



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

Jacqueline Ringersma
Sebastian Drude
Marc Kemps-Snijders



From MDF2LMF

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



MDF

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 ај
\ps interj
\pr межд
\sn 1
\gn1 ај, еј
\gn1lat ај, еј
\gr гэй, эй
\gn2 эго, го
\ge hey
\lv Ай хинар, самагај еке
\lv-ph Ај хинар, самагај еке
\lve V
\lxn1 Ај гыз, бир бура кэл
\lxn1lat Ај қыз, бир бура гөл
\lxr эй, девушка, подойди-ка сюда
\lxе hey girl, come here!
\sn 2
\gn1 ај, ој, вай
\gn1lat ај, ој, вай
\ge woe!!!!
\lv Ай! Без мурелин чахпИи
\lv-ph Ај! Bez murelin џахрї
\lxn1 Ај! Ајағымы тапдаладын
\lxn1lat Ај! Ајағımı tapdaladın
\lxr Наступили мне на ноги
\lxе Ouch! They stepped on my feet!!!
\dt 04/Jan/2010
```



MDF

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



MDF

Example ps orientation (Udi):

```
\lx ай
\ph аж
\ps interj
\pr межд
\sn 1
\gn1 аж, еж
\gn1lat аж, еж
\gr гэй, эй
\gn2 Эго, го
\ge hey
\lv Ай хинар, самагъай еке
\lv-ph Aj xinar, samağaj eke
\lve V
\lxn1 Aj гыз, бир бура кэл
\lxn1lat Aj qız, bir bura gəl
\lxr эй, девушка, подойди-ка сюда
\lxе hey girl, come here!
\sn 2
\gn1 аж, ој, вай
\gn1lat аж, ој, вай
\ge woe!!!!
\lv Ай! Без мурелин чакпІи
\lv-ph Aj! Bez murelin čaxpī
\lxn1 Aj! Ајағымы тапдаладын
\lxn1lat Aj! Ajağımı tapdaladın
\lxr 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
\lfs 205,410; IE 84
\sd plant
\sd material
\lrf Iwa05.Feb2
\lv 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
\lrf Iwaidja Fish Names.xls
