



# LEXUS

## a flexible web-based Lexicon tool



# Background

7 DOBES teams and 12 different lexica (structures, purposes)

Tuvan orthography
Tuvan appendix
German orthography
Russian orthography
Russian appendix
Xakas orthography
Tofa orthography

simple spreadsheet

stem orthography
sense *
lexical sub-entry *

little more complex incl 1:N relations

sense nr
sense
gram cat
gram subcat
Engl Transl
example *

orthography
Engl. Transl
[T pr] nr

entry-type = [stem idiom lexical word]
head
outer-body-L*

headword
citation form
homograph no
phonetic form

inner-body-L
grammar

sense number
variety
meaning
etymology
table
example*
comment*
picture/photo*
housekeeping*

gloss
word-level-gloss
reversal
definition
encyclopedic info
scientific name
semantic domain
semantic index
thesaurus
semantic relation*
cross-ref*

small part of a complex lexicon structure at top level 4 different entry types (only one is shown)



# Problem

- have to use one archival lexicon representation format based on XML
- have to build one archival exploitation framework
- however, receive lexica
  - character encodings
  - in all sorts of formats (var. XML, SBX, CHAT, even Word)
  - in various structures
  - with different terminologies (lexical attributes, values)
- how to do cross-lexical searches?
- how to do lexical merging, linking and comparison?
- how to solve lexicon-corpus interaction?
- etc



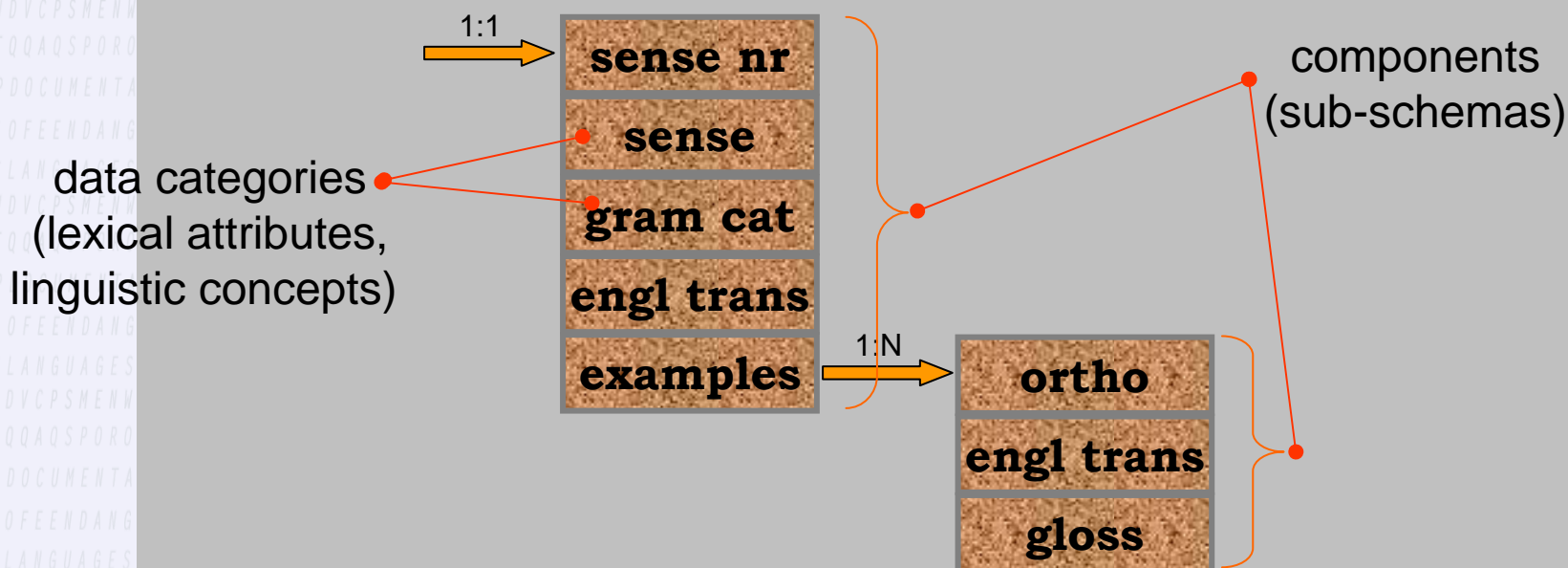
# ISO TC37/SC4 – The solution?

- ISO TC37/SC4 is about standardization in LR Management
  - central is **data category registry**
    - basically a flat list of linguistic concepts
    - will contain is\_a relations that are part of the concept definition
      - *“transitive\_verb” is\_a “verb”*
    - with proper definitions and conceptual space (value range)
    - request for filling DCR (Metadata, morphology, syntax, ...)
  - looking for **abstract models** (frameworks)
    - for lexica
    - for annotation structures
    - for semantic annotations
    - for syntactic annotations
    - ...



