



# The Wichita lexicon in LEXUS

Armik Mirzayan

University of Colorado at Boulder

Jacqueline Ringersma

Max Planck Institute for Psycholinguistics



# Key Issues and Goals

## Workshop Context:

- Importance of Lexical Resources
- Formulation of a Common Framework for Lexica
- Standards for tools and inter-operability

## Our aim is to present and discuss:

- The Current State of the *Wichita* Lexicon
- Two Significant Challenges for a “*Wichita Lexicon*”
- New Approaches/Ideas



# Outline

## Part 1: Contributions of Wichita to Lexicon Structure

- Some key aspects of *Wichita*
- From Wichita Database to XML Lexicon
- Wichita Structure and Lexicon Challenges
  - headword, lexical entry
  - syntactic morphology

## Part 2: Structure of the Wichita Lexicon in LEXUS

- LEXUS and ViCoS
- Wichita XML to LMF
- Wichita XML to ISOcat
- Enhancing inter-operability



# Concerning Wichita



Traditional Style Wichita *Grasshouse*



- Indigenous North American Language
- Caddoan Family
- Northern Caddoan Branch  
(closely related languages: Pawnee and Arikara)

**Highly Endangered:** one elderly fluent speaker, plus a few semi-fluent speakers



# Concerning Wichita Structure

Wichita is Structurally a Polysynthetic Language.

- Arguments and Predicates Associated in Bound Verbal Morphology only
- Noun Incorporation
- No Non-finite Verb Forms

A **minimal** verb contains four morphemes.

tense/mode - argument person marker – **root** - aspect/subord.



other prefix positions {preverb, locatives, dative, noun class, ...}



# Concerning Wichita Structure

- Isolated Noun forms are generally easy to work with, although there are derivational complexities.
- Verbs are very complex (30 position classes of affixes).

Partial  
Example  
for 3<sup>rd</sup>  
person form  
of /ʔarasi/  
("cook")

P R E F I X E S	S U F F I X E S			
	perfective (-∅)	imperfective (-s)	intensive (-staris)	habitual (--ss)
aorist <i>a...ki-</i>	'She cooked it.' <i>ákaʔárasiki</i>	'She was cooking it.' <i>ákaʔarásis</i>	'She was going to cook it, but didn't.' <i>ákaʔarásistaris</i>	'She always used to cook it.' <i>ákaʔarásiki-ss</i>
aorist quotative <i>aʔa...ki-</i>	'I heard that she cooked it.' <i>á-kaʔárasiki</i>	'I heard she was cooking it.' <i>á-kaʔarásis</i>	'I heard she was going to cook it.' <i>á-kaʔarásistaris</i>	'I heard she always used to cook it.' <i>á-kaʔarásiki-ss</i>
future <i>keʔe-</i>	'She will cook it.' <i>keʔárasiki</i>	'She'll be cooking it.' <i>keʔarásis</i>	_____	'It will be her job to cook it every time.' <i>keʔárasiki-ss</i>
future quotative <i>ehe--</i>	'I heard she'll cook it.' <i>ehèʔárasiki</i>	'I heard she'll be cooking it.' <i>ehèʔarásis</i>	_____	'I heard it will be her job to cook it every time.' <i>ehèʔarásiki-ss</i>



# A "Linguist's" XML Wichita Dictionary

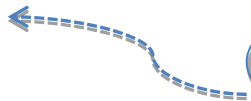
4490 1 a\*raskacic7a iye:\*riwa:ce:s7i i8ssinno:kha:\*r7ih  
 4491 2 +dry meat they had lots of that kind.  
 4492 3 7aras kagic 7a2 i4 iy ru riwa:c re:hi s2 7i issiri7 na wakhahr 7i h xx  
 4493 4 meat dry nounf direc nfocagt poss big beinplacepl impf be causei thatkind ppl activpatq be subp  
 4494 5 bhc61  
 4495 ==7arasi  
 4496 1 wa\*tiwa:rassis  
 4497 2 they are cooking now; they are really ripe  
 4498 3 wa7 ta i2 wa:rassi s2 7arasi wa:rasi wa:  
 4499 4 already pres pfocagt cookpl impf cook ripe distrib xx  
 4500 5 1973-25  
 4501 ==7arasi  
 4502 1 wa\*ta7arass  
 4503 2 it is ripe  
 4504 3 wa7 ta i2 7arasi s2  
 4505 4 already pres pfocagt cook impf causei ripe  
 4506 5 1973-25  
 4507 ==7arasi  
 4508 1 wa\*ke:7aras  
 4509 2 they will get ripe  
 4510 3 wa7 ke7 i2 xx 7arasi  
 4511 4 already fut pfocagt xx cook ripe  
 4512 5 1973-25  
 ... ==7arasi

## Wichita Database (Rood) → XML (2006)

```

<entry>
  <headmorph>&gstop;arasi</headmorph>
  <category>vi</category>
  <entnum>51</entnum>
  <gloss>cook, ripe</gloss>
  <comments>--</comments>
  <examples>
    <exnum>1</exnum>
    <wichita> w&aacute;tiwa:rassis</wichita>
    <free_gloss> they are cooking now; they are really ripe</free_gloss>
    <morphemes> wa&gstop; ta i2 wa:rassi s2 &gstop;arasi wa:rasi wa:</morphemes>
    <morpheme_gloss> already pres pfocagt cookpl impf cook ripe distrib xx</morpheme_gloss>
    <comments> 1973-25</comments>
  </examples>
  <exnum>2</exnum>
  