\so MELP project elicitation
\eb SH
\dt 19/Dec/2006
```



MDF

Example sub-entry (Udi):

\lx биъбестIесун

\ph bibeſtesun

\ns V

\va биъгъиъбестуесун

\va-ph biġibestuesun

\ve N

\ps v

\pr r

\gn1 ағыр етдирмәк, ағырлашдырмаг

\gn1lat aġır etdirmäk, aġırlaşdırmaq

\gr просить (заставить) делать тяжелым,
увесистым

\gn2 ƆaſmƆoſġba

\ge cause to become heavy, loaded!!!

\se быъгъыъбесун

\se-ph bæġəbesun

\se-ve N

\gn1 ағыр еләмәк (чәкидә)

\gn1lat aġır eləmäk (çäkidə)

\gr делать тяжелым, увесистым

\ge make heavy, load

\dt 05/Mar/2010



LMF

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



LMF

Core package:

Structure skeleton for a database

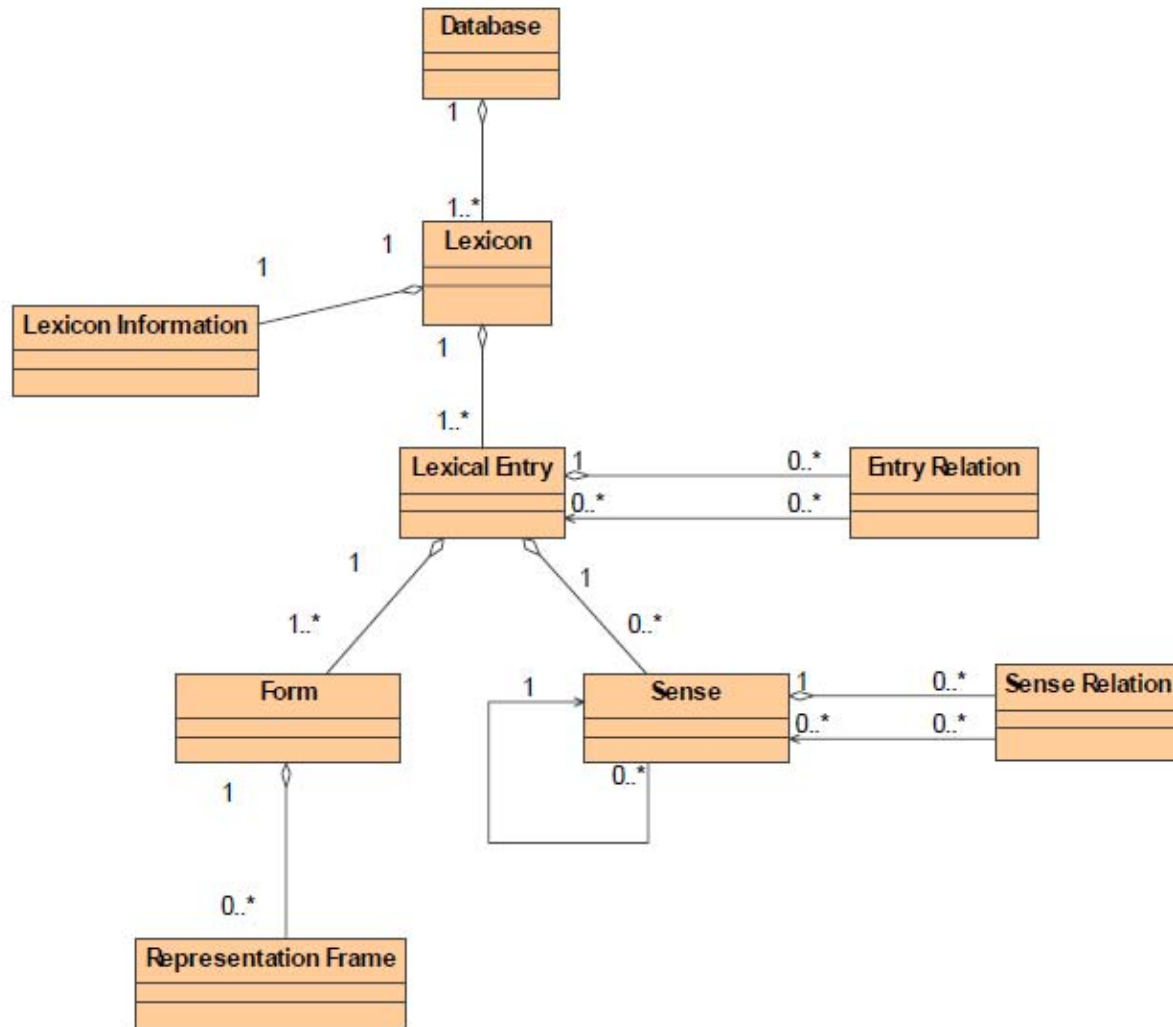
Basic hierarchy of a lexicon, and a lexical entry

Extensions:

