the latter, relations that were formerly represented as holding between entities modelled as endurants now hold between time slices of those same entities modelled as perdurants. So for example if in the original, non-diachronic version of an ontology relation, *employeeOf* was specified as holding between an entity, *p*, of type *Person*, and an entity, *c*, of type *Company*, then according to Welty and Fikes' approach to perdurants we create two new entities *p@t1* and *c@t1* which are temporal parts of *p* and *c* respectively and therefore of type *TemporalPart*, both with the temporal extent of *t1*, with *employeeOf* now holding between *p@t1* and *c@t1* instead of between *p* and *c*. Krieger's approach simplifies this by allowing these time slices to be typed according to the original classes, so that *p@t1* is now also of type Person (as is *p*) and c@t1 of type Company (as is *c*). In this way we don't need to change the type of *employeeOf* to make it a relation between entities of type *TemporalPart*.

In what follows we will apply the perdurantist approach to representing etymological information. Briefly, we will define a class called *Etymon* which consists of time slices of the class *LexicalEntry* (with the latter class belongs to the RDF-based lexicon model *lemon*[1]), where, as per the Krieger approach to perdurant modeling, the class *Etymon* is a subclass of *LexicalEntry*. We will also define a class *Etymology* which represents the evolution through time of the salient properties of a lexeme. Finally we explicitly represent lexicographic attestations using a class called *Attestation*. We hope to show that these new classes together with their related properties permit us to explicitly situate word senses temporally as well as to clearly represent statements and hypotheses *about* senses, within the framework of the RDF data model.

2 Modeling Diachronic/Etymological Lexical Data

The diachronic dimensions of lexical data have been somewhat neglected in models for the representation of lexicons up till now, although they are partly dealt with in the TEI dictionary module [2]. Work has also been carried out in the past on the best way of representing etymological information in LMF (Salmon-Alt (2006)), and detailed proposals for an extension of TEI that deal with etymologies have recently been made (Bowers and Romary (2016)). The de-facto standard for representing lexical datasets as linked data is lemon (McCrae et al. (2011)), the design of which was heavily influenced by LMF, something that is readily apparent from a comparison of lemon with the LMF core model. One of the main ways in which lemon differs from LMF, however, aside from simply having fewer classes and properties in its core model and basic extensions, is in its modeling of lexical semantics. In lemon, members of the class *LexicalEntry* are linked to ontological items which represent the extension (as opposed to intension) of these entries via a *LexicalSense* object. These *LexicalSense* objects are intended as reified pairings of a lexical entry with an ontological item representing one of the meanings of the entry in question.

The core of the lemon model is presented below in Fig. 1.

In contrast to LMF or TEI, lemon's approach to representing word meaning was explicitly designed with linked data in mind. However neither lemon nor its successor Ontolex-lemon have any relations or classes that specifically deal with diachronic information[3]. In previous work we proposed an extension of lemon called lemonDia that allowed the representation of diachronic semantic data (Khan et. al. 2016). At the core of this extension was the idea of modeling senses of lexical entries as perdurants. However, we did not take into consideration the interaction between the meaning of a word and other, non-semantic, properties. This can be a real limitation since it often occurs that multiple properties of a word change at the around the same time. For instance, the pronunciation and/or

[1] For the specifications of this model see https://www.w3.org/2016/05/ontolex/

[2] http://www.tei-c.org/release/doc/tei-p5-doc/it/html/DI.html

[3] Although there is currently a proposal to discuss such an extension on the ONTOLEX mailing list.

