Modelling diachrony in dictionaries

Susanne Salmon-Alt, Laurent Romary, Eva Buchi

Introduction : The variety of lexical structures

Lexical data appear in a wide variety of forms. These can range from basic morpho-syntactic structures (Romary et al., 2004) intended to be used in language engineering application to important editorial projects that cover multiple levels of lexicographic description: morphological information, syntactic constructs, sense related information (definitions, examples, usage notes, etc.) or historical information. Entries can also vary in their internal organization. Among other factors, the fundamental choice between an onomasiological (concept to word) and a semasiological representation (word to sense) directly impacts on the internal structure of entries, as well as on the possible choice of descriptors attached to them. From a computational point of view, this situation prevents the design of one single data structure that would fit all the possible needs, whereas one would like to be able to have uniform access to similar information across heterogeneous lexical sources. This has been the source of strong debates, leading for instance to the ubiquitous Print Dictionary chapter of the TEI (Text Encoding Initiative) that tries to combine structured and unstructured views of lexical entries. Still, we want to show in this paper that it is possible to apply coherent modeling principles to deal with this variety of structures while providing a precise account of complex sub-components such as diachronic information as they appear in dictionaries with wide lexical coverage. Besides, we want to show that such modeling principles can guide the possible evolution of the TEI towards a more flexible data for the concrete representation of dictionaries.

Diachronic information in dictionary entries

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 – tracing out origin and primitive significance of a lexeme in its source language – and historical notes about successive changes of form and meaning once it entered into the target language. This type of information can for instance be found in the *Oxford English Dictionary* (*OED*), in the *Deutsches Wörterbuch* (*DWB*) or in the *Trésor de la Langue Française* (*TLF*), for which Figure 1 illustrates the organization of diachronic information within the micro-structure of a lexical entry ('*pamplemousse*'): here, the *Etymol. et Hist*. section, separated from the synchronic description of the lexeme, consists in two parts, the first one being dedicated to the lexeme's history within the target language (modern French) and the second one to etymology properly speaking, i.e. origin and word sense in the source language (Dutch).

Historical Notes

The main objective of the historical notes is to provide (earliest) written testimony for each of the senses – and possibly different usages of a sense – with respect to the synchronic description of the entry. Therefore, temporal information and quoted source text associated with bibliographical references play a central role in this section. Whereas the OED and the DWB realise the projection from synchronic sense organization explicitly by subordinating historical notes under sense description, the diachronic part of the TLF takes up each of the four synchronic senses by a sense identifier, a date, a quotation and bibliographical reference. The latter might be complex in case of use of secondary literature. One may also notice that differences in word spelling led to two testimonies for sense 1a. Despite of the very strict

application of the sense projection principle – which is far from being applied systematically throughout the dictionary, as mentioned for example in Hausmann et al., 1990 – one may however notice that the synchronization has not been made explicit, for example through the use of the same sense identifiers within the synchronic and diachronic sections.

Etymological Notes

The etymology section of dictionaries is concerned with the origin and development of the lexeme before entering into the target language. As a central task, it informs about one or more etymons and determines the etymological class (inheritance, loan word, word generation) for the oldest sense of the lexeme under consideration. As a consequence, it is not directly related to individual senses in the modern stage of the considered language. In the example, the etymon for the oldest sense of '*pamplemousse*' (1a), is the Dutch '*pompelmoes*', itself being a word generated via composition from '*pompel*' and '*limoes*'. Although there have been attempts to formalize further etymological notes (cf. etymological formulas, Ross 1958), they are generally not subject to well defined organisation principles, at least in current dictionaries. Additionally to core information about etymon, etymological notes may indicate bibliographical sources of the etymological hypotheses and discuss other related issues (phonetic evolution, concurrent hypotheses, confidence statements, secondary etymons, testimony of etymons, intermediate states etc.).