Proposed lexicon structures for different situations

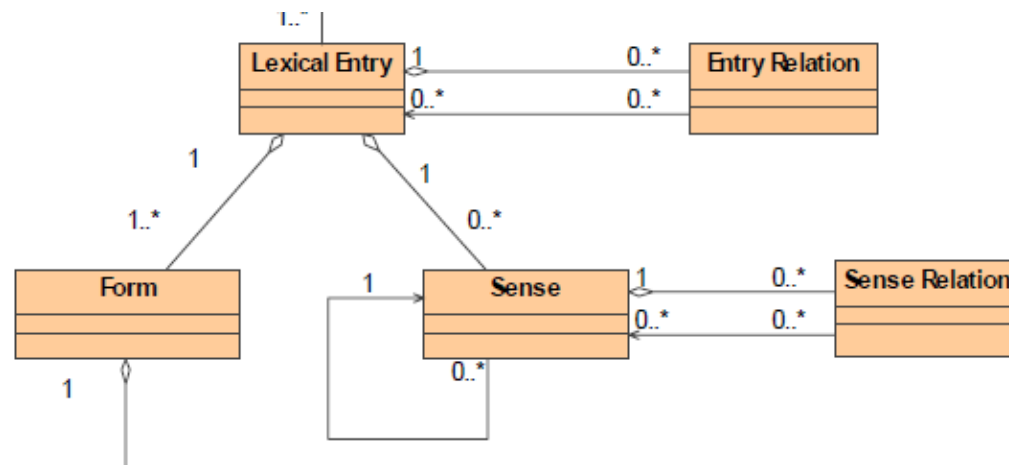


LMF





LMF



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

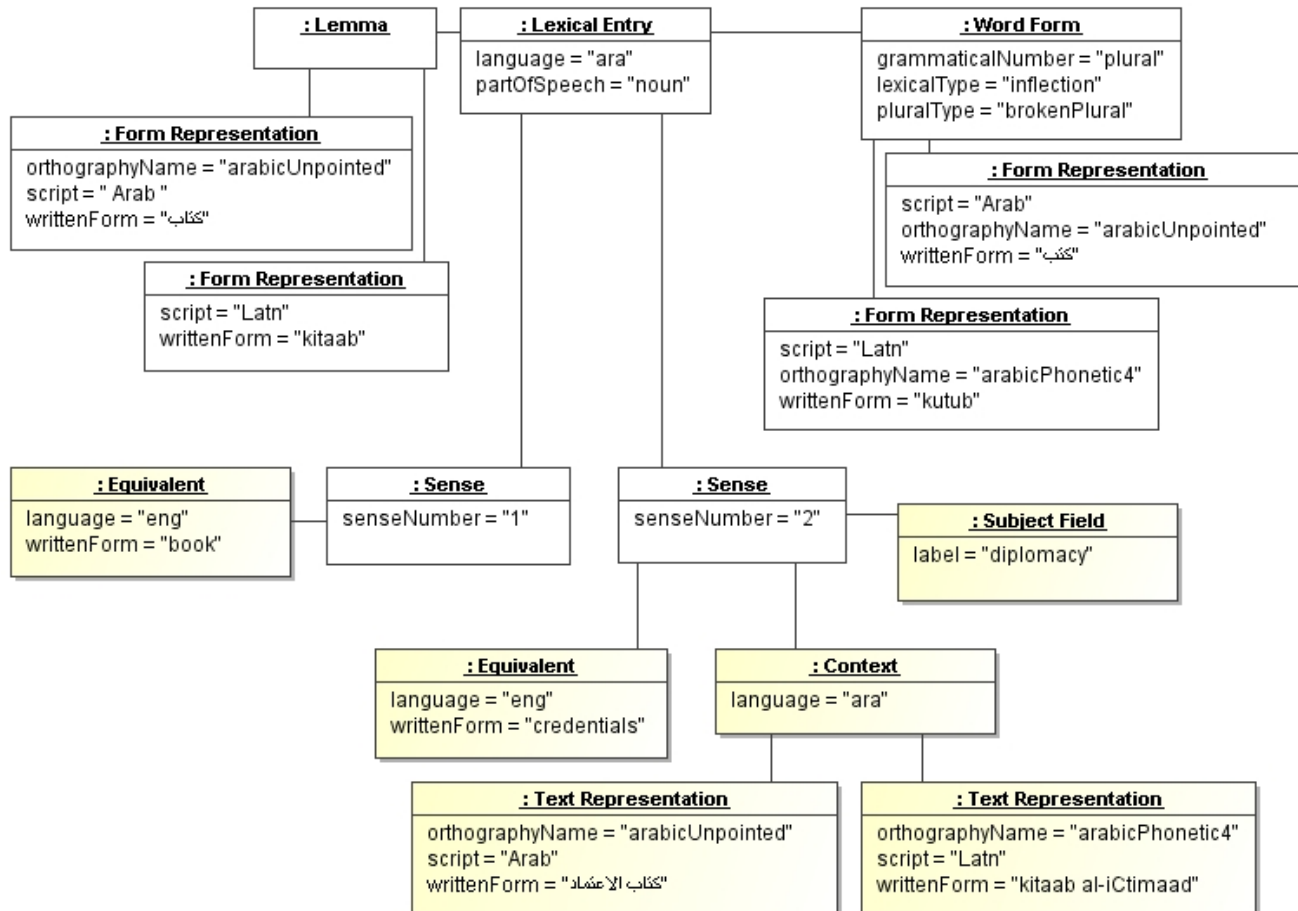
Form: text string representing the word

Sense: specifies the meaning and context



LMF

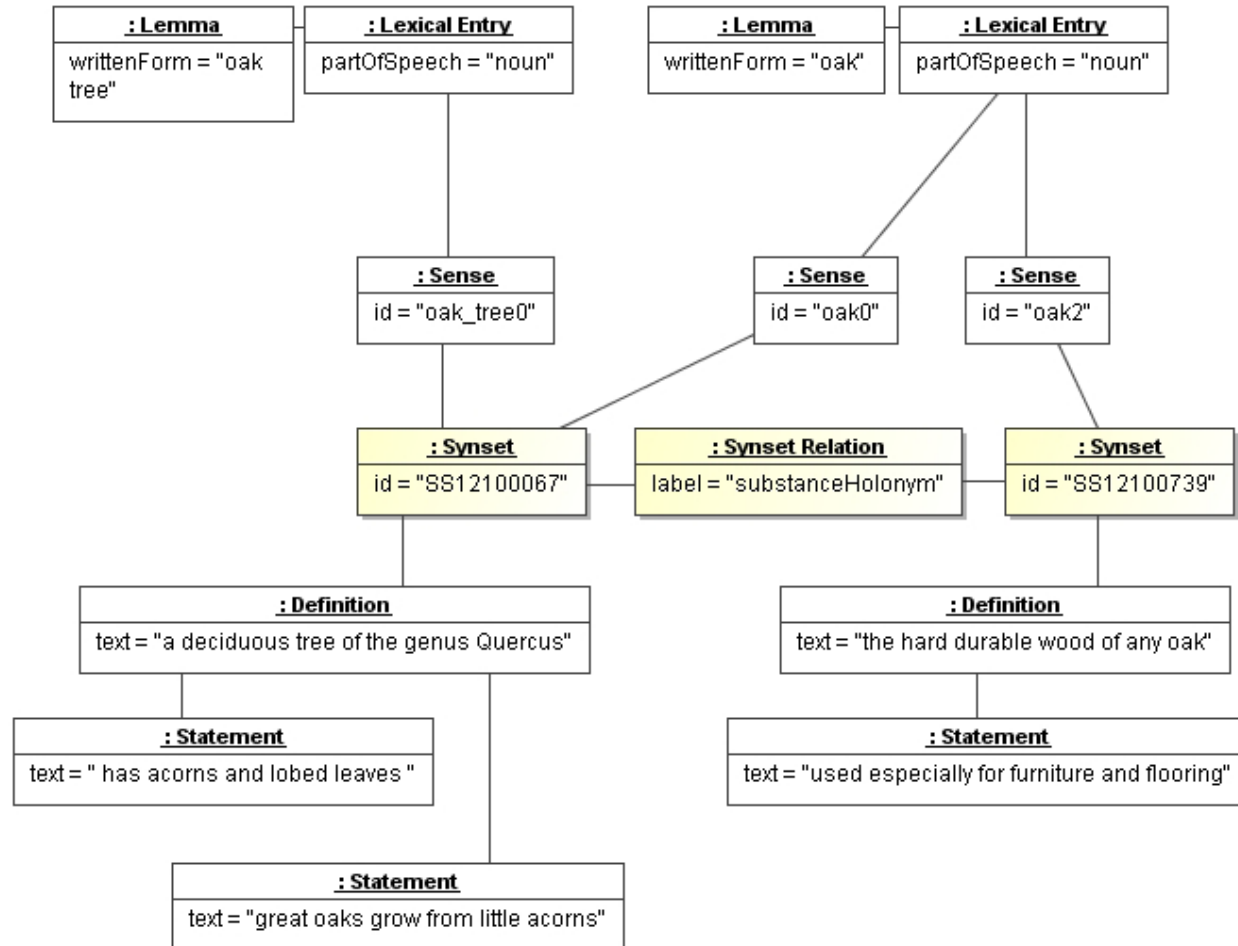
Strict division Form and Sense





LMF

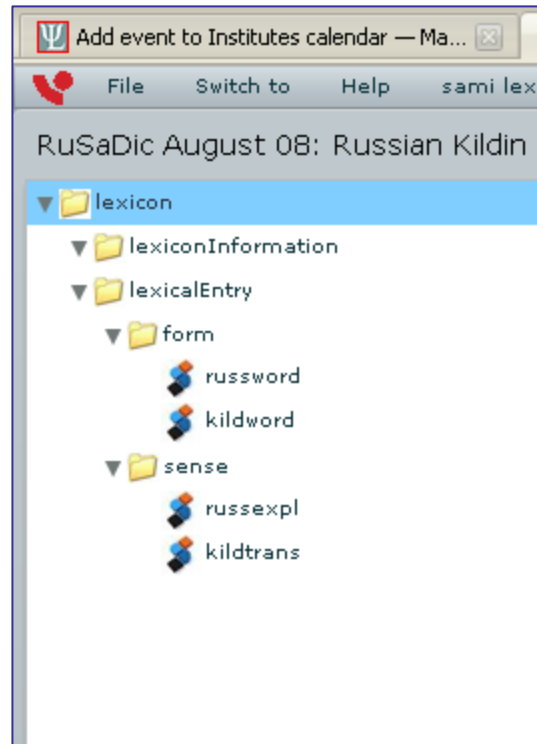
No sub-entries, but relations between LE





LMF in LEXUS

LMF default structure in LEXUS – Form, Sense



Slightly different from LMF:

Data components

Data categories

No attributes



ISOCat

Browser tabs: Add event to Institutes calendar — Ma..., http://corpus1...selectedIndex=0, Presentations — MPI Website, Request for a new user account and co...