```

/ʔarasi/





# Wichita Challenge 1: Headwords?

- Prefix Dominance and Root-Final Words
- Morpheme Boundary Complexities

Root: /tarʔa:ti/ ‘cure, doctor’

ti:ckíciyé:sʔastarʔa:c

“he is doctoring some dogs” (source 1973-20)

ta- i- uc- kiciye:- s- ʔak- **tarʔa:ti** -s  
pres-pfocus-prev.dat-dog-inc-patns-doctor-impf





## Wichita Challenge 1: Headwords?

Given this situation, what do we use as head words for verbs, in a dictionary that is for the community to use?

Solutions (?):

1. Use an inflected form (like indicative) ...  
=> then \*all\* verb entries start with /t/!
2. Use a nominalized form (participle) ...  
=> again, \*all\* verb entries start with /n/!
3. Use another tense/mode form (other complexities ...)
4. Decide on a verb-by-verb basis (?).



## Wichita Challenge 2: Word = Grammar ?

**Syntactic Morphology:** Derivations, Inflections, Incorporation, ... what is part of grammar and what is part of “words in a dictionary”?

Example: (Rood, 2004)

iskiteʔe:ki nackwi:rʔicʔírih

“sit on my shoulder” (source 1973-narratives)

i-	s-	kita-	ʔi:ki	na-t-wi:rʔic-ʔi-hrih
imper-2.sub-	loc.on-	sit		ppl-1.sub-shoulder-be-loc



## Wichita Challenge 2: Word = Grammar ?

### Aspects of Syntactic Morphology:

Should (some) preverbs be part of the verb lexical entry?

Which different prefixes with a given verb root count as separate entries?

How about morphemes like *re:R-* (function of a nominal inflection but coded as a verbal prefix)?

Example:

hancʔa nacé:ra:kʔáskih

“the grass I was talking about” (source Rood-2004)

hancʔa      na-t-re:R-rakʔa-ski-h

grass      ppl-1.sub-shoulder-impf-subord.



# LEXUS and ViCoS

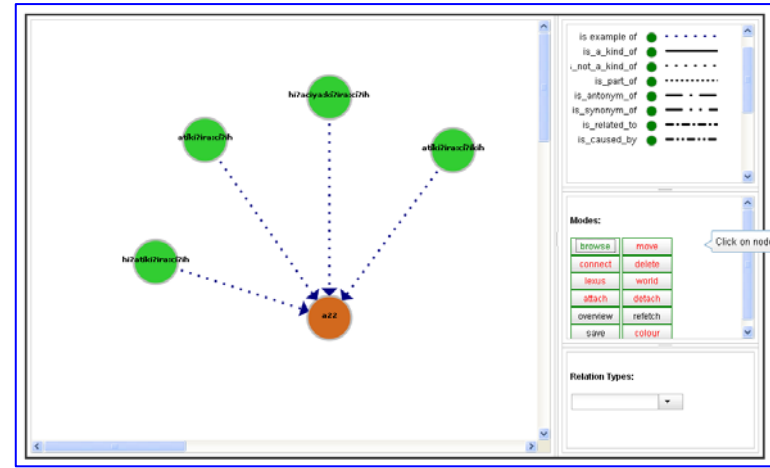
Lexicon: **a2**

Lexical entry: **a2** *non-quotative*

01 **aki:ché:stikiks**  
*he kept putting it in his mouth; answer to 'he made him eat'*  
a2 hi ut: hi ka xx hi B. hi t

02 **tackwa7aki:ʔi**  
*he was old*  
tackwa7ac wa7 a2 hi D (ʔi ʔi)

03 **tackwa7aki:ki**  
*they were (too) big (e.g. cucumbers)*  
tackwa7ac wa7 a2 hi D. hi t xx



## LEXUS

a web based tool for the creation of multi media encyclopedic dictionaries and lexica

## ViCoS

extension of LEXUS for the creation of conceptual spaces



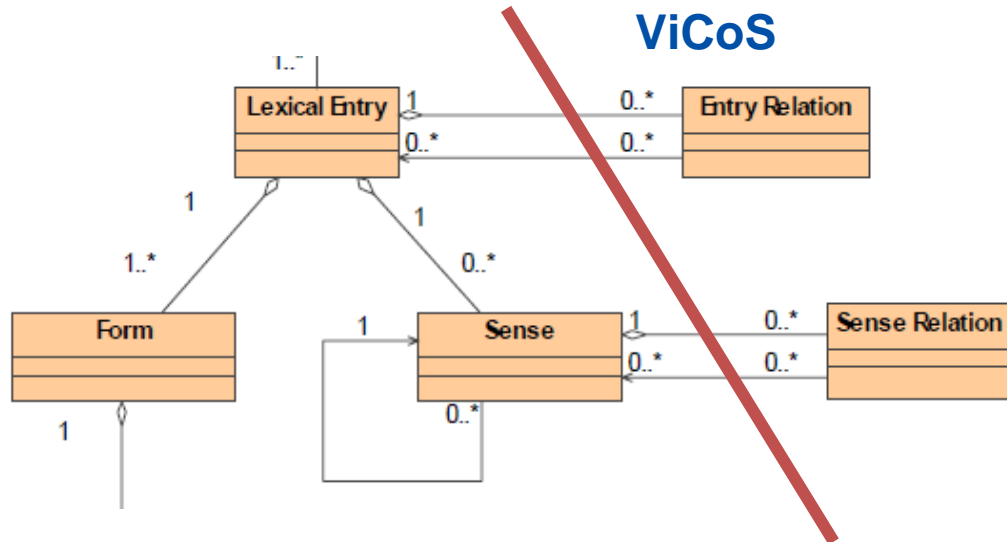
Based on two ISO TC 37 standards for linguistic resources

