Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020), pages 15–22
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Representing Temporal Information in Lexical Linked Data Resources

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

Abstract

The increasing recognition of the utility of Linked Data as a means of publishing lexical resources has helped to underline the need for RDF-based data models with the flexibility and expressivity to be able to represent the most salient kinds of information contained in such resources as structured data; this includes, notably, information relating to time and the temporal dimension. In this article we describe a perdurantist approach to modelling diachronic lexical information which builds upon work which we have previously presented and which is based on the ontolex-lemon vocabulary. We present two extended examples, one taken from the Oxford English Dictionary, the other from a work on etymology, to show how our approach can handle different kinds of temporal information often found in lexical resources.

Keywords: linguistic linked data, diachronic lexical data, perdurantism

1. Introduction

The difficulties of representing relationships that change with time – also referred to as *fluents* or *diachronic relations* in the literature – in RDF have been by now well-rehearsed (Welty et al., 2006). The core problem here, of course, is that the RDF framework does not allow us to simply add an extra temporal parameter to binary and unary properties: something that would otherwise make modelling diachronic relations fairly straightforward. A number of different ‘workarounds’ have been proposed to deal with this situation[1]. There is, however, no single one size fits all solution that will work in every case and different solutions are better suited to different use cases. In this article our focus will be on lexical data, and in particular data that derives from legacy resources including dictionaries and scholarly works on meaning change. In the course of the article we will look at some of the different ways in which such data can carry a temporal dimension, including indirectly, through the use of citations and attestations. We will then propose the use of a perdurantist design pattern for representing this temporal information. Our intention with this submission is to rouse interest in the perdurantist approach to modelling lexical change in light of the work which is going on both in the W3C Ontolex community and in a number of projects and use cases which have recently arisen, and to elicit feedback from the Linguistic Linked Data community in order to help determine the variety of use cases which the approach is able to handle well and those which it cannot.

2. Temporal Information in Lexical Datasets

The increasing recognition of the utility of Linked Data as a means of publishing lexical resources – thanks in large part to projects such as ELEXIS (Krek et al., 2018) and LiLA (Passarotti et al., 2019) – has helped to underline the need for RDF-based data models with the flexibility and expressivity to be able to represent the most salient kinds of information contained in such resources as structured data: this includes, notably, information relating to time and the temporal dimension. Time is a central concern of certain kinds of lexical resource, this is most obviously true of etymological and historical dictionaries, but the inclusion of temporal information in lexical resources is by no means limited to such specialist works, and etymologies in particular are found in a wide range of dictionaries and lexical datasets. In previous related work we have looked at how to represent etymologies, viewed as hypotheses about word histories, explictly both in RDF (Khan, 2018) and in the Lexical Markup Framework (Khan and Bowers, 2020). However, temporal information is not always explicitly included in the form of an etymology. For instance, it is also common for resources to list the senses in each lexical entry in some order of temporal precedence[2]. Other resources include descriptions of the semantic shift processes which led from temporally antecedent senses to subsequent senses; yet others mark senses (or forms, etc) as obsolete and/or give some basic information on the time period in which a sense or form (or grammatical construct) was in use or was most commonly in use. In fact, in these and in other contexts, one frequently finds reference to a particular historical stage of a language, such as Old French or Middle English, something which also helps to group together lexical phenomena in time. Many lexical resources, especially more authoritative or scholarly dictionaries, also give citations for separate senses and even of forms (or, in fact, for any interesting or salient lexical information which has changed over time, such as verb transitivity or the historic existence of now obsolete noun declensions). These citations help to locate senses (and forms etc) in time, and one of the central aims of the current work is to show how such information can be efficiently integrated into the RDF encoding of an entry. It is worth noting that the information originating from citations also tends to be less vague than

[1] See for instance `https://www.w3.org/TR/swbp-n-aryRelations/`

[2] For instance the Oxford English Dictionary describes its entries as being "structured to show the evolution of senses and uses over time". cite `https://public.oed.com/how-to-use-the-oed/glossary/`

other kinds of temporal data present in lexical resources, as will be illustrated by the examples which we will present below. It is a hallmark of most lexical temporal data that it tends to be vague, sometimes very vague (for instance in the case of proto-languages with no written testimony). In the majority of cases it is hard to pinpoint the year, or even the century, that a certain sense or form or word began to be used and/or stopped being used. In other cases it is hard to fix the historical periods in which entire languages were spoken. If we are going to potentially reason with such data we need an approach that takes into consideration the vague nature of such data. We discuss this further and propose a solution in Section 3.1..