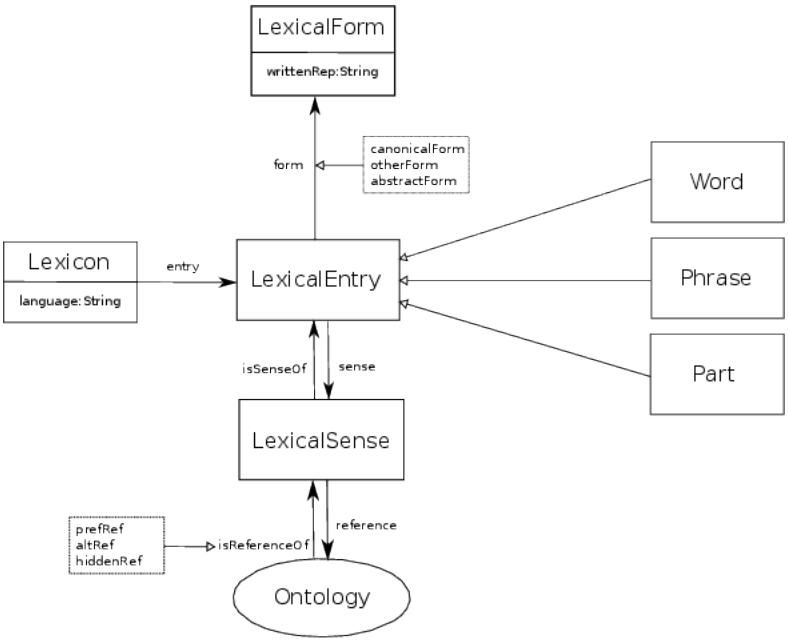

Figure 1: lemon model core diagram.

orthography of a word may change or vary over a span of time during which a new sense of that same word may come into use. In order to remedy this, we present, in the next section, an extension of lemon, lemonEty, that applies the perdurantist approach to the whole lexical entry, instead of just the sense or any other single aspect of an entry's profile (e.g. grammatical, phonetic, etc.). Another issue which quickly comes up in modeling etymological data is how to properly represent textual attestations where these provide strong evidence for the use of a word with a particular sense. Indeed if one views etymologies as scientific hypotheses tracing out a word's origin and the evolution of its many senses, then it is intuitively obvious that we should be able to model attestations as pieces of evidence. We will look at how our approach can help in modeling this too.

3 Representing Etymological Sense Data using lemonEty

Before we move on to describing the model itself, a brief note on working with temporal data. As we have already mentioned, one of the main advantages of publishing datasets as linked data is the ready availability of useful tools and standards. In the case of the knowledge representation language, OWL, this includes open source inference engines that permit us to reason over OWL datasets: something that is especially attractive when it comes to temporal data since there exist a number of vocabularies and SWRL rule sets for working with such data[4]. One difficulty that quickly becomes apparent in this context however is that we rarely have precise dates temporally delimiting a given linguistic phenomenon; instead the intervals we work with tend to be highly underspecified. Although the lack of precise dates can indeed hamper our ability to make temporal inferences on data, we are still able to carry out *qualitative*, as opposed to *quantitative*, reasoning, using Allen relations, e.g., *before*, *meets*, *overlaps*, to make basic inferences as to how different intervals stand with respect to each other (Allen (1983)). Thus even if we do not know *exactly* which years, or even in which centuries a word was used with a certain sense, we will likely have an idea of some broader interval of time in which this interval of use is contained. For example, we might specify a period representing Middle English delimited by the years 1100AD and 1500AD, as its respective start and end point; the time intervals associated with word meanings can then be be specified as overlapping or being contained within this period. Having the possibility of specifying time intervals with respect to their mutual relations instead of being constrained into always giving specific years or centuries, helps us deal with the vagueness and ambiguity of the temporal information that we find

[4]See for instance https://github.com/sbatsakis/TemporalRepresentations

in many pre-existing lexical resources. Any inferences that we run over such data will effectively take this vagueness into consideration. But it's not always enough. For instance how do we represent a preposition such as "circa" when referring to a year or a century, an expression that has a fairly 'fuzzy' definition in natural language? Somewhere along the line we have to make an interpretation of such phases in our formal language and then ensure that this interpretation is consistent across a dataset. Such interpretations should be stated explicitly either in the documentation or in the resource itself, preferably both.

