Journal of the Text Encoding Initiative

Issue 10 | 2016
Selected Papers from the 2015 TEI Conference

Deep Encoding of Etymological Information in TEI

Jack Bowers and Laurent Romary

revues.org

Publisher
TEI Consortium

Electronic version
URL: http://jtei.revues.org/1643
DOI: 10.4000/jtei.1643
ISSN: 2162-5603

Electronic reference
Jack Bowers and Laurent Romary, « Deep Encoding of Etymological Information in TEI », *Journal of the Text Encoding Initiative* [Online], Issue 10 | 2016, Online since 11 July 2017, connection on 03 October 2017. URL : http://jtei.revues.org/1643 ; DOI : 10.4000/jtei.1643

Deep Encoding of Etymological Information in TEI

Jack Bowers and Laurent Romary

1. Introduction

1 This paper aims to provide a comprehensive modeling and representation of etymological data in digital dictionaries. The purpose is to integrate in one coherent framework both digital representations of legacy dictionaries and born-digital lexical databases that are constructed manually or semi-automatically. We propose a systematic and coherent set of modeling principles for a variety of etymological phenomena that may contribute to the creation of a continuum between existing and future lexical constructs, so that anyone interested in tracing the history of words and their meanings will be able to seamlessly query lexical resources.

2 Instead of designing an ad hoc model and representation language for digital etymological data, we will focus on identifying all the possibilities offered by the TEI Guidelines for the representation of lexical information. This will lead us to systematize some existing constructs offered by the existing TEI framework, in particular the use of citation (`<cit>`) for representing etymons in place of `<mentioned>`, and referencing constructs (`<pRef>` and `<oRef>`) for linking etymological

information to existing or putative lexical entries. We also suggest some amendments to the TEI Guidelines that may improve the representation of etymological information as well as lexical entries at large (for instance, deprecation of <oVar> and <pVar>).[1]

3 Since its initial design in the 1990s (Ide and Véronis 1994; Ide and Véronis 1995), the TEI "Dictionaries"[2] chapter (TEI Consortium 2016, chapter 9) has been the basis for a large number of dictionary projects. It has shown its capacity to take into account a variety of perspectives on lexical content, whether one wants to closely follow the original structure of the source material (so-called "editorial view"), or abstract away from it to go closer to a real lexical database ("lexical view"). This has led to an important body of literature (including Erjavec, Tufis, and Varadi 1999; Budin, Majewski, and Moerth 2012; Rennie 2000; Bański and Wójtowicz 2009; Fomin and Toner 2006). Most of these papers have been focused on presenting the general architecture of lexical entries in the corresponding dictionary projects, and on describing the way the various TEI elements have been set up and used over the course of the editorial workflow.

4 Concerning etymological description on the theoretical level within the field of linguistics, as we shall see, phenomena such as metaphor, metonymy, and grammaticalization are very well established, particularly within cognitive linguistics. However, we know of no attempts to represent such processes within any lexical markup system. Additionally, with regard to the theoretical background, very little has been written on the corresponding digital models when such information is being integrated into a lexical database. This is why we will mainly position our work as an elaboration upon the seminal proposals of Salmon-Alt (2006), which represent a unique set of approaches to data modeling for etymological information.

5 Finally, though they are not the primary focus of our paper, we present herein examples of how encoders may make use of linked open data URIs[3] in defining the semantics (sense and/or domain) of a lexical entry; this issue has been discussed recently by Schopper, Bowers, and Wandl-Vogt (2015). The integration of the burgeoning resources of the semantic web with TEI represents a step toward a model of digital lexicography which enables conceptual semantics to play a more prominent role in the representation of linguistic content by grounding such information in the growing networks of ontological knowledge bases.

2. A Quick Overview of the TEI Recommendations for Dictionaries

6 The representation of lexical information is only one of many types of textual forms that are covered by the wide scope of the TEI Guidelines. A dictionary represented in TEI follows all the basic assumptions concerning the general structure of TEI-conformant documents. In particular, all metadata elements related to the identification of the sources used in the document, the various responsibilities in its digital encoding, as well as the possible conditions of publication and reuse can all be described within the TEI header (`<teiHeader>`) element, which is a mandatory component of all TEI documents. In the same way, the actual lexical content of a dictionary document expressed in TEI can be further structured at any depth using the generic division (`<div>`) mechanisms. As a whole, the structural divisions of TEI dictionary entries and their component elements can be seen as analogous to any other type of structured subsection (containing elements such as title, section headers, and paragraph) that may occur in a document containing prose.[4]

7 Besides generic textual constructs, the TEI Guidelines provide a variety of elements to represent dictionary entries, including a general-purpose `<entry>` element for structured content, a specific `<entryFree>` element to provide a flat representation, for instance in the course of a digitization workflow, and a `<superEntry>` container to group together homonyms. Over the course of this paper we will focus on the `<entry>` element, whose organization reflects a standard semasiological model of lexical content.[5] Indeed, the `<entry>` element is mainly organized around two subcomponents:

- a `<form>` element contains the description of the phonetic, orthographic, and morphological characteristics of the headword as well as its possible inflections. This element may, for instance, contain further grammatical constraints (using the `<gramGrp>` element);

- one or more <sense> elements that group together all descriptions related to the various senses that can be associated with the headword. A variety of further descriptors are available in <sense> to provide such information as a definition (<def>), examples or translations (<cit>), various grammatical (<gramGrp>) or usage (<usg>) constraints, and of course etymological information (<etym>).

8 The TEI "Dictionaries" chapter also provides various mechanisms for cross-referencing entries to other components of a dictionary. In particular, we will see in this paper how we can make use of references to the orthographic form (<oRef>) or pronunciation (<pRef>) of a headword within an etymological description.

9 When using the <entry> element, it is particularly important to identify the language information attached to any descriptive element within such representations; in particular, an encoder needs to be able to clearly state the object language of the entry as a whole (the language about which the entry provides a lexical description) and the various working languages (the languages in which various descriptive objects such as definitions, notes, and etymons are expressed). To this purpose, in compliance with ISO standard 16642[6] for terminological data, we recommend using the @xml:lang attribute as follows: @xml:lang is a mandatory attribute for each <entry> and indicates the object language of the whole entry. When not superseded by another indication further down in the entry structure, it also states the working language for all descendant elements within it, when appropriate (i.e., for textual content). When the working language differs locally from the object language of the entry (e.g., a definition expressed in another language than the one being described), a new @xml:lang attribute may be attached to the corresponding element.

10 The representation of etymological information, whether at entry level or for a specific sense, relies on the <etym> element, which we will elaborate upon as we tackle specific phenomena. So far, <etym> has been used as a flat construct where relevant information concerning language (<lang>), etymon (<mentioned>), or source (<bibl>), for instance, would be simply marked up in the flow of a textual etymological description. The purpose of this paper is to deepen the possible usage of <etym> and systematize the way in which specific etymological phenomena can be represented.

3. Past Treatment of Etymological Markup

11 None of the previous attempts either to create a lasting, well-formatted digital corpus of etymological data, or to establish a widely adopted set of recommendations for encoding such information, have ultimately been very successful. Many of the projects which attempted to create such resources have seen the same fate as so many others in the humanities (despite their stated goals of following best practices for interoperability); such problems include obsolescence of formatting and/or encoding scheme, abandonment of project, websites that no longer exist, broken links, and incompatibility with existing software.

12 However, it is useful to review a few publications (and data where possible) that have led to our current understanding of a generic way to represent (digital) etymological information, as it contributes to establishing an understanding of key questions, challenges, and issues that the authors encountered, as well as to recognize what people may be looking for when undertaking such projects. To this end, we will review four main references that have either paved the way for the current status quo in the TEI Guidelines or directly influenced our own understanding of the guidelines and how they should evolve.

13 A major milestone in the digital dictionary era is the work by Amsler and Tompa (1988), which, together with the unifying contributions of Ide and Véronis (1994), led to the earlier TEI "Print Dictionaries" chapter. Focusing here on their contribution to etymology, example 1 shows how they have introduced a highly structured model based on etymons and links implemented as an SGML[7] DTD.[8] The underlying model is based upon a graph of etymons (`<etymon>`) connected with relations (`<rel>`), forming a more global etymological tree (reflected by the `<es>`, etymological segment, element).

Example 1. Etymological representation from Amsler and Tompa (1988).

```
<E>
      <es>
            <etymon   lang=ME>appel</etymon></es>
      <es>
            <rel>fr.</rel>
            <etymon   lang=OE>&aelig;ppel</etymon></es>
      <es>
            <rel>akin   to</rel>
            <eu>
                  <etymon   lang=OHG>apful</etymon>
                  <deftext>apple</deftext></eu>
            <eu>
                  <etymon   lang=OSlav>abl&breve;ko</etymon></eu></es></E>
```

14 The next two examples were early attempts to build corpora capable of representing etymological data using standards.

15 An early pre-XML application of the TEI Guidelines to the systematic recording of etymological information can be found in Good and Sprouse (2000). The work was carried out in the context of the Comparative Bantu Online Dictionary (CBOLD), a complex database for multiple Bantu languages, and used the SGML P3 edition of the TEI Guidelines.[9] The content corresponds to the digitization of existing print dictionaries and word lists, and the authors marked up these texts according to the TEI Guidelines, with some tags added to the standard set. Since the project had to deal extensively with etymological information, it used `<etym>` with a refined recommendation to link etymons (marked up as `<xref>`) to a list of reconstructed historical forms.

16 In the same vein, Jacobson and Michailovsky (2002) used an even simpler approach for their etymological references within a TEI-based encoding of their lexical data. Instead of implementing `<etym>`, they make a plain use of the generic `<ptr>` element, typed as `"cfetym"`, to point to other entries in their dictionary that may be seen as etymological sources.

3.1 Crist (2005)

17 Sean Crist provides some analyses of approaches and a correspondingly precise set of principles to be applied to the Germanic Lexicon Project,[10] which digitized a collection of dictionaries of various Germanic languages whose copyright had expired. The ultimate goal for markup formatting was TEI, but (for reasons unknown) it was apparently never achieved. Notably, Crist mentions the likely need to extend the guidelines in the area of etymology because the `<etym>` element lacks the means for precisely encoding etymological relationships between entries and forms. Key components of this work were the following:

- XML markup of some of the data, while other portions remain as plain text, or simply image scans of the originals;
- formal interrelations among all of the words in an etymology; those specified are cognation, inheritance, and borrowing;
- the use of attribute inheritance for the nodes in the data structure as per Ide, Kilgarriff, and Romary (2000);
- a proposal for the system to require no privileged frame of reference, which would allow data to follow one of three formats.

18 The following excerpt is from the paper, assuming that the structures represent the place in the XML hierarchy in which each data type would occur.

Example 2. Abstracted model of etymological description (Crist 2005, 8–9).

1. (From the vantage point of Modern English) Modern English stone is a reflex of Old English st¯an, which is a reflex of Proto-Germanic *stainaz:

word
  form: stone
  language: Modern English
  etymon
    word
      form: st¯an
      language: Old English
      etymon
        word
          form: stainaz
          language: Proto-Germanic
          attested: no

2. (From the vantage point of Old English) Old English st¯an is an etymon of Modern English stone, and is also a reflex of Proto- Germanic *stainaz:

word
  form: st¯an
  language: Old English
  reflex
    word
      form: stone
      language: Modern English
  etymon
    word
      form: stainaz
      language: Proto-Germanic
      attested: no

3. (From the vantage point of Proto-Germanic) Proto-Germanic *stainaz is an etymon of Old English st¯an, which is an etymon of Modern English stone:

word
  form: stainaz
  language: Proto-Germanic

```
attested:   no
reflex
            word
                        form:            st¯an
                        language:   Old   English
                        form:   stone
                        language:   Modern   English
```

19 Crist outlined a typology of the treatment of etymological markup at the time, while the adoption of standards and the fields of digital humanities and lexicography have been steadily gaining momentum since then. With regard to etymological markup, Crist's typology remains fairly valid. The classifications are as follows:

> "Type I. Markup schemes which make no provision for etymological data" (the majority of lexical markup systems);
> "Type II. Markup schemes where etymological data is delimited as such, but is treated as unstructured prose" (including TEI; Crist points out the need for further structure, possibly reusing structures from other sections of the TEI specifications, especially the dictionary chapter);
> "Type III. Markup schemes where the mathematical relationships recognized in historical/comparative linguistics are somehow embodied in the markup system in (semi-)machine-readable form." These systems, according to Crist, "make some provision for the formal encoding of etymological relationships between words" (Crist 2005, 13, 17).

3.2 Salmon-Alt (2006)

20 The most significant attempt at devising this kind of dynamic system of etymological markup was that of Susanne Salmon-Alt. While etymology is not addressed in the LMF (ISO 24613:2008) standard (ISO 2008),[11] Salmon-Alt attempted to develop an extension of the model for the encoding of etymological markup. The extension module allows for the integration and linking of the etymological information of an entry with the synchronic data and any classifications of a given word/entry within the core module of LMF.

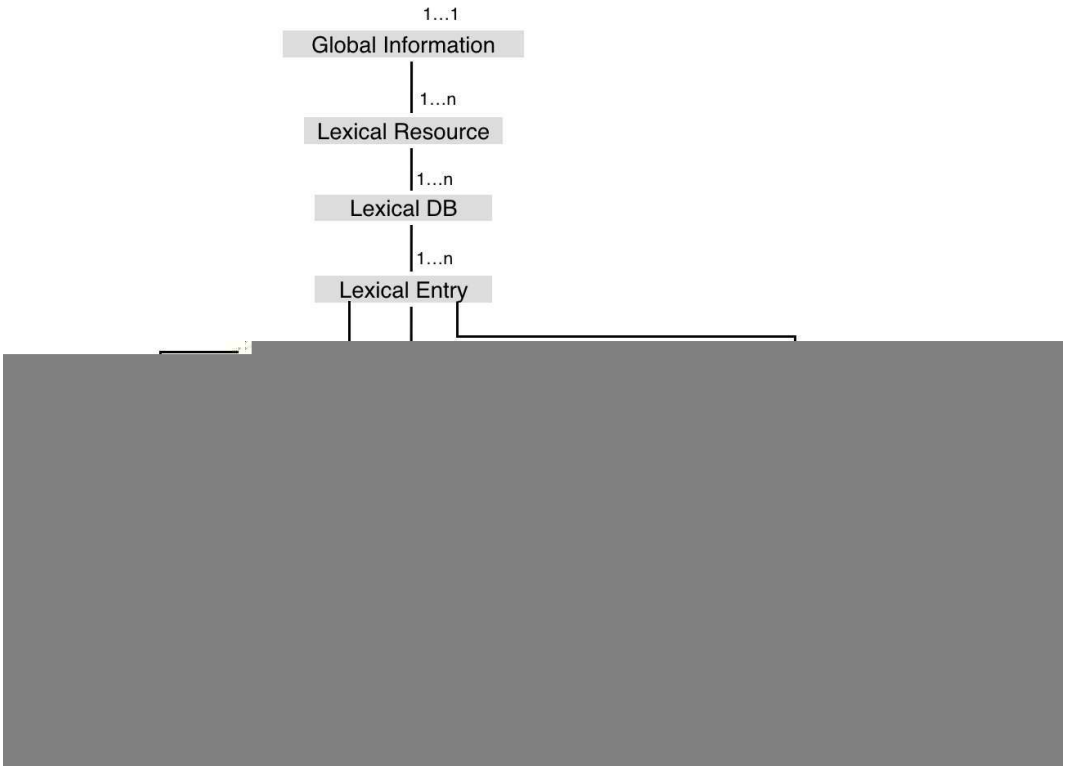
> Our model is based on the overall hypothesis that etymological data might be thought of as a lexical network, i.e., a graph, whose nodes are lexical units (located in space and time) and whose arcs are typed etymological relations.
>
> (Salmon-Alt 2006, 3)

21 The scope of the model was limited to semasiological organizational principles for single lexical entries, such as those outlined in the TEI P5 Guidelines, and did not attempt to support the approaches of many traditional etymological dictionaries in which the structural principles and contents vary significantly from one another.

22 In laying the conceptual and functional foundation for the approach, Salmon-Alt defines etymology proper as concerning the origin and evolution of a lexeme before its entrance into the lexicon of a given language, as it is materialized by one or more etymons.