2.1. Previous Work

The work in this paper builds upon, and in many cases presents in altered form, ideas and proposals which we have published previously. We initially presented the idea of extending the *lemon* (McCrae et al., 2011) and afterwards the *ontolex lemon* (McCrae et al., 2017) models using perdurants in order to represent temporal information in (Khan et al., 2014) and (Khan et al., 2016), using the resulting model to encode a linguistic dataset dealing with the evolution of emotion terms in Old English in (Khan et al., 2018)[3]. Many of the properties and classes in that version have been modified as a result of working on modelling various different lexical datasets. We have also worked on the more specific case of modelling etymologies as linked data (Khan, 2018). However the current work focuses much more on the representation of time and temporal intervals than those previous works, developing our approach by focusing on two extended examples.

2.2. Case Studies

2.2.1. The word *girl*

For our first case study we will look at the Oxford English Dictionary (OED) entry for the word *girl* (OUP, 2008). We chose the OED because of its status as an authoritative work of descriptive, historical lexicography and because of the comprehensive, and therefore challenging (from the modelling point of view), nature of its entries. The OED epitomizes the type of the historical reference dictionary in which individual forms and senses are attested with reference to a historical corpus of citations. In the current case we have chosen the word *girl* because of the somewhat surprising fact (regularly cited in books on etymology, at least in the English language), that it was originally used to refer to a "child of either sex; young person" and not just, as is often the case today, to "a young or relatively young woman". The OED entry starts by giving the standard pronunciation of *girl* in IPA in both British and American English. Next a list of forms is given, we will come back to this shortly. Following this, frequency, origin and etymological information is listed (the etymology of the word is especially obscure and its origin is given as unknown). Then twelve separate senses for the word are given; these are classified into two groups. The first group is labelled as being "[s]enses relating to a person", and the second group (containing only two senses) is labelled as "other senses". Finally a list of phrases and derivatives is presented. Note that the senses are listed in historical order, although in some cases a sense may have a more recent subsense listed immediately below it and before other historically later senses. The forms are listed as follows in the entry:

> ME **garl**, ME **geerl**, ME **gerl**, ME (18– chiefly *Irish English* and *nonstandard*) **gurl** , ME–15 **gerle**, ME– 16 **girle**, ME–16 **gyrle**, 15 **gierle**, 15 **gurle**, 15 **gyrll**, 15–16 **guirle**, 15–16 **gyrl**, 15–**girl**, 16 **garle**, 16 **gerreld**; *Caribbean* 19– **gyal**, 19– **gyul**.

Here ME stands for Middle English, a time period which the OED describes as running from 1150 CE to 1500 CE[4]. The numbers 15, 16, 18, and 19 refer to the 1500s (i.e., 1500-99)[5], the 1600s (i.e., 1600-99), etc. The senses are listed under numbers and, in the case of subsenses, lower case roman letters. Each sense starts with a definition and some other related information before presenting a list of historical citations for that sense. Below we give the first sense in full:

1. Chiefly in *plural*. A child of either sex; a young person. Now *Irish English (Wexford)*. **knave girl** *n.* a boy.

   *C*1300 *St. Thomas Becket* (Laud) 76 in C. Horstmann ***Early S.-Eng. Legendary*** (1887) 108 (*MED*) þe Amirales douȝter was In þe strete þare-oute, And suyþe gret prece of gurles and Men comen hire al-a-boute.

   *C*1400 (→a1376) W. LANGLAND ***Piers Plowman*** (Trin. Cambr. R.3.14) (1960) A. XI. 132 (*MED*) Gramer for girles [*v.rr.* gurles, gerles, childeryn] I garte ferst write, And bet hem wiþ a þaleis but ȝif þei wolde lerne.

   *C*1400 (→?a1300) ***Kyng Alisaunder*** (Laud) (1952) 2798 (*MED*) Men miȝtten seen þere hondes wrynge..Wymmen shrikyng, gyrles gradyng.

   *C*1405 (→c1387–95) G. CHAUCER ***Canterbury Tales Prol.*** (Hengwrt) (2003) l. 664 In daunger hadde he at his owene gyse The yonge gerles of the diocise, And knew hir conseil, and was al hir reed.

   *a*1475 *Bk. Curtasye* (Sloane 1986) l. 328 in ***Babees Bk.*** (2002) I. 308 Ne delf þou neuer nose thyrle With thombe ne fyngur, as ȝong gyrle.

[3] Ours was not of course the first attempt to model diachronic lexical or etymological information in linked data. See for instance (Moran and Bruemmer, 2013), (De Melo, 2014). (Khan, 2018) presents a fuller overview of related work.

[4] `https://public.oed.com/blog/middle-english-an-overview/`

[5] `https://public.oed.com/how-to-use-the-oed/key-to-symbols-and-other-conventions/`

?*a*1475 *Ludus Coventriae* (1922) 171 (*MED*) Here knaue gerlys I xal steke.

*a*1827 J. POOLE *Gloss.* in T. P. Dolan & D. Ó . Muirithe ***Dial. Forth & Bargy*** (1996) 49 *Gurl, gurlès*, a child, a girl.

