Lexicon standards: from de facto standard Toolbox MDF to ISO standard LMF

Jacquelijn Ringersma[1], Sebastian Drude[2] and Marc Kemps-Snijders[1]
[1]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2] Goethe-Universität, Frankfurt, Germany

Abstract
This paper discusses possible solutions for the apparent incompatibility between two standards for lexicon structure and concept naming: the de facto standard MDF, which is part of the widely used lexicon application Toolbox [1] and the newly accepted ISO standard LMF, ISO FDIS 24613:2008 [2], implemented in the online lexicon tool LEXUS [3]. The basic difference between the two standards is that in MDF, the form-related and meaning-related parts of lexical entries are embedded in each other, while in LMF there is a strict separation of the two parts. The difference might be related to the final medium for which the standards have been created; although Toolbox is a tool for digital lexicon creation, the MDF format was created for printed dictionaries, whereas LMF is created for digital presentation of lexicon resources. At first sight the difference seems to be fundamental and impossible to overcome. However, in this paper we would like show possible solutions, and would like to probe them in the LREC2010 workshop on Language Resource and Language Technology Standards, and thoroughly discuss them amongst a wide linguistic public, before implementing a conversion procedure in the Toolbox import module of the LEXUS tool.

Multi-Dictionary Formatter (MDF)

In linguistic field work on minor languages, Toolbox is a widely used data management and analysis tool, designed for maintaining lexicons and for parsing and interlinearizing of text. Toolbox is text-oriented [1]. A lexicon structure is defined as a set of rules which declares the lexicon structure elements (markers), their value domains and their hierarchy. Toolbox delivers a default structure definition file for dictionary formatting: the Multi-Dictionary Formatter (MDF). Lexical entries content can be built following the MDF structure, The hierarchy, however, is not explicitly represented in Toolbox databases (which are in the 'Standard Format', a flat list of feature-value pairs). MDF structures facilitate not only the creation of digital lexicons, but also structured and formatted output of the Toolbox lexicon in a rich text format, which can be imported into Microsoft Word. MDF has become a de facto standard for lexicon structures in field linguistics.

In the MDF hierarchy, there are three main primary markers: `lexeme (\lx)`, `part of speech (\ps)` and `sense number (\sn)`. Lexicon structures are either 'part of speech'-oriented or 'sense'-oriented, and users of the tool are free to choose (in recent versions of Toolbox, the 'part of speech'-orientation is the default). In 'part of speech'-oriented structures, `\lx` is superordinate to `\ps` and `\ps` to `\sn`. One `\lx` can have multiple `\ps` markers and likewise one `\ps` can have multiple `\sn` markers. In 'sense'-oriented structures the hierarchy is `\lx>\sn>\ps`. In both orientations, `\lx` is also superordinate to a set of markers which apply to the lexical entry as a whole, e.g homonym number (`\hm`) or variant form (`\va`). In addition, `\sn` can be followed by a flat set of markers, like english gloss (`\ge`), vernacular gloss (`\gv`), english definition (`\de`) etc. Sense number can also be followed by structured sets of markers, for instance those for example sentences (`\xv\xe\xn`). MDF accommodates sub-entries (`\se`); these are integrated elements of lexical entries, subordinate to `\lx`, and the same hierarchy that applies to a full lexeme entry can also apply to a sub-entry.

Lexical Markup Framework (LMF)

The Lexical Markup Framework model (LMF) [2] was recently accepted as ISO standard for Natural Language Processing lexicons and Machine Readable Dictionaries (ISO-FDIS-24613:2008). LMF prescribes a basic model for lexicon structures elements ('data categories'), and a registry for data category naming and value domains. LMF also defines the constraints on the relations between the data categories. The main goal of LMF is to enhance true content interoperability between all aspects of lexical resources; in specific data exchange between resources, searching across and merging of the resources.

In LMF, the structure of a Lexical Entry consists of three basis components: `Lemma`, `Form` and `Sense`. `Lemma` is the conventional form chosen to represent a lexeme. `Form` manages the

orthographical variants of a lexical entry, as well as any other data category that represents the attributes of the word form (e.g. `writtenForm, inflections`). `Sense` represents one meaning of a lexical entry, with attributes like `definition` or `gloss`. `Part of speech` is considered to be neither form nor sense; in LMF, part of speech is an attribute of the Lexical Entry.