23 The metamodel of the LMF etymology extension from the paper is shown in the diagram below.

Figure 1. Etymological extension to LMF (Salmon-Alt 2006).



24 This diagram reflects the XML representation suggested by Salmon-Alt (2006),[12] with two dedicated elements, `<etymon>` and `<etymologicalLink>`, that each contain a specific set of information. They are defined as follows:

Etymon: `<etymon>`

The basis for describing and encoding etymons in this model is parallel to that of synchronic lexical entries. Specifically, they are characterized by: language (`@xml:lang`); the linguistic form(s) (`<form>`)—orthographic (`<orth>`) and/or phonetic (`<pron>`); sense (`<sense>`); gloss (`<glose>`); grammatical classification (`<pos>`); and inflectional information (if applicable).

Additionally, within the etymon portion of the markup are the optional etymological notes,which serve as a kind of "et cetera" section where one can include other relevant information about the etymon, such as discussions and/or bibliographic references regarding intermediate stages of development, phonetic evolution, concurrent hypotheses, statements of confidence, and secondary etymons. In our system presented herein, we have refined and given more structure to the markup of these datatypes.

Etymological Link: `<etymologicalLink>`

The `<etymologicalLink>` section is intended to be where the relations between the synchronic and diachronic elements, or possibly between multiple stages, etymological relationships, or alternative hypotheses of diachronic components are specified and defined. The main way in which this is done in the data structure is through the pointer mechanisms (`@source`, `@target`) in the attributes which link the lexical entry to an etymological classification specified by means of an `<etymologicalClass>` element which can occur within each `<etymologicalLink>`.

Specified as element values, etymological classes in the model are inheritance, loan word, and word generation, though in the case of disputed word origins, each alternative may have different classifications if need be, and levels of confidence can also be specified (`@confidenceScore`).

25 After reviewing the literature on this topic, we can identify several commonalities, the first of which is that all authors looked to the TEI, but none found it sufficient to adopt without alterations. Additionally, all works reviewed desire the markup system to have:

- a systematic inventory of typed pointers to link between: etymological forms (etymons) and their synchronic descendants; parallel synchronic forms in related languages (e.g., cognates); multiple synchronic forms in a single language;
- structures dynamic and consistent enough to enable automatic processing, manipulation, and evaluation with software applications;
- the ability to classify and assign typological labels to an etymological entry;
- a means of expressing level of certainty of etymological analysis;
- a means of decomposing compounds and components of derivational morphology.

26 In this paper, we elaborate on the sources presented above. We specifically focus on building upon the general model proposed by Salmon-Alt (2006), which is based on a network of etymons and links. Moreover, we identify a more precise group of link types between etymons and explore the consequences in terms of both theoretical implications and possible representations in the TEI framework. Whereas this was not completely stated in Salmon-Alt (2006), we are describing etymological links as the expression of specific etymological processes between etymons (forms), lexical entries, or even senses within entries. The following sections describe the type ontology that we have devised.

4. Basic Mechanisms for Representing Etymological Processes

4.1 An Extended TEI-based Representation of Etymons

27 The current content model of the `<etym>` element and the documentation and examples available in the TEI Guidelines favor a flat annotation of etymological content that does not put forward either the actual nature of etymons as references to dictionary entries or the central role of

etymological links in the diachronic processes. Starting from the following example expressed in current recommendations of the TEI Guidelines, we show in this section a better representation of etymons in etymological description:

```
<entry>
 <form type="headword">
  <orth>Âbend</orth>
 </form>
 <gramGrp>
  <gen>Mask.</gen>
 </gramGrp>
 <!-- sense, other info here -->
 <etym>
  <lang>Ahd.</lang> <mentioned>âband</mentioned>,
  <lang>mhd.</lang> <mentioned>âbent</mentioned>;
  <bibl>zur Etym. s. Kluge Mitzka 18. Aufl. unter ,,Abend'', ferner Schwäb. Wb.
1, 11ff. Schweizdt. Wb. 1,34ff.</bibl>
 </etym>
</entry>
```

28 As we can see, the TEI Guidelines have thus far favored the use of the <mentioned> element as the basis for marking up etymons.

29 The first reason we think this representation is problematic is that it introduces a specific mechanism to refer to lexical items, whereas the TEI "Dictionaries" chapter also provides <oRef> and <pRef> in examples and <ref> in external references. Therefore we suggest that such references be considered a single process of referring to other lexical entries at large, whether within the same entry, within the same dictionary, or potentially to a lexical entry from another dictionary. In the latter case, the dictionary may or may not exist for the corresponding language, and thus the reference may be used as a placeholder for a potential construction. This is typically the case for etymons that refer to other languages or ancient forms thereof, even if such forms are not (yet) part of a real lexical description.

30 To this purpose, we make the recommendation to systematically use <oRef> and <pRef> in all three constructs (examples, etymology, and external references), and thus supersede both <mentioned> and <ref> for such usages.

31 The schematic structure that we propose for this construct is as sketched below:

```
<cit   type="etymon">
<oRef>|<pRef>
<date>|<gramGrp>|<usg>|<gloss>|<ref>
</cit>
```

32 Here we see that we can have <oRef> or <pRef> (or possibly both) to refer to the form of the etymon, to which we add further information or constraints related to dating (<date>), grammatical information (<gramGrp>), semantic domain or register (<usg type="…">), or translational equivalence (<gloss>). There could of course be additional constraints depending on the complexity of the available etymological information, for instance when there is an explicit reference to an externally defined sense, beyond the shallow capacity of <gloss>,[13] as we shall see later in the paper.

33 Moreover, the use of <oRef> and <pRef> here is quite important in our model, since it reflects the vision that an etymon is a potential reference to a lexical entry in a dictionary for the corresponding language either synchronically (e.g., in the case of loan words) or, more often, diachronically.

34 The second flaw with the current TEI proposals for etymology is the lack of mechanisms to group together an etymon with the possible constraints (language, grammar, usage) that may be associated with it. The flat annotation format leaves such pieces of information isolated, as we can see in the previous example, for even the central language information (coded with <lang>).

35 Here again, we take up a construct that already exists in the TEI "Dictionaries" chapter to encompass this new use case, namely <cit>. By definition, <cit> groups together a linguistic segment with additional features that document its usage, and it is currently used for examples and translations in dictionary entries. We suggest extending its scope to make it the central construct for the representation of etymons in combination with the use of <oRef> and <pRef> we have just described.

36 If we take up the preceding example, we can turn it into our suggested representation as follows:[14]

41 Any change that occurs within a lexicon can be labeled in the `@type` attribute of the `<etym>` element in a TEI dictionary;[16] and the fact that a change occurred within the contemporary lexicon (as opposed to its parent language) is indicated by means of `@xml:lang` on the source form.[17]

42 In the TEI encoding, the former two can be respectively labeled as:

```
<etym type="borrowing">…</etym>
```

and

```
<etym type="inheritance">…</etym>
```

Each of the above would represent the top level `<etym>` element, and any other subprocesses can be encoded as embedded `<etym>` elements with `@type` attributes.

43 Alternatively, they can be implicitly encoded as the value of the attribute `@xml:lang` within the source `<oRef>` and/or `<pRef>` form without having to embed one or more `<etym>` elements where this information is redundant or implicitly understood, thus simplifying the data structure. For instance, in an entry denoting changes to the morphological and phonological form of a given Bavarian word inherited from Middle High German, rather than doing the following for every instance of inheritance:[18]

```
<etym type="inheritance">
 <etym type="phonological-processA">
  …
  <pRef xml:lang="gmh">…</pRef>
 </etym>
</etym>
```

it may be desirable to simply specify the source etymon form thus:

```
<etym type="phonological-processA">
 …
 <pRef xml:lang="gmh">…</pRef>
</etym>
```

44 As long as the parent or source language(s) of an entry form are known, and a project-specific or external ontology or schema is declared somewhere, this method allows for a lighter means of expressing the source language of a form.

45 The fact that the origin of the form is borrowed from another language or inherited from a "parent" language can also be implicitly encoded by simply including a language tag distinct from that of the main (synchronic) portion of the entry.

4.3 Using <cit> for Etymons and More

46 Having covered the proposed usage model for <cit type="etymon"> we now move on to two further functions for the <cit> element, namely to represent (sub-)components of an etymological form (e.g., decomposition) and attestations of their usage.

47 Components are significant lexical subunits which editors may want to isolate explicitly within various types of etymological formation. They are indeed objects that are very close to etymons but cover a wider variety of linguistic segments such as morphemes. We will see in the course of the paper how a <cit type="component"> makes sense for this purpose.

48 Attestations of historical forms of an etymon in source context can be included within a citation element as <cit type="attestation">. As the contents of the citation are a quotation, the sampled linguistic content of the attestation is contained within the <quote> element. Within <quote>, the referenced form of the attested etymon can be encoded in the <oRef> element to specify which portion of the text corresponds to the given etymon.

49 Furthermore, where an attestation of an etymological form is in a language other than the entry itself, it can be necessary to include translations of the attestations, which can simply be encoded with <cit type="translation"> embedded within the <cit type="attestation"> (they are embedded within the attestation because they pertain specifically to the language content of the attestation). Thus, the XML structure mirrors that of attestation, with the only differences being the value of @type and @xml:lang.

4.4 Encoding Languages and Representational Aspects of Linguistic Forms

50 Over the course of our research we have constantly faced issues related to the actual encoding of language-related information for headwords in dictionary entries as well as for etymons and similar references. This covers the whole range of TEI elements we are considering here: `<orth>` and `<pron>`, with their referential counterparts `<oRef>` and `<pRef>`.

51 As a basis for our representation, we have of course taken up the use of `@xml:lang` over these elements together with the constraints applicable to its values and the guidelines of BCP 47 (IETF 2009a). We will not go into detail here about the possible limits of this recommendation; instead, we need to explain some of the choices that are applied in our paper.

52 First, there is a general issue with the language coverage offered by BCP 47, which is based upon the IANA registry[19] and only offers language tags for a small number of the historical languages that are needed even from the reduced perspective of Western European etymology: Latin `"la"`, Old French `"fro"`, and Middle French `"frm"`.[20]

53 Additionally, we identify a problem with the abstract context of general language markup BCP 47 recommendations, which specify that both the content language and orthographic script be labeled within the `@xml:lang` attribute (IETF 2009a). This is neither a conceptually accurate (as an orthographic system is of course not a language), nor a functionally pragmatic means of representing this information. Functionally, there is no difference between orthography and phonetic notation (with various degrees of nuance and exceptions for logographic, logosyllabic, and other such systems); they are both representations of the language information at some level. With regards to phonetic information (at least in the dictionary module), in any linguistic data there is an obvious need to distinguish between the various phonetic transcription systems, and the TEI already has a solution in `@notation`.

54 Similarly, we have observed difficulties in relying on the `@xml:lang` attribute alone to cater for the various ways orthographic forms can actually occur beyond simple script variations, such as vocalized/non-vocalized in Arabic, kana/kanji in Japanese, and competing transliteration systems. Whereas BCP 47 at times introduces ways of dealing with such variations or even allows one to define its private subtags to do so, we found it cumbersome to overload `@xml:lang`, the main disadvantage being the loss of systematicity in the way a given language is marked in an encoded text.[21] This is why we have extended the `@notation` attribute to `<orth>` in order to allow for better

representation of both language identification and the orthographic content. With this double mechanism, we intend to describe content expressed in the same language by means of the same language tag, thus allowing more reliable management, access, and search procedures over our lexical content.

55 We are aware that we open a can of worms here, since such an editorial practice could be easily extended to all text elements in the TEI Guidelines. We have actually identified several cases in the sole context of lexical representations (e.g., `<quote>`) where this practice would be of immediate use.

4.5 Encoding Sequence (Diachronic or Order of Presentation in Source)

4.5.1 Maintaining Source Structure vs. Accurate Representation of Etymological Process

56 When encoding multistage diachronic etymologies from existing sources such as attestations from etymological dictionaries or academic papers, it may be the case that the source information is not in chronological order for one reason or another. In such cases it is up to the encoder to decide whether maintaining the format of the source is of any benefit to the data quality.

4.5.2 Sequence

57 Because the etymology being encoded in the example has multiple stages in which the sequence is both known and theoretically relevant, each `<cit>` should be given an `@xml:id` in combination with one or both of the sequential pointer attributes: `@prev`, `@next`. The combination of these within the data structure encodes relative occurrence of the given etymon within the diachrony of any example.

5. Inheritance

58 While inheritance is not itself an etymological process, it identifies lexical items known, or presumed, to be inherited from predecessor or "parent" languages; these forms are sometimes referred to as "native." A simplified view of inheritance is that it is the etymological counterpart to borrowing.[22] In the most basic use of an inheritance etymology, an encoder can simply distinguish the given lexical item as having originated directly from its known parent language, or even theoretical protolanguage.

59 However, within the historical trajectory of most inherited lexical items, any number of different etymological processes may occur on every level of language, including phonetic/phonological, phonotactic, morphological, grammatical, and/or semantic.

60 The basic concepts necessary for a minimal encoding of a simple inheritance etymology are:

- language of the (synchronic) entry;
- parent/predecessor language;
- the synchronic orthographic and/or phonetic form(s), sense(s), and/or grammatical information;
- the diachronic orthographic and/or phonetic form(s), sense(s), and/or grammatical information.

61 The following is a very simple example from Sardinian *semper* "always, still," in which the etymology shows that, at least as far as the orthographic form of the lexical item is concerned, it has not changed from its Latin source. This is perhaps noteworthy in its own right, and sampling an entire lexicon in which such information is included could be useful in measuring how much a language has changed over a given period of time.

Example 3. TEI modeling: `<etym type="inheritance">`

```
<entry xml:id="semper" xml:lang="srd">
 <form type="lemma">
  <orth>semper</orth>
  <gramGrp>
   <pos>temporalAdverb</pos>
  </gramGrp>
 </form>
 <sense>
 ...
 </sense>
 <etym type="inheritance">
  <cit type="etymon">
   <oRef xml:lang="la">semper</oRef>
  </cit>
 </etym>
</entry>
```

62 Note here that we could have a `@corresp` attribute on `<oRef>` if we had a reference Latin dictionary at hand and wanted to point to the actual entry for *semper*.

5.1 Inheritance and Phonetic/Phonological Changes

63 For any lexical item—regardless of other types of etymological processes undergone over a sufficient span of time—it is likely that it will undergo some degree of phonetic change. Such changes may occur either over a span of time during which a descendant language has become distinct from its "parent" (such as Vulgar Latin > French) or within a span of time in which they are regarded as having occurred within the same language. Phonetic and phonological changes often occur in stages and have their own set of classifications and terminology that require their own level of encoding separate from those occurring on the higher levels of language such as morphosyntax and semantics.

64 The basic concepts necessary for a minimal encoding of a phonological etymology within an inherited entry are:

- the language of the synchronic entry;
- the parent/predecessor language at the given stage of etymology;
- the synchronic orthographic and/or phonetic form(s);
- the diachronic orthographic and/or phonetic form(s);
- the relative order of their usage/occurrence (if multiple stages are shown, and their sequence is known).

Additionally beneficial are:

- dates for each of the diachronic forms;
- bibliographic sources for forms, and for the analysis.