1996 T. P. DOLAN & D. Ó. MUIRITHE ***Dial. Forth & Bargy*** 25 *Gurl*, a child of either sex.

Note that the *C* before a date means 'circa', *approximately*, and *a* means 'antes', *before* or *prior to*. The question mark indicates an uncertain date. In cases where there are two dates, one after the other, with the second in parenthesis following an arrow symbol, e.g.,*C*1400 (→a1376), the first date refers to the dating of a manuscript, and the second, the date of composition[6].

2.3. *Sad*

The next example which we will model is adapted from the etymology for the word *sad* given in Philip Durkin's *Oxford Guide to Etymology* (Durkin, 2009). We have chosen this example in order to illustrate how to use our approach to model historical sense shifts (although it can be easily adapted to show the evolution of forms as well for instance). The example regards the meaning shift undergone by the English word *sad* which originally meant 'satisfied' or 'full' in Old English and now has the principal meaning of 'sorrowful, mournful' (this meaning is recorded, as Durkin points out, in a source dated *a*1300). Durkin hypothesises a process of semantic shift that takes place in three stages, via an intermediate sense meaning 'weary or tired of something', as follows:

satisfied, having had one's fill (of something)

[metaphorized and narrowed] > weary or tired (of something)

[broadened] > sorrowful, mournful.

3. Our Approach

In this section we will outline and motivate our particular approach to modelling diachronic lexical data in RDF. The idea, in a nutshell, is twofold: firstly, we propose the use of *qualitative* intervals to model temporal vagueness in lexical data; then, secondly, we define a 'perdurantist' version of certain classes in the ontolex-lemon model in order to allow lexical entries, senses, etc, to each have a 'lifespan' as well as temporal parts.

3.1. Introduction to Qualitative Temporal Intervals and Allen Relations

One way of dealing with the kinds of temporal vagueness which we have previously mentioned is to work with so called qualitative constraints and to focus on the relative temporal positions of different points and intervals on a given timeline, that is, in addition to leveraging whatever exact quantitative information that we might also actually possess. In this case we can make use of Allen relations between temporal intervals (Allen, 1983) in order to reason over such data, and fortunately for us these relations are already encoded within the popular temporal ontology OWL-time[7], see Figure 1. Furthermore there already exists a set of SWRL rules which allow for reasoning over data that describes intervals qualitatively using these relations (Batsakis et al., 2017).

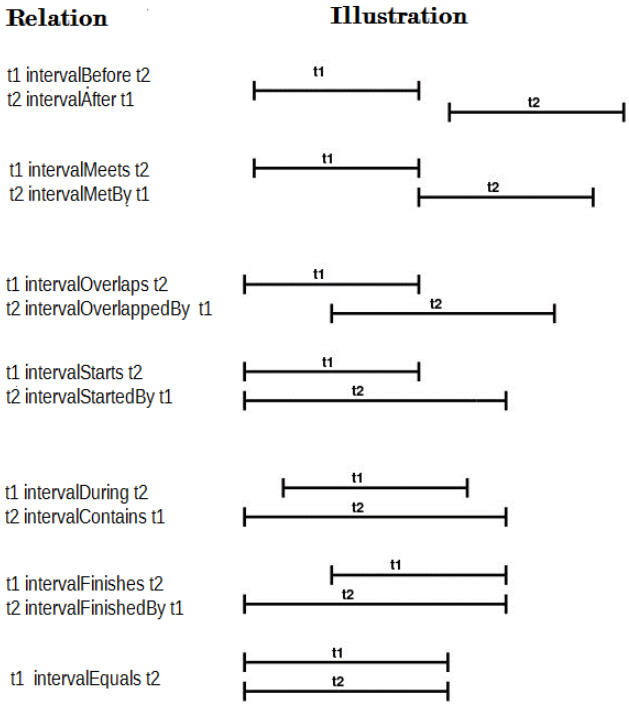

Figure 1: Allen relations used in OWL-time.

As well as defining, or (in cases when they already exist) reusing, standard temporal intervals such as the *13th Century CE* or *1066* with OWL-time[8] we can also define language specific intervals such as *Middle-English* whose definitions are much more subject to variation across different sources: in our case, for Middle-English, we use the definition given by the OED of *(1150, 1500)*. So, for instance, we can define the 14th century, or rather the 1300s as, running from 1300-1399 using the OWL-time vocabulary as follows.

Listing 1: Definition of `13`

```
:13 a owl-time:Interval ;
   owl-time:hasBeginning [
      a owl-time:Instant ;
      owl-time:inXSDDateTimeStamp
      "1300-01-01T00:00:00Z"^^xsd:dateTimeStamp
      ];
   owl-time:hasEnd [
       a owl-time:Instant ;
      owl-time:inXSDDateTimeStamp
      "1399-12-31T23:59:59Z"^^xsd:dateTimeStamp
      ] ;
   owl-time:intervalMeets :14 ;
   owl-time:intervalMetBy :12 ;
   rdfs:label "(1300-1399)"@en.
```

[6] https://public.oed.com/blog/dating-middle-english-evidence-in-the-oed/

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

[8] We are using OWL-time but any other temporal vocabulary/ontology with similarly defined properties and classes can also be used.