3.1 Etymology Examples

In order to explain how our model works and some of the ideas behind it we will look at the encoding of some concrete examples. We will take the entry for the word 'girl' from two different lexical sources. The first is the online etymology dictionary[5]:

> *'girl (n.) c. 1300, gyrle "child, young person" (of either sex but most frequently of females), of unknown origin. One guess [OED] leans toward an unrecorded Old English *gyrele, from Proto-Germanic *gurwilon-, diminutive of *gurwjoz (apparently also represented by Low German gære "boy, girl," Norwegian dialectal gorre, Swedish dialectal gurre "small child," though the exact relationship, if any, between all these is obscure), from PIE *ghwrgh-, also found in Greek parthenos "virgin." But this involves some objectionable philology. Liberman (2008) writes:*
>
> *Girl does not go back to any Old English or Old Germanic form. It is part of a large group of Germanic words whose root begins with a g or k and ends in r. The final consonant in girl is a diminutive suffix. The g-r words denote young animals, children, and all kinds of creatures considered immature, worthless, or past their prime.*
>
> *Another candidate is Old English gierela "garment" (for possible sense evolution in this theory, compare brat). A former folk-etymology derivation from Latin garrulus "chattering, talkative" is now discarded. Like boy, lass, lad it is of more or less obscure origin. "Probably most of them arose as jocular transferred uses of words that had originally different meaning" [OED]. Specific meaning of "female child" is late 14c. Applied to "any young unmarried woman" since mid-15c. Meaning "sweetheart" is from 1640s. Old girl in reference to a woman of any age is recorded from 1826. Girl next door as a type of unflashy attractiveness is recorded by 1953.'*

The second lexical source is the important early 20th century work 'An etymological dictionary of the English language' compiled by Walter Skeat (Skeat 1910):

> *GIRL, a female child, young woman. (E.) ME. gerle, girle, gyrle, formerly used of either sex, and signifying either a boy or girl. In Chaucer, C.T. 3767 (A 3769) gerl is a young woman; but in C.T. 666 (A 664), the pl. girles means young people of both sexes. In Will. of Palerne, 816, and King Alisander, 2802, it means 'young women;' in P. Plowman, B. i.33, it means 'boys;' cf. B. x. 175. Answering to an AS. form *gyr-el-, Teut. *gur-wil-, a dimin. form from Teut. base *gur-. Cf. NFries. gör, a girl; Pomeran. goer, a child; O. Low G. gör, a child; see Bremen Wörtebuch, ii. 528. Cf. Swiss gurre, gurrli, a depriciatory term for a girl; Sanders, G. Dict. i. 609, 641; also Norw. gorre, a small child (Aasen); Swed. dial. gårrä, guerre (the same). Root uncertain. Der. girl-ish, girl-ish-ly, girl-ish-ness, girl-hood*

We will model several of the salient pieces of etymological information contained in the preceding two lexical entries for the word 'girl' – which we will refer to as (a) and (b), respectively – using lemonEty.

3.2 LemonEty

In our perdurantist approach to etymological data we associate objects of the class *Etymon* with a lemon *LexicalEntry*, where *Etymon* objects are modeled as time slices of *LexicalEntry* objects. In other words etymons are bundles of lexical properties that hold throughout a certain interval of time; these properties can be phonetic, phonological, orthographic, morphosyntactic, or semantic. We

[5] "Girl". Online Etymology Dictionary. Retrieved April 14, 2017. http://www.etymonline.com/index.php?term=girl

can then specify the relationships between different *Etymon* objects and their associated properties by including links between them; these links are represented as reified *EtymologicalLink* objects. *EtymologicalLink* objects are associated with a given *Etymology* object which represents one version of the history of a *LexicalEntry*. A *LexicalEntry* can have more than one *Etymology* associated with it. Fig. 2 presents the core of the lemonEty module in diagrammatic form[6].

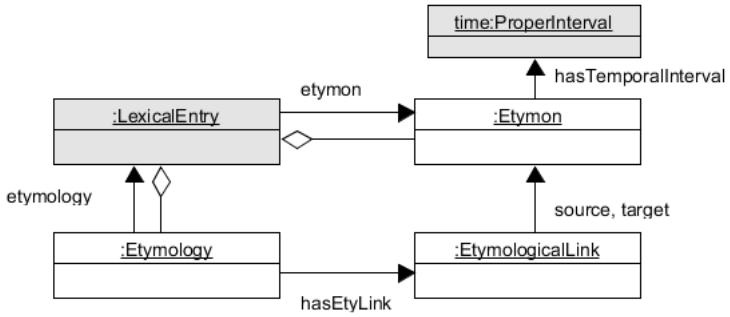

Figure 2: lemonEty core diagram.

