The Wichita lexicon in LEXUS

Armik Mirzayan

University of Colorado at Boulder

Jacquelijn 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 3rd person form of /ʔarasi/ (“cook”)

PREFIXES	SUFFIXES			
	perfective *(-Ø)*	imperfective *(-s)*	intentive *(-staris)*	habitual *(--ss)*
aorist *a...ki-*	‘She cooked it.’ *áka?árasiki*	‘She was cooking it.’ *áka?arásis*	‘She was going to cook it, but didn’t.’ *áka?arásistaris*	‘She always used to cook it.’ *áka?arásiki·ss*
aorist quotative *a·?a...ki-*	‘I heard that she cooked it.’ *á·ka?árasiki*	‘I heard she was cooking it.’ *á·ka?arásis*	‘I heard she was going to cook it.’ *á·ka?arásistaris*	‘I heard she always used to cook it.’ *á·ka?arásiki·ss*
future *ke?e-*	‘She will cook it.’ *ke·?árasiki*	‘She’ll be cooking it.’ *ke·?árasis*	_____	‘It will be her job to cook it every time.’ *ke·?árasiki·ss*
future quotative *ehe-*	‘I heard she’ll cook it.’ *ehè·?árasiki*	‘I heard she’ll be cooking it.’ *ehè·?árasis*	_____	‘I heard it will be her job to cook it every time.’ *ehè·?árasiki·ss*

A “Linguist’s” XML Wichita Dictionary

```
1 a*raskakic7a iye:*riwa:ce:s7i i8ssinno:kha:*r7ih
2 +dry meat they had lots of that kind.
3 7aras kakic 7a2 i4 iy ru riwa:c re:hi s2 7i issiri7 na wakhahr 7i h xx
4 meat dry nounf direc nfocagt poss big beinplacepl impf be causei thatkind ppl activpatq be subp
5 bhc61
==7arasi
1 wa*tiwa:rassis
2 they are cooking now; they are really ripe
3 wa7 ta i2 wa:rassi s2 7arasi wa:rasi wa:
4 already pres pfocagt cookpl impf cook ripe distrib xx
5 1973-25
==7arasi
1 wa*ta7arass
2 it is ripe
3 wa7 ta i2 7arasi s2
4 already pres pfocagt cook impf causei ripe
5 1973-25
==7arasi
1 wa*ke:7aras
2 they will get ripe
3 wa7 ke7 i2 xx 7arasi
4 already fut pfocagt xx cook ripe
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>
            <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) ...
   $\Rightarrow$ 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?

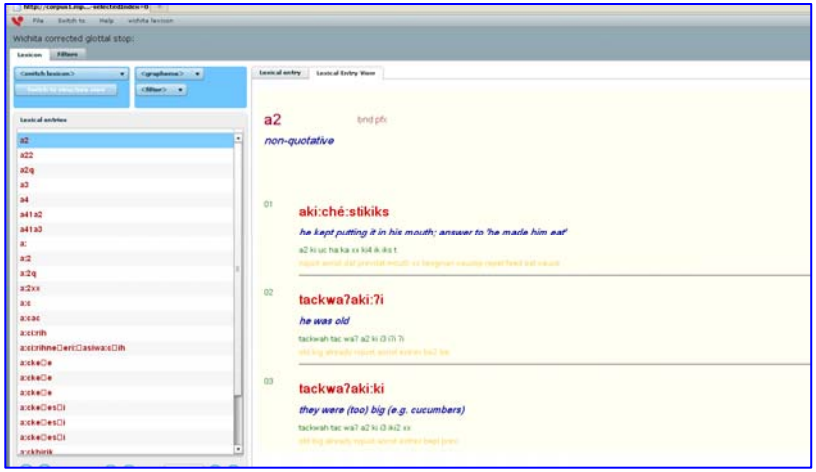
Which different prefixes with a given verb root count as separate entries?

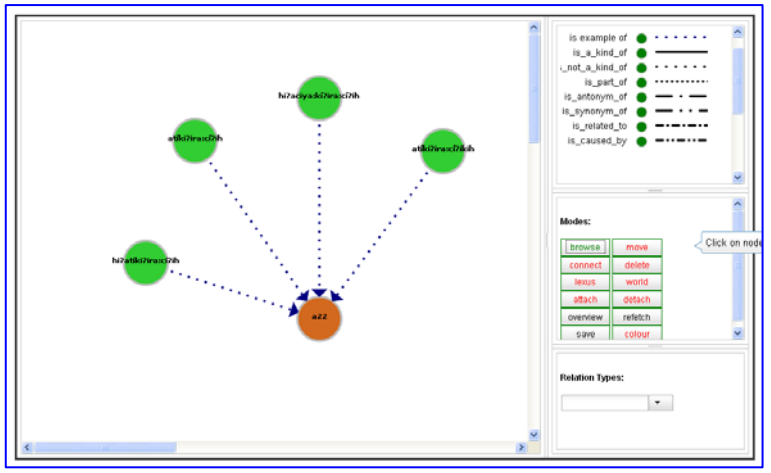
How about morphemes like *re:R-* (function of a nominal inflection but coded as a verbal prefix)?

