Towards the Representation of Etymological and Diachronic Lexical Data on the Semantic Web

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

Abstract

In this article we take a detailed look at a number of issues relating to the publication of etymological data as linked data. We then put forward our proposal for an RDF-based model for representing etymologies that, as we will show, helps to answer at least some of the problems and requirements outlined in the initial part of the paper. We also take a more general look at the representation of diachronic lexical data as linked data.

Keywords: Etymology, Linked data, Ontolex-lemon

1. Introduction

Linked data with its core emphasis on linking together different and sometimes heterogeneous resources seems to be perfectly suited to the representation of etymological data since such data relies on the bringing together of evidence from diverse sources. In the case of etymology these can be primary sources that attest to the appearance, in a text, of a given word or phrase under a specific form or with a particular meaning, or they can be secondary sources that refer to salient hypotheses made by scholars in the past. In this article we take a detailed look at a number of issues relating to the publication of etymological data as linked data. We then put forward our proposal for an RDF-based model built on top of ontolex-lemon for representing etymologies that, as we will show, helps to answer at least some of the problems and requirements outlined in the first part of the paper. In addition we will take a more general look at the representation of diachronic lexical data as linked data.

In the next section, Section 2. we give an overview of some of the main challenges to modelling etymology in linked data. Then in Section 3. we make a first proposal of a model for a model for etymology. Next, in Section 4. we discuss the addition of temporal information to lexcal linked datasets.

2. The Challenges of Modelling Etymology in Linked Data

The word *etymology* has at least two different senses. In the first of these it is a sub-discipline of historical linguistics that concerns itself with the development of individual words (and other lexical entries) over time and attempts to trace their origins as far back as the evidential record will support – and sometimes even beyond. In addition *etymology* can also refer to a single such history of a word (or other lexical item). Etymologies in this latter sense can be found in many dictionaries and lexicons although typically in a condensed or abbreviated form. Note that we will be using both senses of the word in what is to follow, although we will focus predominantly on the latter.

Three important points which, as we argue below, have a significant impact on the modeling of etymologies in RDF, can be seen to immediately follow from the preceding definitions. The first point is that etymologies are essentially diachronic and call for the explicit representation of the unfolding of historical processes. In particular we often need to model the fact that a word w had the sense s during period the t, i.e., that a given property (having the sense s) holds for a certain period of time – something which is notoriously difficult to do, in a human-intuitive way, with a formalism like RDF that is limited to unary and binary predicates. There are several design patterns that can be used to overcome this expressive difficulty, none of which however turn out to be wholly satisfying on all or most accounts. Etymologies can potentially represent more than one kind of change as occurring at (around) the same time, so that as well as showing how a word’s meaning alters over a given period, we might also want to depict the kinds of sound changes which it undergoes along with any shifts in written form and grammatical properties that might have occurred. Furthermore, the temporal information given in etymological sources is frequently underspecified (and of course it cannot be otherwise when it comes to reconstructed roots/words) and in many cases we lack a precise year or even century – or it is the case that whatever dates we do have are qualified with the modifier “circa”. As these issues are very typical of etymological data, both in general purpose dictionaries and in specialist etymological works, we will need to take them into consideration when designing our model.

This leads us onto our next point, which is that etymologies have a marked tendency towards the speculative and in many cases there is no settled consensus as to a word’s origins or the different twists and turns that it might have undergone during its historical development. In fact it’s not unusual to find more than one etymology in a lexical entry and for etymologies to differ substantially for the same word across according to different sources. This is due to the dearth of evidence relating to the earlier stages of modern day languages or to extinct languages and the frequent use of reconstructions in building up etymologies. It is therefore important to have a means of explicitly representing different hypotheses concerning a word's origin and development, as well as an accurate means of citing and, in general, describing the secondary literature. We will discuss this briefly in what follows. For reasons of space, the more general issue of how to represent attestations and citations in RDF versions of lexical/lexicographic resources, will not be covered here, although we do plan to discuss this in forthcoming work.

Another consideration to be borne in mind in the present regard is that, as was mentioned earlier, etymologies encompass different levels of linguistic description, typically the phonological or the semantic levels, and can concern more than one level at the same time. It is therefore an important precondition for an RDF based model for etymologies that there already exist a framework of different modules for representing these levels of linguistic description. In theory linked Data offers us much of the expressivity that we need to represent information at each descriptive level (at least in the case of a large number of etymological examples) but we currently lack specific, specialised, vocabularies; this is especially the case when it comes to representing different kinds of semantic shift.

