Representing Polysemy and Diachronic Lexico-Semantic Data on the Semantic Web

Fahad Khan[1,3], Javier E. Díaz-Vera[2], Monica Monachini[3]

[1] University of Ca' Foscari Venezia, Venice, Italy
[2] University of Castilla-La Mancha, Spain
javierenrique.diaz@uclm.es
[3] Istituto di Linguistica Computazionale "A. Zampolli" - CNR, Pisa, Italy
{fahad.khan, monica.monachini}@ilc.cnr.it

Abstract. In this article we will outline two different vocabularies, both extensions of the *lemon* model, for representing diachronic lexico-semantic data on the Semantic Web. This is especially useful for representing the evolution of scientific terminologies where many terms are polysemous and or imported from other languages. The first vocabulary, *polyLemon*, allows for the representation of data about polysemy; the second, *lemonDIA* the representation of meaning shift over time.

1 Introduction

When it comes to the representation of the evolution of scientific terminologies using formal models such as the Resource Data Framework (RDF), it is important to be able to describe the original non-technical meanings that scientific terms initially had – and which in many cases they continued to have – as well as to represent the process of meaning shift that took place over time and which led to such terms taking on a technical meaning.

In this article we will describe two different vocabularies/models for representing lexico-semantic data on the Semantic Web: in the first case from a synchronic perspective, that is in terms of the polysemic structure of individual lexical entries, and in the second case from a diachronic perspective, namely, by describing the shifts in meaning of scientific or technical terms in a given language, from their pre-theoretical origins in the same or other languages – or indeed their theoretical origins in other languages. Both of these vocabularies build upon the *lemon* model[4]. The first vocabulary which we will discuss, in Section 2, is called *polyLemon* and is intended to handle the sometimes complex relations which can often hold between the various senses of a word; this comes in particulary important when dealing with scientific terms that are still polysemous with regard to pre-scientific/pre-technical meanings. The second vocabulary, *lemonDIA* is discussed in Section 3 and deals with a vocabulary for

[4] Note that in this article we will assume that the reader is already familiar with RDF and the *lemon* model for lexical resources [9]. However for those new to the model, *the Lemon Cookbook* [8] makes an excellent and very accessible introduction.

describing diachronic semantic phenomena, in this case, semantic shifts between the different meanings of a lexical entry over time. Section 4 illustrates the use of *lemonDIA* to represent emotion expressions from a lexicon of Old English (OE).

2 *polyLemon*

We became aware of the strong need for a specialised RDF vocabulary based on *lemon*, to represent the complex interrelationships between different, related, senses of the same word, while working on the modelling into linked data of such legacy resources as the Liddell-Scott Greek-English lexicon [6] and the Lewis-Short Latin-English lexicon. We felt that such a specialised vocabulary would be particularly useful in the context of the study of scientific terminologies, as it allows for analysis of the essentially polysemic nature of scientific and technical language. Towards the general aim of better representing such phenomena of polysemy in RDF, we developed a set of classes and object properties which extend the basic *lemon* model, and help to describe the tree like structure of polysemous lexical entries. We call this tree-like structure, composed of individual senses of the same word, *the sense tree*.

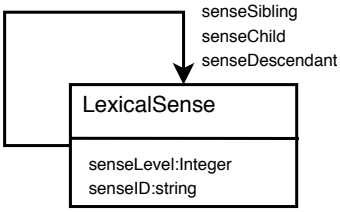

Fig. 1. The *polyLemon* model.

We briefly describe the main different elements of the model. The data property senseLevel serves to specify the level in the sense tree of the sense in question, and the senseID property to specify the particular identifier that a sense has been given in a resource. The object properties senseSibling, senseChild, and senseDescendant help to describe the interrelations between different senses in a sense tree. The senseChild relation connects a more general sense to a less general or more specialised sense. That is, if a sense $s2$ is in a senseChild relation of with $s1$, then $s2$ has a narrower meaning than $s1$ and in addition either $s1$ and $s2$ belong to the same entry or to two entries that are related by a morphological derivation relation.

We will illustrate the use of *polyLemon* with an example from the Middle Liddell (ML), an abridged version of the famous Greek-English Liddell Scott lexicon, which, in its full version, is generally regarded as one of the most comprehensive and authoritative of ancient Greek lexical resources. This variety of nested structure is generally to be found in many specialist lexical resources.