5.2 Stages of Phonological Changes in Inherited Forms

65 The following is our proposal to encode an example of the most significant stages of the phonetic evolution of the French *chef* from the Vulgar Latin[23] *CÁPŬ* as per Laborderie and Thomasset (1994). Each `<cit type="etymon">` element cluster contains the historical phonetic forms posited by the

authors in the <pRef> element, as well as other relevant information pertaining to the given stage in the diachrony of the entry. The etymon clusters begin with the Vulgar Latin form (top) and end with the Middle French form (bottom).[24]

Example 4. Phonological stages in inherited form[25]

```
  <entry xml:id="chef" xml:lang="fr">
   <form type="lemma">
    <orth>chef</orth>
    <pron notation="ipa">ʃɛf</pron>
    <gramGrp>
     <pos>noun</pos>
     <gen>masc</gen>
    </gramGrp>
   </form>
   <sense>
   ...
   </sense>
   <etym type="inheritance">
    <cit type="etymon" xml:id="kápŭ" next="#kábu">
     <pRef notation="private" xml:lang="la">kápŭ</pRef>
    </cit>

    <cit type="etymon" xml:id="kábu" prev="#kápŭ"><!-- intervocalic voicing -->
     <date notBefore="0350" notAfter="0399"/>
     <pRef notation="private" xml:lang="la">kábu</pRef><!-- gallo-latin or (VL-
Gaul) -->
    </cit>

    <cit type="etymon" xml:id="ˌkáβ̥o" prev="#kábu" next="#ˌtáβ̥o">
     <date notBefore="0400" notAfter="0499"/>
     <pRef notation="private">kˌáβ̥o</pRef><!-- late gallo-latin ?-->
    </cit>

    <cit type="etymon" xml:id="ˌtáv̥o" prev="#ˌkáβ̥o" next="t͡sáv̥o">
     <date notBefore="0400" notAfter="0499"/>
     <pRef notation="private">tˌáβo</pRef><!-- late gallo-latin ?-->
    </cit>

    <cit type="etymon" xml:id="t͡sáv̥o" prev="#ˌtáv̥o" next="#t͡šíe̯v̥o">
     <date notBefore="0400" notAfter="0499"/>
     <pRef notation="private">t͡sáv̥o</pRef>
     <!-- late gallo-latin ?-->
```

```
    </cit>

    <cit type="etymon" xml:id="t͡šíe̯v̯o" prev="#t͡šáv̯o" next="#tší̯ef">
     <date notBefore="0450" notAfter="0550"/>
     <pRef notation="private">t͡šíe̯v̯o</pRef>
     <!-- late gallo-latin ?/early gallo-romance-->
    </cit>

    <cit type="etymon" xml:id="tší̯ef" prev="#t͡šíe̯vo" next="#šy̯ef">
     <date notBefore="0600" notAfter="0699"/>
     <pRef notation="private">tší̯ef</pRef><!-- early gallo-romance-->
    </cit>

    <cit type="etymon" xml:id="šy̯éf" prev="#tší̯ef" next="#š̯éf">
     <date notBefore="0700" notAfter="0799"/>
     <pRef notation="private">šy̯éf</pRef><!-- early/Proto Old French (?) -->
    </cit>

    <cit type="etymon" xml:id="š̯éf" prev="#šy̯éf" next="#š´ęf">
     <date notBefore="1500" notAfter="1650"/>
     <pRef notation="private" xml:lang="frm">š̯éf</pRef>
    </cit>

    <cit type="etymon" xml:id="š´ęf" prev="#š̯éf">
     <date notBefore="1500" notAfter="1650"/>
     <pRef notation="private" xml:lang="frm">š´ęf</pRef>
    </cit>
    <bibl>Laborderie, N. and Thomasset, C. (1994). Précis de phonétique
historique. Paris: Nathan.</bibl>
   </etym>
 </entry>
```

66 The diachronic sequence of the forms is encoded in our markup as follows: the `@xml:id` attribute is included for each `<cit>` for which the given language information is available; the ordering of each is encoded in the data structure by the use of the pointing attributes `@prev` and `@next`, the values of which are the unique identifiers of the previous and next `<cit>` block respectively.

67 The <date>[26] element is listed within each etymon block; the values of attributes @notBefore and @notAfter specify the range of time corresponding to the period of time that the given form was in use according to the authors.[27] The attribute values encode the date according to W3C recommendations[28] and must have a four-digit representation of the year.[29] Finally, we have the <pRef> element, which, as always, contains the language of the given etymon as the value of @xml:lang and the notation attribute @notation.

5.3 Morphological and Morphosyntactic Changes in Inherited Forms

68 Changes in morphological inflection paradigms in which there is no difference need not be explicitly represented in a dictionary but instead can be done implicitly.

69 Because an entry in a TEI or other semasiological dictionary represents the etymology of an individual etymon as pertaining to an individual lexical item, diachronic changes in morphological inflection patterning are manifested in the differences in the phonetic and often orthographic forms. These differences are evident when contrasting a large sample of synchronic and diachronic phonological and phonotactic forms with respect to a given historical or contemporary morphological or morphosyntactic feature.

70 When attempting to extract any global information about changes to the grammatical feature inventory of a language from a dictionary, such information can also be inferred through the contrast between the contents of <gramGrp> in the source etymon (in the <cit type="etymon"> cluster) and the resulting (synchronic) form. For example, where the source form has a specific case and the entry in question does not have any case information, the general phenomenon that is the loss of grammatical case is implicitly present.

71 The following Sicilian entry shows an example of such a scenario in combination with a change in grammatical gender. Whereas in Latin, the etymon had the neuter gender, the inherited form in Sicilian is of the masculine gender. Additionally, Sicilian no longer has a neuter gender.[30] Both of these changes are implicit in the data structure we propose:

Example 5. Implicit morphosyntactic changes: Latin > Sicilian.

```
<entry xml:id="mare" xml:lang="scn">
 <form type="lemma">
  <orth>mari</orth>
  <gramGrp>
   <pos>noun</pos>
   <gen>masc</gen>
  </gramGrp>
 </form>
 <sense>
 ...
 </sense>
 <etym type="inheritance">
  <cit type="etymon">
   <oRef xml:lang="la">mare</oRef>
   <gramGrp>
    <pos>noun</pos>
    <gen>neut</gen>
    <case>nom</case>
    <iType>-i stem</iType>
   </gramGrp>
   <gloss>sea</gloss>
  </cit>
 </etym>
</entry>
```

72 As with morphology and phonology, grammatical etymological can be globally inferred. This means that, in a Sicilian TEI dictionary (of sufficient size) containing such synchronic and diachronic information, a search for all entries containing all of the following facets would enable the conclusion that there is no more neuter gender:

- `<pos>noun</pos>`, or `<pos>adjective</pos>`;[31]
- `<etym type="inheritance">` in which value of `@xml:lang` in `<oRef>` and/or `<pRef>` is `"la"`);
- `<cit type="etymon">` in whose `<gramGrp>` there is `<gen>neut</gen>`;
- Get synchronic gender: value of `<gramGrp>`, `<gen>` within `<form type="lemma">`.[32]

5.4 Morphological Inheritance: Inflected Forms

73 Individual morphologically inflected forms of entries may have different etymological histories than the lemma forms with which they are associated. This could be for reasons connected with a separate origin from the lemma, or, more commonly, with the fact that the phonological composition of the lemma and inflected forms were and are distinct and they each underwent unique changes.

74 To encode the etymology of any individual form (when the entry contains more than one that must be distinguished):

- a unique identifier `@xml:id` should be included in the given form;
- the corresponding `<etym>` should point to the form as the value of the attribute `@corresp`.

75 In the following example from Italian, phonotactic changes are reflected in the orthographic form (e.g., *perdidi* > *persí*), and the aspect of an inherited inflected form *persí* has also changed. In particular, it shows that the etymon originates from the Latin past tense, perfective aspect: *perdidi*. In Standard Italian and many (mainly northern) dialects, descendants of the Latin perfective past have come to be used as the remote past tense (i.e., remote aspect).

Example 6. Italian—etymological change in aspect (*perdere*).

```
<entry xml:id="perdere" xml:lang="it">
 <form type="lemma">
  <orth>perdere</orth>
  <gramGrp>
   <pos>verb</pos>
  </gramGrp>
  <!-- rest of inflected forms here -->
  <form type="inflected" xml:id="perdere-1s-rem-pt-indic">
   <orth>persí</orth>
   <gramGrp>
    <per>1</per>
    <number>sg</number>
    <tns>past</tns>
    <mood>indic</mood>
    <gram type="aspect">remote</gram>
    <gram type="voice">active</gram>
   </gramGrp>
  </form>
 </form>
 <sense>
  <!-- sense info here -->
 </sense>
 <etym type="inheritance">
  <cit type="etymon">
   <oRef xml:lang="la">perdere</oRef>
  </cit>
  …
  <etym type="inheritance" corresp="#perdere-1s-rem-pt-indic">
   <cit type="etymon">
    <oRef xml:lang="la">perdidi</oRef>
    <gramGrp>
     <per>1</per>
     <number>sg</number>
     <tns>past</tns>
     <mood>indic</mood>
     <gram type="aspect">perfective</gram>
     <gram type="voice">active</gram>
```

```
      </gramGrp>
     </cit>
    </etym>
   </etym>
  </entry>
```

6. Borrowing

76 Borrowing (a.k.a. loaning, importing, transferring, copying) can generally be described as the process in which a lexical item, phrase, or other linguistic feature from a foreign language or dialect is conventionalized into another language or dialect.

77 Matras and Sakel (2007) distinguish between two primary types of borrowing: *material borrowing* and *structural borrowing.* Material borrowing pertains to the borrowing of sound–meaning pairs (e.g., loanwords, phrases and/or affixes); structural borrowing pertains to the copying of syntactic, morphological, or semantic patterns.

78 A major factor in borrowing is contact between a foreign culture and a given language (Haugen 1950). There are often historical, sociological, and practical explanations for how and why the process occurred and thus such information encoded by a linguist or lexicographer may be relevant to those studying other fields such as history, anthropology, and sociology.[33]

79 Along these lines, distinctions have been made between *cultural borrowings* and *core borrowings* (Myers-Scotton 2002). Cultural borrowings are when a lexical item is borrowed for a new concept; these can appear over a very short period of time. Core borrowings duplicate existing items and are introduced over a longer period of time by bilinguals (Myers-Scotton 2002; Haspelmath 2003).

80 When in some manner a borrowed item is not compatible with that of the recipient language, loanwords undergo one or more adaptations to their orthographic, phonological, and/or morphological forms (Haugen 1950). Our examples will demonstrate encodings for each of these different types of adaptation.

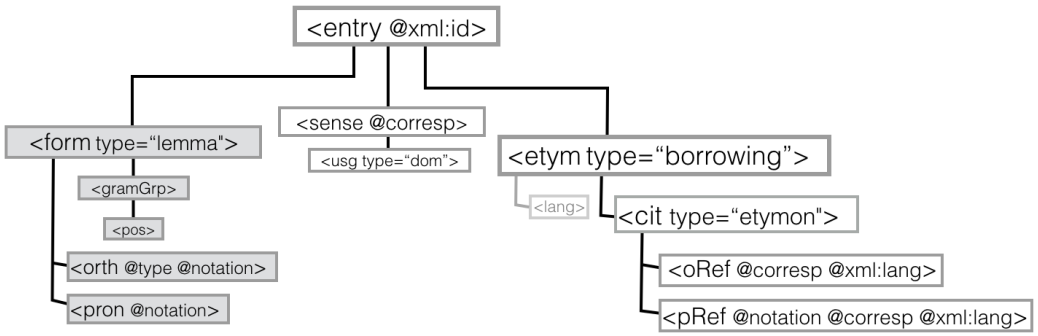
81 The basic concepts necessary for a minimal encoding of a "borrowing" etymology are:

- borrowing language (a.k.a. recipient language);
- source language (a.k.a. donor language);
- source form (orthographic and/or phonetic);

- borrowed form (orthographic and/or phonetic);
- semantic profile/metalinguistic concept or grammatical function of borrowed form (in the case of grammatical borrowing).

82 The combination of these elements in a TEI-based model naturally leads to the overall model depicted in figure 2.

Figure 2. Basic components of the model for etymological borrowing.



83 Borrowing-typed etymologies should be direct children of `<entry>`, as the process of borrowing does not affect the sense of a pre-existing lexical item within the target language. The source language should be labeled with the `@xml:lang` attribute within the `<pRef>` and/or `<oRef>`, and can optionally be labeled explicitly for human readability as the value of the element `<lang>` (see following example). If desired, in addition to the `<lang>` element, other key terms can be included in the `<lbl>` element to modify the language (e.g., `<lbl>source</lbl>`, `<lbl>from</lbl>`, etc.). If the semantics of the item in the source language differ from that of the borrowing language, the original can be included in the element `<gloss>`.

84 Example 7 shows our TEI adaptation from LMF (Salmon-Alt 2006) of the French *pamplemousse*, which is a borrowing from Dutch. The bibliographic source of this entry is the *Trésor de la Langue Française*, which is cited in the `<ref>` element both in the value of the attribute `@target="#TLF"`, and as the content of the `<ref>` element. The former points to a bibliographic entry for the *Trésor de la Langue Française* that is necessarily located within the given document or project; the latter is included for human readability.

Example 7. Simple borrowing *pamplemousse* from Salmon-Alt 2006; converted from LMF.

```
<entry xml:id="LE1" xml:lang="fr">
 <form type="lemma">
  <orth>pamplemousse</orth>
  <gramGrp>
   <pos>commonNoun</pos>
   <gen>masculine</gen>
  </gramGrp>
 </form>
 <sense>
 ....
 </sense>
 <etym type="borrowing">
  <lang>Dutch</lang>
  <cit type="etymon">
   <oRef xml:lang="nl">pompelmoes</oRef>
   <gloss xml:lang="lat">Citrus maxima</gloss>
   <gramGrp>
    <pos>commonNoun</pos>
    <gen>feminine</gen>
   </gramGrp>
   <note>probablement d'origine tamoule, De Vries, Nedrl</note>
   <ref target="#TLF">TLF</ref>
  </cit>
 </etym>
</entry>
```

6.1 Orthographic Adaptation in Borrowed Forms

85 The following example is a revised version of an entry from the TEI Guidelines of the word *biryani* (or *biriani*), which was borrowed into English from Urdu. Herein we observe that in English there is variation in the orthographic adaptation of the borrowed item, which is most commonly spelled "biryani," but also sometimes "biriani." This relative frequency between the two variant forms is encoded in the entry by embedding the less frequent `<orth>` form within a second `<form>` element, labeled `@type="variant"`, which is a direct child of the `<form type="lemma">`.

86 This orthographic variation stems from the fact that there is not a standard 1 for 1 system of transliterating between the Arabic script of Urdu and the Latin script of English. This information is recorded within the <note> element in the etymological section as the origin of this issue lies in the item's etymology.

Example 8. Borrowing and transliteration: *biryani*.

```
<entry xml:id="biryani" xml:lang="en">
 <form type="lemma">
  <form type="preferred">
   <orth>biryani</orth>
   <pron notation="xsampa">%bIrI"A:nI</pron>
  </form>
  <form type="variant">
   <orth>biriani</orth>
  </form>
  <gramGrp>
   <pos>noun</pos>
  </gramGrp>
 </form>
 <sense corresp="http://dbpedia.org/resource/Biryani">
  <def>any variety of Indian dishes…</def>
 </sense>
 <etym type="borrowing">
  <lbl>from</lbl> <lang>Urdu</lang>
  <oRef xml:lang="ur">بریانی</oRef>
  <note>The variation in the English orthographic form of this entry is due to
the fact that there is no standard transliteration between English (Latin) and
Urdu's (Arabic) scripts </note>
 </etym>
</entry>
```

6.2 Phonological Adaptation in Borrowed Forms

87 Example 9 also contains transliteration between the orthographies of the source (in this case English) and borrowing (Japanese) languages. Additionally, Japanese has multiple orthographic systems into which the borrowed form is integrated; each of these systems is identified in the

<orth> element as the value of the @notation attribute. In contrast to the previous example (*biryani*), the presence of a phonetic form in both the source and borrowing languages implicitly shows the difference in the pronunciation between the English source:

Example 9. Borrowing—phonological and orthographical adaptation: *takushī*.

```
  <entry xml:id="taxi" xml:lang="ja">
   <form type="lemma">
    <orth type="transliterated" notation="rōmaji">takushī</orth>
    <orth notation="katakana">タクシー</orth>
    <pron notation="ipa">takushi:</pron>
    <gramGrp>
     <pos>noun</pos>
    </gramGrp>
   </form>
   <sense corresp="http://dbpedia.org/resource/Taxicab">
    <usg type="dom">Transportation</usg>
   </sense>
   <etym type="borrowing">
    <lbl>source</lbl> <lang>English</lang>
    <cit type="etymon" xml:lang="en">
     <oRef corresp="http://en.wiktionary.org/wiki/taxi#English">taxi</oRef>
     <pRef notation="ipa" corresp="http://en.wiktionary.org/wiki/
taxi#Pronunciation">'tæksi</pRef>
    </cit>
   </etym>
  </entry>
```

6.3 Morphological Adaptation in Borrowed Forms

88 In the case that a language whose grammatical system has a feature such as gender borrows a (nominal) item from a language that has no grammatical gender, the borrowing language will generally need to assign a gender to the loanword. In the example below showing the French borrowing of the English *weekend*, the lack of gender is implicitly encoded by the absence of a <gen> element within the <gramGrp> element cluster within the <cit type="etymon">.[34] The adaptation

made in assimilating this item into the French language is visible in the inverse manner (e.g., in the inclusion of the `<gen>masc</gen>`) within the `<gramGrp>` element cluster within the `<form type="lemma">`).

