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

Jacqueline Ringersma
Sebastian Drude
Marc Kemps-Snijders
Lexicon standards:

MDF – Toolbox

LMF – LEXUS

Concept naming

From MDF to LMF

some issues

some examples
Problem introduction

Toolbox (MDF)
- Widely used by (field) linguists
- Freedom for user to rename and restructure
- Form and Sense related are embedded in each other
- De-facto standard

Lexical Mark-up Framework (LMF)
- ISO standard for NLP lexicons and MR dictionaries (24613:2008)
- Basic model for lexicon structures
- Strict Form and Sense separation
- Registry for concept naming
Multi Dictionary Formatter (MDF):

Model for standardized framework for the construction of lexicons

**Structure** is defined as a set of rules declaring
The naming of the element
The hierarchy between the elements
The value domains

**Content** is build on the structure,
but not (explicitly) present in the database
MDF

\+mkr ps
\nam Part of speech
\desc Classifies the part of speech. This must reflect the part of speech of the vernacular lexeme (not the national or English gloss). Consistent labeling is important; use the Range Set feature. Sense numbers are beneath \ps in this hierarchy; don't mark different \ps fields with sense numbers.
\lng English
\rngset adj adv n num v
\mkrOverThis se
\mkrFollowingThis va
\CharStyle
\-mkr

\+mkr sn
\nam Sense number
\desc Where a lexeme has more than one sense, this code is used to mark and number mark the beginning of each section that discusses a new sense. Don't use a sense number to mark a different part of speech; \sn is only used within a given part of speech (in this hierarchy). (Remember to include \sn 1 for records with multiple senses.) Use a Character Range Set.
\lng Default
\mkrOverThis ps
\CharStyle
\-mkr
\lx ай
\ph aj
\ps interj
\pr между
\sn 1
\gn1 aj, ej
\gn1lat aj, ej
\lr гэй, эй
\gn2 ʒjo, jo
\ge hey
\lxv Ай хинар, самағай еке
\lxv-ph Aj xinar, samağaj eke
\lxve V
\ln1 Aj ғыз, бир бура ҝəл
\ln1lat Aj qız, bir bura ɡəl
\lr эй, девушка, подойди-ка сюда
\lxе hey girl, come here!
\sn 2
\gn1 aj, oj, baj
\gn1lat aj, oj, vaj
\ge woe!!!!
\lxv Ай! Без мурелин чахп̣и
\lxv-ph Aj! Bez murelin čaxṛi
\ln1 Aj! Акажымы тапдаладын
\ln1lat Aj! Ajağımı tapdaladin
\lr Наступили мне на ноги
\lxе Ouch! They stepped on my feet!!!!
\dt 04/Jan/2010
Main elements and order:
\lx lexeme
 . \ps part of speech
 . . \sn sense number
 . . . \gloss and definition markers
 . . . \example sentence markers
 . \se subentry
 . . \ps part of speech
 . . \sn sense number
 . . . \gloss and definition markers
 . . . \example sentence markers

Alternative:
\lx lexeme
 . \sn sense number
 . . \ps part of speech
 . . . \gloss and definition markers
 . . . \example sentence markers
 . \se subentry
 . . \sn sense number
 . . \ps part of speech
 . . . \gloss and definition markers
 . . . \example sentence markers
Example ps orientation (Udi):
\lx ай
\ph ай
\ps interj
\pr межд
\sn 1
\gn1 ай, ей
\gn1lat ай, ей
\gr гэй, эй
\gn2 ӡо, ӡо
\ge hey
\lx-v Ай хинар, самагъай еке
\lxv-пх Ай xinar, samağaj eke
\lxve V
\lxn1 Ай гыз, бир бура көл
\lxn1lat Ай qız, bir bura qəl
\lxr эй, девушка, подойди-ка сюда
\lxе hey girl, come here!
\sn 2
\gn1 ай, ой, вай
\gn1lat ай, ой, вай
\ge woe!!!!
\lxv Ай! Без мурелин чахпˌи
\lxv-пх Ай! Bez murelin çaxrˌi
\lxn1 Ай! Аягымы тапдаладын
\lxn1lat Ай! Ajağımı тапdaladın
\lx Наступили мне на ноги
\lxе Ouch! They stepped on my feet!!!!
\dt 04/Jan/2010

Example sn orientation (Iwaidja):
\lx alabanja
\sn 1
\ps n
\de beach hibiscus. Rope for harpoons and tying up canoes is made from this tree species, and the timber is used to make \fv{larrwa} smoking pipes
\ge hibiscus
\re hibiscus, beach
\rfs 205,410; IE 84
\sd plant
\sd material
\rf lwa05.Feb2
\lxv alabanja alhurdu
\lxе hibiscus string/rope
\sn 2
\ps n
\de short-finned batfish
\ge short-finned batfish
\re batfish, short-finned
\sc Zabidius novaemaculatus
\sd animal
\sd fish
\rf lwa05.Fish Names.xls
\ls so MELP project elicitation
\ls eb SH
\dt 19/Dec/2006
Example sub-entry (Udi):

\[x\] биъбест\[gesun\]
\[ph\] bibes\[tesun\]
\[ns\] V
\[va\] биъгиъбесту\[esun\]
\[va-ph\] bi\[gjibestuesun\]
\[ve\] N
\[ps\] v
\[pr\] г
\[gn1\] ағыр етдирмəк, ағырлашдырмəг
\[gn1lat\] a\[gir\] etdirmək, a\[girla\]şdirmaq
\[gr\] просить (заставить) делать тяжелым, увесистым
\[gn2\] ə\[ggə\]dədəgəş
\[ge\] cause to become heavy, loaded!!!
\[se\] биъгъыъбесун
\[se-ph\] bə\[ghəbesun\]
\[se-ve\] N
\[gn1\] ағыр елəмəк (чəкидə)
\[gn1lat\] a\[gir\] elmək (ça\[kidə]
\[gr\] делать тяжелым, увесистым
\[ge\] make heavy, load
\[dt\] 05/Mar/2010
Lexical Markup Framework:
Model for standardized framework for the construction of lexicons