We intend for our model to be used both in the creation of new lexical resources, or at least in cases where a significant amount of source material has yet to be integrated into a meaningful resource-wide organisational structure, as well as for retrodigitised lexicons and in consequence our model needs to be as fairly flexible. However as the conversion of retrodigitised print dictionaries into RDF is likely to be one of the most popular use cases for such a model[1] we have tried, as far as possible, to take the most common conventions of print etymological resources into consideration when designing our model.

2.1. Two Example Etymologies for the Word *girl*

Before we go on to describe our proposed model and in order to make our discussion a little more concrete than it has been up to this point we will take a look at the etymology of the word *girl* from two different sources. The first etymology is taken from Walter Skeat's influential etymological dictionary of English originally published in 1886,(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) *girl* 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 depreciatory 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*.

The second etymology is taken from Eric Partridge's single volume 'Origins: A Short Etymological Dictionary of Modern English' (Partridge, 1966)

> girl
> , whence **girlish**, derives from ME *girle*, varr *gerle*, *gurle*: o.o.o.: perh of C origin: cf Ga and Ir *caile*, EIr *cale*, a girl; with Anglo-Ir *girleen* (dim *-een*), a (young) girl, cf Ga-Ir *cailin* (dim -in), a girl. But far more prob, *girl* is of Gmc origin: Whitehall postulates the OE etymon **gyrela* or **gyrele* and adduces Southern E dial *girls*, primrose blossoms, and *grlopp*, a lout, and tentatively LG *goere*, a young p/erson (either sex). Ult, perh, related to L *puer*, *puella*, with basic idea '(young) growing thing'.

The first entry presents the word *girl* as having undergone a semantic change of narrowing from its original meaning of 'young man or woman' (as attested by a passage in the Canterbury Tales) to its modern meaning of 'young female person'. Skeat offers a number of possible cognates to *girl*, that is words that are probably derived from the same root as *girl*, in other Germanic languages, adding citations to the literature in support. He considers the origin of the word *girl* to be uncertain however, too uncertain, at least, to suggest any plausible hypotheses. Partridge on the other hand – and in spite of the fact that he labels the word as 'o.o.o.' (of obscure origin) – gives three different hypotheses as to the word's origin, citing the literature in support of a postulation from a reconstructed old English etymon.

3. A First Proposal for a Linked Data-Based Model for Etymology

Having prepared the ground in the preceding sections with a discussion of relevant topics, it is finally time to present our proposal for a model, an extension of ontolex-lemon, to represent etymological data in RDF. We do this in Section 3.2.; in the next subsection, Section 3.1., however, we will describe other relevant and/or related work in the area of language resources and technologies.

[1] Indeed this seems to be a very timely moment for the definition of such a model given the growing interest in converting lexicographic resources into formats such as TEI and RDF. C.f. the current European project ELEXIS. The fact that lexicography stands at the crossroads of several different humanistic disciplines – in particular historical linguistics, lexicography and philology – makes it an interesting and salient case study from the point of view of the ongoing development of the digital humanities (as well of course as raising a variety of non-trivial challenges from a computational point of view).

3.1. Related Work

Previous work on defining a framework for representing etymological data in digital lexical resources includes Salmon-Alt's proposal for an LMF based etymology model, (Salmon-Alt, 2006), as well as Bowers and Romary's work on the deep encoding of etymological information in TEI (Bowers and Romary, 2016). We have been influenced by both of these works in the development of our own model, though we will not detail the differences and similarities between their models and ours here.

With respect to modeling etymologies in RDF, previous work includes (De Melo, 2014) and (Moran and Bruemmer, 2013). In (Chiarcos et al., 2016) Chiarcos et al. defined a minimal extension of the lemon model with two properties for encoding and navigating etymological data: these were the symmetric and transitive `cognate` and the transitive `derivedFrom`. The adoption of such a minimal vocabulary for etymological data is likely to be sufficient for a good number of use-cases. Other cases, such as e.g., in the modeling of entries from more scholarly dictionaries, will necessitate a fuller representation of the evolution of a word, taking into consider its various linguistic properties at different points in time as well as the different hypotheses relating to each of them. Our intention in this article is to propose such a model, one that allows for the kind of so called 'deep' etymological modeling as described in bowers2016deep.

3.2. The Core Entities of our Model

Note that as we alluded to above, our proposed model is an extension or module of ontolex-lemon[2], the latest version of the popular lemon model(McCrae et al., 2017).