Below we give the ML entry for the adjective ἀληθής (alethes) meaning "unconcealed, true":

– **ἀληθής** α privat., λήθω = λανθάνω
  - *unconcealed, true:*
    (I) *true*, opp. to ψευδής, **Hom.**; τὸ ἀληθές, by crasis τἀληθές, ionic τὠληθές, and τὰ ἀληθῆ, by crasis τἀληθῆ the truth, **Hdt.**, attic
      (a) of persons, truthful, Il., attic
      (b) of oracles and the like, *true, coming true*, **Aesch.**, etc.
    (II) adv. ἀληθῶς, ionic -θέως, *truly*, **Hdt.**, etc.
      (a) *really, actually, in reality*, **Aesch.**, **Thuc.**, etc.; so, ὡς ἀληθῶς **Eur.**, **Plat.**, etc.
    (III) neut. as adv., proparox. ἄληθες; *itane? indeed? really? in sooth?* ironically,**Soph.**, **Eur.**, etc.
      (a) τὸ ἀληθές *really and truly*, Lat. revera, **Plat.**, etc.; so, τὸ ἀληθέστατον *in very truth*, **Thuc.**

As the example illustrates, the meaning of the lexical entry is divided into three different related senses, all of which are then divided up in their own turn. We can represent this using *polyLemon*, as in Figure 2, which depicts the encoding of the senses of *alèthès* using *polyLemon* in a diagram. Here the horizontal lines represent the symmetric senseSibling relation – that is they connect together two senses that have the same 'parent' sense – and the slanting vertical lines represent the senseChild relation.

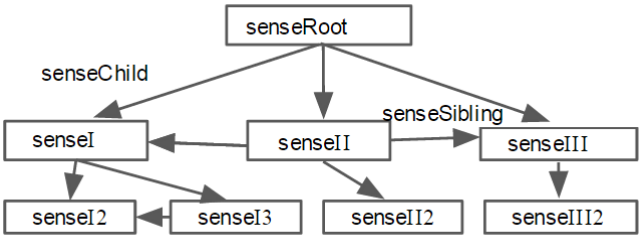

Fig. 2. Diagrammatic representation of the senses of *alèthès* using *polyLemon*

polyLemon allows for a more detailed description of the structure of the senses of individual lexical entries, at least from a *synchronic* point of view; in the next section we will look at how to represent *diachronic* information using the RDF model and *lemon*.

3 *lemonDIA*

The chief difficulty in representing diachronic relations using the RDF model lies in the fact that it constrains us to working with binary and unary relations. This means that if we want to add an extra time argument to, say, the *lemon* sense relation between a lexical entry and a lexical sense then we either adopt a *reification* strategy – that is, we have to create a new class of objects representing instances of this relation – or instead we think in terms of *perdurants* and represent entities as events or processes[5]. This latter is the strategy which we adopted for *lemonDIA* because of how natural a representation it seems in cases like the one we discuss below[6].

We presented an initial version of *lemonDIA* in previous work [5] but since then the experience of working with a number of different diachronic resources has encouraged us to develop the model further and here we present a revised version of the vocabulary.

Recall that in the *lemon* model a lexical sense is modelled as the *reificiation* of the pairing of a lexical entry with an ontological concept, [3]. The core idea with *lemonDIA* is to think of a sense as a process in time; that is the *process* or *event* of a lexical entry l having an extension c, where c is an ontology vocabulary item. We use the term p-sense to refer to lexical senses when viewed as *perdurants*. Each p-sense has an associated temporal extent. In this way we can study how terms change their meaning over time and encode this data at the sense level.

In *lemonDIA* we define a new class of "diachronic" semantic elements called LemonDiaElement, this is a subclass of lemon:LemonElement. Members of LemonDiaElement have a temporal extent (for this purpose we've introduced a new object property temporalExtent):

$$\mathsf{LemonDIAElement} \sqsubseteq \exists \mathsf{temporalExtent.time{:}TemporalEntity}.$$

We define the class LexicalPSense to be a subset of both LemonDIAElement and lemon:LexicalSense. Lexical Entries are associated with p-Senses via the object property pSense.