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





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

LEXUS

Based on two ISO TC 37 standards for linguistic resources

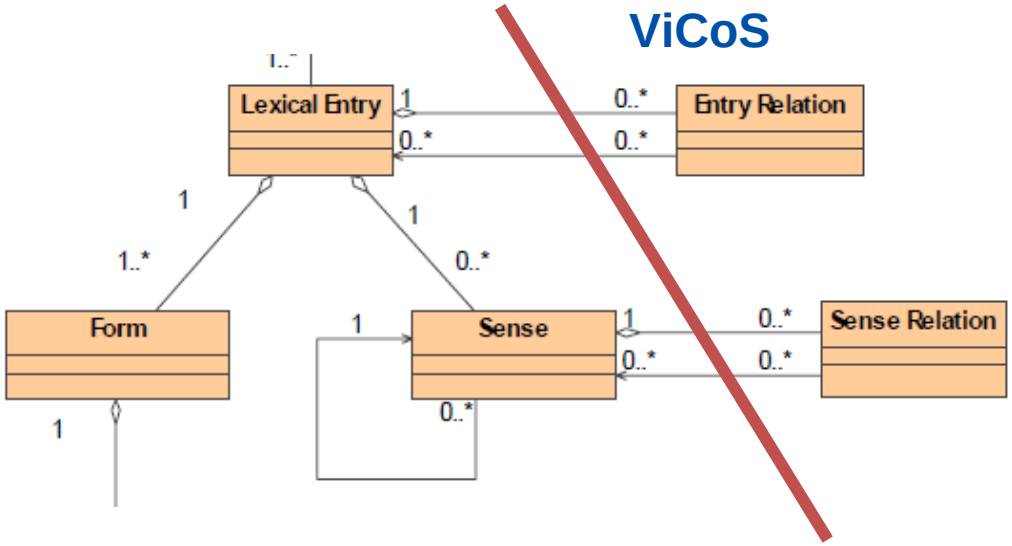
LMF: Lexical Markup Framework (lexicon structure)

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

