HAL
archives-ouvertes.fr

Data structures for etymology: towards an etymological lexical network.

Susanne Salmon-Alt

▶ **To cite this version:**

Susanne Salmon-Alt. Data structures for etymology: towards an etymological lexical network.. BULAG, Presses Universitaires de Franche-Comté, 2006, 31, pp.1-12. <hal-00110971>

HAL Id: hal-00110971

https://hal.archives-ouvertes.fr/hal-00110971

Submitted on 2 Nov 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Bulag n° 31, 2006
Étymologie

DATA STRUCTURES FOR ETYMOLOGY: TOWARDS AN ETYMOLOGICAL LEXICAL NETWORK

Susanne ALT
ATILF-CNRS, Nancy, France
salt@atilf.fr

Résumé

Dans cet article, nous allons présenter une analyse générique de données étymologiques ayant pour but de fournir un cadre uniforme pour modéliser une telle information dans des bases de données lexicales. Basé sur la réification explicite des étymons et de leurs liens entre eux, ce que nous proposons fournit un moyen de formuler des contraintes pour les deux, de même qu'un ensemble préliminaire de descripteurs standards qui pourront être utilisés à cette fin. Le modèle a été abondamment testé sur de nombreux cas concrets tirés du TFLi (Trésor de la Langue Française informatisé), nous permettant d'identifier d'autres mécanismes (alternatives, liens multiples, composition etc.) nécessaires pour les représentations de données étymologiques. En prenant en compte les efforts actuels de standardisation du comité ISO TC 37/SC 4 afin de définir une plateforme spécifique pour les données lexicales (connue sous le nom de LMF Lexical Markup Framework, cadre de marqueurs lexicaux), nous essaierons de montrer que notre étude pourrait être une contribution possible à ce projet.

Mots clefs

Étymologie ; lexique ; standardisation ; Trésor de la Langue Française

Abstract

In this paper, we present a generic analysis of etymological data intended to provide a uniform framework for modeling such information in lexical databases. Based on the explicit reification of etymons and of the links between them, our proposal provides means to state additional constraints for both, as well as a preliminary set of standard descriptors to be used to this end. The model has been extensively tested on a variety of concrete cases extracted from the TLFi (Trésor de la Langue Française informatisé), allowing us to identify further mechanisms (alternatives, multiple links, composition,

etc.) needed for etymological data representation. Taking into account the current standardization efforts within ISO committee TC 37/SC 4 to define a specification platform for lexical data (aka LMF, Lexical Markup Framework), we try to show that our proposal could be a possible contribution to this project.

Key-words

etymology; lexicon; standardization; Trésor de la Langue Française

1. Introduction

Lexical data appear in a wide variety of forms. With regard to the lexicographical content, the data may range from basic morpho-syntactic structures (e.g. *Morphalou*) to important editorial projects that cover multiple levels of synchronic description, such as morphological information, syntactic constructions, sense related information (definitions, examples, usage notes, etc.), but also diachronic information. With regard to the underlying data structures, lexical data range from relatively loosely structured machine readable dictionaries to highly structured lexical databases, primarily intended to be accessed by natural language processing tools or by human users via specific search interfaces.

From a computational point of view, this situation prevented in the past the design of one single data structure that fits all the possible needs. On the other hand, users would benefit from uniform access to similar information across heterogeneous lexical resources. Standardization in lexicography has therefore been subject to strong debates, leading for instance to the Print Dictionary chapter of the TEI (TEI P5, 2005) that tries to combine structured and unstructured views of lexical entries. Some consense has also been achieved on representation models for highly structured lexical data, in particular within the context of exchangeable NLP lexica (cf. the EAGLES, ISLE/MILE or LMF projects). For diachronic information however, the situation is less advanced: there is no current practice to be considered as a good candidate for a standard. The TEI view on diachronic information, for instance, is in fact a loose tagging of some bits of information currently to be found in etymological notes. In practice, even important digitization and encoding projects of national reference dictionaries (e.g. *TLF* or *DWB*) did not tackle the difficult issue of structuring deeply diachronic and etymological information present in the original print versions.

Doing so would require two different steps: first, to analyse carefully the linguistic data and to elaborate a model of the structures underlying etymological information; second, to develop a parser able to analyse the source text of etymological notes in current dictionaries and to annotate the data accordingly.