To begin with we will fix on the most important kinds of entity that we should, ideally, be able to refer to and to describe, i.e., to define predicates over, when modeling etymologies and that we will therefore want to make into classes[3].

The utility of being able to refer to etymologies themselves – their component parts, their provenances, and perhaps even their likelihoods as possible hypotheses – should be clear from the preceding discussion. It will therefore come as no surprise that we have made `Etymology` a class in and of itself[4]. Indeed this seems like an even more obvious move when you consider the frequency with which it is possible to find two or more different etymologies for the same entry in the same dictionary (c.f. the first etymology of *girl* presented above) or to have provenance information associated with individual etymologies (c.f. the citation of secondary literature in the second etymology of *girl*). We have decided not to limit members of the class `Etymology` to being associated with lexical entries only (for instance a sense or a morphological variant can each have their own separate etymologies).

The second main class which we propose is `Etymon`; the name is taken from the term in linguistics referring to words or morphemes from which other words or morphemes derive. The existence of the class `Etymon` enables us to make a distinction between the 'official' lexical entries in a lexicon and other lexemes whose main or only role is to describe etymological information relating to an entry (of course there are also 'official' lexical entries which also play the roles of etymons to other entries but these are still regarded as first-class entries). Why is this a useful distinction to make? Well, most comprehensive monolingual general purpose dictionaries for a language like English will contain etymological information – but we don't necessarily want, in the case of English, thousands of French and Latin words to appear in a list of all the instances of `LexicalEntry` in the resource – or at least not without being able to filter them out and distinguish them in some way. On the other hand it isn't enough to distinguish members of the class `Etymon` from lexical entries by the bare fact of their having been assigned a different language from the language(s) of the lexicon, since this wouldn't allow us to differentiate between cognates and etymons. In fact we also define the class `Cognate` in order to distinguish lexemes that play the role of cognates in an entry[5]. We have chosen to make both `Etymon` and `Cognate` subclasses of `LexicalEntry`.

Returning to the etymologies themselves: how do we relate together an instance of `Etymology` with the `LexicalEntry` whose history it describes and the instances of `Etymon` (or other elements) which it relates together? One option is to represent an `Etymology` as an ordered sequence of elements using one of the data structures provided by RDF, containers or collections or lists. But this might be too restrictive for our purposes since we may want to elaborate on the relationships between the different elements which have been ordered together in the etymology. In order to illustrate this point further we shall take as an example the English word *friar*, with an etymology adapted from Philip Durkin's Oxford Guide to Etymology (Durkin, 2009).

Although the word *friar* ultimately derives from *frāter*, the Latin word for 'brother', it first entered into English from Old French, from the polysemic word *frere* which means both 'brother' (as in the Latin) as well as 'member of a religious fraternity'. This latter sense was borrowed into Middle English as *frere* (with the same pronunciation as in the French) where it meant both 'member of a religious fraternity' as well as the more specialized meaning of 'member

[2] `https://www.w3.org/community/ontolex/wiki/Final_Model_Specification`

[3] With the obvious proviso that this can be done in different ways, and that the proposal we make is only one of several options in accord with the core necessities of describing etymologies.

[4] We are considering making `Etymology` a subclass of the class `Hypothesis` from the Linked Science Vocabulary (`http://linkedscience.org/lsc/ns/`) – but are still undecided on this point.

[5] We will not discuss the class `Cognate` further here, but will develop it in forthcoming work.

of a mendicant order' (but not 'brother' as in sibling), before finally coming, in modern English, to take on the latter sense. We can identify a number of different relationships between the various etymons identified above: Old French ***inherits*** the word *frāter* which, after having undergone a sequence of sound changes in the meantime, becomes *frere*, then Middle English ***borrows*** the word *frere* into its vocabulary, indeed borrows only a single sense of the word, eventually this changes its meaning through a process of ***specialisation***. The following shorthand description for the whole process which uses the '<' symbol [6] is again taken from (Durkin, 2009):

> Latin *frāter* brother<Old French *frere* brother, also member of a religious order of 'brothers'<Middle English *frere*, *friar*<modern English *friar*

By explicitly representing, indeed, reifying the shifts between instances of Etymon (and between an Etymon and a LexicalEntry) we can include important information on the type of etymological process that leads from one element to the other. For this purpose we have defined the EtyLink class that represents an etymological relationship between two elements. The EtyLink class can be seen as equivalent to the etymological symbol <. An instance of Etymology, then, consists of a series of such instances of EtyLink. This leads to a model (part of) which is represented in Figure 1.