89 In modeling this for French, it is especially important to include the gender information because of the implications for the specific forms that will occur in context of and in reference to *weekend* in the context of natural language (e.g., articles, pronouns, determiners, demonstratives, and gender forms of adjectives).

Example 10. Borrowing—morphological adaptation: *weekend*.

```
<entry xml:id="weekend" xml:lang="fr">
 <form type="lemma">
  <orth>week-end</orth>
  <gramGrp>
   <pos>nom</pos>
   <gen>masc</gen>
  </gramGrp>
 </form>
 <sense corresp="http://fr.dbpedia.org/page/Week-end">
  <def>Le week-end (variante weekend, comme en anglais), issu de l’anglais
weekend; ou la fin de semaine est une période hebdomadaire d’un ou deux jours,
généralement le samedi et le dimanche, pendant laquelle la plupart des gens sont
au repos.</def>
 </sense>
 <etym type="borrowing">
  <cit type="etymon">
   <oRef>weekend</oRef>
   <gramGrp>
    <pos>nom</pos>
   </gramGrp>
  </cit>
 </etym>
</entry>
```

7. Metaphor

90 Metaphor is undeniably one of the primary motors of lexical innovation as well as of (synchronic) polysemy in languages. On the cognitive level, metaphor can be roughly defined as the process of understanding one concept from one domain of experience in terms of another concept from a separate domain of experience. This is mirrored lexically, where one concept in a given domain is described or referred to using lexical items (and/or vocabulary) from another (Lakoff and Johnson 1980; Lakoff 1987; Lakoff 1993).

91 While this process is not commonly mentioned as a primary category for linguistic change in lexicographical sources such as etymological dictionaries, it is an essential component of linguistic analysis and is relevant to cognitive sciences and anthropology as well.

92 Metaphorical mappings often follow patterns in directionality; most commonly, the source is a more concrete or primitive domain and the target is a more abstract, less conceptually primitive domain (Lakoff and Johnson 1980; Lakoff 1987; Lakoff 1993). Examples of conceptual primitive domains are space, temperature, olfaction, location, motion, quantity, body, food, and water.

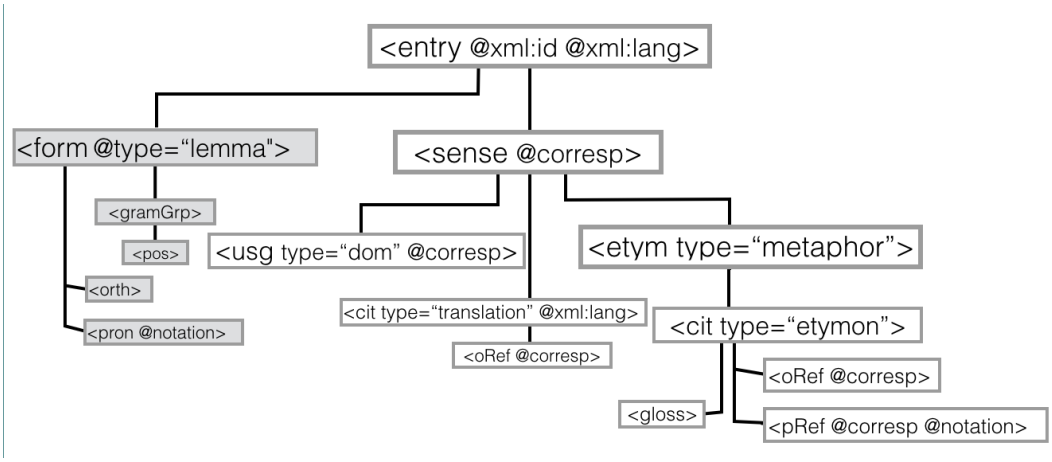
93 The driving forces for this process are ontological correspondences between the attributes or internal structures of entities in the source domain with those of the target domain (Lakoff 1993). Where this is true, the potential for mapping is activated, or is rendered salient on the conceptual level, which in turn enables the linguistic encoding of these structures for communicative purposes. Thus, metaphor in language is a surface level reflection or by-product of operations that occur at the cognitive level. Given the fundamental implication of such ideas, not paying attention to metaphor when documenting language change is missing out on an enormous source of incredibly fascinating information.

94 The basic concepts necessary for a minimal encoding of a "metaphor" etymology are:

- Source form (the same as target form at time of process);
- Target form (the same as source form at time of process);
- Domain of concept (y): source domain;
- Domain of concept (x): target domain.

95 The model below shows an abstract representation of the essential structures for such an entry.

Figure 3. TEI modeling: `<etym type="metaphor">`.



96 Example 11 shows an instance of an encoding of the lexical item for "kidney" in Mixtepec-Mixtec language, which was derived metaphorically from "bean." This metaphorical mapping is clearly due to the physical similarity (color and shape) between a bean and a kidney. While both terms refer to concrete (physical) concepts, the source, which is a primary food source for the speakers of the language, is more conceptually primitive than its target, which is an internal organ and thus seen much less frequently, if at all, by most speakers.

Example 11. Metaphor: Mixtepec-Mixtec "kidney"[35]

```
<entry xml:id="kidney" xml:lang="mix">
 <form type="lemma">
  <orth>ntuchi</orth>
  <pron notation="ipa">ndú.ʧì</pron>
  <gramGrp>
   <pos>noun</pos>
  </gramGrp>
 </form>
 <sense corresp="http://dbpedia.org/resource/Kidney">
  <usg type="dom" corresp="http://dbpedia.org/resource/Human_body">Body</usg>
  <usg type="dom" corresp="http://dbpedia.org/resource/
Human_organs">InternalOrgans</usg>
  <etym type="metaphor">
   <cit type="etymon">
    <oRef corresp="#bean">ntuchi</oRef>
    <pRef notation="ipa" corresp="#bean">ndú.ʧì</pRef>
    <ref type="sense" corresp="http://dbpedia.org/resource/Bean"/>
    <usg type="dom" corresp="http://dbpedia.org/resource/
Category:Edible_legumes">Legume</usg>
    <gloss>bean</gloss>
   </cit>
  </etym>
  <cit type="translation" xml:lang="en">
   <oRef>kidney</oRef>
  </cit>
 </sense>
</entry>
```

97 Metaphor-typed etymologies should always be direct children of `<sense>` as the lexical change affects the sense of an existing lexical item. In order to know that it is a metaphor, it is of course necessary to know a) the original meaning of the source (encoded in `<gloss>`, or `<def>`), and b) the domains of the source and target items, such as source domain (semantic domain of etymon) and target domain (semantic domain of entry). In pursuit of the latter, the TEI `<usg>` element can be used with the attribute–value pair `@type="dom"`, in which the element value represents the domain. Given that the semantic profiles of concepts are not always limited to a single domain, it

may be necessary to include multiple <usg type="dom"> element–value pairs, as in the example. In our model, <usg type="dom"> is used in both the synchronic (<sense>) and the diachronic (<cit>) portions of the entry.

98 However, although such structures, if used consistently, do have the capacity to greatly enhance any data that do not include this information, they inevitably fall short of the extent of knowledge that is truly necessary to create an accurate model of metaphorical processes. In order to do this, it is necessary to make use of one or more ontologies, which could be locally defined within a project, and of external linked open data sources such as DBpedia and Wikidata, or some combination thereof. Within TEI dictionary markup, URIs for existing ontological entries can be referenced in the <sense>, <usg>, and <ref> elements as the value of the attribute @corresp.

99 Within the etymon, the <oRef> and/or <pRef> can be included with a pointer to the source form using the @corresp attribute, the value of which is a reference to the source entry's unique identifier (if such an entry exists within the dataset). In such cases, the etymon pointing to the source entry can be assumed to inherit the source's domain and sense information, and this information can be automatically extracted with a fairly simple XSLT program; thus the encoders may choose to leave some or all of this information out of the etymon section. However, in the case that the dataset does not actually have entries for the source terms, or the encoder wants to be explicit in all aspects of the etymology, as mentioned above, the source domain and the ontologically based sense of an etymon can be encoded within <cit> as <ref> and <usg> respectively.

8. Metonymy

100 Like metaphor, metonymy is another process based in human cognition which is reflected in language and leads to synchronic lexical polysemy. Again like metaphor, metonymy is often central to linguistic etymological analyses, but is also not commonly mentioned as a primary category for language change in lexicographical sources such as etymological dictionaries.

101 In metonymy, one vehicle entity provides mental access to a target entity within a single domain by highlighting different aspects of part–whole relationships (meronymy) (Kövecses and Radden 1998). There are two primary types of relationships that commonly provide the motivation, or

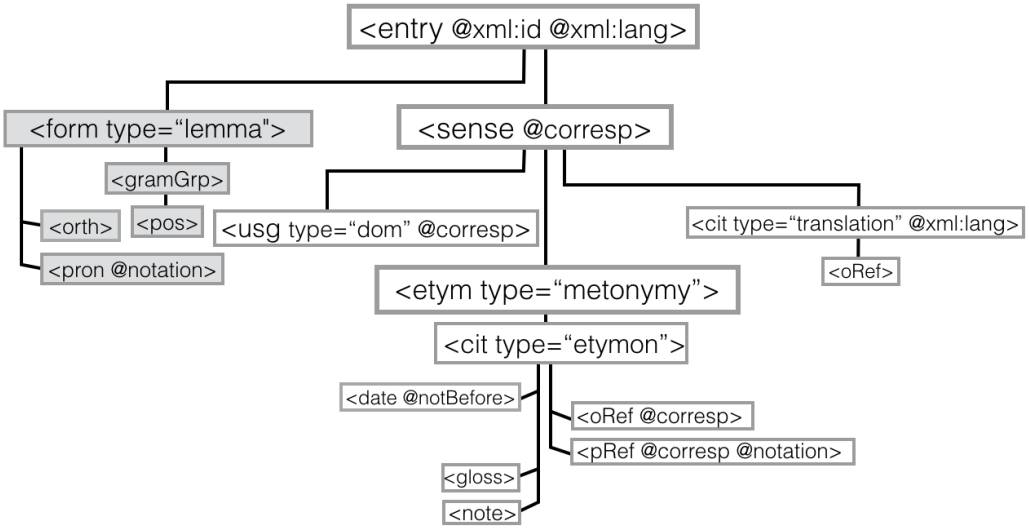
what Vandeloise (2006) refers to as "logical impetus," for the occurrence of metonymy: a) where a part of an entity's profile stands for the whole; b) when a domain's subpart stands for its whole (ibid.).

102 The basic concepts necessary for a minimal encoding of a "metonymy" etymology are:

- synchronic (polysemous) form(s) of item;
- source form(s) (orthographic and/or phonetic);
- semantic profile/metalinguistic concept or entity of borrowed form.[36]

103 The TEI modeling of metonymy is very similar to that of metaphor with the exception that the semantic domain between source and target is the same. Thus the model is as follows: metonymy-typed etymologies should be embedded in the `<sense>` element, which can also contain the `@corresp` with a URI to the conceptual entry in an external ontology; within the `<cit type="etymon">` element block, the source orthographic and/or phonetic forms should be included in `<oRef>` and/or `<pRef>` elements; within these elements, the value of the attribute `@corresp` specifies the unique identifier of the source entry. Often metonymy leads to a change in grammatical role, and thus it is often the case that a `<gramGrp>` element block is necessary, which then should include whatever subelements are necessary.

Figure 4. Abstract TEI modeling for metonymy.



104 In the example below from Mixtepec-Mixtec, the words for “animal” and “horse” are polysemous, the origin of which can clearly be identified as metonymy. Here, the language reflects the history of the people; since there were no horses in Mexico until the arrival of the Spanish, there was no Mixtecan word for “horse,” and the categorical noun for “animal” was used to describe the unnamed animal.

105 In the TEI entry, the synchronic orthographic and phonetic forms are identical to those of the source term “animal.” Within the etymology element block (`<etym type="metonymy">`) and the etymon (`<cit type="etymon">`) the source term’s URI is referenced in `<oRef>` and `<pRef>` as the value of `@corresp` (`@corresp="#animal"`).

106 In `<sense>`, the URI corresponding to the DBpedia entry for “horse” is the value for the attribute `@corresp`. Additionally, the `<date notBefore="…">` element–attribute pairing is used to specify that the term has only been used for the “horse” since 1517 at maximum (corresponding to the first Spanish expedition into Mexico). Within the actual document, the contents of `<note>` discuss the historical context of this word origin (commented out in the example).

Example 12. Metonymy: Mixtepec-Mixtec *kiti*.

```
<entry xml:id="animal-horse" xml:lang="mix">
 <form type="lemma">
  <orth>kiti</orth>
  <pron notation="ipa">kì.˩tí</pron>
  <gramGrp>
   <pos>noun</pos>
  </gramGrp>
 </form>
 <sense corresp="http://dbpedia.org/resource/Horse">
  <usg type="dom" corresp="http://dbpedia.org/resource/Animal">Animal</usg>
  <etym type="metonymy">
   <date notBefore="1517"/>
   <cit type="etymon">
    <oRef corresp="#animal">kiti</oRef>
    <pRef notation="ipa" corresp="#animal">kì.˩tí</pRef>
    <gloss>animal</gloss>
   </cit>
   <note><!-- notes on historical context of term here --></note>
  </etym>
  <cit type="translation" xml:lang="en">
   <oRef>horse</oRef>
  </cit>
  <cit type="translation" xml:lang="es">
   <oRef>caballo</oRef>
  </cit>
 …
 </sense>
</entry>
```

9. Compounding

107 For various reasons, the lexical process of compounding has traditionally been treated in linguistic literature focused on synchronic themes, often in the context of such issues as morphology and word formation. One of the least theoretically controversial reasons for this is that novel compounds can be formed, used, and understood by speakers instantaneously. However, there is significant discord in the linguistics literature as to how compound lexical items and the process

of compounding should be analyzed. This is most distinctly evident in the contrasting approaches between linguistic theories that make strict divisions between lexicon and grammar (traditional generative grammar) and those that do not make such distinctions (cognitive linguistic theories) (Langacker 1987, 1991, 2000; Jackendoff 2009).

108 In attempting to create typologies and/or generalizations about the nature of compound word forms and the process in which they are created, the aforementioned viewpoints emphasize different aspects of the phenomena. Studies by authors adhering to more formal, generative theoretical frameworks often focus on analyzing the morphosyntactic makeup of compound forms in the context of rule patterning: e.g., N+N, A+A, V+V, N+A, P+N. Those adhering to the cognitive-based theory generally focus on the semantic profiles and ontological relationships of the constituent parts, as well as on the cognitive processes (metaphor and/or metonymy) active in the resulting meaning (Goossens 1994; Geeraerts 2002; Benczes 2006a, 2006b; Guevara and Scalise 2008). Examples of such semantic/ontological relationships commonly found in compound forms are type–of (hyponymy), cause–effect, part–whole (meronymy), and location–located (Benczes 2006b).

109 A common means of classifying compounds in linguistics literature (both formal and cognitive) is according to two main typologies: *endocentric* and *exocentric* compounds (Bloomfield 1933).[37] There are numerous different typological interpretations according to which the concepts of endocentric and exocentric compounds are defined in the literature. Many of these differences are attributable to problems with the classification criteria, the types of compounds examined, and the languages analyzed (Scalise and Guevara 2005; Benczes 2006a).

110 Generally, endocentric compound constructions have a "head" element which is the hypernym while the full compound represents a subclassification or hyponym of that head (Benczes 2006a, 2006b). An example of this is the bird called a scrub jay, the head of which is "jay"; the compound "scrub jay" is a type of jay. Exocentric compounds are considered "headless," and while there is no single standard ontological relationship between the compound and/or its components, they can contain any variety of relations other than hyponymy.[38] An example of this from Benczes (2006b) is the English compound "hay fever," which has a cause–effect relation as hay causes the effect of fever.