In order to capture more specific kinds of etymological relationship we have defined a number of subtypes of *EtymologicalLink* such as *SenseShift* and *Inheritance*. In cases of sense shift we can also create a link between the two *Etymon* senses.

At this point we should clarify the different roles that the *Etymon* and *Etymology* classes play in our model since on first glance it may seem redundant or confusing to include both when presumably only one is really necessary: the answer is that they play different roles. The role of *Etymon* is to represent various different periods in the 'lifespan' of a word during which the word is represented as being stable for certain properties such as the fact that it has only one single sense. This *Etymon* object then will have a temporal interval which delimits its validity in time. A lexical entry can have more than one *Etymon* representing different stages of the word's existence. Objects of the class *Etymology* on the other hand serve to describe the arrangement of these different etymons; etymologies, even though they are the etymologies of single lexical entries, can bring together etymons belonging to different lexical entries in describing the origin of a word. But then, why wouldn't it be enough just to have relations between single objects of the class *Etymon* without introducing the class *Etymology*? The answer is that *Etymology* objects serve to describe different hypotheses as to a word's origins and development; each such object reifies the history or evolution of a lexical entry and allows us to refer to it and to predicate different properties of it, such as for example when the etymology was proposed and by whom. We will now look at the lemonEty representation of (part of) the (a) example above.

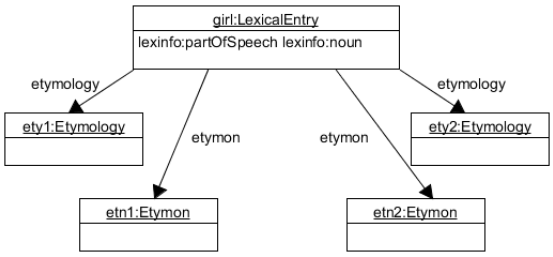

Figure 3: Girl example

In Fig. 3 the word 'girl' has two etymons, *etn1* and *etn2*, representing two different stages in the word's evolution. There are also two different etymology objects, *ety1* and *ety2* which represent two separate versions of the history and development of the word. Note that although we haven't shown

[6]In the diagrams which follow we colour classes and individuals that are from non-lemonEty vocabularies in grey.

this in Fig. 3, we can also associate textual information from the original lexical entry as string values to objects of the class *Etymon* and *LexicalSense* using the *gloss* data property which we have defined.