**LMF:** Lexical Markup Framework (lexicon structure)

**DCR:** set of standardized data categories to be used as a reference for the definition of linguistic annotation schemes or any other formats used in the area of language resources (concept naming)

## **LMF/DCR:**

- A modular structure for content interoperability between lexical resources
- XML based archiving exploitation framework



**LexicalEntry:** container for managing one or several forms and possibly one or several meanings in order to describe a lexeme

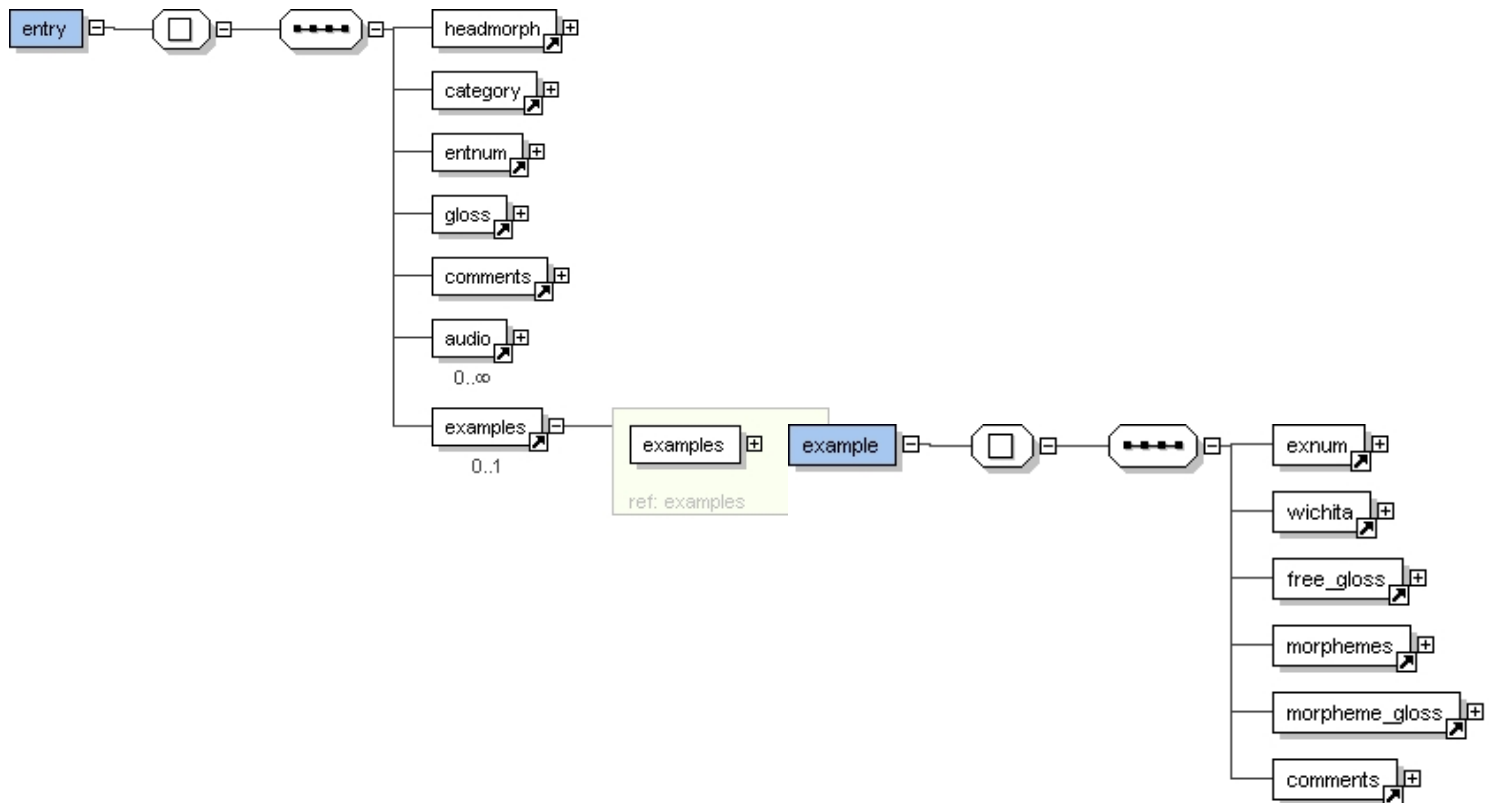
**Form:** text string representing the word

**Sense:** specifies the meaning and context



# From Wichita XML to LMF

Wichita XML elements and structure:





# From Wichita XML to LMF

**Import XML file** [X]

**Step 3. Select LMF class for markers.**

By dragging and dropping the markers into the appropriate LMF classes your lexicon can be made LMF compliant