Goals:
Common model for electronic lexical resources
Manage and exchange data between resources
Enable merging of electronic resources
Core package:
Structure skeleton for a database
Basic hierarchy of a lexicon, and a lexical entry

Extensions:
Proposed lexicon structures for different situations
**LexicalEntry**: container for managing one or several forms and possibly one or several meanings in order to describe a lexeme

**Lexeme**: abstract unit, generally associated with a set of forms sharing a common meaning

**Form**: text string representing the word

**Sense**: specifies the meaning and context
No sub-entries, but relations between LE

- Lemma: writtenForm = "oak tree"
- PartOfSpeech = "noun"

- Sense: id = "oak_tree0"

- Synset: id = "SS12100057"
  - Label = "substanceHolonym"

- Definition: text = "a deciduous tree of the genus Quercus"

- Statement: text = "has acorns and lobed leaves"

- Lemma: writtenForm = "oak"
- PartOfSpeech = "noun"

- Sense: id = "oak0"

- Synset Relation: label = "substanceHolonym"

- Definition: text = "the hard durable wood of any oak"

- Statement: text = "used especially for furniture and flooring"

- Lemma: writtenForm = "oak"
- PartOfSpeech = "noun"

- Sense: id = "oak2"

- Synset: id = "SS12100739"

- Definition: text = "great oaks grow from little acorns"
LMF in LEXUS

LMF default structure in LEXUS – Form, Sense

Slightly different from LMF:

Data components

Data categories

No attributes
New structure for TKL in NGT:

- lexicon
  - lexiconInformation
  - lexicalEntry
    - Number
    - form
      - English
      - Dutch
      - Sign
    - sense
      - Meaning
  - form
  - sense
  - (Written) Form
  - definition
Concept naming
In MDF:

A given set of markers:
Named
Defined
Value range

Users are free to add new markers:
Without naming them
Without defining them
Without a value range
Concept naming

In LMF

ISO 12620:2009

Terminology and language resources

Specification of data categories and management of a Data Category Registry for language resources
Data category

The result of the specification of a given data field

*A data category is an elementary descriptor in a linguistic structure or an annotation scheme.*

Model consists of 3 main parts:

*Administrative part: Administration and identification*

*Descriptive part: Documentation in various working languages*

*Linguistic part: Conceptual domain(s for various object languages)*
Data Category Registry: ISOcat

A free service: anyone can access it or register as an expert and create/share his/her own data categories.

Data categories can be submitted to the standardization process, in which case they are assigned to a Thematic Domain Group which judges it.

At regular intervals, snapshots of the standardized subset of the DCR will be submitted to ISO.

www.isocat.org
LMF default structure in LEXUS – Form, Sense

Slightly different from LMF:
- Data components
- Data categories
- No attributes
Concept naming
Add MDF to ISOCat (RELISH project)
Sub-entries

MDF does allow for sub-entries
LMF does not allow for sub-entries, but does allow for relations

Example sub-entry (Udi):
\lx биъбест\jesun
\ps v
\gn1 аğıр етдирмəк, аğıрлашдырмəг
\se биъгъыъбесун
\dt 05/Mar/2010

Lemma: биъбест\jesun
Lex.Entry
PS: V

RelationType: sub_entry

Sense: аğıр етдирмəк

Lemma: биъгъыъбесун
Lex.Entry

Sense: аğıр елəмəк (чəкидə)
Part-of-speech orientation in MDF:

Entry with one part-of-speech:
No problem in conversion to LMF,
Each separate \\sn set of markers will be a separate Sense Group in LMF

Entry with multiple part-of-speech groups:
Create more than one LMF lexical entry, with relation type homonym
Sense orientation in MDF:

Entry with one part-of-speech under a sense number group:
Each sense number group results in one LMF lexical entry, with relation type homonym
Finally

Proposal MDF2LMF (individual markers):

<table>
<thead>
<tr>
<th>Lemma:</th>
<th>Lexical Entry:</th>
<th>Form:</th>
<th>Sense:</th>
</tr>
</thead>
<tbody>
<tr>
<td>\lexeme</td>
<td>\part-of-speech</td>
<td>\citation</td>
<td>\gloss</td>
</tr>
<tr>
<td></td>
<td>\METADATA</td>
<td>\alternative</td>
<td>\definition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>\underlying</td>
<td>\semantic-domain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>\phonetic</td>
<td>\literally</td>
</tr>
<tr>
<td></td>
<td></td>
<td>\variant</td>
<td>\reversal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>\MORPHOLOGY</td>
<td>\EXAMPLES</td>
</tr>
<tr>
<td></td>
<td></td>
<td>\GRAMMAR</td>
<td>\ENCYCLOPEDIA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>\picture</td>
<td>\SCIENTIFIC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>\video</td>
<td>\ETYMOLOGY</td>
</tr>
</tbody>
</table>
Questions?