This paper is mainly concerned with the first step: we want to show that it is possible to apply coherent modeling principles to currently unstructured lexicographical information, that is etymological data, as appearing in general language dictionaries with wide lexical coverage. The resulting model, based on a careful analysis of etymological data in the TLF, is anchored in the wider context of Lexical Markup Framework (LMF), an ISO initiative intending to elaborate standardizable, while still customizable, data models for synchronic lexical databases. 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. Based on this assumption, we propose a data structure which tries to remain as compatible as possible with the data structures so far developped for synchronic lexical entries.

2. Etymology as a part of diachronics

For the current purpose, we restrict ourselves to general language dictionaries, i.e. dictionaries following *grosso modo* semasiological lexicographical principles such as, for instance, accounted for by the TEI chapter on print dictionaries (TEI P5). In no case, we intend to address here the issue of etymological dictionaries whose both macro- and microstructure rely on different organization principles, such as the *Französisches Etymologisches Wörterbuch* (FEW, cf. Buchi 1996). We consider diachronic information along the lines of its modern, large acceptation as “a word’s biography” (Baldinger, 1959). As such, it covers both etymological information in a restricted sense and historical notes: the first informs about the origin and primitive significance of a lexeme in its source language, whereas the second informs about successive changes of form and meaning in the target language. Such diachronic information in a large sense can for instance be found in the *OED*, in the *DWB* or in the *TLF*.

As a part of diachronic information, etymology properly speaking is concerned with the origin and evolution of the lexeme before it entered into the target language. It is generally presented as a set of one or more etymons, associated

with an etymological class (inheritance, loan word, word generation). Note that etymons in this sense are related to only the oldest sense, but not to all individual senses in the modern stage of the considered language. In the example of Figure 1 (TLF), the etymon for the oldest sense of *pamplemousse* (sense 1a) is the Dutch *pompelmoes*. The etymological class is therefore the class of loan words. The etymon itself is a compound of *pompel* and *limoes*. Additionally to core information about etymons and etymological classes, etymological notes may provide bibliographical references for etymological hypotheses and/or discuss other issues such as phonetic evolution, concurrent hypotheses, confidence statements, secondary etymons and motivations (popular etymology), testimony of etymons or intermediate evolution stages.

PAMPLEMOUSSE, subst. masc.
[…]

Empr. au néerl. *pompelmoes*, fém., au sens 1 a, qui est prob. comp.de *pompel* «gros, enflé» et de *limoes* «citron» (BOULAN, p.148; KÖNIG, pp.159-160). Apparaît d'abord dans des textes fr. qui le donnent comme mot néerl.: 1665 *pompelmoes* (J. LE CARPENTIER, *L'Ambassade de la Compagnie orientale des Provinces Unies...* [trad. d'un ouvrage néerl.], II, p.88 ds ARV.); 1666 *pompelmous* (M. THÉVENOT, *Relation de divers voyages curieux...* t.3 ds KÖNIG).

Figure 1: Etymological information for the entry pamplemousse (TLF)

3. A data model for etymological information

In the following sections, we use the modeling principles of the LMF project, developped within the ISO committee TC 37/SC 4 (Francopoulo & Monte, 2005).

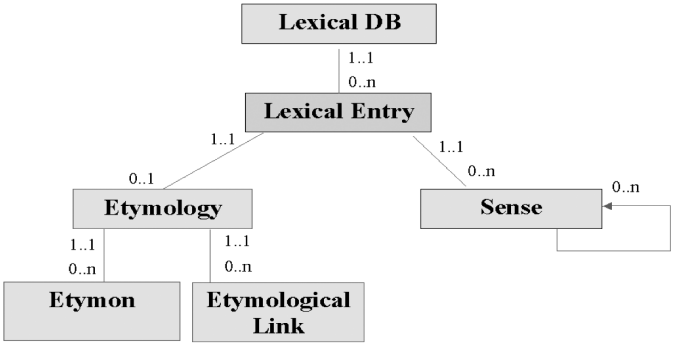

Figure 2: The LMF metamodel and the etymological extension

These principles rely on Ide & Romary (2004), Bunt & Romary (2004) and Salmon-Alt et al. (2005), and allow a user to combine a metamodel, which informs the main components of the data structure, with data categories, corresponding to elementary information units attached to the nodes of the metamodel. In the case of lexical structures, a metamodel is itself the combination of a core metamodel (a fairly simple structure organizing a lexical database into lexical entries, described as a couple of a form and a hierarchy of senses) and lexical extensions, seen as additional modules attached to the core metamodel (Figure 2). In our case, we will consider more in detail the kind of extensions needed for etymological information. In particular, we show how lexicographical data structures initially proposed for synchronic databases might be adapted to outline the structure of etymological information in dictionary entries, while reusing existing components and data categories, and keeping the changes as minimal as possible. The idea behind doing so is to ensure a maximal compatibility with external LMF compliant databases, especially in order to avoid redundancy in a (virtual) network of lexical databases.