# Why not play LEGO ?

- concrete lexicon schema is basically seen as lexical attributes grouped together with others and embedded in a tree structure.





# LMF Lexical Markup Framework

## The goals of LMF are

- to provide a common model for the creation
- and use of very large scale lexical resources,
- to manage the exchange of data between and among these resources, and
- to enable the merging of large numbers of different individual electronic resources to form large global electronic resources.
- The ultimate goal of LMF is to create a modular structure that will enable true content interoperability across all aspects of lexical resources.



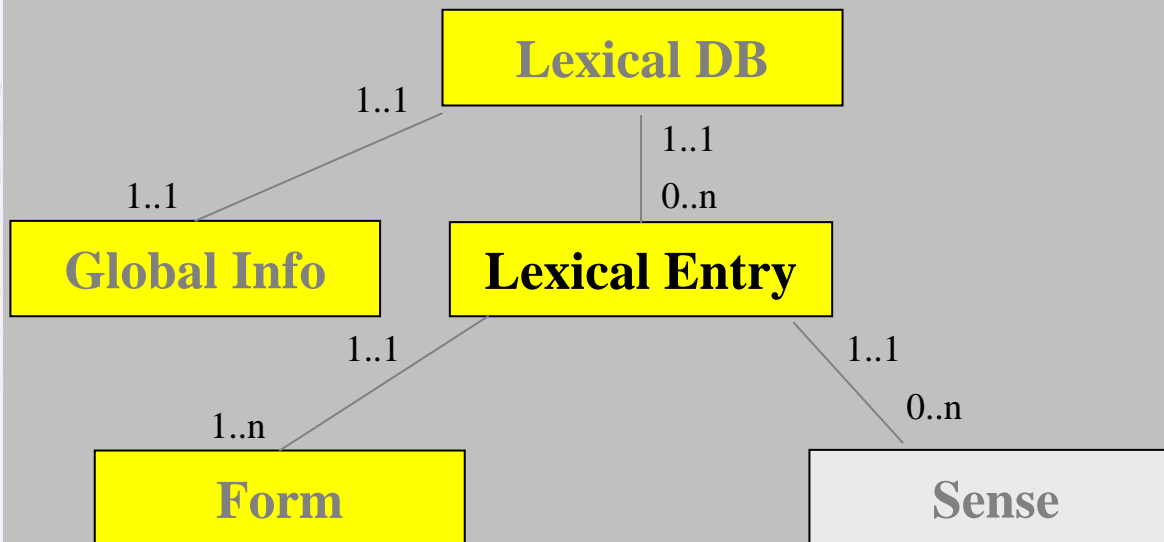
# LMF Core Model

## Metamodel

- Made of *lexical layers*

## Lexical layers

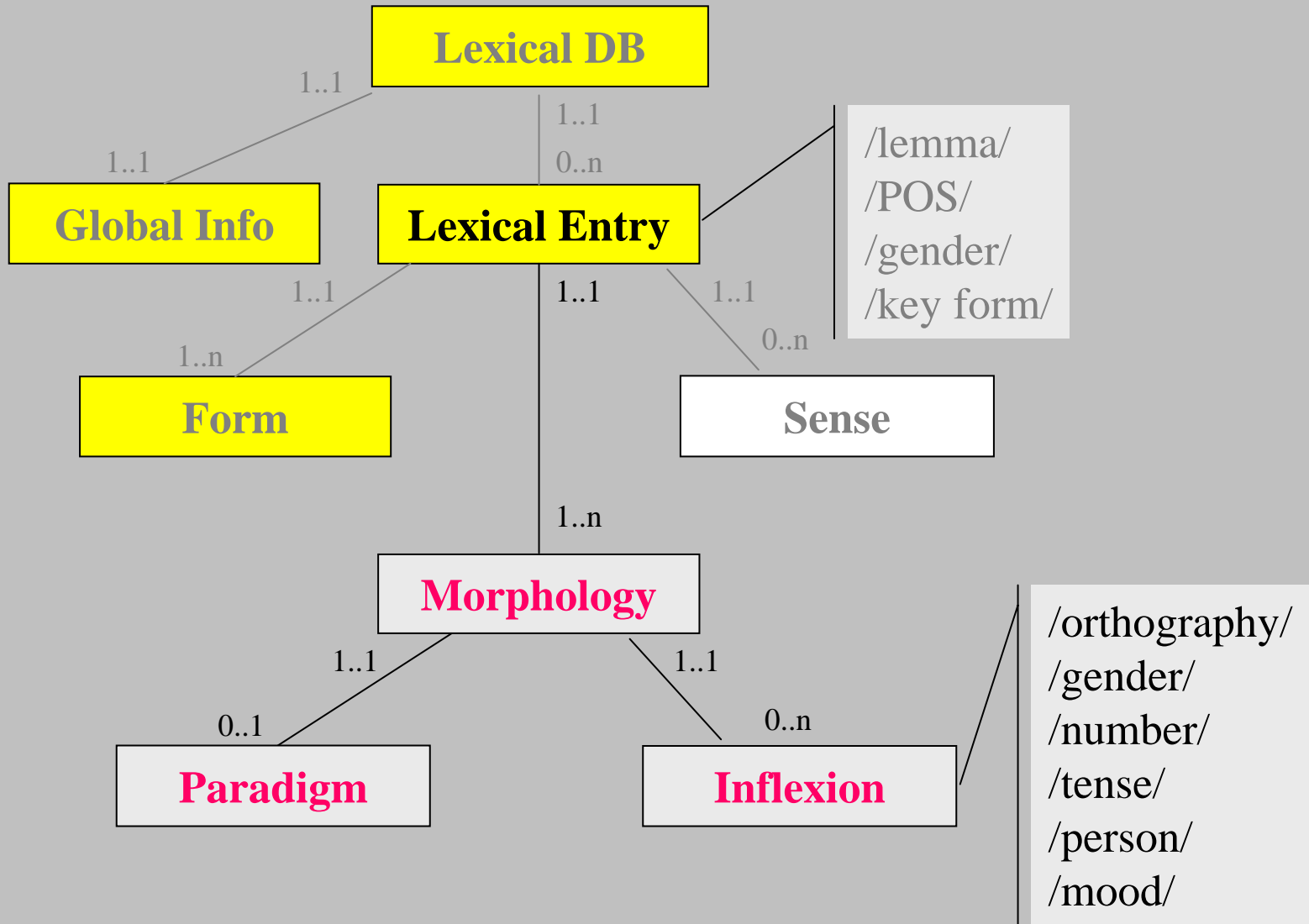
- Made of *lexical components* (or *components*)