111 While it would be impossible to give a full account of all of the various typologies proposed in the extensive literature, our markup framework for compounding is capable of handling the key pieces of information for a wide array of theoretical approaches for explicit and implicit encoding of the relevant information within a single entry.

112 Basic pieces of information for modeling compounds and compounding include:

- the delimitation between the components
- the grammatical profiles of the components
- the meanings (senses) of the components
- the semantic domain of the components.

9.1 Encoding Compounds: Synchronic Portion

113 Modeling compounding in TEI dictionaries can (and ideally should) be done in both the synchronic and diachronic portions of an entry. On the synchronic level, the attribute value `@type="compound"` should always be placed in the `<entry>` element in order to specify that the contents of form (`<orth>` and/or `<pron>`) can be further parsed.

Example 13. Compounding (synchronic encoding): French *merle noir*.

```
<entry xml:id="merle-noir" type="compound" xml:lang="fr">
 <form type="lemma">
  <orth>merle noir</orth>
  <gramGrp>
   <pos>nom</pos>
   <gen>masc</gen>
  </gramGrp>
 </form>
…
</entry>
```

114 From a processing point of view, this is of particular use when the compound form includes whitespace. Optionally, the editor may choose the more thorough method of encoding the individual component strings of a compound form with `<seg>`. Where corresponding entries exist

within a project for each of these portions of the compound, an editor may choose to connect them with the components encased within <seg> using @corresp with the value being the unique identifier for the given entry.

115 The following example shows the encoding of the number thirteen in Mixtepec-Mixtec, which is a case of an exocentric compound, as there is no "head." As is the case with the lexical composition of numbers in many languages, this compound form is the sum of its component parts semantically and phonologically: *utsi* "ten" + *uni* "three."

Example 14. Exocentric compounding (synchronic encoding): Mixtepec-Mixtec *utsi uni*.

```
<entry xml:id="num-13" type="compound" subtype="exocentric" xml:lang="mix">
 <form type="lemma">
  <orth>
   <seg corresp="#num-10">utsi</seg> <seg corresp="#num-3">uni</seg>
  </orth>
  <gramGrp>
   <pos>cardinalNumber</pos>
  </gramGrp>
 </form>
 <!-- sense, etym, translation,etc… -->
</entry>
```

116 Where there is a hyphen in a compound, this can be differentiated from the content portion of the string by using the TEI character element <pc>. The French example below is a type of bird whose name is usually spelled with a hyphen but is also occasionally written as a continuous word without whitespace or a hyphen:

Example 15. Compounding (synchronic encoding): French *rouge-gorge*, *rouge gorge*.

```
<entry xml:id="rouge-gorge" type="compound" xml:lang="fr">
 <form type="lemma">
  <orth>
   <seg>rouge</seg><pc>-</pc><seg>gorge</seg>
  </orth>
  <form type="variant">
   <orth>
    <seg>rouge</seg><pc> </pc><seg>gorge</seg>
   </orth>
  </form>
  <gramGrp>
   <pos>nom</pos>
   <gen>masc</gen>
  </gramGrp>
 </form>
…
</entry>
```

9.2 Encoding Compounds: Diachronic Portion

117 Below is the diachronic portion of the entry of the Mixtepec-Mixtec number “thirteen” discussed above.

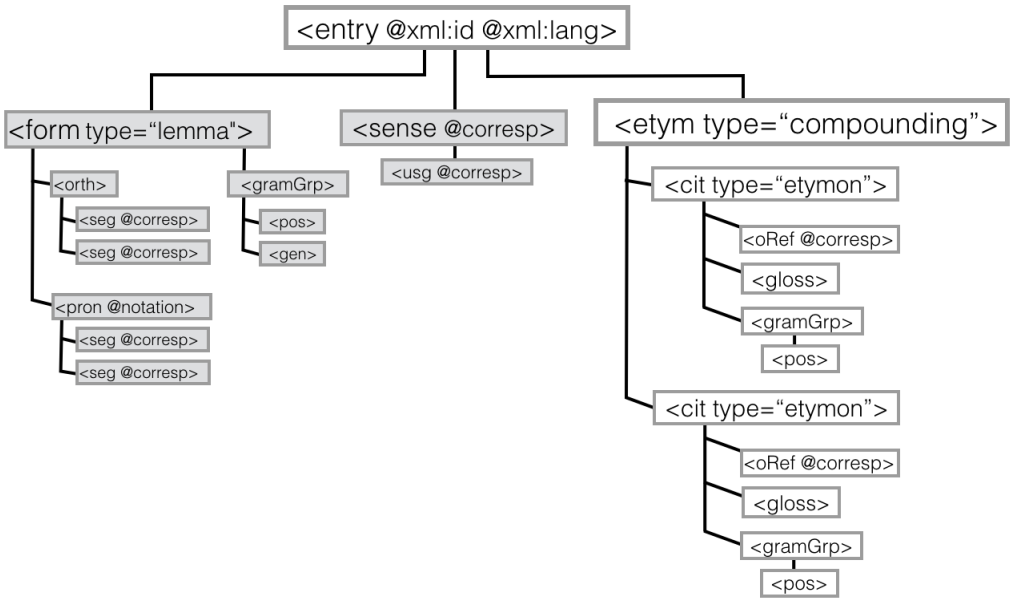
Example 16. Compounding: Mixtec-Mixtec *utsi uni*.

```
<entry xml:id="num-13" type="compound" xml:lang="mix">
 <!-- form, gramGrp, etc. -->
 <sense corresp="http://dbpedia.org/resource/10_(number)">
  <usg type="dom" corresp="http://dbpedia.org/resource/
Category:Cardinal_numbers">CardinalNumbers</usg>
 </sense>

 <etym type="compounding">
  <cit type="etymon">
   <oRef corresp="#num-10">utsi</oRef>
   <gramGrp>
    <pos>cardinalNumber</pos>
   </gramGrp>
   <gloss>ten</gloss>
  </cit>

  <cit type="etymon">
   <oRef corresp="#num-3">uni</oRef>
   <gramGrp>
    <pos>cardinalNumber</pos>
   </gramGrp>
   <gloss>three</gloss>
  </cit>
 </etym>
…
</entry>
```

118 The components of the compound which were encased in the <seg> element within the synchronic forms correspond to the etymons on the diachronic level; as in any other etymological process in our encoding system, each subunit of the compound and its given lexical information is represented as a separate <cit type="etymon">. Each of these citation-etymon crystals is embedded within the compound-typed <etym>. Since compounds do not represent new senses of a *single* existing entry form, the etymology element is placed as a direct child of <entry>. As we have

seen before, the `@corresp` attribute in the `<oRef>` and/or `<pRef>` points to the URI for the entry of each component of the compound; in this example these entries are within the same document. The comprehensive model is depicted in figure 5 below.

Figure 5. TEI modeling diagram: compounding (diachronic portion).



9.3 Co-occurrence of Compounding and Other Etymological Processes

9.3.1 Compounding and Metaphor

119 The German item *Handschuh* "glove" is an example of a compound in which metaphor is clearly present; *Schuh* "shoe," which is an article of clothing specific to the feet (e.g., domain = feet), is used as an article of clothing for the hands (e.g., domain = hand). In the encoding of this item in the entry below, the structure is a combination of everything we have previously seen in examples of compounding- and metaphor-typed etymologies, but with the difference that since `<etym type="metaphor">` is a subpart in the larger word-formation process of compounding, it is embedded within `<etym type="compound">`. Since compounding is the top-level component of the etymology, `<etym>` is placed as a direct child node of `<entry>`, following the guidelines for that element.

120 Also noteworthy is that this example is an "endocentric" compound as the head is *schuh* "shoe."[39]

Example 17. Compounding and metaphor:[40] German *Handschuh*.

```
<entry xml:lang="de" xml:id="handschuh" type="compound" subtype="endocentric">
 <form type="lemma">
  <orth><seg>Hand</seg><seg>schuh</seg></orth>
  <pron notation="ipa"><seg>'hant</seg><seg>ʃu:</seg></pron>
  <gramGrp>
   <pos>subst.</pos>
   <gen>mask.</gen>
  </gramGrp>
 </form>
 <sense corresp="http://dbpedia.org/resource/Handschuh">
 ...
 </sense>
 <etym type="compounding">
  <cit type="etymon">
   <oRef xml:lang="de">Hand</oRef>
   <pRef xml:lang="de" notation="ipa">'hant</pRef>
   <gloss>hand</gloss>
  </cit>
  <etym type="metaphor">
   <cit type="etymon">
    <oRef xml:lang="de">Schuh</oRef>
    <pRef xml:lang="de" notation="ipa">ʃu:</pRef>
    <gloss>shoe</gloss>
   </cit>
  </etym>
 </etym>
</entry>
```

121 The classification of a lexical item as a compound hinges on the semantic transparency of its components to the speakers of a language. An item corresponding to a single conceptual entity that was derived through compounding in a source can be borrowed by another language and the fact that it was a compound is not relevant to the meaning in the target lexicon. In such cases it is up to the encoder whether or not to include the etymological details that occurred in the source language. Below is the full encoding of the earlier example of *pamplemousse* from Salmon-Alt (2006).

Example 18. Borrowing and compounding: *pamplemousse* (Salmon-Alt 2006).

```
<entry xml:id="LE1" xml:lang="fr">
 <form type="lemma">
  <orth>pamplemousse</orth>
  <gramGrp>
   <pos>commonNoun</pos>
  </gramGrp>
 </form>
 <sense>
 ...
 </sense>
 <etym type="borrowing">
  <lbl>source</lbl> <lang>Dutch</lang>
  <cit type="etymon" xml:id="L2">
   <oRef xml:lang="nl">pompelmoes</oRef>
   <gloss xml:lang="lat">Citrus maxima</gloss>
   <gramGrp>
    <pos>commonNoun</pos>
    <gen>feminine</gen>
   </gramGrp>
   <note>probablement d'origine tamoule, De Vries, Nedrl</note>
   <ref target="TLF">TLF</ref>
  </cit>
  <etym type="compounding">
   <cit type="etymon">
    <oRef xml:lang="nl">pompel</oRef>
    <gramGrp>
     <pos>adjective</pos>
    </gramGrp>
    <gloss>gros, enflé</gloss>
   </cit>
   <cit type="etymon">
    <oRef xml:lang="nl">limoes</oRef>
    <gramGrp>
     <pos>commonNoun</pos>
    </gramGrp>
    <gloss>citron</gloss>
   </cit>
```

```
    <ref target="#Boulan-König">Boulan, König...</ref>
   </etym>
  </etym>
 </entry>
```

9.3.2 Lexicalization

122 Compounds, like morphologically-derived coalesced forms and idiomatic forms or other types of phrases, may or may not become integrated into the lexicon of a language. When they do, the process may occur over an extended period of time, during which the original meaning of part or all of the components may become oblique to speakers[41] to the degree that the item is not recognized as the sum of its parts; this phenomenon is called lexicalization (Jackendoff 2009). While lexicalization is indeed a lexico-cognitive process with etymological implications, it is not a process that can be summarized and represented in a dictionary, as there are often different degrees of lexicalization between different speakers.

10. Grammaticalization

123 A major phenomenon of lexical change is the process of grammaticalization in which, over time, a lexical item undergoes changes to its semantic profile, morphosyntactic form, and possibly phonological form via systematic patterns of usage in certain linguistic contexts to serve more grammatical functions within the language.

124 Because of the complexity of the process and the continuing debate as to the details of precisely what it entails, any single definition of grammaticalization inevitably raises theoretical debate, and may potentially be modified down the line within the field. Antoine Meillet (1912), who coined the term, described it as follows:

> The development of grammatical forms by progressive deterioration of previously autonomous words is made possible by ... a weakening of the pronunciation, of the concrete sense of the words, and of the expressive value of words and groupings of words. The ancillary word can end up as an element lacking independent meaning as such, linked to a principal word to mark its grammatical role.

125 Studies of grammaticalization have shown that there are certain tendencies for forms to undergo the same types of functional shifts to their semantic, syntactic, morphological, and phonological properties, which occur in relatively similar orders[42] (Svorou 1994; Croft 2003; Hopper and Traugott 2003). These evolutionary pathways are commonly referred to as "clines" (ibid.). On the semantic level, the most generic characterization of the tendency is for the profile to undergo changes from concrete to abstract (ibid.). Moreover, there is significant evidence that the semantic profiles of a lexical source may motivate and/or constrain the potential structural characteristics and grammatical roles that emerge from a lexical source (Svorou 1994; Hopper and Traugott 2003).

126 Parallel to the semantic shifts are the clines that occur on the syntactic and morphological levels, of which there are a number of different specific possibilities, but which can be generalized as follows:

```
less grammatical → more grammatical
```

127 Hopper and Traugott (2003) discuss a correlation between the degree of grammaticalization (grammatical status) and the loss of certain morphological and syntactic properties that are commonly observed in prototypical members of major grammatical categories (decategorization).

128 `majority category → (intermediate category) → minor category`

It should be noted that because of polysemy, a single form may simultaneously have several grammatical functions in which multiple "stages" in different clines can coexist synchronically (Svorou 1994; Hopper and Traugott 2003). According to Svorou (1994), the study of grammaticalization requires a "panchronic" view of language change, viewing the component evolutionary processes as continuous, rather than applying the artificially discrete concepts of synchronic and diachronic.[43]

129 The negotiation of meaning by speakers, using metaphor and metonymy to extend the contexts in which a lexical item is used, enables such changes as pragmatic/semantic bleaching or enrichment on the cognitive, semantic, and pragmatic levels. On the syntactic and morphological levels, these higher-level changes are manifested in reanalysis[44] and analogy.[45]

130 Grammaticalization is thus very complex and involves multiple possible subprocesses which may or may not co-occur on different levels. As a consequence, encoding such etymologies may require the use of any number of the strategies discussed above.

10.1 Overview of Basic Encoding Recommendations

131 Since grammaticalization potentially involves multiple etymological processes on any level of language at different stages in the diachrony of an item, the formatting recommendations depend on the specifics of a given scenario and involve the integration of concepts and data structures from multiple processes discussed elsewhere in this paper. It is up to the editor to determine what aspects and features of the etymology they want to emphasize within an encoding of grammaticalization (e.g., processes on different levels, bibliographic attestations, discourse, semantics, morphosyntax, or phonology); its nuances and complexities make it impossible to write an entirely comprehensive set of encoding recommendations. However, the constants which can always be included in a grammaticalization etymology are as follows:

- given that grammaticalization inherently entails a shift in the sense of an etymon, `<etym type="grammaticalization">` should be embedded within `<sense>`;
- all forms of the etymon that are hypothesized or attested as having occurred within the span of the time corresponding to the stages of the grammaticalization process should be encoded within `<cit type="etymon">` (as we have previously seen).

132 Conceptual information necessary in the encoding of grammaticalization includes:

- the synchronic form(s);[46]
- the grammatical information.

If the grammaticalized item is one of multiple senses of a polysemous form:[47]

- the grammatical, semantic, and/or pragmatic context in which the given sense occurs;
- the collocates with which it occurs (if any);
- the source form(s);
- whether the source form(s) is/are attested or hypothesized.

If attested:

- bibliographic source of attestation;

- date or date range for each attested form.

If hypothesized:

- semantic profile of source (e.g., semantic domain or metalinguistic concept);
- grammatical profile of source.

133 The placement of the `<etym type="grammaticalization">` is dependent on the specific diachrony of the etymon. For example, if the item was inherited, it can be placed within `<etym type="inheritance">`.

10.2 Traugott (1995): Grammaticalization of the English Discourse Marker "besides"

134 In the following example from Traugott (1995), we show in our encoding a sample of several stages of the grammaticalization process of the English discourse particle "besides." The author makes the case that the following cline (e.g., the process of grammaticalization for a given lexical item) should be added to the scope of grammaticalization phenomena:[48]

```
Adverbial of Extension > Sentence Adverbial > Discourse Particle
```

135 The major stages of this process are presented in a semichronological order and in terms of the following stages:

- Stage 0: Full lexical noun;
- Stage I: Adverbial of extension;
- Stage II: Sentence adverbial;
- Stage III: Discourse marker/particle.