- lexicon
  - lexicalEntry
    - form
      - doc
        - entry
          - headmorph
          - category
          - entnum
          - gloss
          - comments
        - examples
          - example
            - exnum
            - wichita
            - free\_gloss
            - morphemes
            - morpheme\_gloss
            - comments
        - sense
        - lexiconInformation

Wichita XML → LMF

Previous Finish Cancel

**Import XML file** [X]

**Step 3. Select LMF class for markers.**

By dragging and dropping the markers into the appropriate LMF classes your lexicon can be made LMF compliant

- lexicon
  - lexicalEntry
    - headmorph
    - entnum
    - category
    - comments
  - form
  - sense
    - gloss
  - examples
    - example
      - exnum
      - wichita
      - free\_gloss
      - morphemes
      - morpheme\_gloss
      - comments
  - lexiconInformation

Previous Finish Cancel





# From Wichita XML to LMF

**Import XML file** [X]

**Step 3. Select LMF class for markers.**

By dragging and dropping the markers into the appropriate LMF classes your lexicon can be made LMF compliant

- lexicon
  - lexicalEntry
    - form
      - doc
        - entry
          - headmorph
          - category
          - entnum
          - gloss
          - comments
        - examples
          - example
            - exnum
            - wichita
            - free\_gloss
            - morphemes
            - morpheme\_gloss
            - comments
        - sense
        - lexiconInformation

Wichita XML → LMF

Previous Finish Cancel

**Import XML file** [X]

**Step 3. Select LMF class for markers.**

By dragging and dropping the markers into the appropriate LMF classes your lexicon can be made LMF compliant

- lexicon
  - lexicalEntry
    - entnum
    - category
    - comments
    - form
      - headmorph
    - sense
    - gloss
    - examples
      - example
        - exnum
        - wichita
        - free\_gloss
        - morphemes
        - morpheme\_gloss
        - comments
    - lexiconInformation

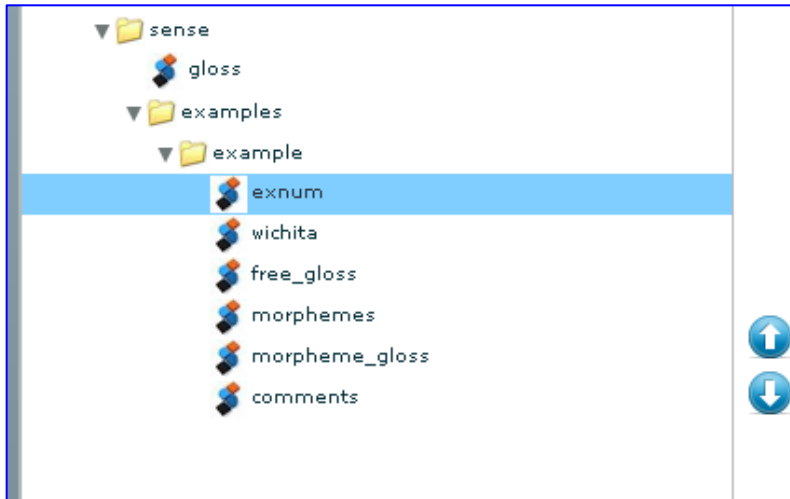
Previous Finish Cancel



# From Wichita XML to LMF

Wichita XML → LMF, points of discussion

1. Example is under Sense, but is all of it sense?

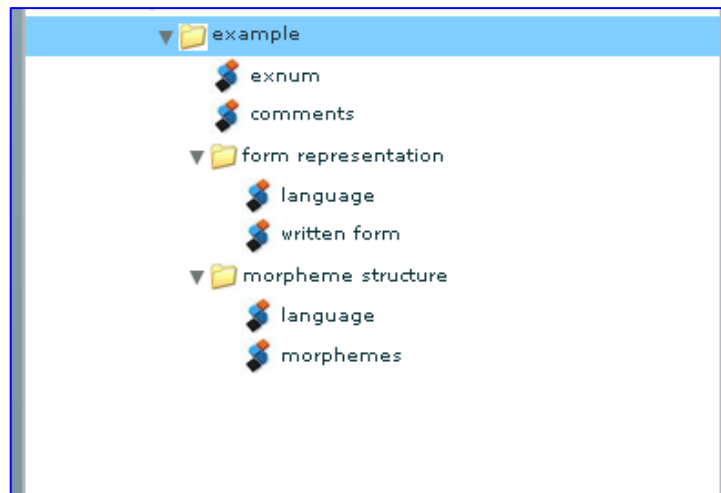
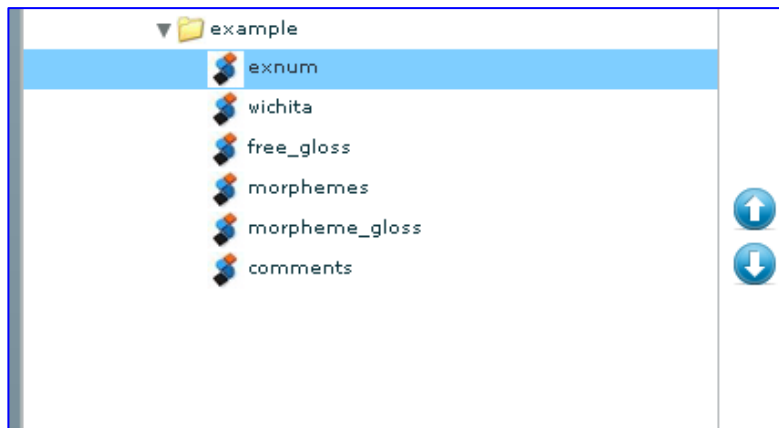