- basis for modeling purposes is UML
- there will be an XML-schema based instantiation



# Extended Model





# Goal LEXUS

To provide a framework capable of handling diverse lexicon structures and formats.

Lexus is based upon Lexicon Markup Framework within ISO TC37/SC4 that defines a blueprint for such a flexible framework.

LEXUS is first test and reference implementation of LMF.

Increase interoperability by offering well accepted data categories (ISO, GOLD, Shoebox MDF)



# Current Status

- supports full LMF core model
- allows for flexible creation of structures and content.
- supports use of well-accepted Data Category Registries (ISO 12620, Shoebox MDF)
- allows for dynamic editing of structures and content.
- supports use of multimedia content.
- import of existing lexica (Shoebox, Chat)
- export( Shoebox/LMF XML)
- customizable layout



# Current Status

- user authentication
- personal workspace for creating and editing lexica
- merging facilities
- simple and advanced search



# Current Status (Technical)

- Implemented in java and using Open Source components
- Uses Spring to 'wire' the application
  - Modular approach avoiding 'hard' links
- Uses Hibernate as the persistence framework
  - Allows use of multiple databases (Postgres, MySQL,...)
- Uses Tomcat as Servlet Container



# Logging onto the application

Please enter your username and password.

Username:

Password:

[register as guest..](#)

## LEXUS lexical resource tool

Users must authenticate before login onto the application.



# User workspace

Workspace

Lexical Resource

Search

DCR

willem wever



Frisian Dictionary



lexus\_RUS.cut

Each user has his/her own personal workspace  
where private lexica are stored



# Lexicon creation

Workspace

Lexical Resource

Search

DCR

willem wever

## Create new lexicon

By supplying a name and a description in the fields below a new lexicon will be created for you.

Lexicon name:	<input type="text" value="My first dictionary"/>
Description:	<input type="text" value="Ths dictionary contains lexemes etc.."/>
<input type="button" value="Save"/>	

New lexica may be created...



# Lexicon import

Workspace Lexical Resource Search DCR

willem wever

Please select the type of lexicon you would like to import

Shoebox  
 Clan  
 Wichita XML

Please select whether you want to add a new lexicon or import your data into an existing one

Create new lexicon  
 Import data into existing lexicon..

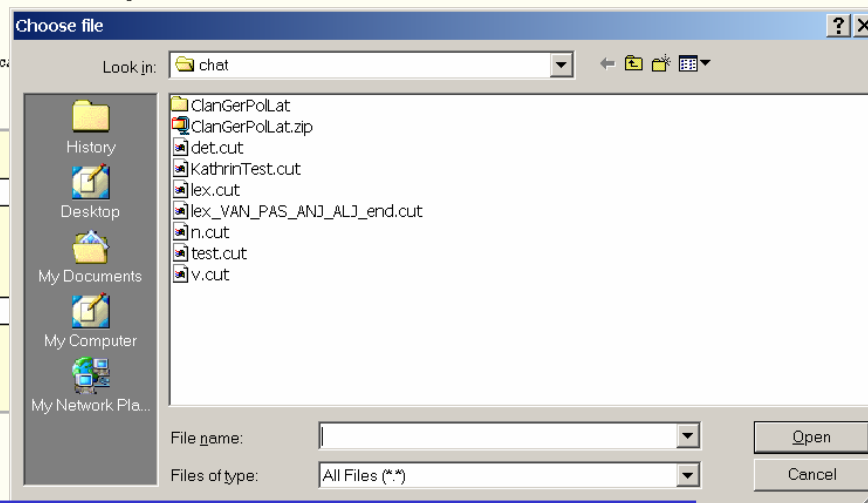
## Shoebox lexicon schema and data import

This page allows you to import a shoebox lexicon. A new Lexicon

Please upload the Shoebox .typ structural file here.