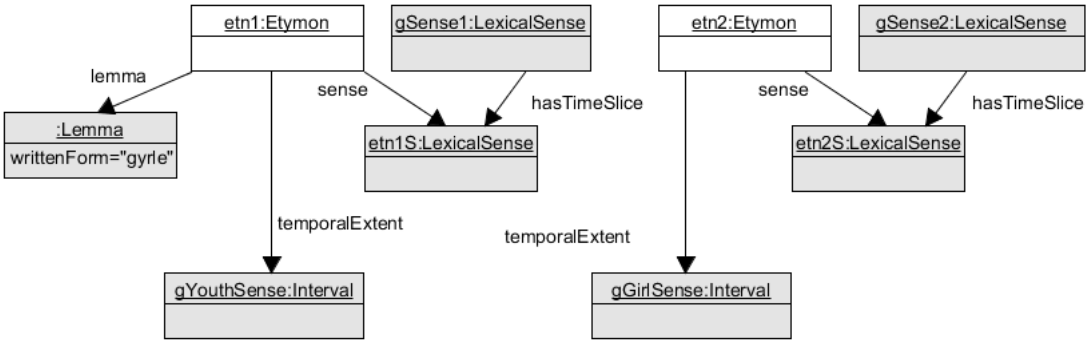

Figure 4: Two *Etymon* individuals

Fig. 4 illustrates the two different objects of class *Etymon*, *etn1* and *etn2* assigned to 'girl' in Fig. 3. The first etymon, *etn1*, corresponds to the first sentence of the entry (The sentence starting with 'c.1300...'). Although we haven't added temporal information to the diagram (for reasons of space) we can specify the *gYouthSense* interval using the OWL Time relation *intervalContains*[7]; we have chosen to represent *gYouthSense* as being contained within an interval with a lower bound of 1250AD and an upper bound of 1350AD (this is our interpretation of 'circa 1300')[8]. The second etymon, *etn2*, represents the narrower sense of 'girl' which is still in contemporary use. We can give the interval *gGirlSense* a lower bound of 1350AD (our interpretation of "late 14th century") and leave it unbounded from the top. Note that each sense of an etymon (in this case *etn1S* and *etn2S*) is a time slice of one of the senses of the lexical entry object (*gSense1* and *gSense2*). Note also that since the original entry did not explicitly mention the existence of a 'narrowing' relationship between the two senses *etn1S* and *etn2S*, we haven't added an explicit *EtymologicalLink* object of type *shiftNarrowing*, even though this would have been a reasonable interpretation to make. Similarly we haven't made a finer division of the second etymon on the basis of the subsequent broadening of the sense of 'girl' in the 15th century, nor have we created new etymons tracking the changes in the written form of the word or additional morphosyntactic changes. The former is for reasons of space; the latter because this information was not included in the original entry which we were modeling. The example reveals how quickly one comes up against multiple ambiguities when representing etymological data from legacy lexical resources in a formal model like RDF.

In Fig. 5 we have given a lemonEty representation of the first etymology given in (a). We represent the etymology using an *Etymology* object, *ety1*, which points to reified etymological links between etymons that in this case belong to different lexical entries. The etymons in Fig. 5 are all linked together with the property *Inheritance*, which is a subproperty of *EtymologicalLink*. In future refinements of our model, we will allow for the addition of confidence measures to members of the class *Etymology* by associating a numerical value or a value from a set of confidence values {low, medium, high} to each etymology object. The second lexical entry, (b), gives a number of attestations for the two different senses of the word 'girl' including a citation apiece from the Canterbury Tales, attestations from Piers Plowman, as well as the translation from French of Guillaume de Palerne ("William of Palerne"). These attestations function as evidence for the use of words and allow us to relate word sense data to information from external bibliographic and historical databases. Again, the graph structure of RDF, along with the fact that RDF requires a universal identifier (URI) for each resource that is dereferenceable using the HTTP protocol, makes this reasonably straightforward. In the present case we would like to link word senses to their attestations which we link in turn to bibliographic and historical datasets. If we can access triples via this second dataset that tell us when the text was (approximately) written, then we have effectively linked up the original lexical sense

[7] https://www.w3.org/TR/owl-time/#time:intervalContains

[8] This is our interpretation of the temporal phrase 'circa'. While it's indeed hard to avoid making such interpretations when representing such data in a computationally useful format, it is important however is to make sure that such assumptions are readily accessible to potential users of the resources, possibly as part of the resource's metadata.

with evidence about when that sense was used. Although there is currently a lack of LOD datasets that contain high-quality, well-curated bibliographic, biographical or historical data, there does seem to be a strong impetus towards the creation of such datasets, especially for use in Digital Humanities' contexts.

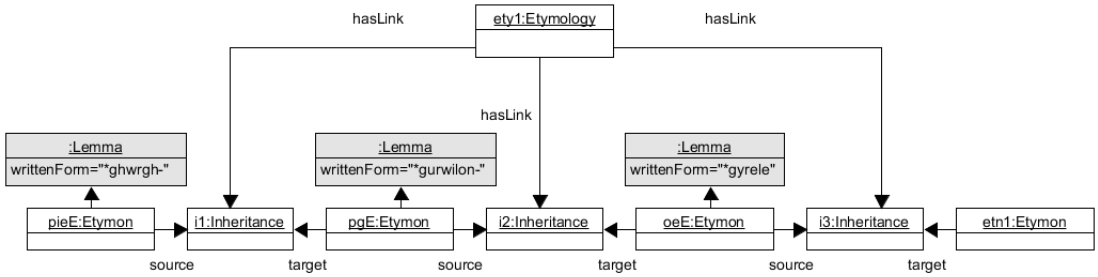

Figure 5: *Etymology* individual.

