Towards a Lexical Standard for the Representation of Etymological Data

Fahad Khan
ILC-CNR/Pisa
fahad.khan@ilc.cnr.it

Jack Bowers
Austrian Center for Digital Humanities/Vienna
Inria - Team ALMAnAC/Paris
Jack.Bowers@oeaw.ac.at

Abstract

English. We introduce a new standard (currently under development), LMF Diachrony-Etymology, which is intended to constitute a common model for the creation of diachronic lexical data, and in particular etymologies, as computation lexical resources. Having situated it whtin in the context of previous developments in this area, we outline the content of the new standard, and in particular the core classes of the model as well as describing our overall approach to modelling etymological data. Finally, we give an example encoding of an entry taking from an etymological dictionary of Latin.

Italiano. Questo paper presenta LMF Diachrony,-Etymology un nuovo standard, (attualmente in fase di sviluppo), volto a costituire un modello comune per la rappresentazione di dati lessicali diacronici e in particolare di etimologie, sotto forma di risorse lessicali computazionali. Situeremo lo standard nel contesto delle tendenze attuali del settore e ne presenteremo il contenuto, con particolare riferimento alle principali classi del modello e alla descrizione dell'approccio complessivo alla modellizzazione diei dati etimpolgici. Infine, presenteremo la modellizzazione formale di una entrata tratta da un dizionario etimologico del latino.

1 Introduction

In this submission we will introduce a new standard, currently in an advanced stage of development[1], for the modelling and publication of etymological data in computational lexical resources[2]. The standard in question, which we will refer to it as LMF-Ety, is the third part of a new multi-part revision of the Lexical Markup Framework (LMF), ISO 24613-4, originally published by the International Standards Organisation (ISO) as a single standard in 2008[3]. We will motivate the need for LMF-Ety by describing some of the main challenges of modelling etymological data in computational lexical resources and showing how our new standard meets these challenges as well as how it differs from other previous models. Subsequently, we will describe the core concepts which we have so far established in our model and illustrate them through the use of an extended example taken from an etymological dictionary. Our intention is both to summarise the work we have carried out in the development of the LMF-Etymology standard as well as to showcase our broader approach to modelling etymologies. This approach entails the representation of etymologies as formal graphs describing simple narratives relating to a given lexicon phenomena; it is an approach that takes account of and consolidates previous attempts at modelling etymologies computationally but that also seeks to extend them in various different directions with a view to obtaining a more robust and expressive model. Additionally both authors are also involved in other initiatives for modelling etymologies in two other frameworks/standards, the Text Encoding Initiative (TEI) (Bowers and Romary, 2016) and Linked Data (Khan, 2018). We will end the submission by

[1] At the time of writing, December 2019, the standard is being prepared for submission to an ISO ballot as a Draft International Standard (DIS). The classes and the approach which we present are now therefore fairly stable.

[2] The two authors are the joint project leaders of this standard.

[3] Note that the original version of LMF did not contain specific provision for modelling etymological/diachronic information.

describing how LMF-Etymology may be rendered inter-operable with work being carried out in these two latter frameworks. This will also be relevant for understanding the practical details of how LMF-Etymology can actually be used (SPOILER ALERT: Part 4 of the LMF standard is a serialisation of all the previous parts in TEI-XML).