File Switch to Help Onno Crasborn

Taalkunde lexicon in NGT:

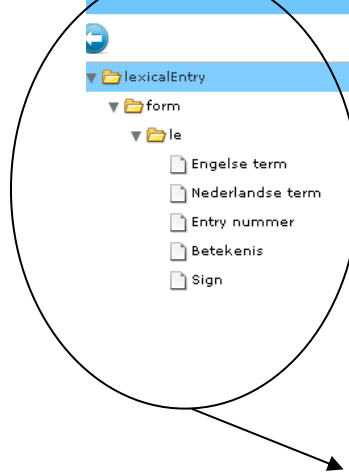

Lexicon Filters

<switch lexicon> <grapheme>
Switch to list view <filter>

- lexicalEntry
 - form
 - le
 - Engelse term
 - Nederlandse term
 - Entry nummer
 - Betekenis
 - Sign

Lexical entry Lexical Entry View

Nederlandse term: Adv. C.
Engelse term: Adverbial constituent



New structure for TKL in NGT:

- lexicon
 - lexiconInformation
 - lexicalEntry
 - Number
 - form
 - English
 - Dutch
 - Sign
 - sense
 - Meaning

New structure for TKL in NGT:

Lexicon structure

- lexicon
 - lexiconInformation
 - lexicalEntry
 - rank
 - form
 - Form representation
 - Language
 - Modularity
 - (Written) Form
 - sense
 - definition

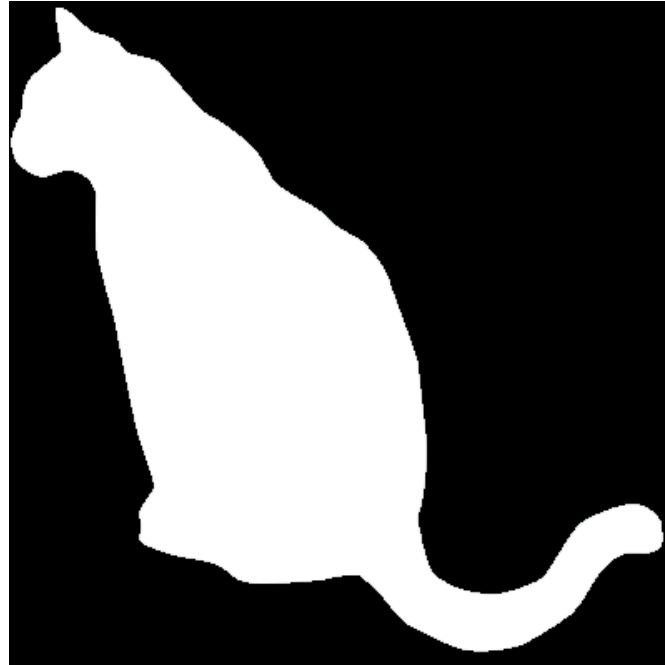
Add features

ituent is

Bijv: Danst mooi.



Concept naming





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



Concept naming

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)



Concept naming

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.

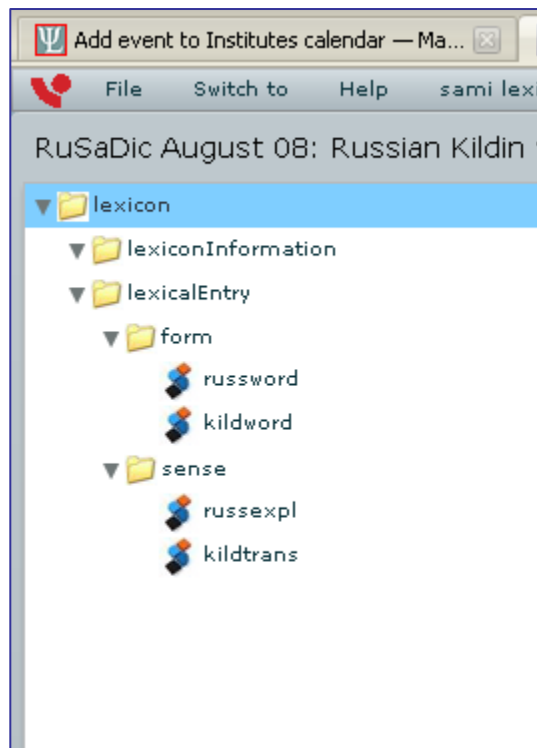
At regular intervals, snapshots of the standardized subset of the DCR will be submitted to ISO.