Please upload the Shoebox lexicon file here.

Import



New lexica may be imported from a lexical resource...



# Lexicon structure

Workspace Lexical Resource Edit Search View DCR willem wever

Copy  
Paste  
Delete  
New Data Category  
New Component

## PACE) LEXUS-Lexicon Scheme Viewer

LEXUS

- LexicalDatabase
  - GlobalInfo
  - LexicalEntry
    - Form
    - Sense
    - New Data Category

New Data Category

user defined  
 shoebox  
 iso 12620

General Layout Open issues

Creation date: Thu Apr 28 15:36:25 ICT 2005 Last modified: Author: willem wever

**Description:**  
to be filled out

**Admin info:**  
to be filled out

**Model name:** user defined

mandatory  
 multiple values allowed

**Reference:**

Save

The LMF core model can be identified in this simple structure. Components and datacategories can be identified using different icons. All may be dynamically created or modified.



# Lexicon structure

Workspace Lexical Resource Edit Search View DCR willem wever

### (WORKSPACE) LEXUS-Lexicon Scheme Viewer

**LEXUS**

- LexicalDatabase
  - GlobalInfo
  - LexicalEntry
    - Lexeme
      - Homonym number
      - Date (last edited)
      - Status
      - Lexeme
      - Citation form
      - Subentry
        - Borrowed word (loan)
        - Morphology
        - Subentry
        - Phonetic form
        - Main entry cross-ref.
      - Etymology (proto form)
        - Etymology comment
        - Etymology source
        - Etymology (proto form)
        - Etymology gloss (E)
      - Part of speech
        - Singular form**
        - Second singular
        - Non-animate plural
        - Third singular
        - Second plural
        - First dual
        - Second dual
        - Non-animate dual

**Singular form**

user defined  
 shoebox  
 iso 12620

General Layout Open issues

Creation date: 2005-04-28 Last modified: Author: willem wever

**Description:**  
This is a special paradigm field used to give the singular form of the lexeme. (It is now better to use the new \pdl field set for this.)

**Admin info:**  
to be filled out

**Model name:** user defined

mandatory  
 multiple values allowed

Reference: sg

Save

Representation of a more complex structure. By selecting a node in the Tree the content of a component or datacategory is shown and may be modified.



# Data category selection

Workspace Lexical Resource Edit Search View DCR willem wever

**(WORKSPACE) LEXUS-Lexicon Scheme Viewer**

Shoebbox data categories are listed below. Please select a data category to be inserted into your scheme.

LEXUS

- LexicalDatabase
  - GlobalInfo
  - LexicalEntry
    - Frisian Word
      - Part of Speech
      - Frisian Word
      - Environment
      - Morpheme Property
      - Comment
      - Frisian Alternate
      - Gloss
      - English Alternate
      - test
      - English
      - Morpheme Cooccurrence Constraint
      - Underlying Form

Data category	Description	Reference
First dual	Used to give the vernacular for this particular paradigm form. (It is better to use the new \pdl field set for this.)	Shbx:MDF:1d
First plural exclusive	Used to give the vernacular for this particular paradigm form. (It is better to use the new \pdl field set for this.)	Shbx:MDF:1e
First plural inclusive	Used to give the vernacular for this particular paradigm form. (It is better to use the new \pdl field set for this.)	Shbx:MDF:1i
First plural	Used to give the vernacular for this particular paradigm form. (It is better to use the new \pdl field set for this.)	Shbx:MDF:1p
First singular	Used to give the vernacular for this particular paradigm form. (It is better to use the new \pdl field set for this.)	Shbx:MDF:1s
Second dual	Used to give the vernacular for this particular paradigm form. (It is better to use the new \pdl field set for this.)	Shbx:MDF:2d
Second plural	Used to give the vernacular for this particular paradigm form. (It is better to use the new \pdl field set for this.)	Shbx:MDF:2p
Second singular	Used to give the vernacular for this particular paradigm form. (It is better to use the new \pdl field set for this.)	Shbx:MDF:2s
Third dual	Used to give the vernacular for this particular paradigm form. (It is better to use the new \pdl field set for this.)	Shbx:MDF:3d
Third plural	Used to give the vernacular for this particular paradigm form. (It is better to use the new \pdl field set for this.)	Shbx:MDF:3p
Third singular	Used to give the vernacular for this particular paradigm form. (It is better to use the new \pdl field set for this.)	Shbx:MDF:3s
Non-animate dual	Used to give the vernacular for this particular paradigm form. (It is better to use the new \pdl field set for this.)	Shbx:MDF:4d
Non-animate plural	Used to give the vernacular for this particular paradigm form. (It is better to use the new \pdl field set for this.)	Shbx:MDF:4p
Non-animate singular	Used to give the vernacular for this particular paradigm form. (It is better to use the new \pdl field set for this.)	Shbx:MDF:4s
Antonym	Used to reference an antonym of the lexeme, but using the \lf (lexical function) field for this is better practice.	Shbx:MDF:an
Bibliography	Used to record any bibliographic information pertinent to the lexeme. MDF adds the label 'Read:' to this field.	Shbx:MDF:bb
Recommended		Shbx:MDF:bw
		Shbx:MDF:ce
Cross-reference	This is a generic reference marker used to link together any two related entries in the lexicon. The content is a vernacular lexeme. If the relationship is known, the lexical function \lf field is a better way	Shbx:MDF:cf