2 Background

The importance of standards for the publication of scientific and scholarly datasets and resources for rendering them more findable, accessible, interoperable and reusable is by now well understood across the board. There have been a number of initiatives for promoting such standards and best practices in the field of language resources. Three of the most notable of these as applied to the case of lexical datasets are: the original version of the Lexical Markup Framework described below; the Dictionaries chapter of the Text Encoding Initiative (TEI) guidelines[4]; and finally the RDF-based Ontolex-Lemon guidelines (McCrae et al., 2017). These standards not only help to ensure a greater measure of interoperability between different computational lexicons, but they also facilitate the representation of lexical information in a way that makes it more amenable to advanced kinds of machine processing. Up until recently all three of these standards have dealt almost exclusively with synchronic lexical data[5]. This neglect of diachronic data is due, in part, to the awkwardness associated with the addition of extra temporal parameters to statements in data frameworks such as UML or RDF, and partly due to the (relatively) slow pace of development in the three standards overall – even if this would seem to constitute a missed opportunity, particularly in the case of etymological data since, at an abstract level, etymologies traditionally describe graph structures. They would therefore be ideally suited for representation in formalisms where this underlying structure can be rendered explicit, making such data easier to query and process. Moreover standards like LMF and especially the RDF-based Ontolex-Lemon would potentially make it easier to link together and query across different etymological datasets and to therefore create extended etymological networks. LMF-Ety is intended both for the creation of etymological datasets *ex novo* as well as for the conversion of legacy print resources as structured data. In this latter respect it should be noted that although our initial use cases have so far been largely concerned with the conversion of legacy dictionaries into structured resources, descriptions of etymological graph structures can be found in, and therefore potentially extracted from, numerous different kinds of texts. These include both scholarly works in linguistics, especially in the sub field of historical linguistics (articles, book chapters, monographs, etc), along with other genres of texts, literary, religious and philosophical[6]. It is clear then that the computational modelling of etymologies stands firmly at the intersection of computational linguistics, e-lexicography and the digital humanities. This inter-disciplinarity can also be appreciated in the fact that etymologies for languages with a sufficiently extensive written tradition will often contain attestations to coprora of historic texts; these texts can sometimes be reconstructions or have disputed interpretations as to particular word senses, (bringing to bear issues concerned with textual criticism and philology/literary criticism more generally).

2.1 The Lexical Markup Framework

The original 2008 version of LMF, ISO 24613: 2008, was intended as a "standardized framework for the construction of computational lexicons" (Francopoulo, 2013) an was conceived of as a common model both for lexicons for use in NLP applications as well as for computational versions of print or legacy dictionaries. Regarded as one of the most important standards in the field of lexical resources, LMF was enormously influential in the definition of the Ontolex-Lemon model and its predecessor*lemon*. A review of the Lexical Markup Framework undertaken by ISO in 2016 resulted in a decision to revise the standard, and to publish it as a multi-part standard (Romary et al., 2019) to render it more modular; the decision was also made to broaden the applicability of the new version of LMF to capture more kinds of lexical information. In consequence it was decided that one of the new parts of LMF should

[4] `https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html`

[5] The TEI guidelines do permit the representation of etymological data in a structured way but in a relatively shallow way. A recent proposal to allow for more salient kinds of etymological annotation can be found in (Bowers and Romary, 2016)

[6] See (Khan, 2018) which presents an example taken from Hobbes'*Leviathan*.

be a specialised module dealing with diachronic lexical information. The different parts of the new LMF standard are: the **Core Model** (ISO 24613-1); the **Machine Readable Dictionary Module** (ISO 24613-2); the **Diachrony-Etymology Module** (ISO 24613-3); Serialisations in both TEI (ISO 24613-3) and LBX (ISO 24613-3). Two new modules dealing with syntax and semantics and morphology have also been proposed.

2.2 Related Work and Standards

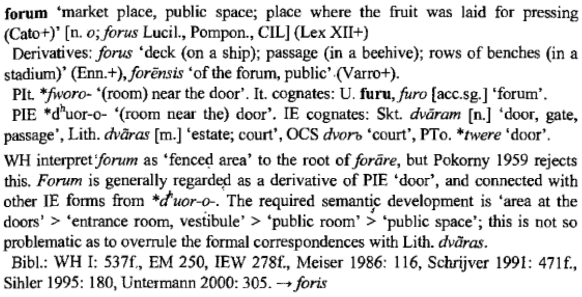
A number of proposals have been made in the past to try and redress the lack of provision for encoding structured etymological data in lexical resources. These have included suggestions for extensions of both for TEI (Bowers and Romary, 2016) and LMF (Salmon-Alt, 2006), as well as proposals for new Ontolex-lemon classes and properties(Chiarcos et al., 2016), (Khan, 2018). The standard described in the current work is influenced by the aforementioned works, and in particular it is informed by the approach in taken in (Khan, 2018) while abstracting away from specific details mentioned in that work and which pertain to the Resource Description Framework (RDF). Moreoever, it is the result of an attempt to converge towards a set of high level concepts, abstractions, that are sufficiently expressive to encode the main kinds of phenomenon and information which tend to be included in etymologies, as well as being simple enough to be usable by a wide community of potential users. The main concepts which we have determined upon are described in the next section. As we mentioned above the fourth part of the new standard is a TEI serialisation of all the other parts (due to be published at the same time as LMF-Etmyology). This should ensure that at a high level both TEI and LMF are interoperable and in particular the etymological apprpach taken by LMF-Etymology is compatible with TEI. It also means that LMF-Etymology should ultimately be accessible to digital humanists who are more used to working with TEI.