Note that we have defined the interval both in terms of its beginning and end points (using the xsd:dateTimeStamp datatype property) as well as by relating it to two other intervals using the object properties intervalMeets and intervalMetBy. In addition we can also define the time interval, ME, mentioned in the OED entry and corresponding to the time interval in which Middle English was spoken, by specifying its start and end years and its relationships with other (named) intervals. So in this case we state that it overlaps the interval 11 (the 10th century), contains the interval 12 (the 11th century), and is finished by the interval 13.

Listing 2: Definition of the Middle English interval

```
:enm_interval a owl-time:Interval ;
   owl-time:hasBeginning
      [ a owl-time:Instant ;
      owl-time:inXSDDateTimeStamp
      "1150-01-01T00:00:00Z"^^xsd:dateTimeStamp
      ] ;
   owl-time:hasEnd
      [ a owl-time:Instant ;
      owl-time:inXSDDateTimeStamp
      "1399-12-31T23:59:59Z"^^xsd:dateTimeStamp
      ] ;
   rdfs:label "Middle English (1150-1300)"@en ;
   owl-time:intervalOverlappedBy :11 ;
   owl-time:intervalContains :12 ;
   owl-time:intervalFinishedBy :13 .
```

We can, on the basis of these prior 'building block' intervals, once again use Allen relations to define further interval combinations and thereby capture other salient intervals such as, for instance, the interval ME-16 mentioned in the list of historical forms given above for *girl*.

Listing 3: Definition of ME-16

```
:ME-16 a owl-time:Interval;
   owl-time:intervalContains :12 , :13 ,:14 , :15 ;
   owl-time:intervalFinishedBy :16 ;
   owl-time:intervalStartedBy :enm_interval .
```

The case of the interval 18- above is a little bit trickier in that it requires the specification of one point as the present. Here we have chosen a point in the current year (although the point in question could be the date of the publication of the resource being modelled for instance).

Listing 4: Setting a present point

```
:present a owl-time:Instant;
owl-time:inDateTime [
   a owl-time:DateTimeDescription ;
   owl-time:unitType owl-time:unitYear ;
   owl-time:year "2020"^^xsd:gYear ;
  ] ;
rdfs:comment "The Present"@en .
```

Assuming then that we have already defined the intervals 18, 19, and 20, we can define 18- as in interval that is started by the 17th century, contains the 19th century and is overlapped by the 21st century.

Listing 5: Modelling *18-*

```
:18- a owl-time:Interval ;
   owl-time:hasEnd :present ;
   owl-time:intervalContains :19 ;
   owl-time:intervalStartedBy :18 ;
   owl-time:intervalOverlaps :20 ;
   rdfs:label "(1800-)"@en .
```

One of the biggest challenges in the present context relates to the use of the preposition 'circa' as in *C1300* or *C1400*, which obviously is used to codify 'fuzziness'. The essential thing, however, is to pick a modelling approach and to remain with it consistently throughout a dataset. For instance the option which we have chosen is to model Cd, where d is a date, as being contained in an interval of 10 years before and after d.

Listing 6: Modelling *circa*

```
:circa_1300 a owl-time:Interval ;
   owl-time:intervalContains [
   a owl-time:Interval ;
   owl-time:hasBeginning [
      a owl-time:Instant ;
      owl-time:inXSDDateTimeStamp
      "1290-01-01T00:00:00Z"^^xsd:dateTimeStamp ];
   owl-time:hasEnd [
      a owl-time:Instant ;
      owl-time:inXSDDateTimeStamp
      "1310-01-01T00:00:00Z"^^xsd:dateTimeStamp ];
   ].
```

It will also be useful to define an interval EnglishInterval corresponding to the time during which the English language was spoken and which is ended by the present and 'contains' other, previously defined, intervals such as enm_interval; every English lexical entry in the resource can then be interval contained by EnglishInterval. We can further temporally locate this interval by adding statements to the effect that English language was spoken *after* proto-Germanic, while also taking into account an intervening proto-English period by defining the appropriate intervalBefore Allen relation.

4. Perdurantism v. Endurantism

Although the perdurantist approach has recently become popular amongst computer scientists and knowledge engineers for reasons that are in large part practical, it has its basis in a well established philosophical theory. *Perdurantism*, also known as *four dimensionalism* (in its is most common formulations), argues that time should be treated analogously to the spatial dimensions so that objects can have temporal extension just as they can have length, breadth, and width: this means that they can have (spatio-)temporal parts in the same way that we usually describe them as having spatial parts. Naturally the notion of *temporal part* is fundamental to the perdurantist approach; we will use the following definition, given by (Sider, 1997).