Data categories can easily be selected from data category registries.



# Lexical entry overview

## LEXUS-Lexicon Scheme Viewer

### Overview of lexical entries(134 entries)

Lexicon: Frisian Dictionary

Description: Frisian Dictionary for Interlinear Tutorial

Number of entries on page 20

<< [previous](#) Page number: 4 [next](#) >>

By clicking on an entry you will be directed to a detailed view of the entry.

Frisian Alternate	Morpheme Property	Underlying Form	Gloss	Frisian Word	English	Environment	default	Part of Speech	Comment	Morpheme Cooccurrence Constraint	English Alternate
			probable	wierskyndlik	probable			Adj			
			stratum	xxcx	stratum			N			strata plural +/ PLUR
			wide	wide	wide			Adj			
in			accessible	tagonklik	xxcx			Adj			
			quorum	xxcx	quorum			N			
			while	wylst	while			Conj			
			his	syn	his			Pron			
			faithful	trou	faithful			Adj			
			back	werom	back			Adv			
			un	ûn-				Neg			
			datum	xxcx	datum			N			data plural +/ PLUR
			we	wry	we			Pron			
			2	-st				Pers			
			be	wêze	be			V			
			3	-t				Pers			
											errata plural +/ PLUR
											shone past +/ PAST
			welcome	wolkom	welcome			Adj			
			certain	wis	certain			Adj			

Overview of lexical entries. By selecting a lexical entry the details will be revealed.



# Lexical entry details

Details of a lexical entry. Entry structure modifications are bound to schema definition, e.g. cardinality.



# Lexical entry details

Lexus - Microsoft Internet Explorer provided by MPI Nijmegen

LexicalEntry Edit View willem wever

**(WORKSPACE) LEXUS-Lexical Entry Viewer**

**LEXUS**

- LexicalDatabase
- GlobalInfo
- LexicalEntry
  - Lexeme
    - Lexeme: ba
    - Date (last edited): 03/Jul/1990
    - Status
    - Subentry
      - Part of speech
        - Part of speech: prep
        - Second singular
        - Non-animate dual
        - Non-animate singular
        - Sense number
          - Gloss (n): keluar
          - Antonym: ma3
          - Synonym: ei, kita, ti1
          - Gloss (E): away
          - Reversal (n): luar, ke
          - Notes (general): The "fv:ba\_ti" of
          - Notes (grammar): Used with mot
          - Usage (n)

Example (v):

Creation date: 2005-04-28 Last modified: Author: willem wever

**Description:**  
Used to give an example or illustrative sentence in the vernacular to legitimate or exemplify each separate sense. Should be short and natural.

**Admin Info:**  
**Model name:** user defined  
**Reference:** XV  
 mandatory  
 multiple values allowed

Save

Example free trans. (E): l'n  
Example (v): Ktulis suratke  
Example free trans. (n): Sa

Introduction  
LEXUS 1.0  
May 2005

4:2

Attribute values can be easily modified. Various value types are supported( text, video, audio, image or file)



# Lexical entry details

The screenshot shows a web browser window titled "Lexus - Microsoft Internet Explorer provided by MPI Nijmegen". The browser's address bar and menu bar are visible. The main content area displays the "LexicalEntry" details for the word "ba". The details include:

- Lexeme: ba
- Date (last edited): 03/Jul/1990
- Status
- Subentry
- Part of speech: prep
- Second singular
- Non-animate dual
- Non-animate singular
- Sense number
- Gloss (n): keluar
- Antonym: ma3
- Synonym: ei, kita, ti1
- Gloss (E): away
- Reversal (n): luar, ke
- Notes (general): The "f"
- Notes (grammar): User
- Usage (v)
- Usage (n)

An "Upload of video file" dialog box is overlaid on the page. It contains the following text:

**Upload of video file**

*NOTE: To be able to view these files an appropriate viewer must be present on the user's system.*

*NOTE: The maximum file size is currently set to 1000000 bytes (1Mb).*

Please upload your file here.

