Lexicon Formats

This guide describes a number of lexicon formats that are in use, the problems that are associated with them and where users have to take care of them.

General Framework:

- Lexicon creation is one of the most important tasks that linguists face when documenting a language or processing a text document. Lexica come in many different formats and encodings which make it often very difficult to access them and to integrate them in interoperability scenarios.
- Therefore it is seen as important to discuss the properties of a number of well-known formats to make users aware of choices.

1. General Remarks

When creating or manipulating a lexicon exclusively with one specific tool, the only question that used to be relevant for linguists was whether the tool suited the needs. This traditional view has changed since people now want to re-use lexica, compare lexica of different languages, merge and/or link lexica or combine them with other types of documents and maintain accessibility for many years. These requirements make it necessary to store the lexicon content in open formats and to take care that the encodings are standardized or documented. Traditionally the lexicon formats were not relevant to the users – only the tools were evaluated. This view has changed completely since new tools with more functionality have become available. The format is now the essential anchor for capturing all the work of linguists. Therefore, linguists should be aware of the format issue.

The table below may give an impression about some of the major formats that are used.
It indicates some relevant criteria that users should be aware of when deciding about a format: (1) is it a widely accepted standard or best practice; (2) does it supported structuring, hierarchies and cross-references; (3) is it flexible to adapt to the user needs and does it support changes; (4) does it allow to check for validity; (5) does it support an open and human readable format; (6) does it support UNICODE; (7) is it supported by good tools and will it be supported in future by good tools;

Type	Major Properties
Shoebox	plain text; human readable; nicely structured by tags; flexible structure; hierarchies are possible; no validation possible; arbitrary character encoding; supports cross-references; no media linking;
Toolbox	same as Shoebox; but support for UNICODE and XML;
CHAT	plain text; human readable; flat structure; limited expressional power; special symbols used to encode properties; no explicit validation; media linking possible; UNICODE support;
WORD	document format is proprietary and not human readable, no tag structure, no consistency control; mixture of visualization and representation aspects; RTF export possible but not easy to process; XML output is of limited use; no reference concept; no media linking;
EXCEL	flat table structure, field labels can be seen as tags; only simple lexica; xls format is proprietary; mixture of visualization and representation aspects; no reference concept;
Relational Databases	requires a logical design for related tables; flexible structure and hierarchies possible; validation of content by database mechanism; content can be transferred to XML in different ways; mostly no schema after XML output – no validation; character encoding often unclear after export; no reference concept; often fixed setups;
XML	plain text; human readable; standard for structuring documents; UNICODE as character encoding is strongly recommended; any tree structures are possible; to define a specific format an XML schema is necessary, with schema a validation is possible; a mechanism for expressing typed references is available;
LMF	Lexical Markup Framework is a flexible XML-schema based standard for lexica; the underlying idea is to allow constructing lexica like working with LEGO bricks but nevertheless to work within one family of structures;

3. Résumé and Recommendations

Based on what was said we can make the following recommendations:

- Only tools that support UNICODE should be chosen as character encoding problems are in general difficult to solve, i.e. the conversion is time and money consuming.
- Only tools that generate clearly structured lexical documents with explicit tags should be used, in particular if they support constraints and controlled vocabularies.
- XML schema based formats are optimal in many dimensions since they allow others to parse the structure, to build tools and to check validity.
- Even more optimal are such XML formats that are based on abstract lexicon models such as LMF.

In this sense, and due to its flexibility and useful linguistic functionality, Shoebox is still one of the best programs around. In particular, the new version of Toolbox that supports UNICODE and XML is an excellent tool.

We expect a set of new tools that are based on the latest flexible ISO LMF format that allow for the new types of functionality indicated above. Importantly, such tools can import and export Shoebox/Toolbox files, for example the LEXUS tool from MPI is such a tool.

Tools based on relational databases have their great advantage in the ease with which users can create complex table structures and simple user interfaces. However, they encapsulate all data, i.e. if the tool changes or if it is not available anymore it will be difficult to extract the content. Also it is the experience in many cases that it is not possible to generate correct XML formats from the database contents, i.e. again expensive conversion is necessary. There are many lexica created with the help of relational database systems and it is likely that many of them will not be accessible indefinitely, so we cannot recommend the use of such tools for all tasks.

This document is not intended to give a comprehensive overview about tools.