> x is a **temporal part** of y during interval $T =_{df}$ (i) x is a part of y at every moment during T; (ii) x exists during T, but only during T; and for any sub-interval t of T, x overlaps every part of y at t.

Perdurantism by treating people, animals, and things in general like processes, only parts of which exist at different times, has the clear shortcoming that it can be unintuitive and difficult to understand. For a perdurantist, the temporal part of a person that *perdured* (to use the technical term) through, say, the month of January, is a part of them in the same or in an exactly analogous way as their limbs or their organs, things which constitute a physical part of them. However, the vast majority of people simply don't think in this way about the world. Perdurantism constrasts with the philosophical approach known as *endurantism*, or *three dimensionalism*, and which many philosophers argue is a more natural way of thinking about existence through time. Indeed, according to most endurantist acccounts, things, objects, etc, instead of being only partially present at one single point of time, are wholly present at each instant of their existence[9] (Sider, 1997). The perdurantist approach, however, has other features which compensate for its relative conceptual oddity. Most importantly, it helps to resolve a number of longstanding metaphysical conundrums relating to change over time. In addition, it provides a very useful way of modelling vagueness, and is also able to meet several of the challenges raised against more traditional theories of time and change by the theory of special relativity (Effingham, 2012). Note that although the names 'three-dimensionalism' and 'four-dimensionalism' might suggest that we are only dealing with 'concrete' objects, namely those that occupy a continuous physical portion of space, this is not in fact the case and perdurantism has in fact been applied to musical works (Caplan and Matheson, 2006) and institutional objects (Hansson Wahlberg, 2014). Our proposal in this work is to apply it to language and linguistic phenomena. We can explain the precise kind of perdurantist approach which we take in this paper through the provision of a simple (non-linguistic) example.

The relation capitalCity, which links together an urban location with a state, can be modelled as a diachronic relation or fluent since it can change over time, i.e., a country can have different capital cities at different points in time. Indeed this is the case with the nation of Italy which has had three separate capital cities, or seats of government, since its unification in 1861. These were/are: Turin from 1861 to 1865 (period t1), Florence from 1865 to 1871 (the period t2), and Rome from 1871 to the current day (the interval t3). One perdurantist approach to modelling this situation (and indeed the one which we will propose for lexical data below) is the following:

- We define separate time slices (or temporal parts) TurinCapital, FlorenceCapital, RomeCapital of each of the Italian cities mentioned above, Turin, Florence, Rome respectively. Note that each of these time slices are also typed as cities, e.g., city(Turin) and city(TurinCapital)

- We relate the temporal parts of these cities together with their wholes using the property temporalPartOfwhich relates a perdurant together with another perdurant of which it is a temporal part, i.e., temporalPartOf(TurinCapital, Turin), temporalPartOf(FlorenceCapital, Florence), and temporalPartOf(RomeCapital, Rome)

- Each of these timeslices is associated with its lifespan using the temporalExtent property which relates a perdurant together with the interval during which it exists, i.e., temporalExtent(TurinCapital, t1), temporalExtent(FlorenceCapital, t2), temporalExtent(RomeCapital, t3).

- We also create time slices of Italy for each of these periods, Italyt1,Italyt2 and Italyt3 where: temporalExtent(Italyt1,t1),temporalExtent(Italyt2,t2), and temporalExtent(Italyt3,t3)

- Finally we relate these timeslices of Italy using the capitalCity relation: capitalCity(TurinCapital,Italyt1), capitalCity(FlorenceCapital,Italyt2), capitalCity(RomeCapital,Italyt3).

This is a version of the kind of perdurantist approach initially proposed for RDF in (Welty et al., 2006); our version is more directly based on (Krieger, 2014). However it should be noted that when it comes to creating diachronic versions of relationships between a lexical entry and its lexical senses and forms we can make a simplification. These latter are usually defined as being dependent on lexical entries, not least in the ontolex-lemon model where a form or a lexical sense cannot be shared by more than one entry. That is although the relation sense which relates a lexical entry together with each of its senses is a diachronic relation, each of the senses in question is parasitic on that entry, that is, the lifespan of a sense is necessarily contained in the lifespan of an entry, and similarly with forms. Our proposal then is, when it comes to perdurants versions of these classes, to define sense as holding between lexical entries and senses (and form as holding between lexical entries and forms) rather than timeslice of these, as in the case of capitalCity.