3 LMF Etymology

The definition of LMF Ety (ISO Standard 24613-3) is dependent on the two preceeding ISO standards in the new multi-part LMF standard, i.e., the Core Module, recently published by ISO, and the Machine Readable Dictionaries Module, due to be published in 2020 (Romary et al., 2019). Both of these standards contribute foundational concepts (for modelling lexicons) to LMF Ety such as **Lexical Entry**, **Lemma**, **Form**, and **Sense**: all of which keep their (fairly intuitive) meaning from the previous version of LMF and all of which share the meaning of similarly titled concepts in TEI and Ontolex-Lemon. On the basis of these foundational concepts then LMF Ety defines a number of additional classes which enable us to associate temporal/historical information with lexical data encoded in LMF. The strategy we adopt is that suggested in (Khan et al., 2014) of modelling linguistic elements such as words, senses, forms, etc as *perdurants*, that is, as entities associated with a lifespan, which in the present case represents the interval of time in which they are considered to have been part of common usage within a given linguistic community[7]. This enables us to situate lexical entries etc in a temporal dimension and also to relate them together via diacrhonic linguistic processes. Our model then represents etymologies as simple narratives, or as rather simple *narrative graphs*, in which different linguistic phenomena (each of which can be potentially associated with a lifespan and situated on a timeline) are linked together using special etymological link elements, individuals of the class **EtyLink**, which can represent different kinds of historical linguistic processes such as *inheritance* or *borrowing* or *semantic shift*. The other new elements in the standard are the described below:

- **Etymon** and **Cognate**: Two elements modelled as subtypes of **Lexical Entry**. What differentiates them from other lexical entries in a lexicon is their (specialised) role: they are used in describing the etymologies of other lexical entries: **Etymon**s are lexical entries from which a given lexical entry is derived via some historical process; **Cognate**s are lexical entries which share a common ancestor with a given a lexical entry; Additionally **Cognate Set** represents the reification of a set of cognates.
- **Etymology**: An element that represents a single history of a lexical entry or other element. We associate **Etymology** individuals with an ordered series of **EtyLink** instances; this allows us to

[7] This approach makes it easier to represent such information in RDF.



forum 'market place, public space; place where the fruit was laid for pressing (Cato+)' [n. *o*; *forus* Lucil., Pompon., CIL] (Lex XII+)

Derivatives: *forus* 'deck (on a ship); passage (in a beehive); rows of benches (in a stadium)' (Enn.+), *forēnsis* 'of the forum, public' (Varro+).

PIt. **fworo-* '(room) near the door'. It. cognates: U. **furu**, *furo* [acc.sg.] 'forum'.

PIE *d^{h}uor-o- '(room near the) door'. IE cognates: Skt. *dvāram* [n.] 'door, gate, passage', Lith. *dvāras* [m.] 'estate; court', OCS *dvorъ* 'court', PTo. **twere* 'door'.

WH interpret *forum* as 'fenced area' to the root of *forāre*, but Pokorny 1959 rejects this. *Forum* is generally regarded as a derivative of PIE 'door', and connected with other IE forms from **d^{h}uor-o-*. The required semantic development is 'area at the doors' > 'entrance room, vestibule' > 'public room' > 'public space'; this is not so problematic as to overrule the formal correspondences with Lith. *dvāras*.

Bibl.: WH I: 537f., EM 250, IEW 278f., Meiser 1986: 116, Schrijver 1991: 471f., Sihler 1995: 180, Untermann 2000: 305. → *foris*

Figure 1: Entry for *forum*.