Based on our previous work on modeling lexicographic resources (Khan et. al. 2017) we decided to add an *Attestation* class to our model. Objects of *Attestation* are linked to objects of the class *LexicalSense* using the object property *attestation*; these *Attestation* objects can then be linked to texts, textual fragments, or corpora, using already existing vocabularies such as CITO (Peroni and Shotton, 2012). So then, to return to our examples, and assuming that we've already created the *LexicalEntry* and *Sense* objects for the entry, similar to those we showed in Fig. 3, we can create an etymon representing the period of time in which 'girl' had both senses, girl and youth, using lemonEty as in Fig. 6. Note that we have specified that the interval of time in which the word *girl* had both meanings is included in the interval of Middle English[9], this is our interpretation based on the text of the entry for the purposes of the encoding, although one might argue that this information isn't actually specified in the text itself.

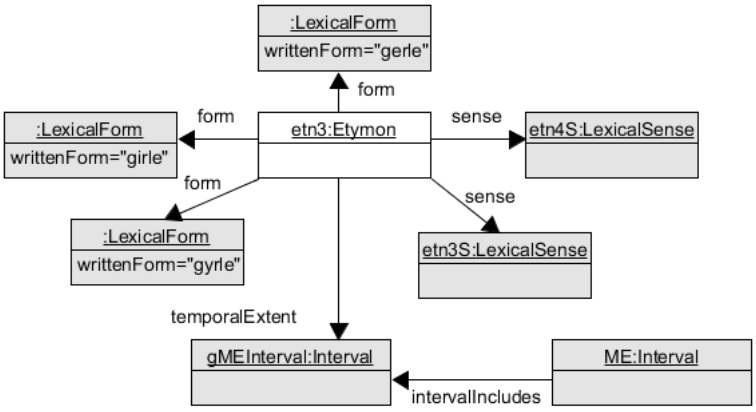

Figure 6: Attested *Etymon*.

We can represent the fact that each of the two senses has an attestation in the Canterbury Tales using the attestation relation which links each sense to a specific object of the type *Attestation* as in Fig. 7.

Here we can specify the immediate context of the word in question, assuming that we have already specified that *etn3S* and *etn4S* are time slices of senses with references *dbpedia:girl* and *dbpedia:youth* respectively. We also assume that the CITO property *citesAsEvidence* points to a

[9]We can use the time span associated with the ISO 639-3 language code for Middle English: http://www-01.sil.org/iso639-3/documentation.asp?id=enm

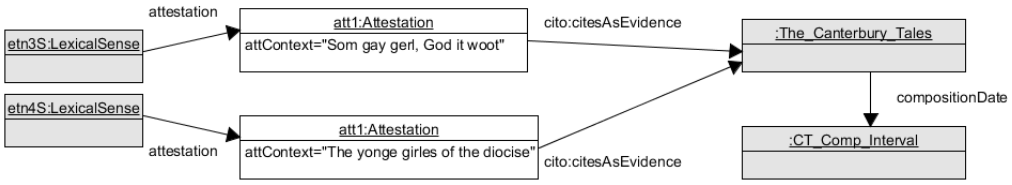

Figure 7: An attestation in the Canterbury Tales.

dataset containing information about the text and its date of composition as in the diagram. Finally, in Fig. 8, we show how to link a lexical entry with its cognates, taking three of the cognates from the girl (a) example. Our model contains the symmetric *hasCognate* object property which links together an object of type *LexicalEntry* with objects of type *Cognate*, which is a subtype of *LexicalEntry* itself[10].

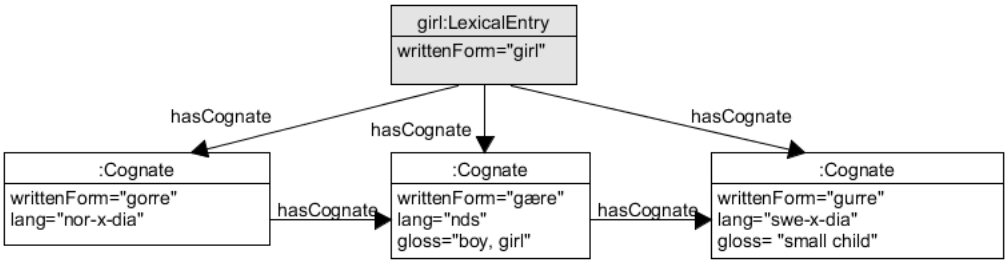

Figure 8: Cognates of girl.