Sometimes we would like to specify the semantics of a lexical entry, or a root, or even of a cluster of terms without necessarily invoking a sense element. In these cases we use the class LexicalDomain, another subclass of LemonDIAElement, to describe one of the general lexical domains into which a given entry can fall. Elements of LexicalDomain are linked to ontology items using the object property lexicalDomainConcept.

Crucially, we would like to describe the actual shift that takes place between various senses (or lexical domains) of a word. For this purpose we have created

[5] The temporal span of a lexical sense can be specified to some degree within the original *lemon* model, via the usedSince object property which takes a xsd:date value to specify the date from which a sense was first used. *lemonDIA* is an attempt to develop a more flexible and powerful way of dealing with time in such situations.

[6] For more details of the different strategies for representing temporal information in RDF and OWL, including in depth discussions of the use of perdurants see [10] and [7].

the class SemanticShift as a subclass of LemonDIAElement. Each semantic shift object links together a pair of senses (or lexical domains), using the two object properties, shiftSource and shiftTarget.

In addition there are certain, more complicated instances where it is a root word plus a privative affix that changes meaning. To take an example from the OE-Emotion dataset, which we will describe below, the OE noun *arleas* meaning 'honourless', has, as a root lexeme, the OE noun *ar(e)*, which refers to the concept of honour, and includes the privative suffix *-leas*. On the other hand *arleas* also took on the secondary meaning *ashamed*; it therefore exemplifies the shift of a lexeme from a certain type of event (in this case *loss of honour*) to its result, *shame*, via the addition of a privative suffix. To capture these cases we have created a subclass of SemanticShift, NegatedShift.

There were another couple of classes which we also felt it necessary to add to *lemonDIA* in order to properly model datasets like the OE-Emotion lexicon. The class Expression, a subclass of lemon:LexicalEntry, is intended to cover word clusters encompassing lexical roots and their morphological derivations, as well as other variants. The class Lemma on the other hand covers individual spelling and inflectional variants of a lexeme. We link objects of the class Expression to members of LexicalEntries using the object property explex, and the other way round with lexexp.

4 Using *lemonDIA* to Represent OE Emotion Expressions

The rest of this paper will be dedicated to an expository example, namely, in this case an entry taken from the translation into RDF of a lexicon of Old English (OE) Emotion expressions using *lemonDIA*. The lexicon was originally compiled by the second author and lists all the terms used to describe or refer to emotions in a representative corpus of Old English texts, also constructed by the second author. In the majority of cases the terms in the lexicon are polysemic and have a clear derivation based on an expression in another semantic field; see [4] for more details of the dataset. The conversion of the dataset into RDF is currently incomplete, but we plan on completing it within the next few months.

We have made use of the *lemonDIA* vocabulary to represent four different scenarios in the OE-Emotion lexical dataset itself:

- Scenario 1: The meaning of the expression, remains unchanged throughout the period of interest[7];
- Scenario 2: From the expression meaning $s1$ (emotion meaning) we get the meaning $s2$ (new emotion meaning). The original meaning element is also maintained (polysemy/coexistence of meanings);
- Scenario 3: From a non-emotion meaning $s1$, we get an emotion meaning $s2$.$s1$ is lost during or before the period in question (dead metonymy/metaphor);
- Scenario 4: As in Scenario 3, but $s1$ is kept by speakers (polysemy/coexistence of meanings)

In this article we will look at an example of Scenario 4.

[7] In this dataset our period of interest is the period during which OE was spoken.

4.1 The Classification of Shifts and Old English Time Periods

The lexical entries in the dataset are either classified as *literal*, or as non-literal, and thus owing their status as emotion terms to a process of meaning shift. These non-literal lexical entries are afterwards further classified in terms of the kinds of semantic shift involved. The different semantic shifts in the dataset have been grouped and ranked according to the degree of literalness of the shifts as in Table 4.1 (the plus sign here represents the most literal and the minus sign the least)[8].