# From Wichita XML to LMF

Wichita XML → LMF, points of discussion

1. Example is under Sense, but is this sense?
2. Keep as one component? Or create sub-components?





# From Wichita XML to ISOcat

## Renaming data categories to ISOcat names in LEXUS:

The screenshot shows the LEXUS interface in Mozilla Firefox. The browser address bar displays the URL: `http://corpus1.mpi.nl/mpi/lexusDojo/Lexus.html#app=ae628bd0-selectedIndex=0&3997-selectedIndex=1&f6ca-selectedIndex=0`. The page title is "Wichita-exemple-group:". The left sidebar shows a "Lexicon structure" tree with "language" selected. The main content area displays the "ISOcat 12620 Data category registry" dialog box.

**ISOcat 12620 Data category registry**

ISO 12620 provides a framework for defining data categories compliant with the ISO/IEC 11179 family of standards. According to this model, each data category is assigned a unique administrative identifier, together with information on the status or decision-making process associated with the data category. In addition, data category specifications in the DCR contain linguistic descriptions, such as data category definitions, statements of associated value domains, and examples. Data category specifications can be associated with a variety of data element names and with language-specific versions of definitions, names, value domains and other attributes.

language name

identifier  name  data element name  definition  explanation  example  note

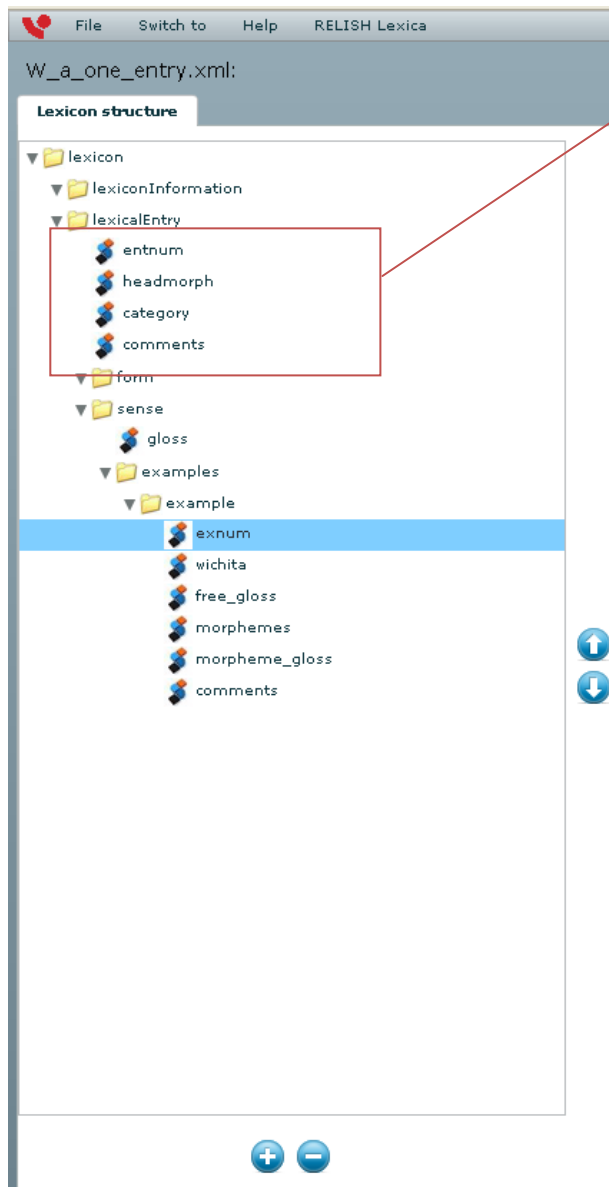
**Data categories found**

Registration authority: Max Planck Instituut voor Psycholinguïstiek, Nijmegen, The Netherlands

Name	Description	Version	Owner
Manobo languages	the alpha-3 (bibliographic) and alpha-2 (terminologic) code	1:0	SEBIRE, Philippe
Tupi languages	the alpha-3 (bibliographic) and alpha-2 (terminologic) code	1:0	SEBIRE, Philippe
Wakashan languages	the alpha-3 (bibliographic) and alpha-2 (terminologic) code	1:0	SEBIRE, Philippe
orthography name	Precision concerning the orthography	1:0	Francopoulo, Gil
Sami languages (Other)	the alpha-3 (bibliographic) and alpha-2 (terminologic) code	1:0	SEBIRE, Philippe
Nubian languages	the alpha-3 (bibliographic) and alpha-2 (terminologic) code	1:0	SEBIRE, Philippe
Australian languages	the alpha-3 (bibliographic) and alpha-2 (terminologic) code	1:0	SEBIRE, Philippe
thesaurus name	The title of a thesaurus from which a descriptor is taken.	1:0	Wright, Sue Ellen
Apache languages	the alpha-3 (bibliographic) and alpha-2 (terminologic) code	1:0	SEBIRE, Philippe
Mayan languages	the alpha-3 (bibliographic) and alpha-2 (terminologic) code	1:0	SEBIRE, Philippe
meta language	Name of the language that is used as a metalanguage in ...	1:0	Wittenburg, Peter
language name	A human understandable name of the language that is us...	1:0	Wittenburg, Peter
originating database name	A database treated as a document for the purpose of bibli...	1:0	Wright, Sue Ellen
Organisation	the name of an organisation	1:0	van Eerten, Laura
Person	the name of a person	1:0	van Eerten, Laura
language identifier	A unique identifier in a language resource entry that indica...	1:0	Wright, Sue Ellen
source language	Indicates if a language is a source language. (boolean)	1:0	Wittenburg, Peter
dominant language	Specifies the most frequently used language in a resource.	1:0	Wittenburg, Peter
tagset language	Indicates the language of the tag set itself, expressed in t...	1:0	Wittenburg, Peter