LEXUS

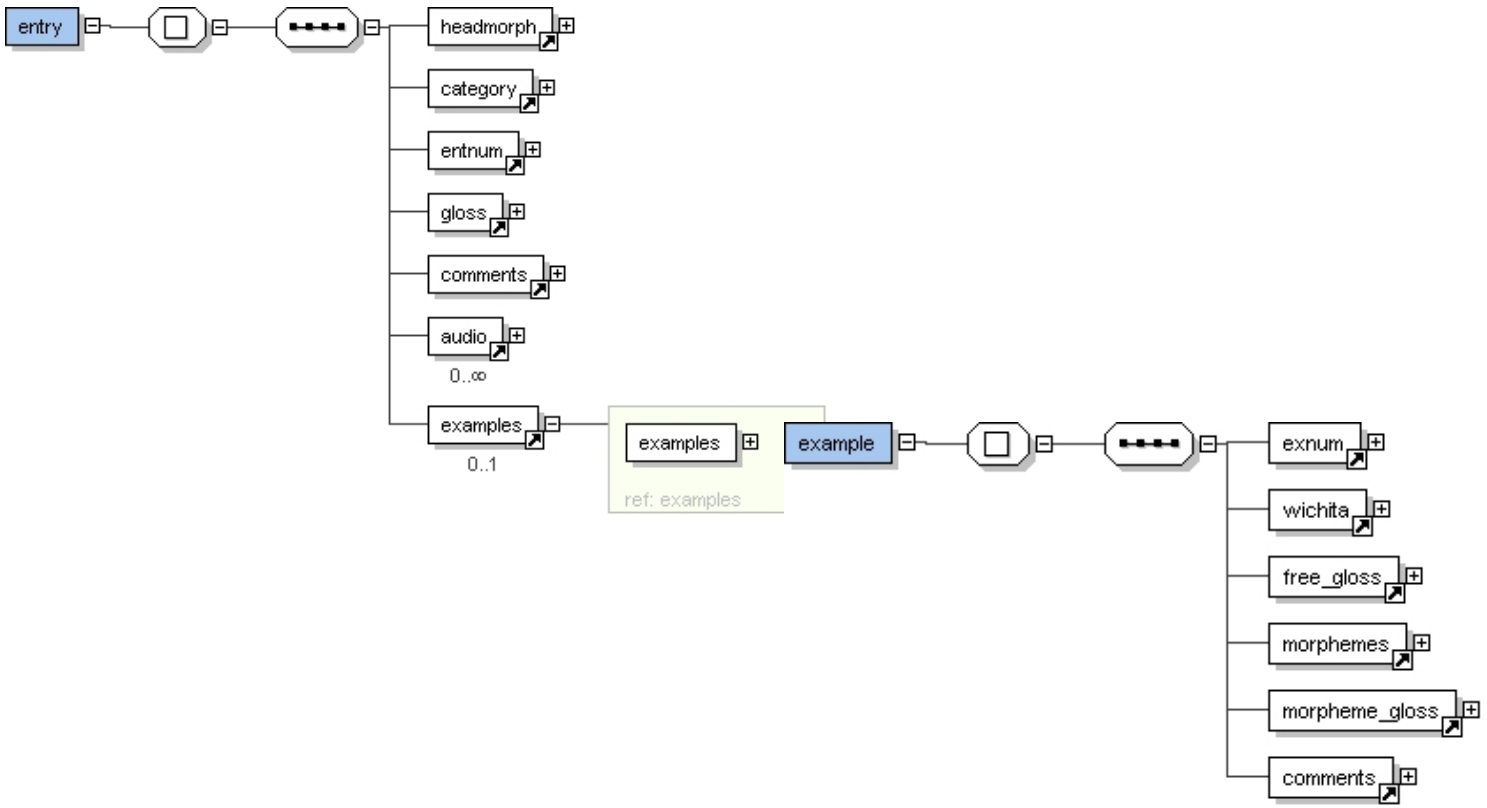


LexicalEntry: container for managing one or several forms and possibly one or several meanings in order to describe a lexeme

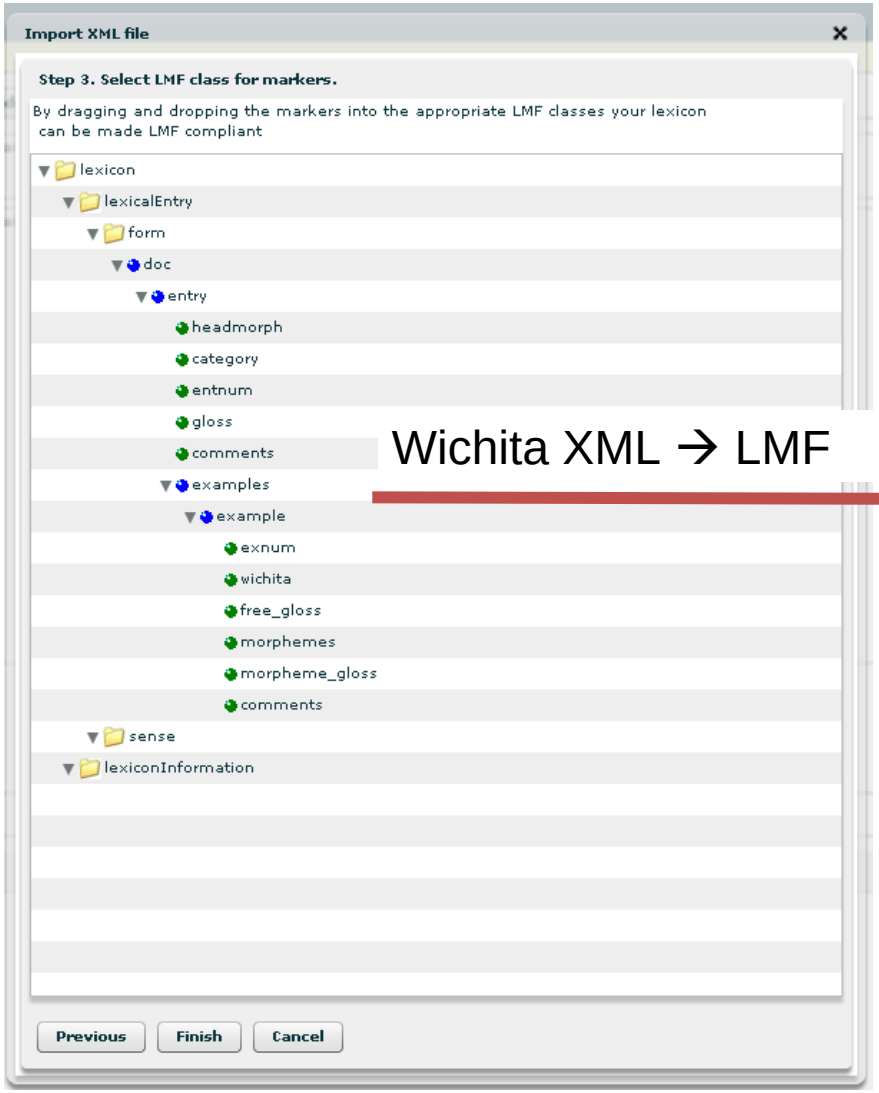
Form: text string representing the word

Sense: specifies the meaning and context

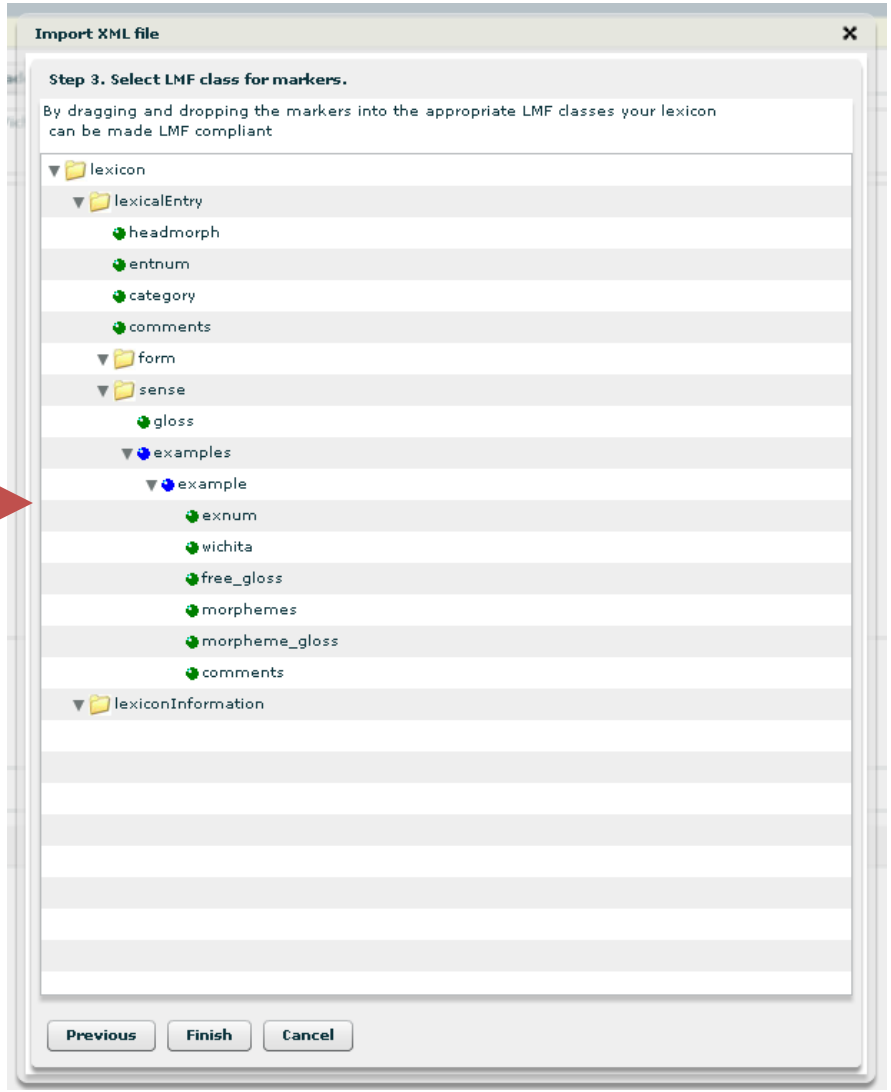