3.1. Etymology

We start with the assumption that the fundamental function of etymological notes is to assign (at least) one etymon to the lexical unit under consideration, and to provide information about the type of the relation holding between them. Therefore, we propose a basic lexical extension for etymological notes (*Etymology*) that accounts for the description of etymons and links. Under

the assumption that lexical entries are purely polysemous (do not contain homonyms), the *Etymology* component occurs at most once for a given lexical entry. It is further structured by means of *Etymon* and *Etymological Link* components.

3.2. Etymons

Etymons are basically words, located in time and space, which stand in a particular diachronic relation to other words. As such, it seems convenient to describe etymons in the same way as lexical entries. Hence, in terms of LMF, they might be characterized by any existing data category used in the description of synchronic lexical entries. Among those basic descriptors, *language, orthography, pronunciation, glose, part-of-speech* and *inflectional information* are able to capture most of the linguistic features associated with etymons in current dictionaries. For example, language, orthography, part-of-speech, but also inflectional information are necessary to describe the etymon of *luire*, i.e. the future tense of old French *luisir* (Figure 3). However, a careful analysis of data in the TLF points to some important differences to be taken into account:

- The coverage of the *language* attribute should be extended to more fine-grained geographical and diachronic variants than those currently available from the ISO 639 series. Figure 3 illustrates this for languages such as *a. fr.* ("ancien français"), *gaulois, picard* or *a.b. frq.* ("ancien bas francique").
- As opposed to the description of synchronic lexical data, etymons might be reconstructed, or hypothesized word forms. Those are graphically indicated by an asterisk (cf. **werra*). Moreover, the information about the word form might be totally unknown, like for the etymon of latin ENCAUSTUM (Figure 3). As a consequence for the data model, we propose to encode the status of word forms by a newly defined data category *testimony* with an appropriate range of values.
- In case of composition (cf. *pamplemousse*), it should be possible to refer to a set of (possibly embedded) etymons, one for each part of the compound, and to factorize common descriptors, such as *language*.
- As opposed to composition, derivation (cf *ronfler*, Figure 3) rises the question of the lexicographical status of derivational affixes and non autonomous roots: the hypothesis of etymons as lexical entries leads *in fine* to the assumption that those morphemes are to be treated as lexical objects in the same way as autonomous lexical entries. This

assumption is conformant to in-depth theoretical work on derivational morphology (Corbin, 1987). Practically, the TLF treats derivational prefixes and suffixes indeed as separate lexical entries.

– For practical and documentary reasons, we allow for one ore more free text elements (*note*), in order to capture less structured information and/or to keep track of the original source text of the etymological notes. Those element are indeed compatible with the model, but they are not considered as mandatory parts of it.

MODERN FRENCH	DIRECT ETYMON	SECONDARY ETYMON
guerre	< *a.b. frq.* *werra	
encre	< *lat.* ENCAUSTUM	< gaulois (?)
luire	< *a.fr.* luisir remplacé par formes du futur	< *lat.* LUCERE
ronfler	< *lat.* RONFL- [depuis 12e siècle, aussi italo- et ibéroroman]	
albarelle	< *it.* albarello/albarella	
amour	< a.fr. ameur remplacé par formes dialectales picardes	< lat. AMOR

Figure 3: Some examples of etymological data (TLF)

3.3 Etymological links

Etymological Link components stand for etymological relations between linguistics units. They might be thought of as directed and typed arcs between etymons, i.e. lexical entries. A link is basically characterized by two elementary data categories: an *etymological target* (the synchronic lexical unit) and an *etymological source* (the etymon). Both point to structures similar to lexical entries. In practice, the linguistic material is mainly recovered from the current dictionary, and corresponds either to lexical entries of the dictionary or to etymological units as described in the previous section. However, in principle, it should be envisageable to point also to existing external resources, e.g. for another language or another time span.

Etymological links are typed by an *etymological class* (loan word, inheritance, word generation). Additionally, they may bear information about the bibliographical reference, confidence level or other notes. This simple data model accounts for different cases of etymological structures as found in current dictionaries, including etymological chains, multiple etymons (standing in conjunctive or disjunctive reading), and multiple evolutions of a same etymon.