# From Wichita XML to ISOcat



entnum → Id, Identification of an element

headmorph → lemma Base form a word or term that is used as the formal entry in a dictionary

category → part of speech, Term used to describe how a particular word is used in a sentence.

comments → note, A statement that provides further information on any part of a language resource entry.



# From Wichita XML to ISOcat

File Switch to Help RELISH Lexica

W\_a\_one\_entry.xml:

**Lexicon structure**

- lexicon
  - lexiconInformation
  - lexicalEntry
    - entnum
    - headmorph
    - category
    - comments
  - form
  - sense
    - gloss**
  - examples
    - example
      - exnum
      - wichita
      - free\_gloss
      - morphemes
      - morpheme\_gloss
      - comments

entnum → Id, Identification of an element

headmorph → lemma, Base form a word or term that is used as the formal entry in a dictionary

category → part of speech, Term used to describe how a particular word is used in a sentence.

comments → note, A statement that provides further information on any part of a language resource entry.

gloss → gloss, A phrase or word used to provide a gloss or definition for some other word or phrase



# From Wichita XML to ISOcat

File Switch to Help RELISH Lexica

W\_a\_one\_entry.xml:

Lexicon structure

- lexicon
  - lexiconInformation
  - lexicalEntry
    - entnum
    - headmorph
    - category
    - comments
    - form
    - sense
      - gloss
      - examples
        - example
          - exnum
          - wichita
          - free\_gloss
          - morphemes
          - morpheme\_gloss
          - comments

exnum → rank, Reference to one specific element in an ordered list of elements

morphemes → morpheme, A morpheme is the smallest meaningful unit in the grammar of a language

comments → note, A statement that provides further information on any part of a language resource entry



# From Wichita XML to ISOcat

W\_a\_one\_entry.xml:

Lexicon structure

- lexicon
  - lexiconInformation
  - lexicalEntry
    - entnum
    - headmorph
    - category
    - comments
  - form
  - sense
    - gloss
  - examples
    - example
      - exnum
      - wichita
      - free\_gloss
      - morphemes
      - morpheme\_gloss
      - comments

↑

↓

+

-

http://corpus1....selectedIndex=0

Wichita corrected glottal stop:

Lexicon structure

- lexicon
  - lexiconInformation
  - lexicalEntry
    - id
    - part of speech
    - note
  - form
    - lemma
  - sense
    - gloss
  - examples
    - example
      - rank
      - wichita
      - free\_gloss
      - morpheme
      - morpheme\_gloss
      - note

Schema element

General informati

Reference:

Name:

Description:

Admin Info:

Mandatory:

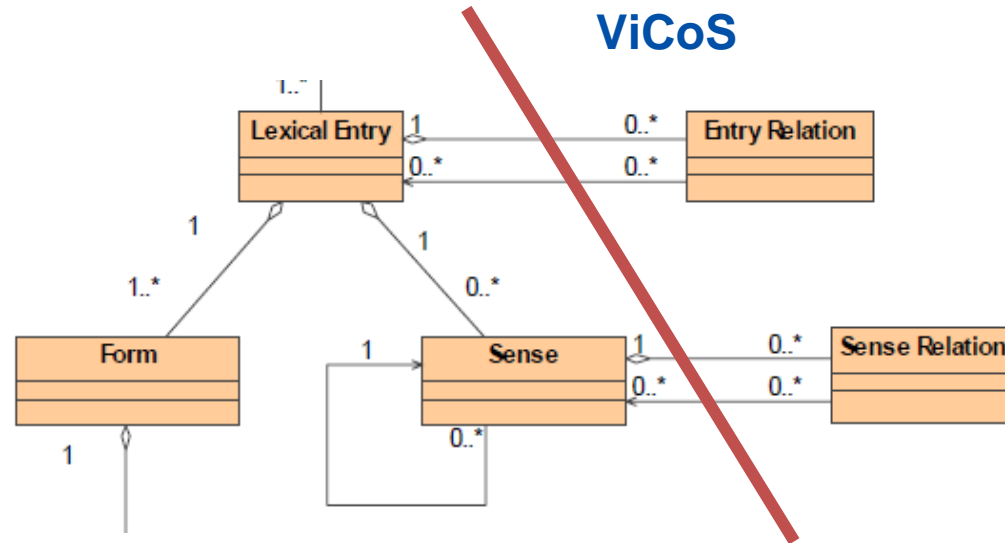
Multiples allowed:

