Lexicon Schemas and Related Data Models: when Standards Meet Users

Thorsten Trippel, Michael Maxwell, Greville Corbett, Cambell Prince, Christopher Manning, Stephen Grimes, and Steve Moran

Universität Bielefeld, Germany; University of Maryland, USA; University of Surrey, United Kingdom; SIL, Thailand; Stanford University, USA; Indiana University, USA; University of Washington, USA

Looking at conventions from the customers perspective

- Fieldworkers
  - Lexicographers
  - Anthropologists
  - Language documentation
  - Economically less relevant languages (Long Tail?)
- Tool providers
  - Supporters for fieldworkers
  - Lexicon creation based on (fieldwork-) corpora
  - Language documentation

Problem

- Extremely “expensive” data
  - Scarce funding
  - Few institutional sponsors
  - Academic domain
- Preservation of data
  - No data centres as driving force
  - Academic interest for typology and language change
  - Long term portability required

Standardization for “them”

- Standardization in the researchers interest
  - (Re-)use of tools
  - Sharing interpretable data
- Use of standard
  - Provided by a usable tool
  - Will use recommendations
  - Require maximum of flexibility
- Reading standards low priority
- Easy and useful and easy to encode MY lexicon

What people use

- Text processors:
  - Schema implied in layout
  - Portability, compatibility, adequacy issues
- Spreadsheets:
  - Columns imply lexical data categories
  - Easy to convert for RDBMS
  - Exchange with other applications
- Shoeboxes
  - File cards
  - Boxes

More linguistically motivated tools

- Praat, Wavesurfer, Transcriber:
  - Motivated by signal processing
  - If recordings are available
  - No standardized data format
  - Interoperability with other tools for lexicon building?
- Shoebox, Toolbox, Fieldworks, LAMUS, ...
  - “orthographic” transcription with lexicon support
  - IGT for lexicon

Data formats used by linguists

- CSV: Character separated value - Spreadsheet
- DATR: character based for inference purposes
- LIFT: XML format for SIL-Tool compatibility
- FSR: Feature Structure Representation
- TEI: XML dictionary encoding/
- LMF: Lexical Markup Framework

Conversion table

From ↓ to →	DATR	FSR	LMF	LIFT	TEI	CSV
DATR	-	Naming conventions of DATR theories used for hierarchies; inheritance structures can be expressed	Depends on the used data categories; LMF requires at least one form property.	Depends on the use of data categories; some data categories predefined in LIFT; tag misuse possible	Representation of fields possible; inheritance rules labelled as some kind of grammatical rules, danger of tag abuse	Full form lexicon: see DATR to FSR comment; for inheritance lexicons: similar but inheritance not explicit
FSR	Types of feature structures refer to inheritance, feature names to data categories; lossless	-	Depends on the used data categories; LMF requires at least one form property.	See DATR to LIFT comment	See DATR to TEI comment	Each data category is one column; multiple occurrences of same data category requires repetition of
LMF	Hierarchy of data categories representable in DATR category names; else simple	Hierarchy of data categories representable in FSR hierarchy names; else simple	-	Depends on the concrete implementation of LMF; examples in LMF are subject to the same problems as DATR conversion into LIFT	Depends on the concrete implementation of LMF; examples in LMF are subject to the same problems as DATR conversion into TEI	see LMF to FSR comment
LIFT	See FSR to DATR comment	See LMF to FSR comment	LIFT can be seen as one implementation of LMF	-	Different data category hierarchies	see LMF to FSR comment
TEI	Hierarchy of data categories representable in DATR category names or by abstract entries; lossless	See LMF to FSR comment	TEI can be seen as one implementation of LMF	Different data category hierarchies	-	see LMF to FSR comment
CSV	Each column is one data category in DATR, inheritance of DATR not used; lossless	Simple binary structure, lossless	See DATR to LMF comment	See DATR to LIFT comment	See DATR to TEI comment	-

Result

- Modulo data categories: all schemas implementations of LMF (!)
- LMF: Framework only
- Prerequisits for interchange
  - Mapping of data categories
  - Format conversion
- Interchange results in loss of implied information
- Tools lack support for interchange

Summary

- Looked at different lexicon schemas used
- Tried to evaluate interchange between them
- Recommendation for the fieldworker
  - Work with a data center
  - Use one of the specialized tools
- Recommendation for the tool provider
  - Implement the standards
  - Provide export to other formats
- Recommendation for standardizers
  - Provide modules/plugins/instances of LMF
  - Include the tool providers in standadization

Thank you

... and also we acknowledge

- Organizers of the workshop "Toward the Interoperability of Language Resources" organized at Stanford University in July 2007
- Sponsors
  - National Science Foundation (USA),
  - University of Kansas, USA
  - Eastern Michigan University, USA
  - The Linguist List

Trippel et al.: Lexicon schemas: standards and users meet