Wichita XML elements and structure:



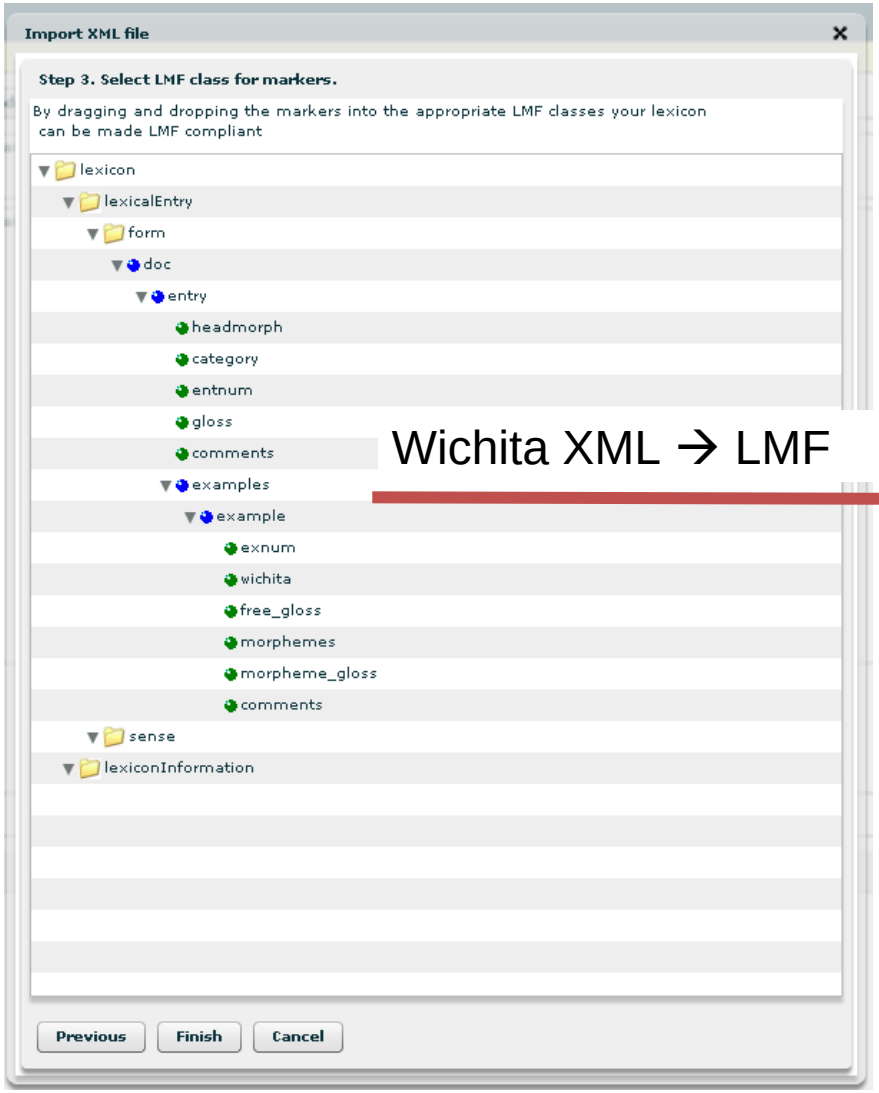
From Wichita XML to LMF



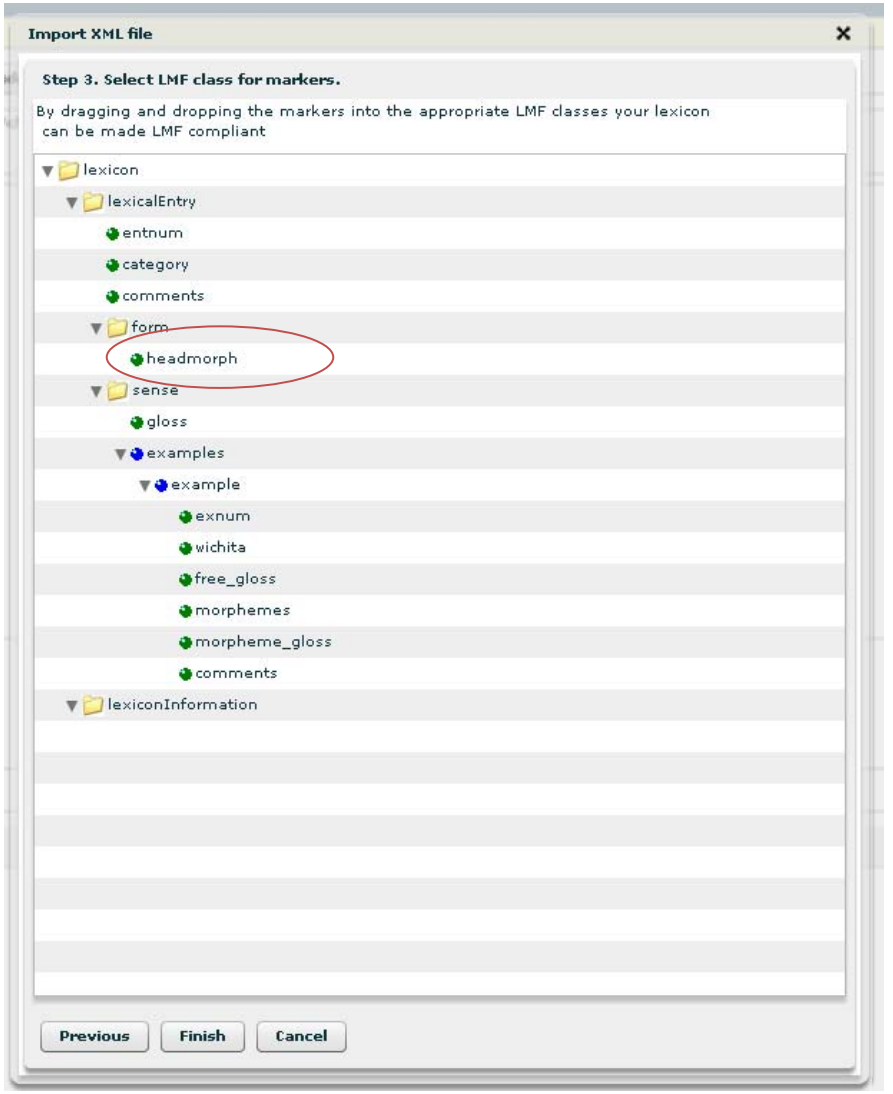
Wichita XML → LMF



From Wichita XML to LMF



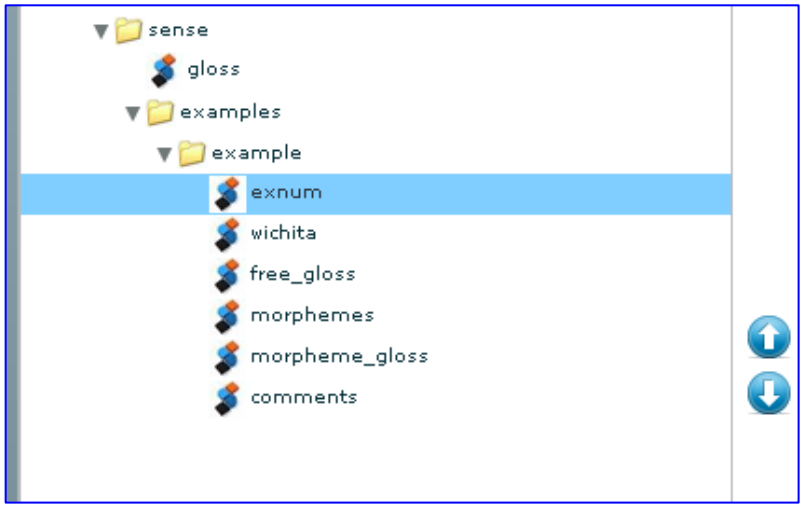
Wichita XML → LMF



From Wichita XML to LMF

Wichita XML → LMF, points of discussion