A "Choose file" dialog box is also open, showing the "resources" directory. The file "PeterW.mpg" is selected. The dialog box shows the file name "PeterW.mpg" and the file type "All Files (\*.\*)".

**Example of uploading a video file.**



# Lexical entry details

Lexus - Microsoft Internet Explorer provided by MPI Nijmegen


LexicalEntry Edit View willem wever

**(WORKSPACE) LEXUS-Lexical Entry Viewer**

LEXUS

- LexicalDatabase
  - GlobalInfo
  - LexicalEntry
    - Lexeme
      - Lexeme: ba
      - Date (last edited): 03/Jul/1990
      - Status
      - Subentry
        - Part of speech
          - Part of speech: prep
          - Second singular
          - Non-animate dual
          - Non-animate singular
          - Sense number
            - Gloss (n): keluar
            - Antonym: ma3
            - Synonym: ei, kita, ti1
            - Gloss (E): away
            - Reversal (n): luar, ke
            - Notes (general): The "fv:ba\_ti" of
            - Notes (grammar): Used with mot
          - Usage (v)
          - Usage (n)
          - Reference
            - Refer...
            - Examp
            - Exa
            - Example (v): <<video>>
            - Example free trans. (n): S...

**Example (v):**



- text
- file
- video
- audio
- image

Creation date: 2005-04-28 Last modified: Thu Apr 28 16:35:27 ICT 2005 Author: willem wever

**Description:**  
Used to give an example or illustrative sentence in the vernacular to legitimate or exemplify each separate sense. Should be short and natural.

**Admin Info:**  
**Model name:** user defined

**Reference:** XV

mandatory

multiple values allowed

Save

Viewing multimedia content.

Introduction LEXUS 1.0 May 2005

D:\marc... 29 april ... lexus-sh... Comman... Lexus - ... Tomcat Lexus - ... 4:30



# Alternative entry view

Lexus - Microsoft Internet Explorer provided by MPI Nijmegen

LexicalEntry Edit View willem wever

**(WORKSPACE) LEXUS-Lexical Entry Viewer**

The workspace displays five alternative views for the word 'hoef':

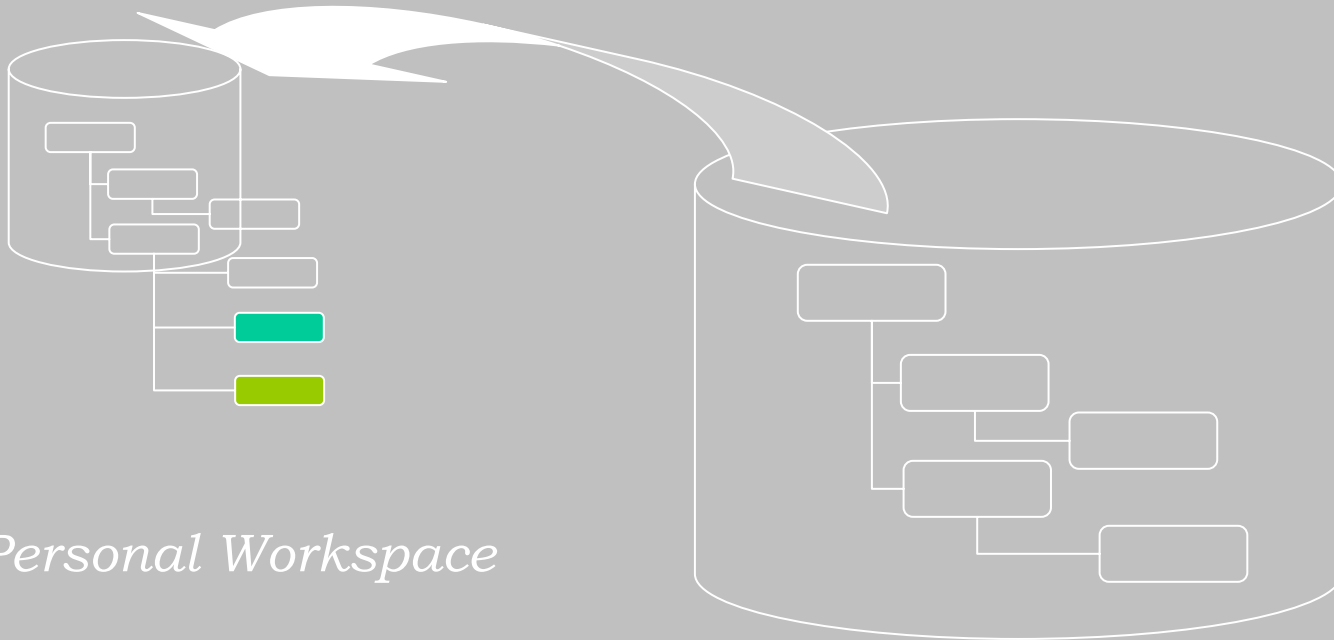
- Frisian Word**: Shows the word 'hoef' with a document icon.
- Part of Speech**: Shows a document icon with a vertical bar.
- Gloss**: Shows the word 'hoef' with a document icon.
- Comment**: Shows a video frame of a person gesturing while speaking.
- English**: Shows the word 'hoef' with a document icon.
- English Alternate**: Shows the word 'hoove / \_ s' with a document icon.