www.isocat.org



LEXUS

LMF default structure in LEXUS – Form, Sense



Slightly different from LMF:

Data components

Data categories

No attributes



From MDF2LMF

Concept naming

Add MDF to ISOCat (RELISH project)

The screenshot displays the ISOCat web interface. The main content area shows the details for a concept named "first_dual - 1:0". The interface is divided into several sections:

- Administration Information Section**
 - 1.1 Administration Record**

Identifier	first_dual
Version	1:0
Registration Status	private
Administration Status	private
Justification	Used to give the vernacular for this particular paradigm form.
Origin	MDF
Explanatory Comment	combined of two categories: person and number
Until Date	2010-04-20
<i>1.1.1 Creation</i>	
Creation Date	2010-02-16
Change Description	first_dual
<i>1.1.2 Last Change</i>	
Last Change Date	2010-04-20
Change Description	this category is combined of two categories: person and number
- 2. Description Section**

The interface also includes a search bar at the top, a navigation pane on the left, and a table at the bottom with columns for #, Name, Version, Administration stat, Registration status, Check, Type, Owned by, and Scope.



From MDF2LMF

Sub-entries

MDF does allow for sub-entries

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

Example sub-entry (Udi):

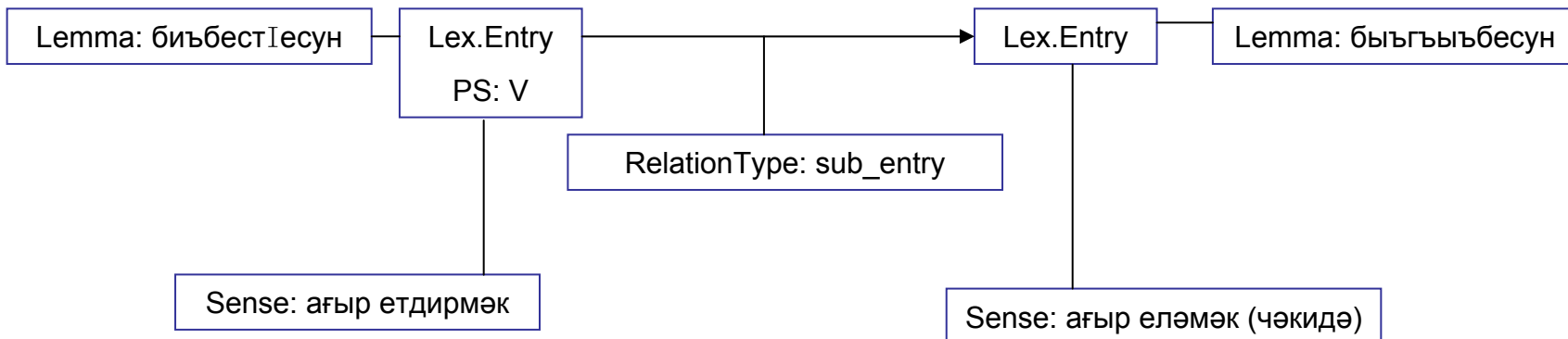
\lx биъбестIесун

\ps v

\gn1 ағыр етдирмәк, ағырлашдырмаг

\se быъгъыъбесун

\dt 05/Mar/2010





From MDF2LMF

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



From MDF2LMF

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):

Lemma:

\lexeme

Lexical Entry:

\part-of-speech

\METADATA

Form:

\citation

\alternative

\underlying

\phonetic

\variant

\MORPHOLOGY

\GRAMMAR

\picture

\video

Sense:

\gloss

\definition

\semantic-domain

\literally

\reversal

\EXAMPLES

\ENCYCLOPAEDIC

\SCIENTIFIC

\ETYMOLOGY

Linguistic Frames:

\CROSS-REFERENCE

\lexical-function



Questions?