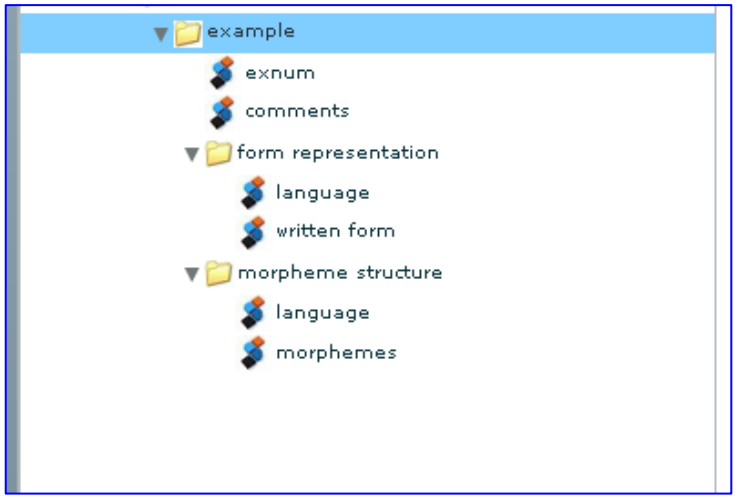
1. Example is under Sense, but is all of it sense?



From Wichita XML to LMF

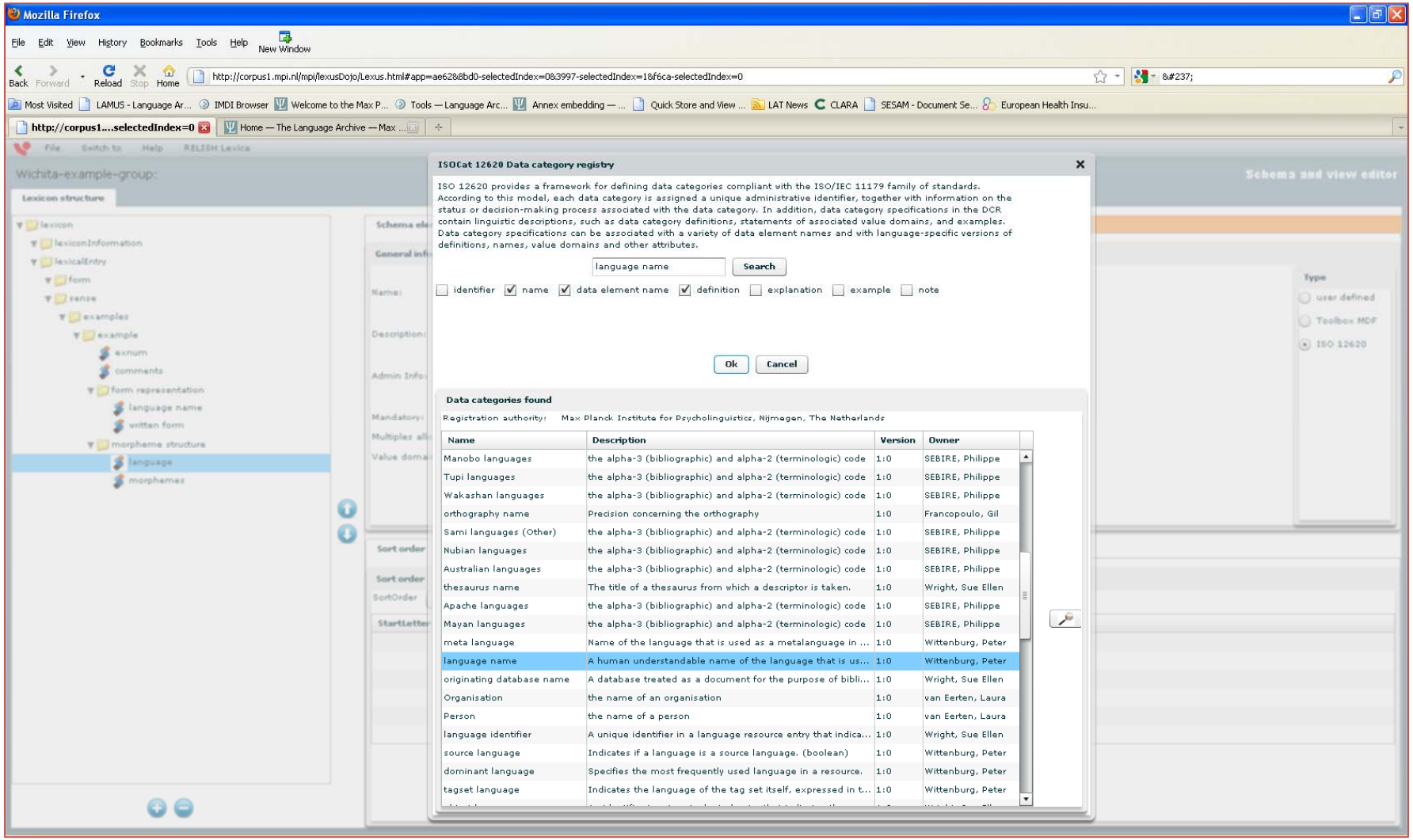
Wichita XML → LMF, points of discussion

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

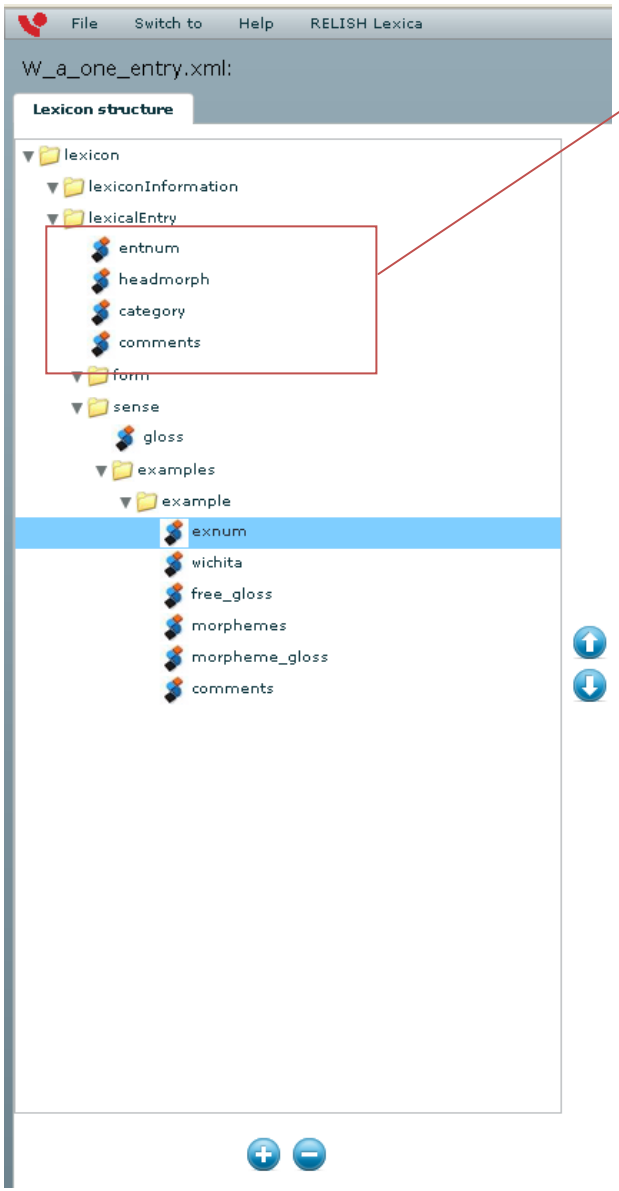


From Wichita XML to ISOcat

Renaming data categories to ISOcat names in LEXUS:



From Wichita XML to ISOcat



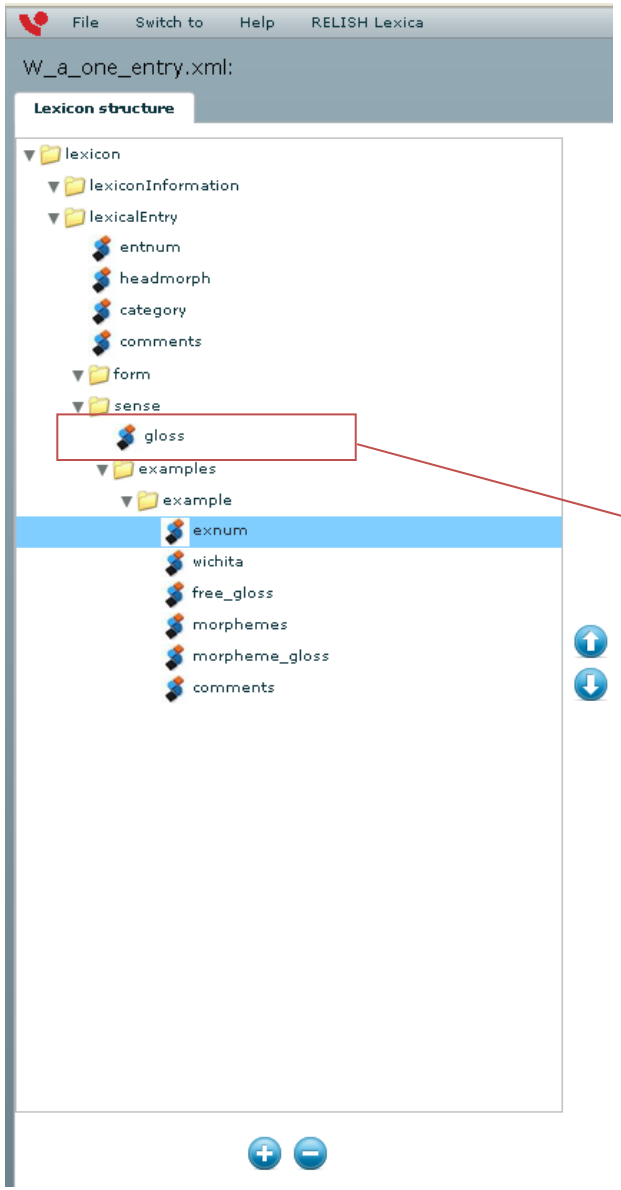