4.1. First Definitions

Our first definitions are, as anticipated, perdurantist subclasses of the ontolex-lemon classes Lexical Entry, Lexical Sense, and Lexical Form; these are pLexical Entry, pLexical Sense, and pLexical Form respectively.

$$\text{pLexical Entry} \sqsubseteq \text{Lexical Entry}$$

$$\text{pLexical Sense} \sqsubseteq \text{Lexical Sense}$$

$$\text{pLexical Form} \sqsubseteq \text{Lexical Form}$$

In order to define these classes we will make use of the new object property mentioned above, temporalExtent, whose range is the owl-time class time:Interval, and whose purpose is to relate a perdurant to its temporal dimension; thereafter we impose the restriction that each member of the p- classes is related to exactly one time:Interval individual via the property temporalExtent.

$$\begin{aligned}&\text{pLexical Entry}, \text{pLexical Sense}, \text{pLexical Form}\\ &\sqsubseteq \;\; = 1\ \text{temporalExtent.}(\text{time:TemporalEntity})\end{aligned}$$

$$\exists\text{temporalExtent.}\top \sqsubseteq \text{time:Interval}$$

4.1.1. Modelling the *girl* Example

In this section we will model the OED entry for *girl* using the classes and properties just defined. Note that for reasons of space and clarity of explanation we will focus on those parts of the entry that show the use of our new classes and properties rather than giving a comprehensive encoding of the entry using elements already available in

[9] T he standard approach to ontology modelling and knowledge engineering can be described as endurantist.

the ontolex-lemon module and the recently published lexicographic extension of the latter[10](Bosque-Gil et al., 2017) (so for instance we leave out the canonical form and phonetic forms below).

We start by defining the entry as a pLexicalEntry with an associated temporal extent (girl_time) and relating the entry with its forms using the (non-fluent) ontolex-lemon property lexicalForm

Listing 7: The entry *girl* and its forms.

```
:girl rdf:type :pLexicalEntry ;
    :temporalExtent :girl_time;
    ontolex:lexicalForm :garl_form ,
                 :garle_form ,
                 :geerl_form ,
                 :gerl_form ,
                 :gerle_form ,
                 :gerreld_form ,
                 :gierle_form ,
                 :girle_form ,
                 :guirle_form ,
                 :gurl_form ,
                 :gurle_form ,
                 :gyal_form ,
                 :gyrl_form ,
                 :gyrle_form ,
                 :gyrll_form ,
                 :gyul_form .
```

Note that each form is contained within the temporal extent of the entire entry girl_time.

The first two forms which we will look at from the list given in Listing 7 are *garl* and *girle*, the first of which the OED tells us was in use during the Middle English period and the second of which was in use from the Middle English period through to the 15th century. We model these two as in Listing 8: that is after we introduce them as elements of type pForm we thereafter associate them with a given temporal interval using the property temporalExtent. In this case the two periods are those defined above, namely emn_interval and ME-16.

Listing 8: The forms *garl* and *girle*

```
:garl_form rdf:type :pForm ;
   ontolex:writtenRep "garl";
         :temporalExtent :enm_interval .

:girle_form rdf:type :pForm ;
   ontolex:writtenRep "girle";
   :temporalExtent :ME-16 .
```

In the case of the forms *gurl*, *gyal*, and *gyul* we have extra dialectal and geographical information to take into consideration, that is alongside the purely temporal information which has been provided. Indeed the form *gurl* has two separate temporal parts. The first part perdures over the ME period; the second part, which perdures over an interval which starts in the 18th century and ends in the present, is described as being "chiefly *Irish English* and *nonstandard*"). In order to model cases such as the latter, we have added a new datatype property hasUsageNote to our model; this allows for the encoding of geographical and dialectal constraints on the usage of a lexical element as free text. Our choice in this regard was informed by the fact that such information is often difficult to encode as structured data using formalisms such as RDFS and OWL (for instance in the present case how would we encode the adverb *chiefly* in the description of the form *gerl*? It would be tricky to come up with a general way of encoding such descriptions that would please everyone). However this also leaves the door open to encoding such information in other ways and using other properties, for instance when it comes to purely geographical variations (a specific case which we plan to look at in future work). In Listing 9, therefore, we model the temporal interval associated with gurl_form as consisting of two separate temporal part: relating it to the interval enm_interval using intervalStartedBy and the interval 18- using intervalFinishedBy.

Listing 9: *gurl*

```
gurl_form a :pForm ;
   :hasTemporalPart :gurl_form_IEN , :gurl_form_ME ;
   :temporalExtent [
      a owl-time:Interval;
      owl-time:intervalStartedBy :enm_interval;
      owl-time:intervalFinishedBy :18-;
       ];
   ontolex:writtenRep "gurl"@en .

:gurl_form_ME a :pForm ;
   :temporalExtent :enm_interval .

:gurl_form_IEN a :pForm ;
   :temporalExtent :18- ;
   :hasUsageNote
   "chiefly Irish English and nonstandard"@en .
```

Now we will move onto the temporal modelling of the information contained in the senses of the entry[11]. We can use this information to delimit the temporal interval associated with the sense in time. In what follows we focus on the first sense of the word, the one with the definition '[a] child of either sex; a young person. Now *Irish English (Wexford).*' This usage of the word is attested in texts such as *Piers Plowman* and the *Canterbury Tales* and the 13th century text *Ludus Conventriae*. The entry for this sense also identifies a temporal part of this sense with a particular quality not shared by the whole: that of being limited to a certain geographically defined dialect (Irish English, or more precisely the dialect of Wexford). And in fact this sense continues to be used up till the present day. We can model this as in Listing 10.