Literalness	Classification	Conceptualisations
+	LITERAL	TTHE EMOTION IS AN EMOTIONAL EXPERIENCE
	METONYMIC	THE EMOTION IS A CAUSE OF THE EMOTIONAL EXPERIENCE
		THE EMOTION IS A RESPONSE TO THE EMOTIONAL EXPERIENCE
	SYNESTHETIC	THE EMOTION IS A SENSORIAL EXPERIENCE
	METAPHORIC	THE EMOTION IS A LIVING ENTITY
		THE EMOTION IS A SUBSTANCE
		THE EMOTION IS AN OBJECT
		THE EMOTION IS A FORCE
-		THE EMOTION IS A PLACE

Table 1. A Classification of the Semantics of Expressions.

In the RDF version of the dataset we created a SKOS taxonomy to capture this classification.

Now we come to the treatment of the temporal aspect of the original dataset. The dataset makes reference to the following periods in classifying the development of OE:

- OE1 (before 850)
- OE2 (850-950)
- OE3 (950-1050)
- OE4 (1050-1150) .

Each of these intervals is encoded as a time:ProperInterval in the RDF version: we have created the OWL individuals `OE1, OE2, OE3, OE4` whose dates have been specified using the OWL-TIME object properties hasBeginning and hasEnd; in the case of `OE1` we have only an end date. By making use of OWL-TIME encoded Allen relations[1] such as before, intervalStarts, intervalFinishes and the SWRL rules axiomatising these relations[9] [2], we also also able to define new intervals combing and incorporating those previously defined such as for example `OE` that combines all the Old English periods,.e.g.,

[8] Note that the conceptualisations given below represent just a few of the many options (mappings) that have been discovered so far in the world's languages.

[9] `https://github.com/sbatsakis/TemporalRepresentations`

```
:OE rdf:type owl:NamedIndividual ,
            <http://www.w3.org/2006/time#ProperInterval> ;
            <http://www.w3.org/2006/time#intervalStarts> :OE1 ;
            <http://www.w3.org/2006/time#intervalContains> :OE2 ;
            <http://www.w3.org/2006/time#intervalContains> :OE3 ;
            <http://www.w3.org/2006/time#intervalFinishes> :OE4 .
```

We have also defined two other periods, `PrIE` and `PrGrm`, representing the time periods during which Proto-Indoeuropean and Proto-Germanic respectively were hypothetically spoken. These periods were defined *qualitatively*: that is not in terms of exact dates but by specifying that texttttPrGrm came before `OE1`, and *PrGrm* before `PrIE`. Again this was achieved through the use of Allen relations.

4.2 Example: Representation of the *scyld* OE expression

The example which we will look at comes from the list of GUILT terms included in the dataset. The OE expression *scyld* exemplifies the transition of a term that initially stood for *debt* into one meaning *guilt*. As mentioned above, we use the term *expression* here to refer to a lexical root *plus* its related derivations, and so we start with the entry for the expression `SCYLD` [10].

```
:SCYLD a lemon:LexicalEntry, :Expression ;
        lexinfo:etymology :SKULD ;
        lemond:lexicalDomain :scyld_debt_domain, :scyld_guilt_domain;
        lemond:explex :FOR_SCYLDIG_ADJ, :FOR_SCYLDIGIAN_VB, :DEA_PLUS_T_SCYLDIG_ADJ, :SCYLD_N,  ...
        lemon:language "ang"  .
```

Here we specify the language at the level of the entry itself ("ang" is the tag for Anglosaxon or Old English) rather than at the level of the lexicon; this is because many of the entries in the dataset belong to other, different languages. This is the case with the lexical entry `:SKULD`, mentioned in the entry for `SCYLD`, and representing the Proto-Germanic root **skuld-*. `SKULD` is itself related to the Proto-IndoEuropean root **(s)kel-*, represented in the lexicon by the entry `SKEL`.

The expression `SCYLD` is linked to its individual lexemes using the explex object property (here for reasons of space we present only 4 out of a total of 13 lexical entries associated with the entry). It also has two associated lexical domains: these are `scyld_debt_domain` and `scyld_guilt_domain`, the first relating to debt and the second to the domain of guilt.