By rendering explicit many of the etymologically (and in general the diachronically) salient aspects of the original textual data we are able to represent etymological lexical information in a way that makes it easier to query and, in general, more computationally actionable and accessible than it might otherwise have been as free text. We chose to make certain theoretical entities explicit classes in our model (such as the classes *Etymon* and *Etymology*) because we felt potential consumers of such lexical datasets would find it useful to have the possibility of querying for these classes and searching for different attributes belonging to them (and here it should be pointed out that the second author is a domain expert in lexicography). The fact that we have have chosen the popular RDF data framework also ensures interoperability for datasets encoded using our new model.

4 Conclusion

In this article we have presented an RDF-based model for representing lexical data using a perdurantist approach in order to make the diachronic aspects of the data more accessible. We have applied the model to some concrete examples from the domain of etymology to illustrate its viability. We are currently testing the lemonEty model by using it to encode a large and varied number of test examples from various sources. Although certain parts of the model will likely undergo subsequent changes, the core elements, as we have presented them in this article, are intended to be stable. The model we have presented is specifically based on RDF, and takes advantage of several features of the RDF framework that may not be as relevant for other standards, and so we would like, in further work, to identify a core set of classes and properties for encoding etymological/diachronic data that could be common to lemonEty as well as LMF and TEI.

[10]The language tags for Norwegian and Swedish (“nor-x-dia”) and (“swe-x-dia”) respectively combine the IS0 639-3 tags with a private tag denoted by “-x-”, followed by “dia” as per BCP 47 (Phillips and Davis (2009)). This is necessary in order to express the underdefined dialect varieties of these languages mentioned by the author of the entry.

References

Allen, J. F. (1983). Maintaining Knowledge About Temporal Intervals. *Communications of the ACM 26 November 1983*, 832–843.

Bellandi, A., F. Boschetti, A. M. Del Grosso, A. F. Khan, and M. Monachini (2017 to appear). Provando e riprovando modelli di dizionario storico digitale: collegare voci, citazioni, interpretazioni. In *AIUCD 2017 – Book of Abstracts, 24-28 January 2017*, Rome, Italy.

Bowers, J. and L. Romary (2016). Deep encoding of etymological information in TEI. *TEI. Journal of the Text Encoding Initiative* (10). Available at `https://doi.org/10.4000/jtei.1643`.

Khan, F., J. E. Díaz-Vera, and M. Monachini (2016). Representing Polysemy and Diachronic Lexico-Semantic Data on the Semantic Web ? In *Proceedings of the Second International Workshop on Semantic Web for Scientific Heritage co-located with 13th Extended Semantic Web Conference (ESWC 2016), Heraklion, Greece, May 30th, 2016.*, Volume V-1595, pp. 37–46.

Krieger, H.-U. (2014). A Detailed Comparison of Seven Approaches for the Annotation of Time-Dependent Factual Knowledge in RDF and OWL. In *Proceedings of the 10th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (held in conjunction with LREC 2014)*. European Language Resources Association.

McCrae, J., D. Spohr, and P. Cimiano (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *The semantic web: research and applications*, Extended Semantic Web Conference, pp. 245–259. Springer.

Peroni, S. and D. Shotton (2012, December). FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web 17*, 33–43.

Phillips, A. and M. Davis (2009). *BCP 47, RFC 5646 – Tags for Identifying Languages*. IETF. Published: BCP 47 Standard, see http://www.rfc-editor.org/rfc/bcp/bcp47.txt.

Salmon-Alt, S. (2006). Data structures for etymology: towards an etymological lexical network. *Bulletin de linguistique appliquée et générale 31*, 101–112.

Skeat, W. W. (1910). *An etymological dictionary of the English language. (4th ed.)*. Oxford Clarendon Press.

Welty, C. and R. Fikes (2006). A Reusable Ontology for Fluents in OWL. In *Proceedings of the 2006 Conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006)*, Amsterdam, The Netherlands, The Netherlands, pp. 226–236. IOS Press.