Decoding Lexical Resources

Susanna Allés Torrent
Milà i Fontanals Institution - CSIC, Spain;

This paper outlines a proposal for a modelling of lexical structures, in Latin medieval dictionaries, and points out several problems in the encoding of printed dictionaries.

The Guidelines of the Text Encoding Initiative defines indeed a chapter for encoding lexical resources entitled “Dictionaries”. However, as it has been underlined, print dictionaries are extremely complex typographically and structurally, mostly because of two main issues, that is to say, the diversity of the structure of dictionary entries and the need to declare the implicit and highly compressed information (such as the use of numerous abbreviations). One of the mark-up languages most used to encode this kind of text has been the use of XML standards and in fact, as Budin, Majewski and Mörth have recently pointed out, “in modeling lexicographic data, it has become common practice to conceptualize the underlying structures as tree-like constructs, which makes XML an ideal syntax for expressing the data” (2012).

The process for encoding lexical structure is intricate and this paper wants to underscore some case studies, which will contribute to the semantics of the TEI conceptual model and also semantics of lexical structures and lexical data. We will present examples of encoding dictionary entries in order to illustrate our concepts. The cases will be taken from Latin medieval dictionaries, a typology of lexicographical works that present a very rich variety of problems, mainly because they belong to a lexicographical tradition that dates back to nineteenth century – even though many effort have been taken to systematize their drafting method-. More precisely, our encoding will be focused on the work carried out with the Glossarium Mediae Latinitatis Cataloniae (GMLC), a dictionary of the Latin and Romance terms documented in the lands corresponding to the Catalan linguistic domain between the ninth and twelfth centuries.

Our contribution starts from a deep analyse of the primary source, the idiosyncrasy of its methodology and drafting, the entries structure, the nature of the lexical information and the distribution of the information inside the articles. Moreover, we also take into account the author’s intention and the purpose of the specific layout. In fact, the relevance of the typographical design and typographical features of the dictionary have been widely highlighted. The line between typography features and semantics is not always clear, because the first one is most of the times a representation of the second ones. In order to organise our paper, we differentiate three basic elements that concern lemmatisation, meanings and citations. First of all, we will focus on the model used by us to encode the general entry

structure and its constituents parts, in order to determine the level of granularity that convey to its structure. The following will be a detailed explanation of the casuistry of lemmas and their encoding: the treatment between principal lemmas and variants, that is to say, the different forms associated with a given lexical entry. Other questions will be treated more generally, such as the quantity of grammatical and morphological information to be given in order to differentiate data details, which will be really useful for further processing and will permit interchange between heterogenic lexical resources.

As far as the meaning is concerned, central to the entire encoding model is the concept of the “sense” element, in our case also understood as “as a generic container organizing the further description of a signifier" (Romary – Wegstein, 2012). Thus, and always following the TEI Guidelines, a particular way to conceive the granularity of the sense element will be proposed. Some related aspects concerning definitions and translations in multiple languages will be also called into question.

Another of the main questions addressed in this paper will be the link between electronic dictionaries and corpora through the citations, quotes and bibliographic information. We understand the encoding of the GMLC dictionary within a framework of interconnected digital resources. Thus, we have conceived a specific method of retrieval between our dictionary and a corpus of Latin medieval texts, called CODOLCAT, a lexical database of texts written in Latin in Catalonia in the early mediaeval period (IX – XII centuries). This corpus is permanently enriched and updated and it represent the raw material for the compilation of the GMLC, besides the fact that it is a powerful instrument to browse primary sources written in Latin.

We are convinced that, once again, the efforts devoted to the use of mark-up standards will ensure the interoperability and durability of digitized texts and will facilitate exchanges between different resources. We believe it is worthwhile to dedicate efforts to make the lexical data TEI P5 compliant in order to insert this material into an interoperability context between heterogeneous sources and to optimize the data processing and to facilitate the re-usability in electronic form.

Bibliography

- BUDIN, G., MAJEWSKI, S., MÖRTH, K., «Creating Lexical Resources in TEI P5. A Schema for Multi-purpose Digital Dictionaries», Jounal of the Text Encoding Initiative , Issue 3, November 2012 : TEI and Linguistics.
http://www.chass.utoronto.ca/epc/chwp/merrily2/
- GLORIEUX, F., «Pourquoi informatiser un vieux glossaire. Présentation du Du Cange en linge», Ela. Études de linguistique appliquée, 2009/4 (no 156), pp. 417-415.

URL: http://ducange.enc.sorbonne.fr/doc/Glorieux2010_ela.pdf
- GLORIEUX, F. – THUILLER, S., «Grec ancien, latin médiéval, balisage comparé de deux dictionnaires, vers des ressources linguistiques», ALMA 68, 2010, pp. 161-181.
URL: http://ducange.enc.sorbonne.fr/doc/DC-DGE-2010.pdf
- IDE, N - VERONIS J., «Codage TEI des dictionnaires électroniques », Cahiers GUTenberg, 24, Rennes, 1996, pp.170-176.

URL : http://cahiers.gutenberg.eu.org/cg-bin/article/CG_1996___24_170_0.pdf
MERRILEES, B., The Shape of the Medieval Dictionary Entry, Toronto, Editors of CHWP, 1996.

URL: http://journals.sfu.ca/chwp/index.php/chwp/article/view/B.15/94
- ROMARY, Laurent – WEGSTEIN, Werner, «Consistent Modeling of Heterogeneous Lexical Structures», Journal of the Text Encoding Initiative, Issue 3, November 2012 URL : http://jtei.revues.org/540 ; DOI : 10.4000/jtei.540
- TUTIN, A.- VÉRONIS, J., «Electronic Dictionary Encoding: Customizing the TEI Guidelines», Eighth Euralex International Congress (EURALEX'98), Liège, 1998, p.4- 8.
http://www.up.univ-mrs.fr/veronis/pdf/1998euralex.pdf
- VALETTE, Mathieu et al., «Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens. », Verbum ex machina, Actes de la 13ème conférence sur le traitement automatique des langues

naturelles (TALN 06), 2006.

URL: http://www.revue-texto.net/Corpus/Publications/Valette_Estacio.pdf