We present the *lemonDIA* version of the lexical entry for the noun *scyld* which is linked to the expression `SCYLD` using the object property lexexp:

```
:SCYLD_N a lemon:LexicalEntry ;
        lemond:lexexp :SCYLD ;
        wordnet:synset_member :GUILT_N ;
        lexinfo:partOfSpeech lexinfo:noun;
        lemond:lemma :scylde, :scylde_n, :scylde_n_a, :scylde_n_d, :scylde_n_g;
        :corpusFreq "441"^^xsd:nonNegativeInteger;
        :totalFreq "0"^^xsd:nonNegativeInteger ;
        lemon:language "ang" .
```

[10] Note we use `lemond` as namespace for the lemonDIA vocabulary.

In the original lexicon each of the different lexical entries associated with an expression belonged to a specially-defined synset. We represented these synsets using the Princeton Wordnet vocabulary and linked each of these synsets to the corresponding WordNet 3.0 sysnset.

The lexical entry `SCYLD_N` is associated with five lemmas. The entries for these lemmas describe the frequency of the lemmas, their locations in the corpus, as well as the syntactic context of each lemma.

To return to the two lexical domains given above, `scyld_debt_domain` and `scyld_guilt_domain`: in order to specify the ontological content of the lexical domains and lexical senses in the dataset – which we have done using the object properties reference and lexicalDomainConcept – we have provisionally chosen to link to the DBpedia dataset. Both of these lexical domains have a temporal extent that encompasses the whole of the Old English period:

```
:scyld_debt_domain a lemond:LexicalDomain ;
        lemond:lexicalDomainConcept dbpedia:Debt ;
        lemond:temporalExtent :OE .

:scyld_guilt_domain a lemond:LexicalDomain ;
        lemond:lexicalDomainConcept dbpedia:Guilt ;
        lemond:temporalExtent :OE .
```

We can then specify the semantic shift that occurred between them:

```
:debt_to_guilt a lemond:SemanticShift;
        lemond:shiftType :Resultative_metonomy ;
        lemond:shiftSource :scyld_debt_domain ;
        lemond:shiftTarget :scyld_guilt_domain ;
        lemond:temporalExtent :BeOE.
```

Here the shiftType is defined as a `Resultative_metonomy`.

5 Conclusion

In this article we have presented two vocabularies based on the *lemon* model for representing lexical resources in RDF each of which deals with an important aspect of lexical semantics, and each of which is salient for the study of the development of scientific and technical lexica.

References

1. Allen, J.F.: Maintaining knowledge about temporal intervals. Communications of the ACM 26(11), 832–843 (1983)
2. Batsakisa, S., Petrakisb, E., Tachmazidisa, I., Antonioua, G.: Temporal representation and reasoning in owl 2.0
3. Cimiano, P., McCrae, J., Buitelaar, P., Montiel-Ponsoda, E.: On the role of senses in the ontology-lexicon. In: New Trends of Research in Ontologies and Lexical Resources, pp. 43–62. Springer (2013)
4. E Díaz-Vera, J.: From cognitive linguistics to historical sociolinguistics: The evolution of old english expressions of shame and guilt. Cognitive Linguistic Studies 1(1), 55–83 (2014)

5. Khan, F., Boschetti, F., Frontini, F.: Using lemon to Model Lexical Semantic Shift in Diachronic Lexical Resources. Proceedings of the Workshop on Linked Data in Linguistics 2014 (LDL-2014) (2014)
6. Khan, F., Frontini, F., Boschetti, F., Monachini, M.: Converting the liddell scott greek-english lexicon into linked open data using lemon. In: Digital Humanities 2016. Kraków, Forthcoming (2016)
7. Krieger, H.U.: 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. p. 1 (2014)
8. McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Pérez, A.G., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., et al.: The lemon cookbook (2010)
9. McCrae, J., Spohr, D., Cimiano, P.: Linking Lexical Resources and Ontologies on the Semantic Web with lemon. In: Proceedings of the 8th extended semantic web conference on The semantic web: research and applications - Volume Part I. pp. 245–259. ESWC'11, Springer-Verlag, Berlin, Heidelberg (2011), http://dl.acm.org/citation.cfm?id=2008892.2008914
10. Welty, C., Fikes, R.: 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). pp. 226–236. IOS Press, Amsterdam, The Netherlands, The Netherlands (2006), http://dl.acm.org/citation.cfm?id=1566079.1566106