Listing 10: First *gurl* sense

```
:girl ontolex:sense :sense_I1 ,
   :sense_I1_irish_english .

:sense_I1 a :pLexicalSense ;
   :hasTemporalPart :sense_I1_irish_english ;
   :temporalExtent :sense_I1_interval .

:sense_I1_irish_english :pLexicalSense ;
   :temporalExtent :sense_I1_irish_english_interval ;
   :hasUsageNote "Irish English (Wexford)"@en .
```

This brings us to the question of how to integrate the temporal information included in the illustrative examples in the OED entry. Note that in this article we will not discuss how to model the bibliographic information included with each illustrative example – something which we can do using pre-existing vocabularies as well as potentially new classes and properties – even though it would be (scientifically) valuable to have this kind of information in a structured

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

[11] Note that whenever we can assume that the ordering of senses is given in temporal order we can define precedence relations between them using, you guessed it, Allen relations.

form; this is purely for reasons of space but it is something we hope to address in further work. Instead we will use the temporal information included in the examples to enrich the description of the intervals associated with each sense.
In terms of the illustrative examples given in the entry we take those that allow us to give some kind of a (vague) temporal outline to the use of that sense. We therefore utilize the periods $C1300$, $a1475$ and 1996 for the main sense and the two periods/intervals[12] $a1827$ and 1996 for the second Irish English interval. Putting everything together we can define the two intervals in question as in Listing 11.

Listing 11: First *girl* sense intervals.

```
:sense_I1_interval a owl-time:Interval ;
   owl-time:hasEnd :present ;
   owl-time:intervalFinishedBy
      :sense_I1_irish_english_interval;
   owl-time:intervalContains
   :a1475 , :circa_1300 , :year_1996 .

:sense_I1_irish_english_interval a owl-time:Interval;
   owl-time:intervalFinishes :sense_I1_interval ;
   owl-time:intervalContains :a182 .
```

Another way of integrating this information into the entry is to view each of the illustrative examples as describing a very restricted sub-sense of the main sense (the main sense being in this case 'A child of either sex; a young person'): indeed **the** sub-sense of the main sense restricted to that particular use of the word. This allows us to create a new pSense for each example to which we can attach all the temporal information included in the illustrative example, and which can then be related to the main sense using the appropriate Allen relations. Finally it is important to note that a word or any lexical entry will always have some minimal temporal information associated with it thanks to the fact that it belongs to a language or a language stage to which we can usually associate a minimal of temporal information.

4.2. Modelling semantic shifts: the *sad* example.

The *sad* example given in Section 4.1.1. could be modelled explicitly using etymologies (and indeed the more extended history of the word presented in (Durkin, 2009) would be better modelled this way). However in cases where we want to track and describe a change in the sense of one word or the phonological changes in the pronunciation of a single form, it is often more convenient and efficient not to represent this via an etymology. Summarising the problem then our task is, given two lexical senses l_1, l_2 which have undergone a process of semantic shift of type s, to model the typed sense shift relationship *semanticShift* between them:

semanticShift (l_1, l_2, s)

We can also choose to make this a fluent, a diachronic relationship and add a temporal parameter, i.e., semanticShift (l_1, l_2, s, t). However in this case, as in the case of relationships such as parentOf, we believe that both modelling choices are natural, and we have decided therefore not to make it a fluent. We also introduce a new object property between senses senseShiftsTo which enables us to model the fact that one sense 'gives birth to' another. We use the following property chain to relate this to the properties mentioned previously:

$$\text{shiftSource} \circ \text{shiftTarget} .$$

We define the *sad* entry as having three separate senses

Listing 12: The entry for *sad*

```
:sad a :pLexicalEntry ;
   :temporalExtent [a owl-time:Interval ;
         :intervalFinishes :EnglishInterval ];
   ontolex:sense
   :sad_sense_1 , :sad_sense_2 , :sad_sense_3 .
```

Moreover we also define two shift objects relating together the second and third senses of the words. (We have used rdfs:comment here to describe each shift however we are currently working on developing a taxonomy of semantic shifts which we will present in future work and to which we can link such shifts).