- Etymological chains are built from lexical units standing in direct etymological filiation relations (cf. *amour, ameur, AMOR* in Figure 3). Sometimes, etymological information in dictionaries is limited to the direct etymon, i.e. the most recent item of the chain, fulfilling the conditions for being considered as an etymon (for instance, the Latin word for French inheritants). However, dictionaries contain often more than one etymon, trying to provide also the etymon of the etymon and so on. Etymological chains can be easily represented by encoding more than one etymon and link under the same lexical entry.
- Multiple etymons may occur in a disjunctive reading. This is frequently the case for more than one hypotheses about etymological filiation. As an example, the French *gley* has been related in the litterature alternatively to the Russian *glej* or the Ukrainian *hlej*. In case there is a preferred solution, the alternative links might be weighted with confidence scores or attributed to scientific authority (bibliographic sources).
- Multiple etymons also occur in conjunctive readings. As an example, one may think of derivational or inflectional variants (cf. *albarello/albarella*, Figure 3), or of re-motivation in popular etymology, such as for French *choucroute*. Initially, it is an Alsacian loan word (corresponding to German *Sauerkraut*), that has been influenced phonetically by French *chou*, due to semantic proximity. In those cases, it is plausible to consider both *Sauerkraut* (i.e. its Alsacian variant) and *chou* as contributing etymons for *choucroute*. The link between *chou* and *choucroute* may be typed as a secondary or re-motivating link.
- Multiple evolution of a same etymon is a frequent case, especially for languages with a recent and well documented parent language, like Latin for French. As an example, Latin DIRECTIARE is considered as a possible etymon for French *adroit* as well as for *adresser*. In our proposal, those cases might be treated without redundancy: the etymon has to be described only once (say, in the entry for *adresser*), and the link to the target *adroit* can refer to this etymon as its source.

4. An integrated example

The application of the previously introduced principles to the TLF example of Figure 1 leads to the sample data structure in Figure 4. The nodes in the graph are lexical entries in a virtual network for *pamplemousse*, *pompelmoes* and (*pompel+moes*), respectively.

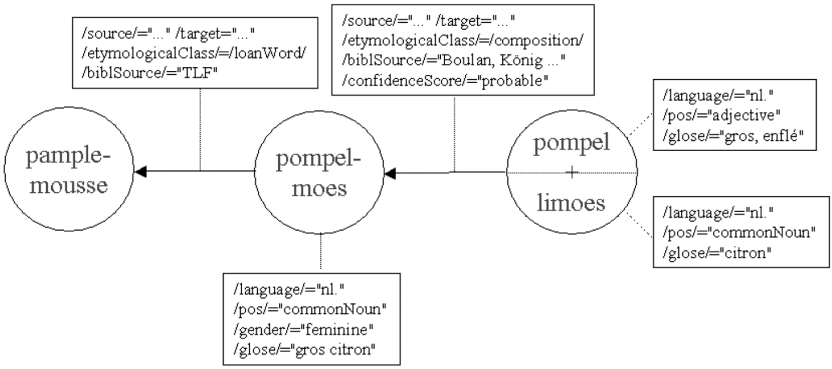

Figure 4: Etymological data structure for *pamplemousse* (TLF)

The first one has not to be described furthermore, since it is the anchoring entry for the etymological information (and as such supposed to be described out of the etymological part). The second one corresponds to the direct etymon *pompelmoes*, whose descriptors (language, part-of-speech, gender and glose) are able to capture all the information provided in the source. The TLF goes still further in the etymological chain, and characterizes *pompel* and *moes* as parts of the compound. We consider this as a complex etymon, built from two subparts, each with its proper set of descriptors. The three lexical entries are part of a common etymological chain. Therefore, we need two links in order to express the relationships between them: a first link, with *pompelmoes* as the source and *pamplemousse* as the target, is typed as a loan word relation. The second one relates *pompel* and *moes* to the compound *pompelmoes*. Note that we are also able to record in an appropriate way the bibliographical reference and confidence indication.