136 The example below is an encoded representation of the arguments put forth in the source. We have attempted to represent the linguistic information, terminology, and etymological processes as accurately as possible without adding or subtracting from the original based on our own interpretation or external data.

10.2.1 Stage 0: Full lexical noun

Example 19. Grammaticalization of Old English *sidan*[49]

```
  <entry>
   <!-- form, gramGrp, etc. -->
   <sense>
    ...
    <etym type="grammaticalization">

     <cit type="etymon" xml:id="at-850-950" next="#at-1225-a">
      <date notBefore="0850" notAfter="0950"/>
      <oRef xml:lang="ang">
       <seg xml:id="e1s1">sid</seg>
       <seg xml:id="e1s2">an</seg>
      </oRef>
      <gloss>side</gloss>

      <gramGrp><!-- inherently applies to the top level structure -->
       <pos>locativeNoun</pos>
      </gramGrp>

      <cit type="component" corresp="#e1s1">
       <gramGrp>
        <pos>noun</pos>
       </gramGrp>
      </cit>

      <cit type="component" corresp="#e1s2">
       <gramGrp>
        <case>dative</case>
       </gramGrp>
      </cit>
      <cit type="attestation" xml:lang="ang">
       <quote>&amp;amp; þonne licge on ða swiðran <oRef>sidan</oRef> gode hwile</
quote>
       <cit type="translation" xml:lang="en">
        <quote>and then lie on the right side for a good while</quote>
       </cit>
```

```
      </cit>
      <bibl>(850-950 Lacnunga Magicand&amp;amp;Med., p. 120 [HC])</bibl>
     </cit>

     <!-- <cit type="etymon"> elements from examples 20-27 -->
    ...
    </etym>
   ...
  </sense>
 </entry>
```

137 This first stage *sidan* is from Old English (ISO 639–3: `"ang"`); as usual the `<cit>` element is typed as `@type="etymon"`. Within this first etymon, the noun is inflected in the dative case. This is represented in the encoding, with the nominal root "sid-" and the dative inflection "-an" each being given unique `<seg>` tags with their own `@xml:id`s. These components are annotated grammatically within the `<gramGrp>` element blocks embedded in `<cit type="component">`, each of which points to the given segment with the `@corresp` attribute.

138 The source of this is dated sometime between 850 and 950. As we have discussed previously, a range of dates can be expressed using the attributes `@notBefore` and `@notAfter` in the `<date>` element. While the bibliographic source (the last direct child element of `<cit type="etymon">`) does contain the date itself, using the `<date>` element helps keep the consistency of the data structure, which will likely have benefits to any automatic retrieval process.

- a partial example of the context in which the etymon form was observed is given in the `<quote>` element whose parent is an embedded `<cit type="attestation">`. Also specified here is the language attribute and tag, which as always are expressed in the form `@xml:lang="ang"`. The form itself is encased in an `<oRef>` with no attributes.
- finally, embedded within the attestation is the translation of the Old English source sentence into Modern English: `<cit type="translation" xml:lang="en">`

10.2.2 Stage 0.5: Dative nouns with prefix

139 These next two attested etymons are both from 1225 (Middle English), and from the same bibliographic source, yet they have different grammatical functions; by including the range of dates for each with their contrasting grammatical usage, the encoded information is able to capture the emerging state of polysemy for the lexical item:

- each has the prefix bi-;
- each is dative case.

Example 20. Grammaticalization of Middle-English *bisiden* 1225 (a)[50]

```
<entry>
 <!-- form, gramGrp, etc. -->
 <sense>
  ...
  <etym type="grammaticalization">
   ...
   <!-- <cit type="etymon"> element from example 19 -->

   <cit type="etymon" xml:id="at-1225-a" prev="#at-850-950" next="#at-1225-b">
    <date when="1225"/>
    <oRef xml:lang="enm">
     <seg xml:id="e2s1">bi</seg><pc>-</pc>
     <seg xml:id="e2s2">sid</seg>
     <seg xml:id="e2s3">en</seg>
    </oRef>
    <gloss>at his side</gloss>
    <gramGrp>
     <pos>locativeAdverbial</pos>
    </gramGrp>
    <cit type="component" corresp="#e2s1">
     <gramGrp>
      <pos>adverb</pos>
     </gramGrp>
    </cit>
    <cit type="component" corresp="#e2s2">
     <gramGrp>
      <pos>noun</pos>
     </gramGrp>
    </cit>
    <cit type="component" corresp="#e2s3">
     <gramGrp>
      <case>dative</case>
     </gramGrp>
    </cit>
    <cit type="attestation" xml:lang="enm">
     <quote>His pic he heold <oRef>bi-siden</oRef></quote>
     <cit type="translation" xml:lang="en">
```

```
      <quote>He held his staff at his side</quote>
     </cit>
    </cit>
    <bibl>(1225 Lay. Brut 30784 [MED])</bibl>
   </cit>

   <!-- <cit type="etymon"> elements from examples 21-27 -->
  ...
  </etym>
 ...
 </sense>
</entry>
```

Example 21. Grammaticalization of Middle-English *bisiden* 1225 (b)[51]

```
<entry>
 <!-- form, gramGrp, etc. -->
 <sense>
  ...
  <etym type="grammaticalization">
   ...
   <!-- <cit type="etymon"> elements from examples 19-20 -->

   <cit type="etymon" xml:id="at-1225-b" prev="#at-1225-a" next="#at-c1300">
    <date when="1225"/>
    <oRef xml:lang="enm">
     <seg xml:id="e3s1">bi</seg><pc>-</pc>
     <seg xml:id="e3s2">sid</seg>
     <seg xml:id="e3s3">en</seg>
    </oRef>
    <gloss>in addition</gloss>
    <gramGrp>
     <pos>conjunctiveAdverbial</pos>
    </gramGrp>
    <cit type="component" corresp="#e3s1">
     <gramGrp>
      <pos>preposition</pos>
     </gramGrp>
    </cit>
    <cit type="component" corresp="#e3s2">
     <gramGrp>
      <pos>nominalAdposition</pos>
     </gramGrp>
    </cit>
    <cit type="component" corresp="#e3s3">
     <gramGrp>
      <case>dative</case>
     </gramGrp>
    </cit>
    <cit type="attestation" xml:lang="ang">
     <quote>Heo letten forð <oRef>bi-siden</oRef> an oþer folc riden, ten þusend
kempen</quote>
```

```
      <cit type="translation" xml:lang="en">
       <quote>They sent another army forth in addition, 10,000 warriors</quote>
      </cit>
     </cit>
     <bibl>(1225 Lay Brut 5498 [MED])</bibl>
    </cit>

    <!-- <cit type="etymon"> elements from examples 22-27 -->
   ...
   </etym>
  ...
  </sense>
 </entry>
```

10.2.3 Stage I: Adverbial of extension/verbal adverbs

140 A major feature of interest in adverbials of extension is where they occur at the end of a clause; in these cases, the adverbs are often an oblique syntactic argument. This is the case in the section of the example below in which the tag "extensionAdverb" is used in the source for adverbials of extension.[52]

- prefix *be-* now fused onto *side*;
- adverb oblique argument, labeled as value of `@ana` in the `<oRef>` in the attestation.

Example 22. Grammaticalization of Early Modern English “beside” 1514–1518.

```
  <entry>
   <!-- form, gramGrp, etc. -->
   <sense>
    ...
    <etym type="grammaticalization">
     ...
     <!-- <cit type="etymon"> elements from examples 19-21 -->

     <cit type="etymon" xml:id="at-1514-1518" prev="#at-1450"
next="#at-1535-1543">
      <date notBefore="1514" notAfter="1518"/>
      <oRef xml:lang="emodeng">beside</oRef>
      <gramGrp>
       <pos>extensionAdverb</pos>
      </gramGrp>
      <cit type="attestation"><!-- early modern english -->
       <quote>In whiche albeit thei ment as muche honor to hys grace as wealthe
to al the realm <oRef ana="#Oblq">beside</oRef>, yet were they not sure howe hys
grace woulde take it,</quote>
      </cit>
      <bibl>(1514-18 More, History of Richard III, p. 78)</bibl>
     </cit>

     <!-- <cit type="etymon"> elements from examples 23-27 -->
    ...
    </etym>
   ...
   </sense>
  </entry>
```

Example 23. Grammaticalization of Early Modern English *beside* 1535–1543.

```
  <entry>
   <!-- form, gramGrp, etc. -->
   <sense>
    ...
    <etym type="grammaticalization">
     ...
     <!-- <cit type="etymon"> elements from examples 19-22 -->

     <cit type="etymon" xml:id="at-1535-1543" prev="#at-1514-1518"
next="#at-1567">
      <date notBefore="1535" notAfter="1543"/>
      <oRef xml:lang="emodeng">beside</oRef>
      <gramGrp>
       <pos>extentionAdverb</pos>
      </gramGrp>
      <cit type="attestation">
       <quote>The toune of Chester is chiefly one streate of very meane building
yn lenght: ther is <oRef>beside</oRef> a smaul streat or 2. about the chirch; that
is collegiatid, .... </quote>
      </cit>
      <bibl>(1535-43 Leland, Itinerary I p. 74 [HC])</bibl>
     </cit>

     <!-- <cit type="etymon"> elements from examples 24-27 -->
    ...
    </etym>
   ...
   </sense>
  </entry>
```

10.2.4 Stage II: Sentential Adverb

141 One of the two contextual conditions in which sentential adverbials (sententialAdverb) occur is that they need to immediately follow a complementizer. This is represented in the data by the encoding of the relevant item within the <seg> element, and the @ana attribute is used to label it with the needed data category (complementizer).

Example 24. Grammaticalization of Early Modern English *beside* 1552–1563.

```
  <entry>
   <!-- form, gramGrp, etc. -->
   <sense>
    ...
    <etym type="grammaticalization">
     ...
     <!-- <cit type="etymon"> elements from examples 19-23 -->

     <cit type="etymon" xml:id="at-1552-1563" prev="#at-1535-1543"
next="#at-1554">
      <date notBefore="1552" notAfter="1563"/>
      <oRef xml:lang="emodeng">besides</oRef>
      <gramGrp>
       <pos>sententialAdverb</pos>
      </gramGrp>
      <cit type="attestation">
       <quote>…when the end is knowen, all wil turne to a iape ['trick,deceit'],
        Tolde he not you <seg ana="#Complimentizer">that</seg> <oRef>besides</
oRef> she stole your Cocke that tyde?</quote>
      </cit>
      <note>The adverbial of extension which signals the extension of a list
of referents as per (sense in example “at-#at-1535-1543”) is presumed to be
the source of the clause-initial, focused sentence adverbial, which extends the
propositional content with additional, non-central material:</note>
      <bibl>(1552-63 Gammer Gurton, p. 61 [HC])</bibl>
     </cit>

     <!-- <cit type="etymon"> elements from examples 25-27 -->
    ...
    </etym>
   ...
   </sense>
  </entry>
```

10.2.5 Stage III: Discourse Marker/Particle

142 In this next stage, *besides* serves to refocus the attention on the purposes of the discourse, and occurs on the left periphery of the sentence structure serving a pragmatic function. As in the previous example, the context is labeled using <seg> with the attribute @ana.

Example 25. Grammaticalization of Early Modern English *besides* 1554[53]

```
<entry>
 <!-- form, gramGrp, etc. -->
 <sense>
  ...
  <etym type="grammaticalization">
   ...
   <!-- <cit type="etymon"> elements from examples 19-24 -->

   <cit type="etymon" xml:id="at-1554" prev="#at-1552-1563" next="#stage3-26a">
    <date when="1554"/>
    <oRef xml:lang="emodeng">besides</oRef>
    <gramGrp>
     <pos>discourseMarker</pos>
    </gramGrp>
    <cit type="attestation">
     <quote>...<seg ana="#LPeriph">And <oRef>besides</oRef></seg>, it is very
unlike, that </quote>
    </cit>
    <bibl>(bef. 1554 Trial Throckmorton I,66.C1 [HC])</bibl>
   </cit>

   <!-- <cit type="etymon"> elements from examples 26-27 -->
  ...
  </etym>
 ...
 </sense>
</entry>
```

143 In these examples from 1619 and 1872 the usage of “beside” extends the discourse with afterthoughts:

Example 26. Grammaticalization of Early Modern English *beside* 1619.

```
  <entry>
   <!-- form, gramGrp, etc. -->
   <sense>
    ...
    <etym type="grammaticalization">
     ...
     <!-- <cit type="etymon"> elements from examples 19-25 -->

     <cit type="etymon" xml:id="at-1619" prev="#at-1554" next="#at-1872">
      <date when="1619"/>
      <oRef xml:lang="emodeng">beside</oRef>
      <gramGrp>
       <pos>discourseMarker</pos>
      </gramGrp>
      <cit type="attestation">
       <quote>..and <oRef>beside</oRef>, my complexion is so blacke, that..</
quote>
      </cit>
      <bibl>(1619 Deloney, Jack of Newbury, p.70 [HC])</bibl>
     </cit>

     <!-- <cit type="etymon"> elements from example 27 -->
    ...
    </etym>
   ...
   </sense>
  </entry>
```

Example 27. Grammaticalization of Modern English “besides” 1872.

```
<entry>
 <!-- form, gramGrp, etc. -->
 <sense>
  ...
  <etym type="grammaticalization">
   ...
   <!-- <cit type="etymon"> elements from examples 19-26 -->

   <cit type="etymon" xml:id="at-1872" prev="#at-1619">
    <date when="1872"/>
    <oRef xml:lang="en">besides</oRef>
    <gramGrp>
     <pos>discourseMarker</pos>
    </gramGrp>
    <cit type="attestation">
     <quote>The whooping cough seems to be a providential arrangement to force
you to come, as the expense will be little greater than going anywhere else;
      <oRef>besides</oRef> if you put a trusty female at Ravenscroft… </quote>
    </cit>
    <bibl>(1872 Amberley Ltrs, p. 513 [CLME])</bibl>
   </cit>

   ...
  </etym>
 ...
 </sense>
</entry>
```

144 Finally the source of the etymological data is cited as a child element of <etym>:

```
<entry>
 <!-- form, gramGrp, etc. -->
 <sense>
  ...
  <etym type="grammaticalization">
   <!-- <cit type="etymon"> elements from examples 19-27 -->
   ...
   <bibl>(Traugott, 1995)</bibl>
  </etym>
 </sense>
</entry>
```

145 This example is obviously an extremely complex yet informative representation of the history of the lexical item in question (English: “besides”).

146 Despite its importance within the field of linguistics, grammaticalization is rarely if ever identified as such in any kind of etymological or print dictionary. Moreover, to our knowledge, there have not been any attempts to encode such processes within XML, or any other digital markup standards. Grammaticalization is thus a major untapped resource of highly advanced etymological data. Establishing a practice in which this kind of markup is done may help linguists and lexicographers begin to more regularly make maximum use (and reuse) of the research being done.

11. Problematic and Unresolved Issues

147 For the issues regarded as the most fundamentally important to creating a dynamic and sustainable model for both etymology and general lexicographic markup in TEI, we have submitted formal requests for changes to the TEI GitHub, and will continue to submit change requests as needed. While this work represents a large step in the right direction for those looking for means of representing etymological information, there are still a number of unresolved issues that will need to be addressed. These remaining issues pertain to: (i) expanding the types of etymological information and refining the representation of the processes and features which are covered; and (ii) the need for continued progress in a number of issues within the body of international standards on which lexical markup relies.

148 Some examples of issues from (i) are as follows:

- encoding onomasiological etymological information, which groups and represents the converging and diverging histories of related forms across multiple related languages;
- relatedly, the markup of existing non-semasiological etymological dictionaries, which can be extremely long and which can be organized in extremely complex ways;
- sense shifts not involving metaphor or metonymy, but subtle, often usage- and context-based patterns such as pejoration–amelioration or narrowing–widening;
- changes to a lexical item which stem from corresponding changes and functional linguistic pressures occurring within other parts of the lexicon, such as semantic bleaching.

149 An example of issues from (ii) is:

- the need for expansion of the inventory of categories in ISOcat[54] to include well-established etymological processes, such as metaphor and metonymy, as well as other features not specific to etymology.