Note the presence of the object property etySource which relates an EtyLink to a LexicalEntry/Etymon as its source, similarly with etyTarget. The two properties etySubsource and etySubtarget are designed to further specify the source and targets of an etymological relation between two entities. This is useful in case we want to elaborate on the sense or form which a lexical entry derives from. Using this model we can represent the Durkin *friar* example as in Figure 2.

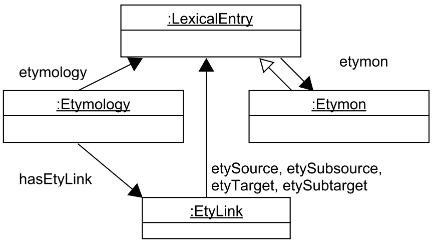

Figure 1: The relationship between some of the classes in our proposed model.

Here we've given the first Etymon in the series the special status of root as the earliest Etymon to which we can trace the word back to. If we wish to further specify the fact that the word *frere* in Early Modern English derives from the sense of the word in Old French in which it meant 'member of a religious order of brothers' then we can proceed as in Figure 3.

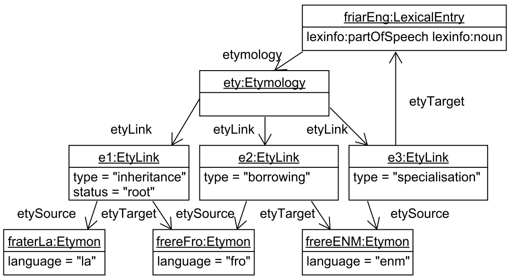

Figure 2: Modelling the *friar* example.

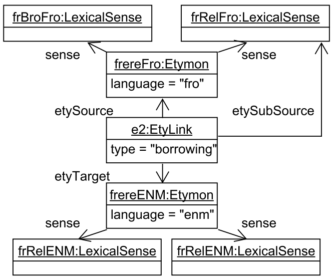

Figure 3: Modelling the *friar* example.

We can also specify, in a similar way, that the word *frere* in Old French derives from the accusative singular form of the Latin *frāter*, namely, *fratrem*.

As we stated above etymologies often include reconstructed words/root forms of words in reconstructed languages (as well as in historical languages for which there is a lack of relevant attestations) such as proto-Indo-European and proto-Germanic for which we have no surviving written attestations. Often etymologists will assign a meaning to these reconstructions on the basis of the evidence of words in other, attested languages; these reconstructed meanings however should be distinguished from other lexical entries for which there actually exists direct evidence. As Watkins (quoted by Durkin(Durkin, 2009)) points out 'reconstructed words are often assigned hazy, vague or unspecific meanings...The apparent haziness in meaning of a given Indo-European root often simply reflects the fact that with the passage of several thousand years the different words derived from this root in divergent languages have undergone semantic changes that are no longer recoverable in detail.'(Watkins, 2000).

In such cases the use of the ontolex-lemon LexicalSense class (and the sense relationship) would usually be inappropriate – on the other hand though we **would** like to be able to include semantic information

[6] Note that '<' is overloaded because it stands both for the development of one word from another or of a word borrowing from another language

associated with the root or reconstructed word in question. Therefore, and given that this issue is an especially pertinent one in the encoding of etymologies, we have defined a new class in our model, `LexicalDomain`, in order to provide a weaker notion of meaning than that of `LexicalSense`, although as with the latter class `LexicalDomain` is intended to link a `LexicalEntry` with an ontology concept. So for instance the reconstructed root **ker-tā-* which has been assigned the meaning 'fire' and which is hypothesised to be the root of the English word *hearth* can be modelled as in Figure 3.2.. Note that the object relations `lexicalDomain` and `domainField` play a role that corresponds to `sense` and `reference`, respectively, in ontolex-lemon.

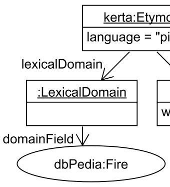

Figure 4: An example using the `LexicalDomain` class.

4. Adding Temporal Information to lexical data in RDF

Up until now we have avoided the issue of how to include temporal information in RDF etymologies and for good reason too. That is, as we mentioned above, it is not immediately obvious what the best way of doing this in RDF actually is. However in this section we will discuss one particular strategy for doing this.

In previous work(Khan et al., 2014), (E Díaz-Vera, 2014) we have opted for a perdurantist/four-dimensionalist(4D) approach when modeling sense shift[7], along with other diachronic lexical information, in RDF, and this is also what we propose in the present work. What, then, does this approach entail in the current case? Simply put, the idea is to treat elements such as senses, forms, and even whole lexical entries as having an inherent temporal extension. And so by making temporal extent a property of these elements we do not need to reify the original relation in order to introduce a temporal parameter[8].