Introduction  
LEXUS 1.0  
May 2005

Alternative views are provided which may be customized in look and feel.



# Synchronization of lexica



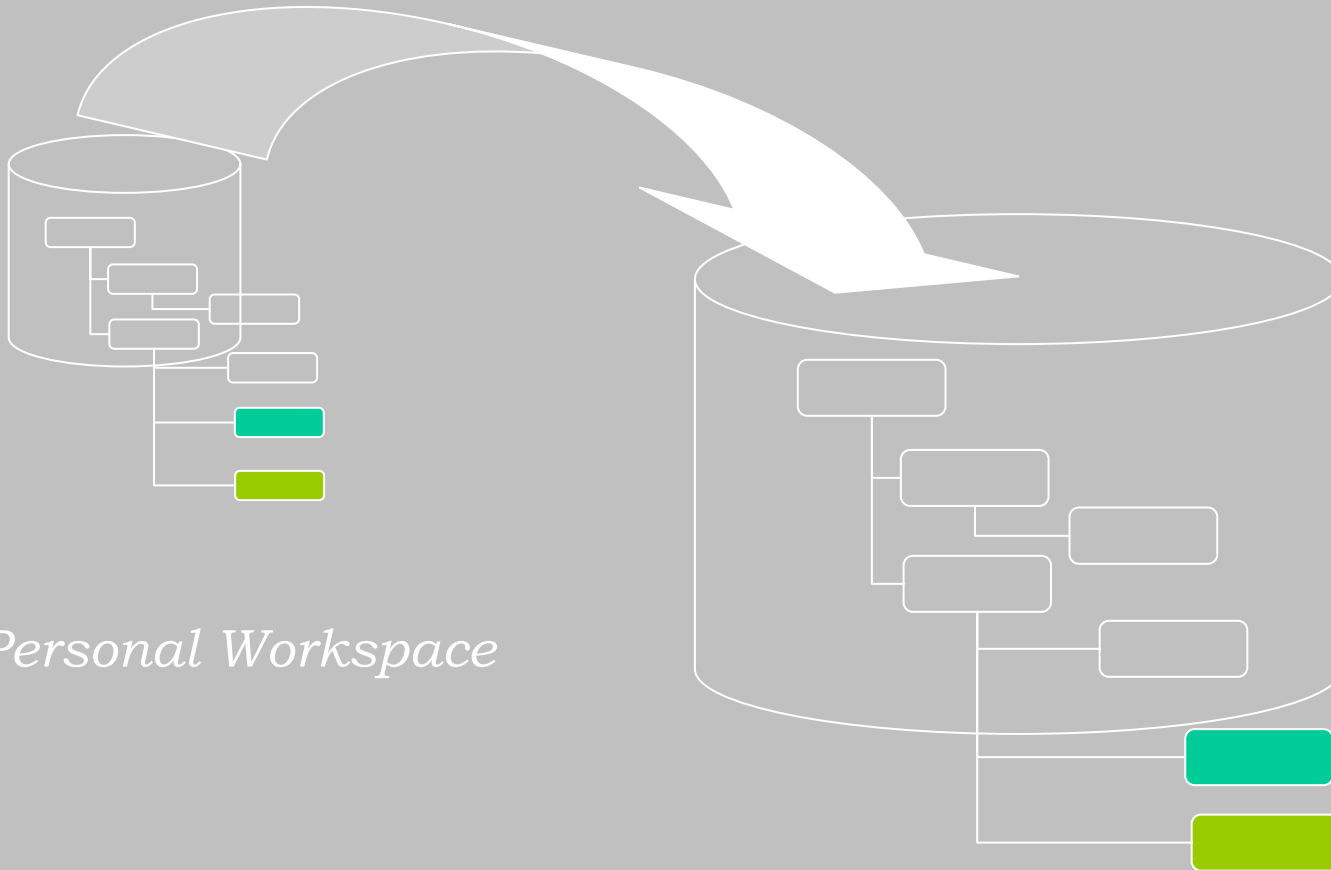
*Personal Workspace*

*Main Lexicon*

*Lexica may be copied to and modified in personal workspace*



# Synchronization of lexica



*Personal Workspace*

*Main Lexicon*

*Lexica may be synchronized with main lexicon*



# Synchronization of lexica

**workspace** **master**

**STEP 2: Synchronize schemas**

List of dataCategories (and maybe components) which have been modified. The selected dataCategories will be modified in the master. All unselected dataCategories will be modified in your workspace copy.