A representational model for etymological and diachronic information

In the following sections, we apply the main modeling principles of the LMF (Lexical Markup Framework) project within ISO committee TC 37/SC 4 to outline the structure of diachronic information in dictionary entries. Those principles (Ide & Romary, 2004) allows one to combine a meta-model, which informs the main agreed upon practices within a given field, 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 simple structure organizing a lexical entry with form related information and a hierarchy of senses) and lexical extensions, seen as additional modules attached to the core meta-model. In our case, we will consider what kind of lexical extensions are needed for both etymological and historical information.

A Lexical Extension for Etymological Structure

We propose a basic lexical extension for etymological notes (*Etymology*), i.e. a structure that accounts for the description of links to etymons. The *Etymology* component may occur at most once for a given lexical entry, under the assumption that lexical entries are purely polysemous, excluding homonyms. This etymological information is further structured by means of *Etymological Unit* and *Etymological Link* components. *Etymological Unit* components are word forms playing the role of etymons. As such, they might be characterized by any existing data category defined for the description of lexical entries, i.e. lemmata and inflected forms (*language, orthography, sense, part-of-speech, inflectional information* etc.). Two points have to be noticed. First, the coverage of *language* should be extended to more fine-grained geographical and diachronic variants as those currently available from the ISO 639 series. Second, depending on available resources, all or part of this information could be recovered by a pointing mechanism. *Etymological Link* components stand for the etymological relation between linguistics units. A link is basically characterized by an *etymological target* and an *etymological source*, i.e. pointers to external resources, including lexical entries of the current dictionary and etymological units previously described. Etymological links are typed by the *etymological class* (loan word, inheritance etc.). They may additionally bear information about the bibliographical source, confidence level or other

type of notes. The full paper will show how this data structure accounts for different types of etymological notes in current dictionaries, including cases of concurrent, popular, secondary and multiple etymons.

A Lexical Extension for Historical Notes

The modeling of historical notes can actually be seen from two complementary and somehow sequentially organized perspectives. Firstly, we have identified that historical notes are organized as a hierarchy of sense like objects, which leads to the simple historical extension depicted in Figure 3. This extension takes up the sense component that already exists in the core LMF meta-model, while further characterizing it with specific dating (/date/) and bibliographic (/bibliography/) information. Such an extension accounts for the situations where there is no *a priori* editorial coherence between the sense organization in the lexical entry and their possible counterparts in the historical notes as encountered in, e.g., the TLFi or the OED. In that case, we can see that we keep open the possibility to actuate links (/synchronic reference/) between components of the historical notes and senses in the main entry. If we want to model more controlled editorial project, we suggest to move from the previous extension to an integrated view (cf. Figure 4), which directly anchors historical descriptions on the corresponding senses. Doing so, it is always possible to externalize the corresponding information, to derive an autonomous representation conformant to Figure 3.

Implementation in the framework of the TEI

The final paper will precisely show how the two types of structures described above can be implemented using the latest version of the specification platform of the TEI (ODD — One Document Does it all; Burnard & Rahtz, 2004). In particular, we will show that, on the one hand, we can extend the scope of the existing <etym> element from the P4 guidelines, and, on the other hand, it is necessary to introduce a new element dedicated to the representation of historical notes, which mimics the behavior of related entries (sub-structure with a strong structural analogy to a full entry), combined with dating and bibliographical descriptors. Depending on the feedback we will receive from the lexicographic community, these extensions could be incorporated into the next version (P5) of the TEI guidelines.

References

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

Burnard L., Rahtz S. (2004). Relaxing with Son of ODD, or What the TEI did Next. Extreme Markup Languages, Montréal (Canada) , 2-6 August 2004.

Ide N., Romary L. (2004), International Standard for a Linguistic Annotation Framework. *International Journal on Natural Language Engineering*, forthcoming.

Romary L., Salmon-Alt S., Francopoulo G. (2004). Standards going concrete: from LMF to Morphalou, Coling Workshop on Enhancing and Using Electronic Dictionaries, Geneva, 29 août 2004.

Wörterbücher. Ein internationales Handbuch zur Lexikographie. Hausmann F. J., Reichmann O., Wiegand H. E., Zgusta L. (eds.). Walter de Gruyter, Berlin / New York, 1990.