entnum→ Id, Identification of an element

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



entnum→ Id, Identification of an element

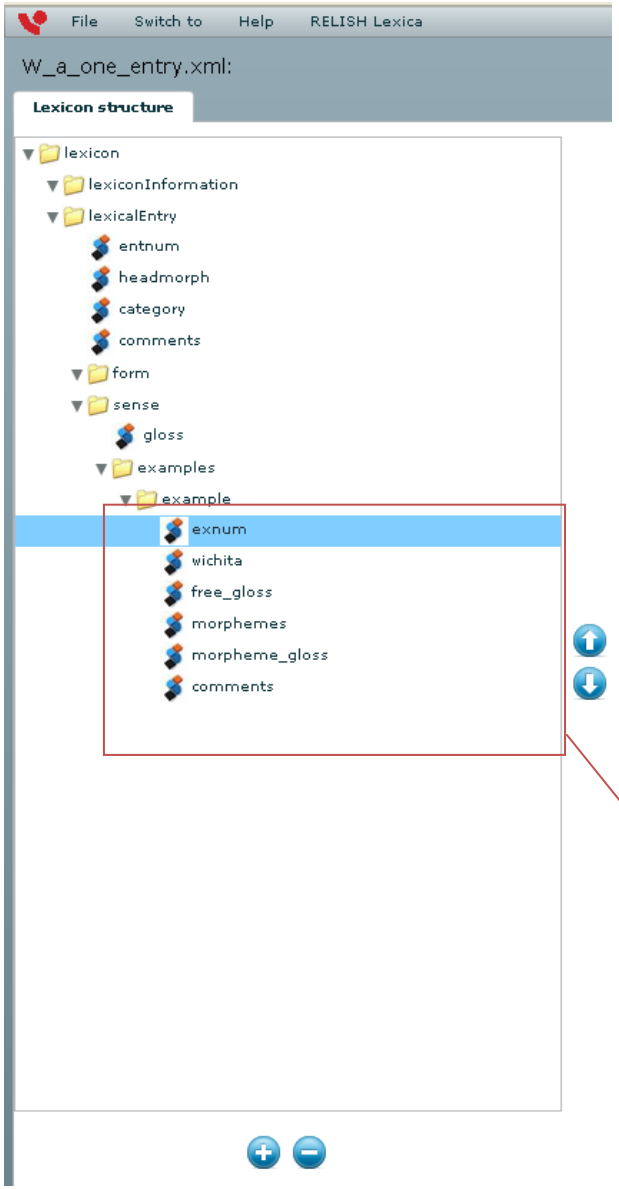
headmorph → lemma, Base form a word or term that is used as the formal entry in a dictionary

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

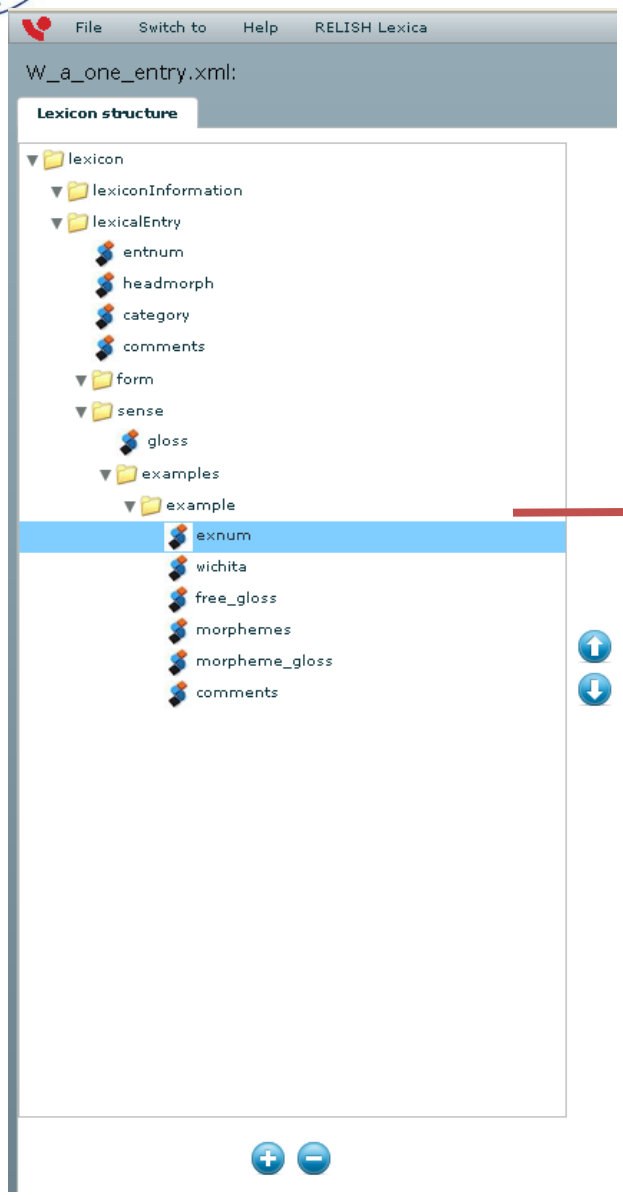


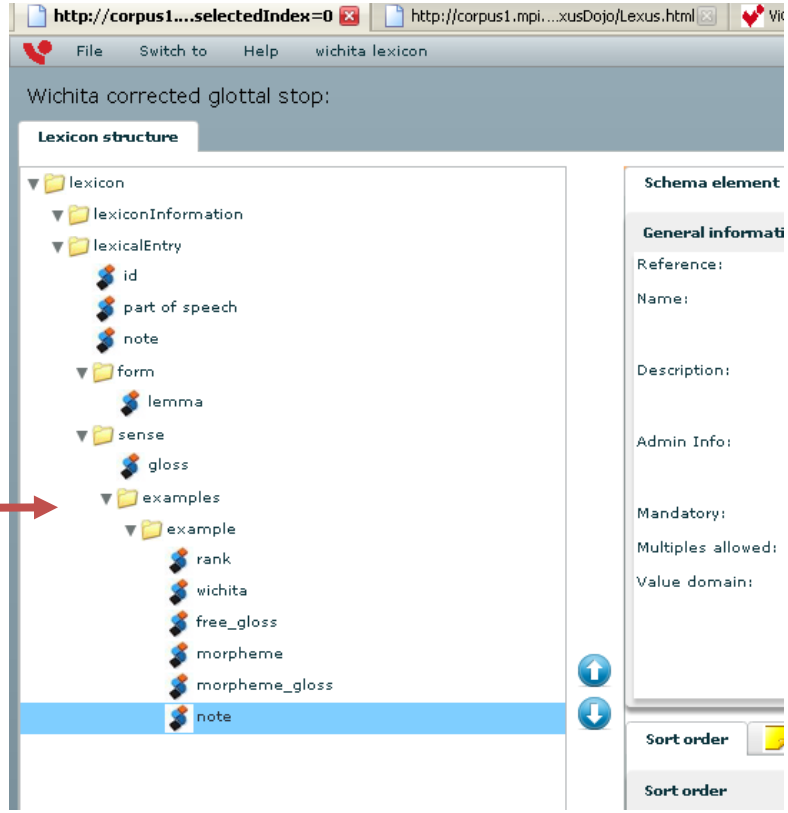
exnum→ rank, Reference to one specific element in an ordered list of elements