150 Several significant issues regarding language identification are as follows:

- the need for continued expansion of the inventory of the IANA registry, probably in conjunction with the maintenance of future versions of ISO 639;
- the need for more granularity in the means by which `@xml:lang` is specified without using the private tag (`-x-`) (see BCP 47 [IETF 2009a]);[55]
- the lack of existing codes within any body of standards for historical places such as Gaul and Carthage;
- the lack of a way to label when the language data is from an intermediate, transitional period for which there are neither IANA entries nor even commonly used terms within the body of literature.

12. Conclusion

151 In this paper, we have proposed a number of markup scenarios for encoding different types of etymological information within TEI dictionaries, drawing both from traditional lexicographic practice and from analytical approaches from functional and cognitive linguistics. Our examples

have shown a number of cases in which, in order to represent the lexical data as accurately as possible, it has been necessary to alter the content models of certain elements and attributes in ways we think should actually be considered necessary evolutions for the TEI Guidelines in the future. Relevant to both etymological and synchronic lexical markup, we have also touched upon how encoders can make use of linked open data URIs as a means of linking the sense and semantic domain(s) of an item to multilingual knowledge bases. Future work that remains to be undertaken is the encoding of onomasiological lexical data such as that found in dialectal and traditional etymological dictionaries.

BIBLIOGRAPHY

Amsler, Robert A., and Frank Wm. Tompa. 1988. "An SGML-Based Standard for English Monolingual Dictionaries." In *Proceedings of the 4th Annual Conference of the UW Centre for the New Oxford English Dictionary*, 61–80. http://projects.oucs.ox.ac.uk/teiweb/Vault/AI/aiw04.ps.

Bański, Piotr, and Beata Wójtowicz. 2009. "FreeDict: An Open Source Repository of TEI-Encoded Bilingual Dictionaries." http://www.tei-c.org/Vault/MembersMeetings/2009/files/Banski+Wojtowicz-TEIMM-presentation.pdf.

Bazin-Tacchella, Sylvie. 2001. "Rupture et continuité du discours médical à travers les écrits sur la peste de 1348." In *Air, miasme et contagion: les épidémies de l'Antiquité au Moyen Âge*, edited by S. Bazin-Tacchella, D. Quéruel, and E. Samama, 105–56. Langres: Guéniot. Author version available at https://hal.archives-ouvertes.fr/hal-00526759/.

Benczes, Réka. 2005a. "Metaphor- and Metonymy-based Compounds in English: A Cognitive Linguistic Approach." *Acta Linguistica Hungarica* 52(2–3): 173–98.

———. 2005b. "Creative Noun–Noun Compounds." *Annual Review of Cognitive Linguistics* 3: 250–68.

———. 2006a. *Creative Compounding in English: The Semantics of Metaphorical and Metonymical Noun-Noun Combinations*. Amsterdam: John Benjamins.

———. 2006b. "Analysing Metonymical Noun–Noun Compounds: The Case of Freedom Fries." In *The Metaphors of Sixty: Papers Presented on the Occasion of the 60th Birthday of Zoltán Kövecses*, edited by Réka Benczes and Szilvia Csábi, 46–54. Budapest: Eötvös Loránd University.

———. 2009. "What motivates the production and use of metaphorical and metonymical compounds?" In *Cognitive Approaches to English: Fundamental, Methodological, Interdisciplinary and Applied Aspects*, edited by Mario Brdar, Marija Omazić, and Višnja Pavičić Takač, 49–67. Newcastle-upon-Tyne: Cambridge Scholars.

Bisetto, Antonietta, and Sergio Scalise. 2005. "The Classification of Compounds." *Lingue e Linguaggio* 4(2): 319–32. doi:10.1418/20728.

Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt.

Botne, Robert. 2006. "Motion, Time, and Tense: On the Grammaticization of *Come* and *Go* to Future Markers in Bantu." *Studies in African Linguistics* 35(2): 127–88.

Brinton, Laurel J., and Elizabeth Closs Traugott. 2005. *Lexicalization and Language Change*. Cambridge: Cambridge University Press.

Budin, Gerhard, Stefan Majewski, and Karlheinz Mörth. 2012. "Creating Lexical Resources in TEI P5: A Schema for Multi-purpose Digital Dictionaries." *Journal of the Text Encoding Initiative* 3. http://jtei.revues.org/522. doi:10.4000/jtei.522.

Crist, Sean. 2005. "Toward a Formal Markup Standard for Etymological Data." Paper presented at the LSA Annual Meeting. Accessed April 16, 2017. http://www.sean-crist.com/professional/publications/crist_etym_markup.pdf.

Croft, William. 2003. *Typology and Universals*, 2nd ed. Cambridge: Cambridge University Press.

Erjavec, Tomaž, Dan Tufiş, and Tamás Váradi. 1999. "Developing TEI-Conformant Lexical Databases for CEE Languages." In [*Proceedings of the 4th International Conference on Computational Lexicography], COMPLEX'99*, edited by Ferenc Kiefer, Gábir Kiss, and Júlia Pajzs, 205–209. Budapest: Hungarian Academy of Sciences. http://www.racai.ro/media/Erjavec-TV-COMPLEX1999.pdf.

Fauconnier, Gilles. 2007. "Mental Spaces." In *The Oxford Handbook of Cognitive Linguistics*, edited by Dirk Geeraerts and Hubert Cuyckens, 351–76. Oxford: Oxford University Press.

Fauconnier, Gilles, and Mark B. Turner. 1996. "Blending as a Central Process of Grammar." In *Conceptual Structure, Discourse, and Language*, edited by Adele Goldberg, 113–29. Stanford: Center for the Study of Language and Information.

———. 1998a. "Conceptual Integration Networks." *Cognitive Science* 22(2): 133–87. doi:10.1207/s15516709cog2202_1.

———. 1998b. "Principles of Conceptual Integration." *Discourse and Cognition: Bridging the Gap*, edited by Jean-Pierre Koenig, 269–83. Stanford: Center for the Study of Language and Information.

———. 2000. "Compression and Global Insight." *Cognitive Linguistics* 11(3/4): 283–304. doi:10.1515/cogl.2001.017.

———. 2003. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.

Fomin, Maxim, and Gregory Toner. 2006. "Digitizing a Dictionary of Medieval Irish: The eDIL Project." *Literary and Linguistic Computing* 21(1): 83–90. doi:10.1093/llc/fqh050.

Geeraerts, Dirk. 2002. "The Interaction of Metaphor and Metonymy in Composite Expressions." In *Metaphor and Metonymy in Comparison and Contrast*, edited by René Dirven and Ralf Pörings, 435–65. Berlin: Mouton de Gruyter.

Good, Jeff, and Ronald Sprouse. 2000. "SGML Markup of Dictionaries with Special Reference to Comparative and Etymological Data." Paper presented at the workshop on Web-Based Language Documentation and Description, December 12–15, Philadelphia.

Goossens, Louis. 1994. "Metaphtonymy: The Interaction of Metaphor and Metonymy in Figurative Expressions for Linguistic Action." In *By Word of Mouth: Metaphor, Metonymy and Linguistic Action in a Cognitive Linguistic Perspective*, by Louis Goossens, Paul Pauwels, Brygida Rudzka-Ostyn, Anne-Marie Simon-Vandenbergen, and Johan Vanparys, 159–74. Amsterdam: John Benjamins.

Guevara, Emiliano, and Sergio Scalise. 2008. "Heads in Compounding: Variation and Patterns." Paper presented at the 13th International Morphology Meeting, Main Theme: Variation and Change in Morphology, Vienna, February 3–6. Abstract available at https://pdfs.semanticscholar.org/d2b5/652c01aca9306dc3f7f7e70cb38a6587ce0f.pdf, p. 43.

Harris, Alice C., and Lyle Campbell. 1995. *Historical Syntax in Cross-linguistic Perspective*. Cambridge Studies in Linguistics 74. Cambridge: Cambridge University Press.

Haspelmath, Martin. 2003. "The Geometry of Grammatical Meaning: Semantic Maps and Cross-linguistic Comparison." In *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, vol.2, edited by Michael Tomasello, 211–42. Mahwah, NJ: Lawrence Erlbaum.

———. 2009. "Lexical Borrowing: Concepts and Issues." In *Loanwords in the World's Languages: A Comparative Handbook*, edited by Martin Haspelmath and Uri Tadmor, 35–54. Berlin: De Gruyter Mouton.

Haspelmath, Martin, and Uri Tadmor, eds. 2009. *World Loanword Database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Accessed April 6, 2015. http://wold.clld.org.

Haugen, Einar. 1950. "The Analysis of Linguistic Borrowing." *Language* 26(2): 210–31. doi:10.2307/410058.

Hopper, Paul J., and Elizabeth Closs Traugott, 2003. *Grammaticalization*. Cambridge: Cambridge University Press.

Ide, Nancy, Adam Kilgarriff, and Laurent Romary. 2000. "A Formal Model of Dictionary Structure and Content." In *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*, edited by Ulrich Heid, Stefan Evert, Egbert Lehmann, and Christian Rohrer, 113–26. http://euralex.org/category/publications/euralex-2000/.

Ide, Nancy, and Jean Véronis. 1994. "Machine Readable Dictionaries: What Have We Learned, Where Do We Go?" In *Proceedings of the Post-COLING '94 International Workshop on Directions of Lexical Research*, 137–46. Beijing.

Ide, Nancy. 1995. "Encoding Dictionaries." *Computers and the Humanities* 29(2): 167–79.

IETF (Internet Engineering Task Force). 2009a. *BCP [Best Current Practice] 47: Tags for Identifying Languages*. Edited by Addison Phillips and Mark Davis. N.p.: IETF, September. https://tools.ietf.org/html/bcp47.

IETF (Internet Engineering Task Force). 2009b. *RFC (Request for Comments) 5645: Update to the Language Subtag Registry*. Edited by Doug Ewell. N.p.: IETF, September. https://tools.ietf.org/html/rfc5645.

ISO (International Organization for Standardization). 2007. *Presentation/representation of entries in dictionaries — Requirements, recommendations and information*. ISO 1951:2007. Geneva: ISO.

———. 2008. *Language resource management — Lexical markup framework (LMF)*. ISO 24613:2008. Geneva: ISO.

Jackendoff, Ray. 2009. "Compounding in the Parallel Architecture and Conceptual Semantics." In *The Oxford Handbook of Compounding*, edited by Rochelle Lieber and Pavol Štekauer. Oxford: Oxford University Press, 105–28.

Jacobson, Michel, and Boyd Michailovsky. 2002. "Linking Linguistic Resources: Time Aligned Corpus and Dictionary." Paper presented at the International Workshop on Resources and Tools in Field Linguistics, Las Palmas, Canary Islands, Spain, May 26–27. http://www.mpi.nl/lrec/2002/papers/lrec-pap-27-JACMICv2.pdf.

Kövecses, Zoltán, and Günter Radden. 1998. "Metonymy: Developing a Cognitive Linguistic View." *Cognitive Linguistics (includes Cognitive Linguistic Bibliography)* 9(1): 37–77. doi:10.1515/cogl.1998.9.1.37.

Laborderie, Noëlle, and Claude Thomasset. 1994. *Précis de phonétique historique*. Paris: Nathan.

Lakoff, George. 1987. *Women, Fire and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago Press.

———. 1993. "The Contemporary Theory of Metaphor." In *Metaphor and Thought*, 2nd ed., edited by Andrew Ortony, 202–51. Cambridge: Cambridge University Press.

Lakoff, George, and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Langacker, Ronald W. 1977. "Syntactic Reanalysis." In *Mechanisms of Syntactic Change*, edited by Charles N. Li, 57–139. Austin: University of Texas Press.

Langacker, Ronald W.. 1987. *Foundations of Cognitive Grammar. Vol. 1, Theoretical Prerequisites.* Stanford: Stanford University Press.

———. 1991. *Foundations of Cognitive Grammar. Vol. 2, Descriptive Application.* Stanford: Stanford University Press.

———. 2000. *Grammar and Conceptualization*. Berlin and New York: Mouton de Gruyter.

Lehmann, Winfred P. 1962. *Historical Linguistics: An Introduction*. New York: Holt, Rinehart & Winston.

Lemnitzer, Lothar, Laurent Romary, and Andreas Witt. 2013. "Representing Human and Machine Dictionaries in Markup Languages." *Dictionaries: An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Computational Lexicography*, edited by Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard, and Herbert Ernst Wiegand, 1195–1208. HSK 5.4. Berlin: De Gruyter Mouton. Author version available at https://hal.inria.fr/inria-00441215.

Matras, Yaron, and Jeanette Sakel, eds. 2007. *Grammatical Borrowing in Cross-Linguistic Perspective*. Berlin: Mouton de Gruyter.

Meillet, Antoine. 1912. "L'Evolution des Formes Grammaticales." *Scientia (Rivista di scienza)*, XII:XXVI,6. Reprinted (1965) in *Linguistique Historique et Linguistique Cimerale*. Paris: Librairie Honoré Champion.

Myers-Scotton, Carol. 2002. *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford: Oxford University Press.

Pagliuca, William, ed. 1994. *Perspectives on Grammaticalization*. Amsterdam: Benjamins.

Rennie, Susan. 2000. "Encoding a Historical Dictionary with the TEI (with Reference to the Electronic Scottish National Dictionary Project)." In *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*, edited by Ulrich Heid, Stefan Evert, Egbert Lehmann, and Christian Rohrer, 261–71. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. http://euralex.org/category/publications/euralex-2000/.

Romary, Laurent. 2001. "An Abstract Model for the Representation of Multilingual Terminological Data: TMF - Terminological Markup Framework." Paper presented at TAMA 2001, February 1–2, Antwerp, Belgium. https://hal.inria.fr/inria-00100405.

———. 2009. "Questions & Answers for TEI Newcomers." In *Jahrbuch für Computerphilologie* 10. Paderborn, Germany: Mentis. Author version available at https://hal.archives-ouvertes.fr/hal-00348372.

———. 2011. "Stabilizing Knowledge through Standards: A Perspective for the Humanities." In *Going Digital: Evolutionary and Revolutionary Aspects of Digitization*, edited by Karl Grandin. Stockholm: Center for History of Science at the Royal Swedish Academy of Sciences.

———. 2015. "TEI and LMF Crosswalks." *Journal for Language Technology and Computational Linguistics* 30(1): 47–70. Author version available at http://hal.inria.fr/hal-00762664; http://www.jlcl.org/2015_Heft1/3Romary.pdf.

Romary, Laurent, and Werner Wegstein. 2012. "Consistent Modeling of Heterogeneous Lexical Structures." *Journal of the Text Encoding Initiative* 3. http://jtei.revues.org/540. doi:10.4000/jtei.540.

Romary, Laurent, and Andreas Witt. 2014. "Méthodes pour la représentation informatisée de données lexicales/Methoden der Speicherung lexikalischer Daten." *Lexicographica: International Annual for Lexicography* 30: 152–86. Author version available at https://hal.inria.fr/hal-00991745. doi:10.1515/lexi-2014-0006.

Salmon-Alt, Susanne. 2006. "Data Structures for Etymology: Towards an Etymological Lexical Network." *BULAG* 31: 101–12. Author version available at http://hal.archives-ouvertes.fr/hal-00110971.

Salmon-Alt, Susanne, Laurent Romary, and Eva Buchi. 2005. "Modeling Diachrony in Dictionaries." Paper presented at ACH/ALLC Conference 2005, Victoria, BC, Canada, June 15–18.

Scalise, Sergio, and Emiliano Guevara. 2005. "The Lexicalist Approach to Word-formation and the Notion of the Lexicon." In *Handbook of Word-formation*, edited by Pavol Štekauer and Rochelle Lieber, 147–87. Dordrecht: Springer. doi:10.1007/1-4020-3596-9_7.

Scholz, Johannes, Thomas J. Lampoltshammer, Norbert Bartelme, and Eveline Wandl-Vogt. 2015. "Spatial-Temporal Modeling of Linguistic Regions and Processes with combined Indeterminate and Crisp Boundaries." In *Proceedings of the 1st ICA European Symposium on Cartography*, edited by Georg Gartner and Haosheng Huang, EuroCarto 2015, 10–12 November 2015, Vienna, Austria.