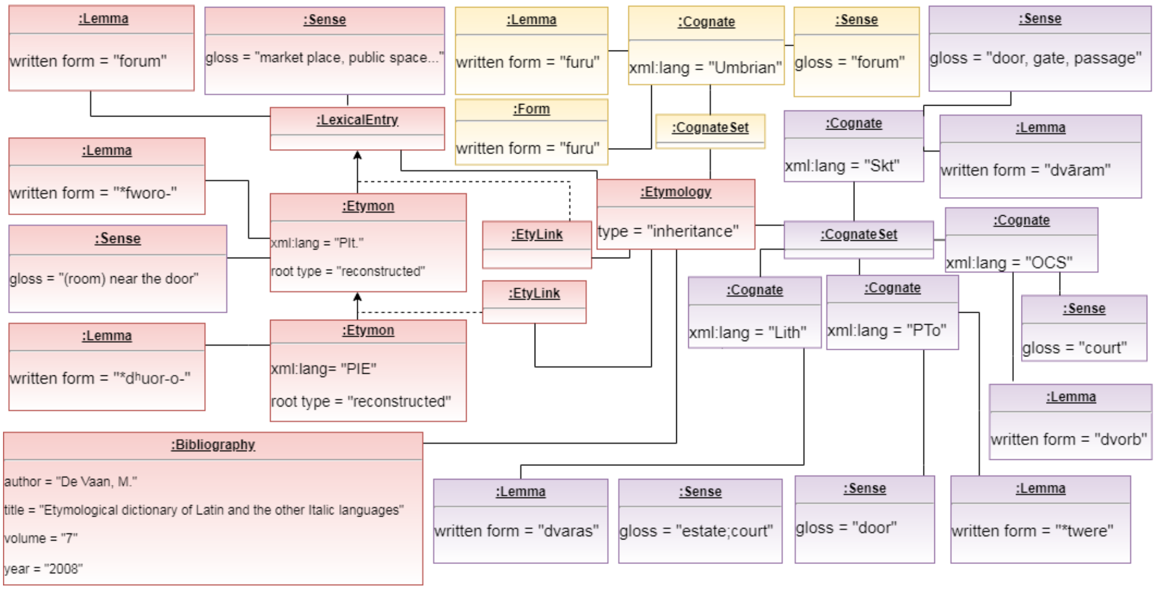


Figure 2: Encoding of the entry for *forum*.

define different etymologies featuring shared elements. In addition **Etymology** instances can be recursive, they can also be typed to define the changes undergone according to any number of linguistic processes.

Although we have had to leave out a number of details in this brief summary, the classes which we have enumerated above are the fundamental ones for understanding and using the standard. We were able to establish these classes over the course of numerous iterative design cycles during which draft proposals were reviewed against a large and diverse number of use cases: evaluating them on the basis of their salience, expressivity, and understandability. In Figure 1 we present the entry for the Latin word *forum* from De Vaan's *Etymological dictionary of Latin and the other Italic languages* (De Vaan, 2018), and in Figure 2 a partial encoding of this entry using LMF. Here we have focused chiefly on the information in the written entry which concerns etymons and cognates[8]. Note the relationship between the lexical entry and its two etymons (both of which have been categorized as reconstructed lexemes). We have also added two **Cognate Set** elements which, although we haven't shown it in the diagram, can be linked to their associated etymons. Note that these elements are linked to the LMF lexical entry for *forum* via an **Etymology** element which is in reality a container for an ordered set of **EtyLink** elements. It is important also to note that not all of the information in an etymology can be easily represented in a graph like structure and this we can instead represent in additional textual elements.

[8] There exists provision in LMF-Ety for representing information concerning attestations, references to secondary literature and for adding textual information as notes attached to entries or etymologies although we haven't presented it here. We also haven't added explicit temporal important to this example either. Full details will be available in the final version of the standard.

4 Acknowledgements

The first author was supported by the EU H2020 programme under grant agreements 731015 (ELEXIS - European Lexical Infrastructure).

References

Jack Bowers and Laurent Romary. 2016. Deep encoding of etymological information in TEI. *Journal of the Text Encoding Initiative* 10. https://jtei.revues.org/1643

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

Michiel De Vaan. 2018. *Etymological dictionary of Latin and the other Italic languages*, volume 7. LEIDEN·BOSTON, 2008.

Gil Francopoulo. 2013. *LMF lexical markup framework*. Wiley Online Library.

Anas Fahad Khan. 2018. Towards the representation of etymological data on the semantic web. *Information* 9(12):304.

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

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. *The OntoLex-Lemon Model: Development and Applications*, pages 587–597. https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf

Laurent Romary, Mohamed Khemakhem, Monte George, Jack Bowers, Fahad Khan, Mandy Pet, Stephen Lewis, Nicoletta Calzolari, and Piotr Banski. 2019. Lmf reloaded. In *Asialex 2019*.

Susanne Salmon-Alt. 2006. Data structures for etymology: towards an etymological lexical network. .