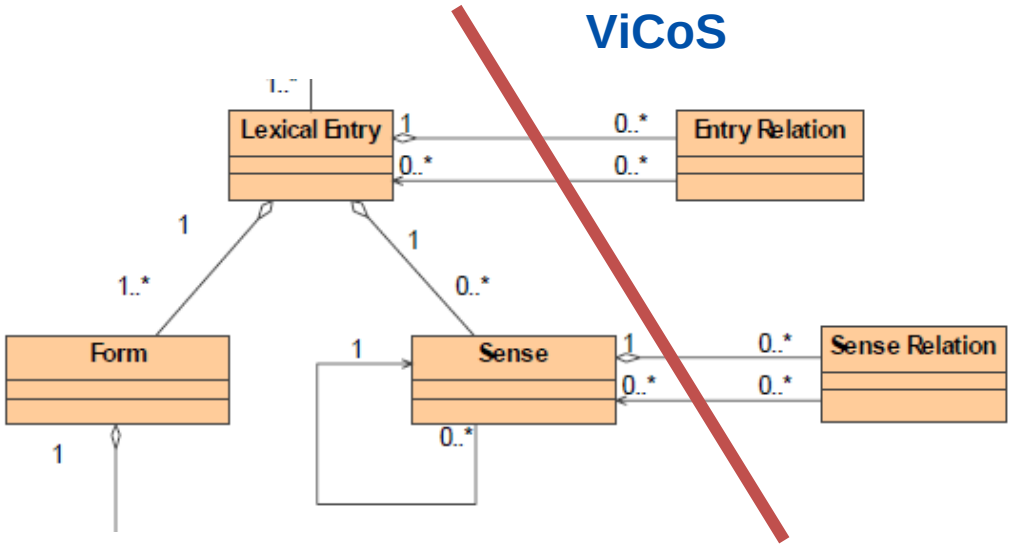
morphemes → morpheme, A morpheme is the smallest meaningful unit in the grammar of a language

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

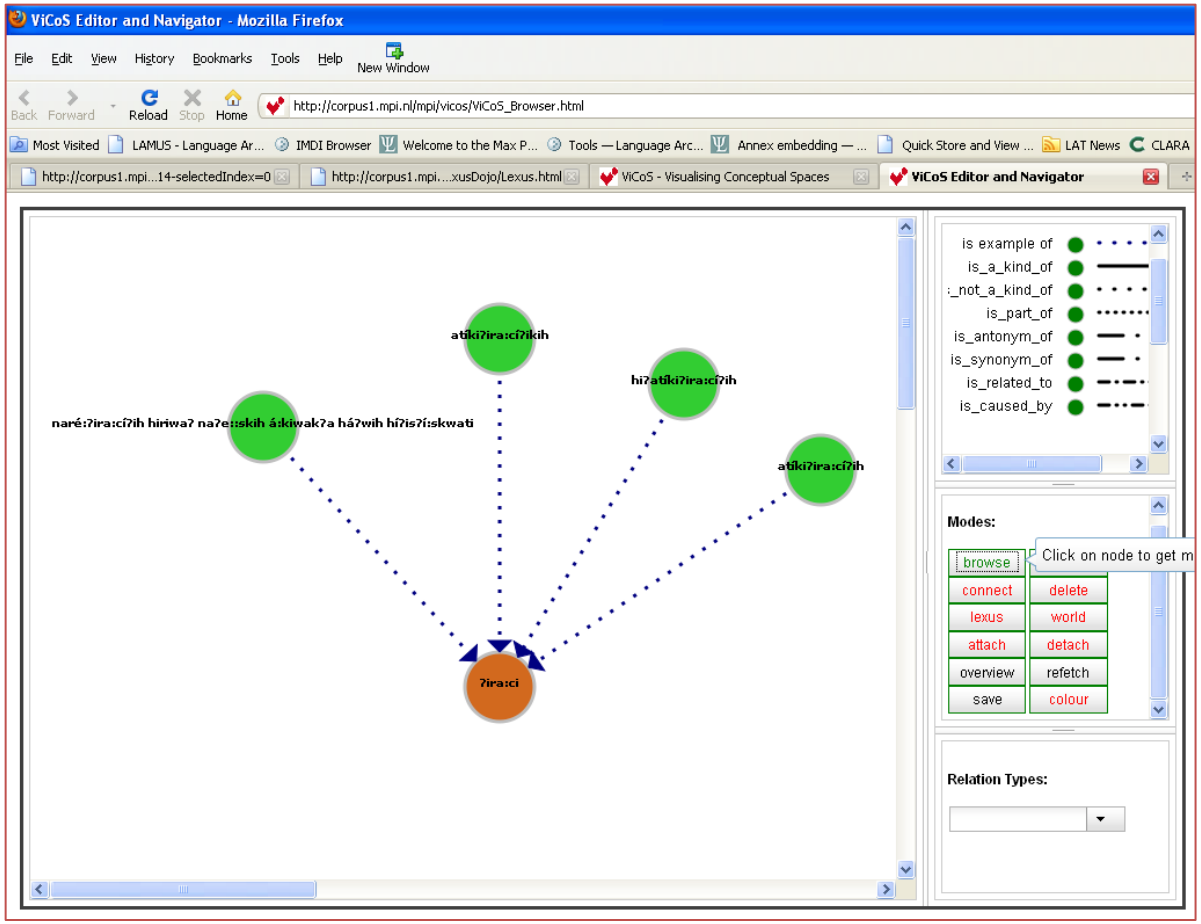
From Wichita XML to ISOcat



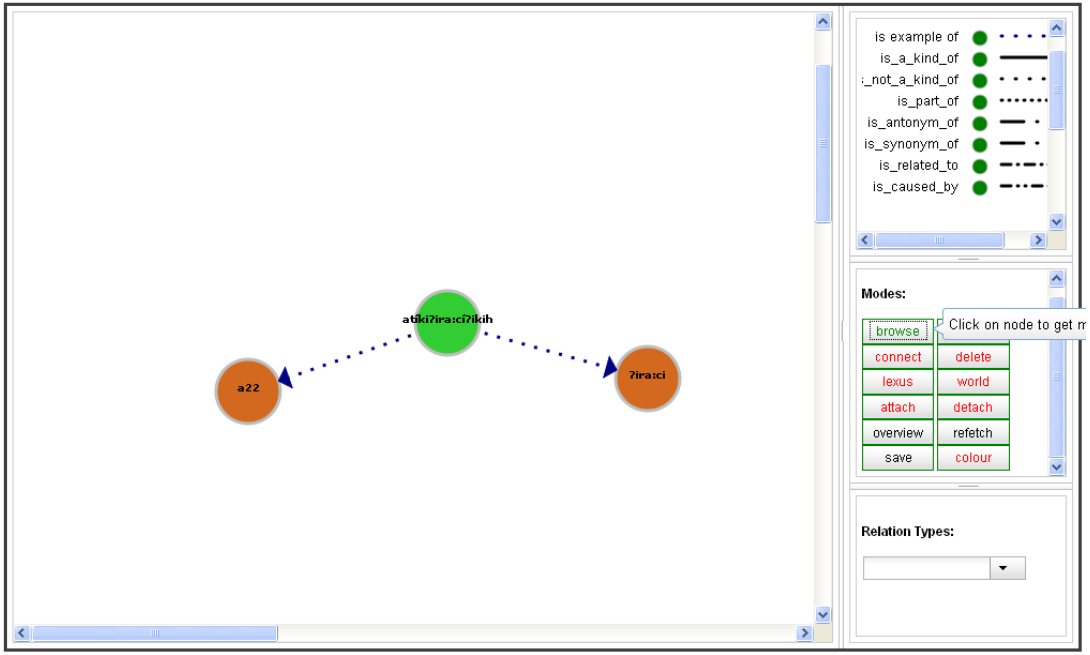



ViCoS
Lexical Entry
Entry Relation
Form
Sense
Sense Relation
1
0..*
1..*

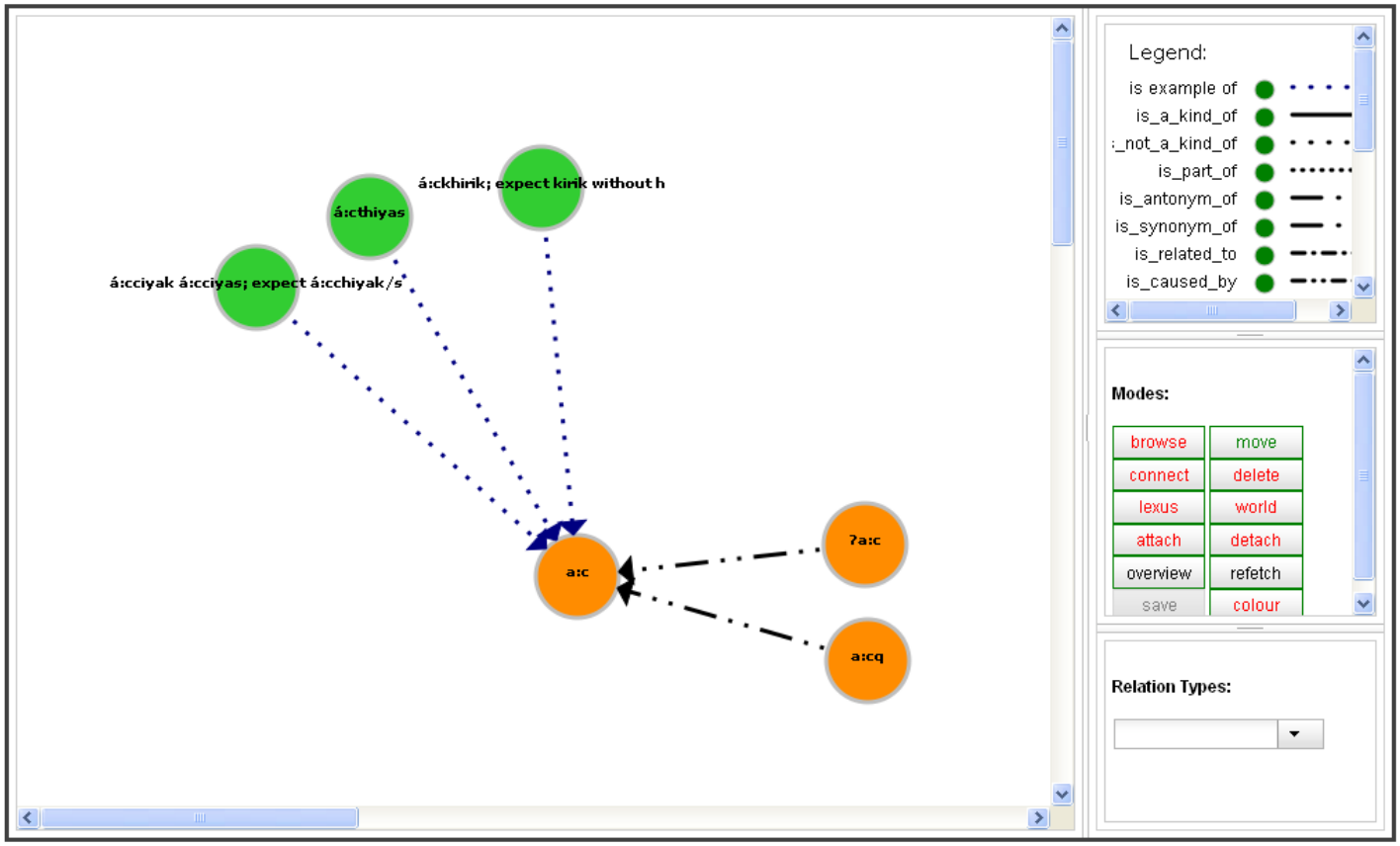