Data categories:

	name	description	cardinality
<input checked="" type="checkbox"/>	my datacategory	to be filled out	0..n
<input checked="" type="checkbox"/>	Frisian Alternate	to be filled out	0..n

OK

When synchronizing lexica the user is notified of structural changes and is in total control of the synchronization proces.

**workspace** LEXUS LexicalDatabase GlobalInfo LexicalEntry Form Sense Frisian Word Frisian Alternate my datacategory Comment Part of Speech Gloss Environment English Underlying Form Morpheme Property Frisian Word Morpheme Cooccurrence C English Alternate Gloss Part of Speech English Underlying Form Environment Morpheme Property

**master** LEXUS LexicalDatabase GlobalInfo LexicalEntry Form Sense Frisian Word Morpheme Cooccurrence C English Alternate Comment Morpheme Property English Part of Speech Frisian Word test Underlying Form Gloss Environment Frisian Alternate Part of Speech Underlying Form Gloss Environment Morpheme Property English

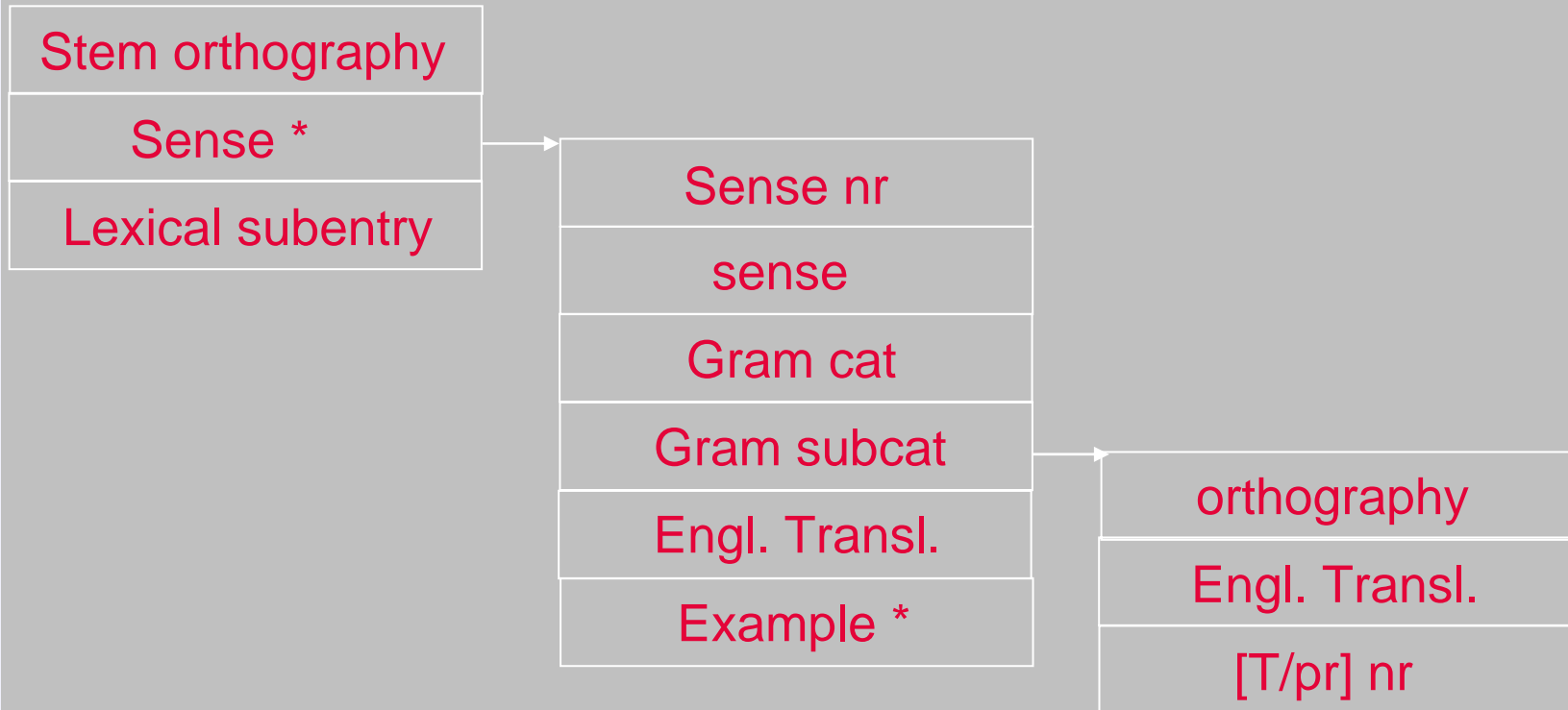


# Future directions

- Support for various types of relations
- Import of data from other sources
- Support for other Data Category Registries, e.g. GOLD
- Integration with MPI archive
- Integration with exploitation tools (ELAN, ANNEX)
- Miscellaneous user requests



# Example lexical structure



Example lexical structure used in the TEOP project within DOBES