Value domain:

Sort order

Sort order







## Relation between headmorph and examples

ViCoS Editor and Navigator - Mozilla Firefox

File Edit View History Bookmarks Tools Help New Window

Back Forward Reload Stop Home [http://corpus1.mpi.nl/mp/vicos/ViCoS\\_Browser.html](http://corpus1.mpi.nl/mp/vicos/ViCoS_Browser.html)

Most Visited LAMUS - Language Ar... IMDI Browser Welcome to the Max P... Tools - Language Arc... Annex embedding - ... Quick Store and View ... LAT News CLARA

<http://corpus1.mpi...14-selectedIndex=0> <http://corpus1.mpi...xusDojo/Lexus.html> ViCoS - Visualising Conceptual Spaces ViCoS Editor and Navigator

is example of ● ● ● ● ●  
is\_a\_kind\_of ———  
:not\_a\_kind\_of ● ● ● ● ●  
is\_part\_of ● ● ● ● ●  
is\_antonym\_of ● ● ● ● ●  
is\_synonym\_of ● ● ● ● ●  
is\_related\_to ● ● ● ● ●  
is\_caused\_by ● ● ● ● ●

Modes:

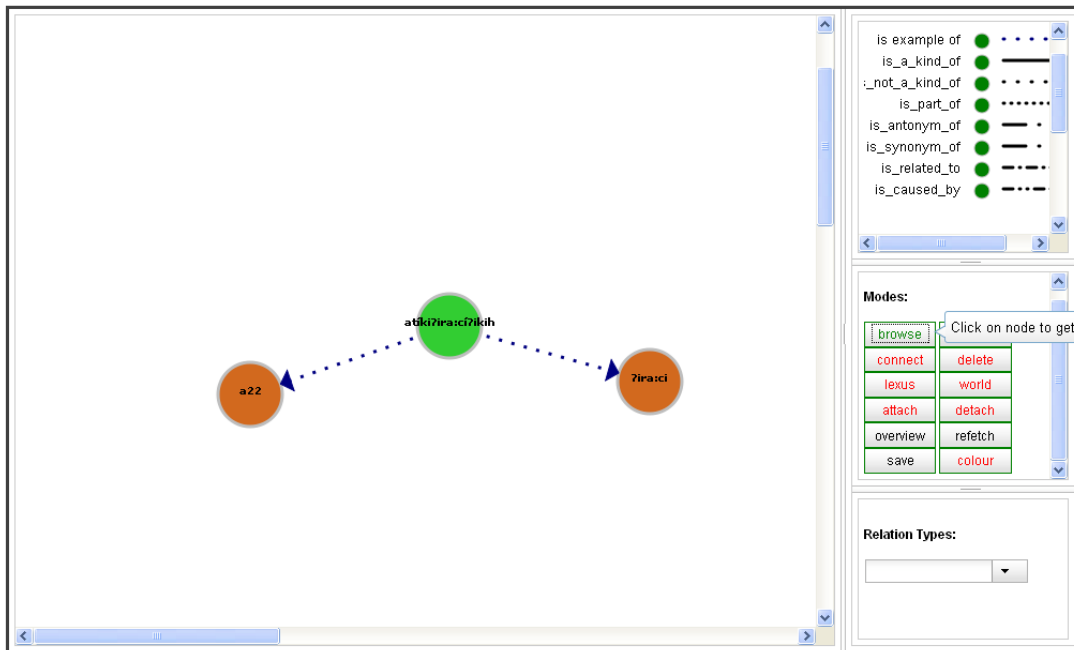
browse Click on node to get m  
connect delete  
lexus world  
attach detach  
overview refetch  
save colour

Relation Types:

▼

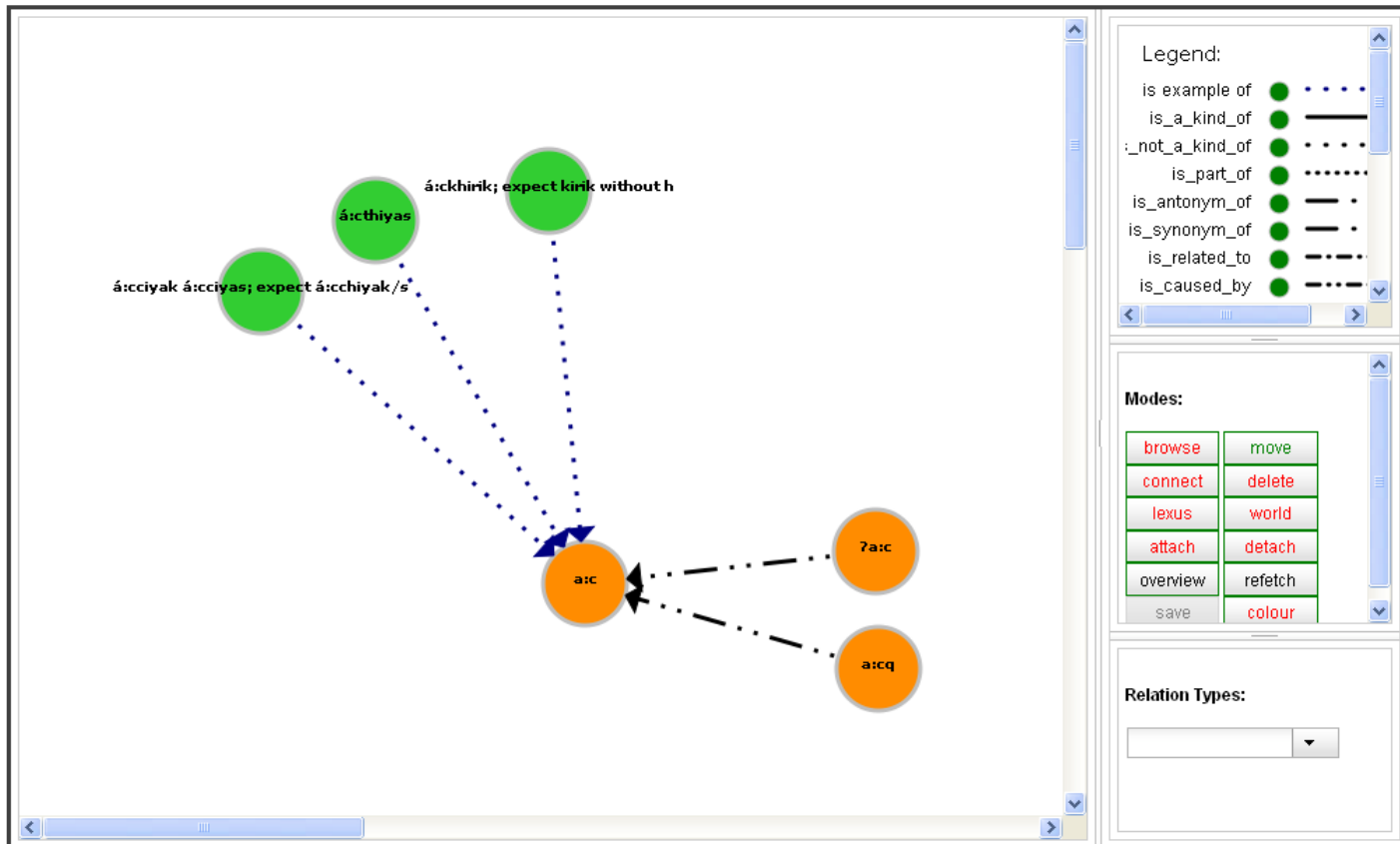


## Relation between and example and lemmas (headmorph)





## Relation between and headmorphs (lemmas)



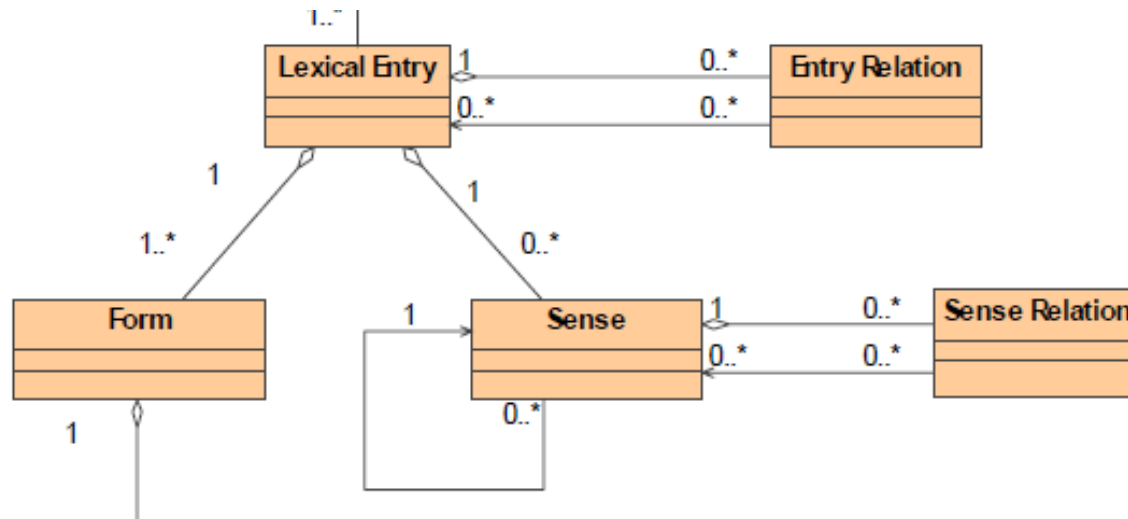


## Points of discussion





# How to handle the Wichita Challenges in LEXUS?



**LexicalEntry:** What should be used as the “entry”?

**Form:** text string representing the word ... but which word?

**Sense:** specifies the meaning and context ...



### Wichita Challenge 1: Headword for verbs?

1. Does LMF offer a solution?

Not really .... (Because it is a linguist dilemma)

2. Does LEXUS offer a solution?

Wordlist are user definable

ViCoS browsing by example sentences, or senses



### Wichita Challenge 2: Word = Grammar ?

1. Does LMF offer a solution?

Grammar and meaning separated, but can be related

2. Does LEXUS offer a solution?

Different components for Grammar and Sense

Its up to the linguist to decide what is the “headword”

ViCoS browsing!





Enhance interoperability through:

1. Standardizing structure
2. Harmonizing element naming, and referencing

Why interoperable?

1. Cross lexica search on equal data categories
2. Merging

Interoperable with what:

1. Other LMF/ISOCat lexica