LEXUS

LEXUS [3] implements an instantiation of LMF. It is the online lexicon tool of the Language Archiving Technology suite (LAT) developed by the Max Planck Institute for Psycholinguistics (MPI) in The Netherlands. With LEXUS, users may create, manipulate and visualize lexicons and enrich the lexical entries with multimedia fragments. The default lexicon structure in LEXUS for new lexica is based on LMF. LEXUS offers the ISOcat data category registry for data category naming and value domain specifications (ISOcat is the ISO implementation of the ISO 12620:2009 standard and offers standard linguistic concepts to be used in linguistic resources). LEXUS has been operational since 2007 and currently has about 450 registered users, of which some 20 are active. The active users have developed around 60 lexica. Most of these lexica were imported in LEXUS, initially created in XML or Toolbox. For both formats LEXUS provides an import and export facility. However, in LEXUS it is possible to avoid the LMF structure; this means that when Toolbox lexica are imported into LEXUS it is possible to maintain the structure defined in the Toolbox typ file in the LEXUS lexicon.

From Toolbox MDF to LEXUS LMF

A first difference between MDF and LMF is in the naming of the concepts. However, in a recently created working group of the RELISH project [4] on lexicon standard interoperability, it was proposed to add the MDF markers to the ISOcat data category registry. MDF is thereby acknowledged as an important de facto standard in lexicography, and its data elements can be related to data elements used elsewhere.

One principal difference between MDF and LMF structures is that MDF does allow for sub-entries, whereas LMF does not. Since LMF is created for digital formatting of lexicons, this gap seems to be easy to overcome: for every sub-entry within a Toolbox lexeme, create a new Lexical Entry in LMF and attribute it with a cross reference and pointer to the Lexical Entry of the lexeme (and vice versa). Lexicographers might argue that the status of the two Lexical Entries is not equal, but also this difference can be covered with an attribute at the Lexical Entry level.

For ‘part of speech’ oriented MDF structures, the conversion from MDF to LMF is not too problematic. Lexical Entries in LMF can have multiple senses, so for each group of markers under `\ps\sn`, there will be a separate Sense container in the LMF structure. In case `\lx` contains multiple `\ps`, the option is again the creation of multiple Lexical Entries for each `\ps` block, possibly with several sense blocks within, with cross reference attributes.

For ‘sense number’ oriented MDF structures the situation is more complicated, since in this orientation, one `\sn` can have more than one `\ps`. But again the gap can be bridged by splitting the Toolbox lexical entries in multiple LMF Lexical Entry’s. An algorithm for this will not be too hard to define, but it is not trivial to define the multiple cross referencing attributes which indicate the relations among the different entries.

In our paper we will describe the possible conversion from MDF to LMF, on the basis of examples taken from the Marquesan lexicon Dico général - tekao tapapatīa [5] and the Iwaidja lexicon [6]. These lexica were initially created in Toolbox, with an MDF structure. We will discuss lexical entries with and without sub-entries and we will discuss both the part of speech and the sense number orientation. We will make a qualitative description of the required algorithms. On the basis of the examples we show how the conversion from MDF to LMF could be realized in the future import module of LEXUS.

The MDF format is not a suitable format for interoperability because it is based on a textual database system and because it is very prone to inconsistencies. Since the trend is that more resources will become digital, the need for interoperability will increase. However, when LMF is to become one new major standard for lexicon structures, it is important to suggest to the research community which concerns exist when converting MDF to LMF. Our paper is meant to initiate the discussion.

References

[1] Toolbox, http://www.sil.org/computing/toolbox/ Visited on January 13, 2010

[2] ISO technical committee ISO/TC37 (2008) Language resource management - Lexical Markup Framework (LMF) http://www.lexicalmarkupframework.org/ Visited on January 13, 2010
[3] Ringersma, J., & Kemps-Snijders, M. (2007). Creating multimedia dictionaries of endangered languages using LEXUS. In H. van Hamme, & R. van Son (Eds.), Proceedings of Interspeech 2007 (pp. 65-68). Baixas, France: ISCA-Int.Speech Communication Assoc.
[4] RELISH project: http://www.neh.gov/ODH/Default.aspx?tabid=111&id=117 Visited on February 1, 2010
[5] Cablitz, G. (2007-2009) Marquesan lexicon Dico général - tekao tapapatīa, published in the online lexicon tool LEXUS (http://corpus1.mpi.nl/mpi/lexusDojo). Accessible after permission from the owner.
[6] Birch, B. and others (2006-2009) Iwaidja lexicon, published in the online lexicon tool LEXUS (http://corpus1.mpi.nl/mpi/lexusDojo). Accessible after permission from the owner.