Relation between headmorph and examples



Relation between and example and lemmas (headmorph)

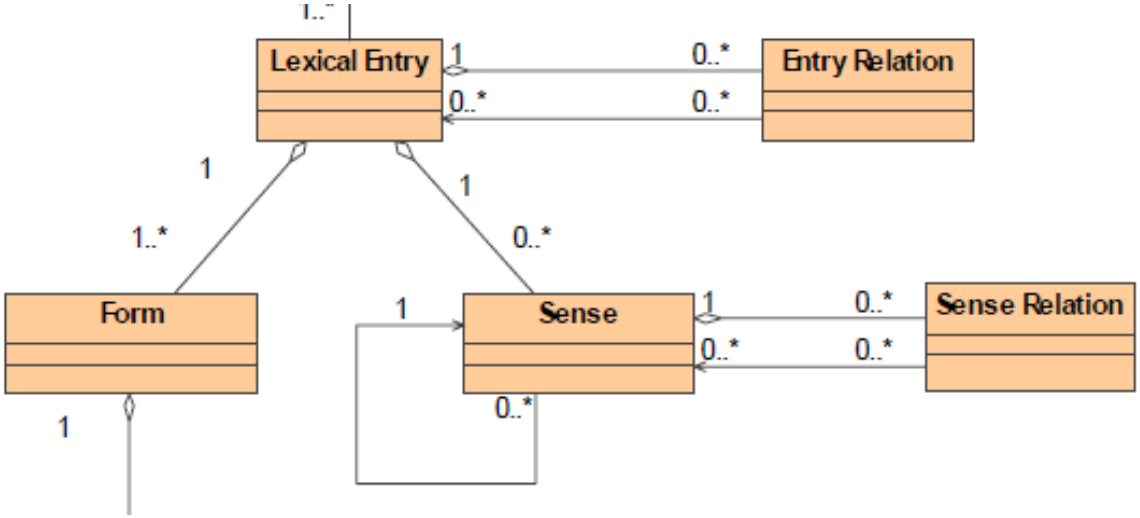


Relation between and headmorphs (lemmas)



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: LMF and ISOcat

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: LMF and ISOcat

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!

Summary LMF and ISOCat:

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