We can think of it as follows. In ontolex-lemon the relation `sense` holds between a lexical entry l and each one of its lexical senses s. Now if it were an ideal world we could simply add a temporal parameter, specifying the interval, t, in which the `sense` relation holds, i.e., `sense`(l, s, t). Obviously we can't do this in RDF. On the other hand, however, since a sense is already a reification of the meaning relation between a lexical entry and a reference (representing the extension of the entry) we *can* 'attach' this temporal information to s itself, that is rather than adding an extra parameter to the `sense` relation, without wreaking too much conceptual havoc as a result[9]. To reiterate then, we can represent a lexical sense as a entity with an extension in time that can be associated with a lexical entry and that describes one of its meanings as if it were a process in time, i.e., `sense`(l, s) and `hasTime`(s, t) . It may be useful to distinguish senses that have temporal extent from 'normal' senses by referring to them as *p-Senses* (the 'p' here stands for *perdurant*) and creating a new subclass of `LexicalSense` called `LexicalpSense`. We can do something similar with the class `Form` and the object property `lexicalForm` in ontolex-lemon. Indeed one can go further and define an `Etymon` as a perdurant. This would give us much more expressivity in representing etymologies.

One of the advantages of explicitly representing temporal information in RDF is that it becomes much easier to query for such data. By making use of OWL axioms we can also reason over such data. The fact that the temporal information in etymological datasets is often vague and underspecified need not necessarily prove to be an insurmountable barrier to the use of OWL-based reasoning over such data. As we demonstrated in (Khan et al., 2016) it is fairly straightforward to reason with and query over such data by using Allen relations to describe the relationships between temporal intervals and by using other Semantic Web standards such as e.g., the Semantic Web Rule Language and the Semantic Query-Enhanced Web Rule Language.

5. Conclusion

Our intention in this article has been to present a first proposal on how to model etymologies as well as diachronic lexical data more generally in RDF through an extension of the RDF-native lexical model ontolex-lemon. Some of our proposals will, no doubt, be controversial but we hope the present work will serve to stimulate discussion on this issue and thereby help to contribute towards the definition of a standard (or recommendation) for modeling such data, one that will gain some measure of acceptance within the various communities that find themselves working with etymological data as part of their research.

6. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

[7] An good introduction to the 4D perspective can be found in (Welty and Fikes, 2006). We favour the slightly altered formulation given in (Krieger, 2014).

[8] C.f. `https://www.w3.org/TR/swbp-n-aryRelations/`

[9] In contrast a statement like 'Rome is the capital of Italy' is a little bit more difficult to model, and in the 4D view we would have to define a relation 'isCapitalOf' between a time slice of the referents of 'Rome' and 'Italy', i.e., we have to create two new entities that separately encode this temporal aspect. Other RDF temporal representation strategies have their own specific drawbacks.

7. Bibliographical References

Bowers, J. and Romary, L. (2016). Deep encoding of etymological information in tei. *arXiv preprint arXiv:1611.10122*.

Chiarcos, C., Abromeit, F., Fäth, C., and Ionov, M. (2016). Etymology meets linked data. a case study in turkic. In *Digital Humanities 2016. Krakow*.

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

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

E Díaz-Vera, J. (2014). From cognitive linguistics to historical sociolinguistics: The evolution of old english expressions of shame and guilt. *Cognitive Linguistic Studies*, 1(1):55–83.

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, A. F., Bellandi, A., and Monachini, M. (2016). Tools and instruments for building and querying diachronic computational lexica. *LT4DH 2016*, page 164.

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. 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.

Partridge, E. (1966). *Origins : a short etymological dictionary of modern English / by Eric Partridge*. Routledge and Kegan Paul London, 4th ed. (with numerous revisions and some substantial additions). edition.

Salmon-Alt, S. (2006). Data structures for etymology: towards an etymological lexical network. *BULAG*, 31:1–12.

Skeat, W. W. (1910). *An etymological dictionary of the English language / Rev. Walter W. Skeat*. Oxford University Press London, 4th ed. revised, enlarged and reset. edition.

Watkins, C. (2000). *The American Heritage Dictionary of Indo-European Roots*. Houghton Mifflin Harcourt, second edition edition, September.

Welty, C. and Fikes, R. (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)*, pages 226–236, Amsterdam, The Netherlands, The Netherlands. IOS Press.