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicalEntry id="LE1">
    <form>
        <orth>pamplemosse</orth>
        …
    </form>
    <sense>
        …
    </sense>
    <etymology>
        <etymon id="LE2">
            <form>
                <orth xml:lang="dutch">pompelmoes</orth>
                <pos>commonNoun</pos>
                <gender>feminine</gender>
            </form>
            <sense>
                <glose>Citrus maxima</glose>
                <note>probablement d'origine tamoule, De Vries, Nederl</note>
            </sense>
        </etymon>
        <etymologicalLink source="LE2" target="LE1">
            <etymologicalClass>loan word</etymologicalClass>
        </etymologicalLink>
    </etymology>
</lexicalEntry>
```

Figure 5: XML version of the etymological data structure for *pamplemousse*

Finally, Figure 5 shows a simplified XML implementation of the first subgraph (i.e. the lexical entry *pamplemousse* and its direct etymon *pompelmoes*) of the data structure in a TEI like format. Note that the basic building blocks for the characterization of a lexical entry (<form> and <sense>) are completely reusable for the description of an etymon, and the language has been implemented as a standardized xml:lang attribute.

5. Perspectives: the etymological network

To summarize, the current model of etymological structures as presented here has two important characteristics: it is based on the hypothesis that the quintessence of etymological information can be expressed as a directed and acyclic graph, whose nodes and arcs are labelled with various descriptors. Furthermore, it relies on the assumption that the nodes of the graph are autonomous lexical entities, i.e. lexical entries, whose status as etymons is only a functional one. On the practical side, this assumption has allowed us to reuse basic building blocks and descriptors already defined in current initiatives on the standardization of lexical content representation (TEI, LMF), with only a small number of adjustments.

On the theoretical side, this perspective opens the way towards an etymological network, in which etymological links are seen as relations between entries recovered from one ore more external lexical databases providing access to lexical material of any language across time and space, including reconstructed forms and affixes. In this view, the physical description of etymons as subordinate to lexical entries should be understood as a provisional solution, only intended to supply the lack of external lexical databases. Note that nothing prevents the implementer from externalizing from now onwards the descriptions of etymons, if he/she wants to do so. In practice, this is what we intend to do with the information annotated automatically in the etymological notes in the TLF, especially to avoir the redundancy in the description of etymons. The idea of such an etymological network is also the main reason for us to advocate in favor of shared, reusable and ideally standardized data models for diachronic data, which still remains as close as possible to existing recommendations and practice in the field of lexicography.

References

BALDINGER K. (1959). « L'étymologie d'hier et d'aujourd'hui. » In: Etymologie. Schmitt R. (ed.), Darmstadt, 1977.

BUCHI E. (1996). „Les structures du Französisches Etymologisches Wörterbuch". *Recherches métalexicographiques et métalexicologiques*, Tübingen, Niemeyer, Beihefte zur ZRP, 268.

BUNT H., ROMARY L. (2004). "Standardization in multimodal content representation: Some methodological issues". *Fourth International Conference on Language Resources and Evaluation - LREC 2004*, 2219-222.

CORBIN D. (1987). « Morphologie dérivationnelle et structuration du lexique », 2 vol., Niemeyer, Tübingen.

FRANCOPOULO G., GEORGE M. (2005). ISO/TC 37/SC 4 N130 Rev.7. Language resource management - Lexical Markup Framework (LMF). http://www.tagmatica.fr/doc.

IDE N., ROMARY L. (2004), "International Standard for a Linguistic Annotation Framework." *International Journal on Natural Language Engineering.*

ROMARY L., SALMON-ALT S., FRANCOPOULO G. (2004) "Standards going concrete : from LMF to Morphalou". *Workshop on Electronic Dictionaries, Coling 2004*, Geneva, Switzerland.

SALMON-ALT S., ROMARY L., BUCHI E. (2005). "Modeling Diachrony in Dictionaries". *ACH-ALLC 2005*, Vancouver, Canada.

TEI P5 - Guidelines for Electronic Text Encoding and Interchange, edited by C.M. SPERBERG-MCQUEEN and Lou BURNARD. *TEI Consortium* Oxford , Providence, Charlottesville, Bergen. January 2005.

"Wörterbücher. Ein internationales Handbuch zur Lexikographie" (1990). HAUSMANN F. J., REICHMANN O., WIEGAND H. E., ZGUSTA L. (eds.). Walter de Gruyter, Berlin / New York, 1990.

Dictionaries

DWB - Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm. 16 Bde. [in 32 Teilbänden]. Leipzig: S. Hirzel 1854-1960 *: http://germa83.uni-trier.de/DWB/welcome.htm*

FEW - Wartburg, Walther von: *Französisches Etymologisches Wörterbuch*, 25 vol., Bonn, Leipzig et al. 1928–2002.

Presentation.html

MORPHALOU - Lexique morphologique ouvert du français : http://actarus.atilf.fr/morphalou/

OED - Oxford English Dictionary, 2nd edition, by John Simpson and Edmund Weiner, Clarendon Press, 1989, 20 volumes.

TLF - Trésor de la Langue Française. CNRS, Editions Gallimard,. 16 volumes : http://atilf.atilf.fr/tlf.htm