Schopper, Daniel, Jack Bowers, and Eveline Wandl-Vogt. 2015. "dboe@TEI: Remodelling a Database of Dialects into a Rich LOD Resource." Paper presented at the Text Encoding Initiative Conference and Members' Meeting 2015, Lyon, France, October 28–31. Abstract available at http://tei2015.huma-num.fr/en/papers/#146.

Svorou, Soteria. 1994. *The Grammar of Space*. Amsterdam: J. Benjamins.

Sweetser, Eve E. 1988. *Grammaticalization and Semantic Bleaching*. In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, edited by Shelley Axmaker, Annie Jaisser, and Helen Singmaster, 389–405. Berkeley, CA: Berkeley Linguistics Society.

TEI Consortium. 2016. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.1.0. Last updated December 15. N.p.: TEI Consortium. http://www.tei-c.org/Vault/P5/3.1.0/doc/tei-p5-doc/en/html/.

Traugott, Elizabeth C. 1995. "The Role of the Development of Discourse Markers in a Theory of Grammaticalization." Paper presented at the Twelfth International Conference on Historical Linguistics (ICHL XII), Manchester, England, August 13–18.

Vandeloise, Claude. 2006. "Are There Spatial Prepositions?" In *Space in Languages: Linguistic Systems and Cognitive Categories*, edited by Maya Hickmann and Stéphane Robert, 139–54. Typological Studies in Language 66. Amsterdam: J. Benjamins.

Windhouwer, Menzo, and Sue Ellen Wright. 2012. "Linking to Linguistic Data Categories in ISOcat." In *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, edited by Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, 99–107. Berlin: Springer.

Yin, Hui. 2010. "The So-called Chinese VV Compounds—A Continuum between Lexicon and Syntax." In *Proceedings of the 2010 Annual Conference of the Canadian Linguistic Association*, edited by Melinda Heijl. Toronto: CLA. http://homes.chass.utoronto.ca/~cla-acl/actes2010/actes2010.html.

NOTES

1 Some of these amendments have been validated by the TEI Council. See https://github.com/TEIC/TEI/issues/1512.

2 Originally titled “Print Dictionaries” before it made an appropriate digital turn to cover lexical resources at large.

3 Uniform Resource Identifier: A naming mechanism to identify a resource on the Internet in a univocal way.

4 For non-expert readers interested in a quick overview of general encoding possibilities offered by the TEI Guidelines, we recommend looking at the TEI by Example initiative, http://www.teibyexample.org/, or Romary 2009.

5 See Romary and Witt 2014 for an overview of onomasiological and semasiological models; Lemnitzer, Romary, and Witt 2013 and Romary and Wegstein 2012 for an in-depth analysis of the TEI dictionary model; and Romary 2015 for a discussion of the relation between the TEI dictionary model and the ISO 24613 (LMF) standard (ISO 2008).

6 ISO 16642:2003: *Computer applications in terminology — Terminological markup framework* (https://www.iso.org/standard/32347.html); see also Romary 2001.

7 Standard Generalized Markup Language, ISO standard ISO 8879 published in 1986, which is the direct ancestor of XML.

8 Document Type Definition, the grammar of an SGML document.

9 The P4 edition of the TEI Guidelines, which was completely based upon XML, was not published until 2001.

10 Germanic Lexicon Project, accessed June 16, 2017, http://lexicon.ff.cuni.cz/.

11 In the context of the ongoing revision of the LMF document as a multipart standard, there is now provision for a specific part on diachrony and etymology.

12 Since the paper was written before the XML serialization of LMF was stabilized, Salmon-Alt (2006) construed an XML representation partially informed by the then-ongoing discussions and partially inspired by the TEI Guidelines.

13 A possible replacement for `<gloss>` could be a construct such as `<ref corresp="…" type="sense">`.

14 For the sake of conciseness we have not added the bibliographic description (`<bibl>`) to this example, although it would also fit into the `<cit>` construct outlined here.

15 It may of course be the case that the source of a lexical item is unknown.

16 Currently the use of `@type` in the `<etym>` element is not permitted in the TEI schema as the aforementioned element is not a member of the `att.typed` class. However, a proposal for this use has been accepted by the TEI Council and will be implemented in a future version of the TEI Guidelines; meanwhile, the ticket is available at https://github.com/TEIC/TEI/issues/1512. In our proposal, the `<etym>` element has to be made recursive in order to allow the fine-grained representations we propose here. The corresponding ODD customization, together with reference examples, is available on GitHub.

17 There may also be cases in which it is unknown whether a given etymological process occurred within the contemporary language or parent system; in such cases the encoder can just use the main language tag for both the diachronic and synchronic portions of the entry as a default (see, for instance, example 11).

18 The language information of an etymon can also be specified within the `<cit type="etymon">` in which the `<pRef>` or `<oRef>` is embedded (as explained above).

19 IANA Language Subtag Registry, accessed June 16, 2017, http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry.

20 Romance dialectologists estimate the range of dates that Vulgar Latin began to become distinct in Gaul/France anywhere between the second century CE and the fall of the Roman Empire around the end of the fifth century CE (Bazin-Tacchella 2001). Vulgar Latin gave way in the ninth century to Old French (ISO 639–3 `"fro"`), which is designated by the registration body as dating between circa 842 and 1400; however, in the example, the antepenultimate form in the phonological development *šyęf* is dated as having been used in the eighth century, which means anywhere from 700 to 799 CE, the later end of which is just 43 years prior to the early portion of Old French. This issue raises the question of whether it is really useful to designate a separate language for a period of roughly forty-three years.

21 The interested reader may ponder here the possibility to also encode scripts by means of the `@notation` attribute instead of using a cluttering of language subtags on `@xml:lang`. For more on this issue, see the proposal in the TEI GitHub (https://github.com/TEIC/TEI/issues/1510).

22 The theoretical line between the two (*inheritance* and *borrowing*) becomes blurry when the scope of a language's history is expanded; depending on how far back one looks, an item that was inherited from a direct ancestor may have been borrowed at an earlier time. Where such cases are known, it is possible to encode both etymological lineages within the same entry.

23 Despite the fact that it is widely accepted among researchers that the actual language spoken by most Roman peoples in everyday life was a non-standardized and non-literary language referred to as Vulgar Latin (VL), there is no ISO 639 language code for this. Instead, there only exists a single tag for Latin (ISO 639–3: `"la"`). Not distinguishing at least between Classical, Vulgar, and Medieval Latin is just not an accurate depiction of the etymological information. Needless to say, this issue needs to be addressed and one or more proposals made for the creation of these tags within the ISO 639 system.

24 The final form given is the Middle French (and not Modern) because according to the source, it was the final stage in the phonological evolution of that item and is identical to the present-day form.

25 In other examples, the values of `@notation` used in `<pron>` and `<pRef>` have been well-known standard systems with conventional names, e.g., `"ipa"` and `"xsampa"`. However, in this example, while in the source the author explains the phonetic correlate for the transcription notation used in the work, and there are some characters that are used in the IPA, this system has no proper name, thus we have chosen to label it `"private"`.

26 The element `<date>` as a child of `<cit>` is another example which does not adhere to the current TEI standards. We have allowed this within our ODD document. A feature request proposal will be made on the GitHub page and this feature may or may not appear in future versions of the TEI Guidelines.

27 In the (French language) source of this example (Laborderie and Thomasset 1994, 97) the dates were given in a combination of roman numerals with superscripted numbers and letters to indicate the century, e.g., IVe2, which corresponds to deuxième moitié du 4e siècle, second half of the fourth century (CE). Optionally, the encoder could include the original date as the value of the `<date>` element for human readability. In such a case, for the purposes of quality data structuring, compatibility, and retrievability the dating information should nonetheless also be included as attribute value(s): e.g., `<date notBefore="0350" notAfter="0399"> IVe2 </date>`.

28 See Misha Wolf and Charles Wicksteed, "Date and Time Formats" (W3C Note), 1997, accessed April 26, 2017, http://www.w3.org/TR/NOTE-datetime.

29 One of our anonymous reviewers kindly noted that we should bear in mind that these attributes are for Gregorian dates only. If we are ever encoding anything more specific than a year, it is necessary to use `@datingMethod` and custom dating attributes to clarify that the Julian calendar is the one we are using here, rather than the proleptic Gregorian.

30 This would be referred to as case loss from the perspective of the item's morphosyntactic etymology.

31 Or any terminological equivalent to either.

32 Which one could easily retrieve by means of a simple XPath instruction such as `./cit[@type='etymon']/gramGrp[gen='neut']`

33 The question "Why do languages borrow?" is examined by Haspelmath (2009). This work also gives an excellent overview of the various typological differences discussed in past works.

34 Alteration to the gender property in the borrowed form is also evident in the first example *pamplemousse*, as in French it is a masculine noun, whereas in Dutch it was feminine.

35 In the digital source quoted here, entries are associated with an identifier (`@xml:id`) named from the English gloss. Besides, different senses correspond to different entries. The identifier `"#bean"` thus refers to the entry for *ntuchi* with the meaning "bean." In many cases, such a reference would normally point to senses instead of separate entries.

36 Sometimes the semantic, grammatical, or other aspects of the source item and those of the importing language are not equivalent. In such a case, the contrast can be inherently represented through the differing information within the synchronic portion of the entry and that in the `<cit type="etymon">` section. Additionally, the encoder can also include a prose description of this with a `<note>` element.

37 While our examples do not show the use of the exocentric/endocentric typology, if so desired, the basic components of this could be encoded in the synchronic portion of an entry using the attribute `@subtype`, e.g., `<entry type="compound" subtype="endocentric">`.

38 Studies using a cognitive linguistic framework have additionally shown that metaphor and metonymy are very often active in both exocentric and endocentric compounds across languages and regardless of the grammatical subcomponents (Langacker 1987, 1991, 2000; Benczes 2005a, 2005b; Benczes 2006a, 2006b).

39 While we do not show an example here, if the editor desires to label the head of the compound it could be done using the attribute `@ana`, the value of which can correspond to a feature structure that is defined and declared within the project. Where tagging within the synchronic portion of the entry, `@ana` should be included in the `<seg>` of the head component. Where tagging within the diachronic portion, `@ana` should be included on the `<cit type="etymon">` of the head component of the compound. The ISOcat has an entry for the category "head" whose persistent identifier is http://www.isocat.org/datcat/DC-2306.

40 Note that the structure for modeling *compounding + metaphor* applies in the exact same manner to *compounding + metonymy*, which is not shown in these examples.

41 An example of this in English is "horseradish," which speakers can parse into the components of "horse" and "radish"; however, the antiquated figurative sense of horsewhich meant "large, strong, or coarse" and which gave rise to the compound is no longer used in modern English (*OED Online*, June 2017, s.v. "horse, n." [Compounds, section 2, sense c], Oxford University Press, accessed June 19, 2017, http://www.oed.com/view/Entry/88583?rskey=ODVZef&result=1).

42 Examples of some of the most common grammaticalization patterns ("clines") include the following (from Croft 2003): noun > adposition; adposition > case affix; verb > classifier; demonstrative or article > gender/noun class marker; main verb > tense, aspect, mood marker.

43 Svorou's panchronic view of grammaticalization is supported by Hopper and Traugott (2003), who state that clines are both diachronic (schema of evolution) and synchronic (all/multiple forms of evolutionary stages may co-exist).

44 According to Hopper and Traugott (2003), reanalysis modifies underlying representations (semantic, syntactic, morphophonological) and brings out rules changes. Langacker (1977, 58) has provided the basis for the definition of reanalysis, describing it as "change in the structure of an expression or class of expressions that does not involve any immediate or intrinsic modification

of its surface manifestation." Harris and Campbell (1995) and Hopper and Traugott (2003) have added to the description of the phenomenon, pointing out that reanalysis can only be manifested through analogy.

45 Analogy involves paradigmatic organization, changes in surface collocations, and changes in patterns of usage, which in turn make unobservable changes from reanalysis observable (Hopper and Traugott 2003).

46 Given the prominent role of phonological change in grammaticalization (e.g., coalescence, phonological reduction, loss), inclusion of `<pron>` is recommended though it is not always possible to present such forms with full certainty (in which case there is always the TEI attribute `@cert`).

47 While not entirely essential, including any known contextual, pragmatic, or collocational information greatly enhances the usefulness of the synchronic data modeling and enhances the precision of the etymological account.

48 In Traugott (1995), discourse markers are a subtype of discourse particle.

49 Additional information related to this etymon could be added here, for instance that sid is a stem and the head of the inflected noun form. This would be a typical application of `@ana` on the `<cit>` element.

50 In this example, the adverbial component of the etymon, the additional feature of prefix, which would be specified as the entry type within a synchronic entry, could be labelled within a `@ana` attribute, e.g., `<cit type="component" corresp="#e2s1" ana="#prefix">`.

51 One etymon from the source has been skipped in this example: *beside* in the prepositional sense attested ca. 1450.

52 "Extension adverbials" are also referred to by Traugott as "verbal adverbials."

53 Etymon given in the source but not encoded here: *besides* in the sense of an adverbial of extension attested 1567.

54 ISOcat is currently stalled, with two initiatives from ISO TC 37 and CLARIN taking up the legacy. We shall be working on making the two initiatives converge again.

55 Currently, labeling language varieties while remaining interoperable is limited to the specification of the following information: (language-country-region).

ABSTRACT

In this paper we provide a systematic and comprehensive set of modeling principles for representing etymological data in digital dictionaries using TEI. The purpose is to integrate in one coherent framework both digital representations of legacy dictionaries and born-digital lexical databases that are constructed manually or semi-automatically.

We provide examples from many different types of etymological phenomena from traditional lexicographic practice, as well as analytical approaches from functional and cognitive linguistics such as metaphor, metonymy, and grammaticalization, which in many lexicographical and formal linguistic circles have not often been treated as truly etymological in nature, and have thus been largely left out of etymological dictionaries.

In order to fully and accurately express the phenomena and their structures, we have made several proposals for expanding and amending some aspects of the existing TEI framework.

Finally, with reference to both synchronic and diachronic data, we also demonstrate how encoders may integrate semantic web/linked open data information resources into TEI dictionaries as a basis for the sense, and/or the semantic domain, of an entry and/or an etymon.

INDEX

Keywords: TEI, dictionary, etymology

AUTHORS

JACK BOWERS

Jack Bowers is a research assistant at the Austrian Academy of Sciences (ÖAW)—Austrian Center for Digital Humanities (ACDH). He is the primary curator of the DBÖ (Datenbank der bairischen Mundarten in Österreich) corpus and is working on cleaning, converting, and enhancing the lexical and metadata in order to facilitate reuse, and to bring it in line with international standards and best practice. With a background in cognitive and functional approaches to all levels of linguistics and their interfaces (semantics, morphosyntax, phonetics, phonology, etymology), he applies linguistic expertise to best represent and integrate lexical data (spoken or textual) with metadata and semantic knowledge/information resources. He is also interested in working towards interoperability between standards for lexical markup (TEI, LMF, ONTOLEX, TBX) and in the emerging prospects offered by semantic web/ontological resources in the integration of human

knowledge across academic and scientific fields. Jack holds a B.A. in history and French from San Francisco State University (2009) and an M.A. in linguistics and a certificate in computational linguistics from San Jose State University (2012). He is a member of the DARIAH-GiST project, which works to promote the use of digital standards in the humanities, and became involved in the digital humanities while working on his documentation/multimedia resource/corpus creation of the Mixtepec-Mixtec language variety (spoken Juxtlahuaca district, Oaxaca, Mexico) (2011–).

LAURENT ROMARY

Laurent Romary is Directeur de Recherche at Inria, France, director general of the European infrastructure DARIAH, and guest scientist at the Centre Marc Bloch and the Academy of Sciences in Berlin. He carries out research on the modeling of semi-structured documents, with a specific emphasis on texts and linguistic resources. He received a PhD in computational linguistics in 1989 and his Habilitation in 1999. During several years he launched and directed the Langue et Dialogue team at Loria in Nancy, France, and participated in several national and international projects related to the representation and dissemination of language resources and on human–machine interaction. He is the chairman of ISO committee TC 37 and has been a member (2001–2007), then chair (2008–2011), of the TEI (Text Encoding Initiative) Council and now member of the TEI board (2017–2018). Beyond his research activities, he has been responsible for defining and implementing the scientific information and open access policies of major research institutions in Europe, namely CNRS, Max Planck Society, and Inria.