Listing 13: Two semantic shifts for the senses of *sad*

```
:sad_shift_1 a :SemanticShift ;
   :shiftSource :sad_sense_1 ;
   :shiftTarget :sad_sense_2 ;
   rdfs:comment "metaphorized and narrowed"@en .


:sad_shift_2 a :SemanticShift ;
   :shiftSource :sad_sense_2 ;
   :shiftTarget :sad_sense_3 ;
   rdfs:comment"broadened"@en .
```

The three senses of sad which we have singled out (and which are of course far from being exhaustive for the word can be defined as follows.

Listing 14: The senses of *sad*

```
:sad_sense_1 a :pLexicalSense ;
   :senseShiftsTo :sad_sense_2 .

:sad_sense_2 a :pLexicalSense ;
   :senseShiftsTo :sad_sense_3 .

:sad_sense_3 a :pLexicalSense;
   :temporalExtent [a owl-time:Interval ;
         :intervalDuring :a1300 ].
```

5. Conclusions and Further Work

In this article we have attempted to consolidate and build upon work which we have previously introduced, with the aim, this time, of making a detailed and convincing case for the expressive advantages of the perdurant approach to modelling diachronic lexical information. A secondary, though related aim has been to demonstrate how naturally certain, common, kinds of temporally-enriched dictionary data can be modelled in this way. In future work, as we have mentioned above, we aim to integrate bibliographic information more generally, making use of vocabularies such as CiTO as well as making use of the new Frequency Corpus and Attestations extension of ontolex lemon currently being developed by the W3C ontolex group. We also plan to investigate the efficiency of temporal reasoning using the perdurantist approach along with vocabularies like OWL-time and the rules developed by (Batsakis et al., 2017) and to test it on different sizes of dataset; something which we had previously begun to do in (Khan et al., 2018). We are also planning, with colleagues, to undertake a more detailed

[12] Note that we have been modelling years as intervals throughout.

study of the different kinds of temporal information that are common found in dictionaries and legacy lexical resources in order to test the robustness and expressivity of our model.

6. Bibliographical References

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Batsakis, S., Petrakis, E. G., Tachmazidis, I., and Antoniou, G. (2017). Temporal representation and reasoning in owl 2. *Semantic Web*, 8(6):981–1000.

Bosque-Gil, J., Gracia, J., and Montiel-Ponsoda, E. (2017). Towards a module for lexicography in ontolex. In *LDK Workshops*, pages 74–84.

Caplan, B. and Matheson, C. (2006). Defending musical perdurantism. *The British Journal of Aesthetics*, 46(1):59–69.

De Melo, G. (2014). Etymological wordnet: Tracing the history of words. Citeseer.

Durkin, P. (2009). *The Oxford guide to etymology*. Oxford University Press.

Effingham, N. (2012). Endurantism and perdurantism. *The continuum companion to metaphysics*, pages 170–197.

Hansson Wahlberg, T. (2014). Institutional objects, reductionism and theories of persistence. *dialectica*, 68(4):525–562.

Khan, F. and Bowers, J. (2020). Towards a lexical standard for the representation of etymological data. In *Convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale*.

Khan, F., Boschetti, F., and Frontini, F. (2014). Using lemon to Model Lexical Semantic Shift in Diachronic Lexical Resources. Proceedings of the Workshop on Linked Data in Linguistics 2014 (LDL-2014).

Khan, F., Bellandi, A., and Monachini, M. (2016). Tools and instruments for building and querying diachronic computational lexica. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 164–171, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Khan, F., Díaz-Vera, J., and Monachini, M. (2018). Representing meaning change in computational lexical resources: The case of shame and embarrassment terms in old english. *Formal Representation and the Digital Humanities*, page 59.

Khan, A. F. (2018). Towards the representation of etymological data on the semantic web. *Information*, 9(12):304, Nov.

Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen, B. S., Tiberius, C., and Wissik, T. (2018). European lexicographic infrastructure (elexis). In *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts*, pages 881–892.

Krieger, H.-U. (2014). A detailed comparison of seven approaches for the annotation of time-dependent factual knowledge in rdf and owl. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, page 1.

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

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. pages 587–597, September.

Moran, S. and Bruemmer, M. (2013). Lemon-aid: using lemon to aid quantitative historical linguistic analysis. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 28 – 33, Pisa, Italy, September. Association for Computational Linguistics.

OUP. (2008). *Oxford English Dictionary, Third Edition*. Oxford: Oxford University Press.

Passarotti, M. C., Cecchini, F. M., Franzini, G., Litta, E., Mambrini, F., and Ruffolo, P. (2019). The lila knowledge base of linguistic resources and nlp tools for latin. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, pages 6–11. CEUR-WS. org.

Sider, T. (1997). Four-dimensionalism. *The Philosophical Review*, 106(2):197–231.

Welty, C., Fikes, R., and Makarios, S. (2006). A reusable ontology for fluents in owl. In *FOIS*, volume 150, pages 226–236.