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
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

- 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.


other prefix positions \{preverb, locatives, dative, noun class, \ldots\}
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")

<table>
<thead>
<tr>
<th>PREFIXES</th>
<th>SUFFIXES</th>
<th>perfective (-Ø)</th>
<th>imperfective (-s)</th>
<th>intensive (-staris)</th>
<th>habitual (-ss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>aorist</td>
<td></td>
<td>‘She cooked it.’</td>
<td>‘She was cooking it.’</td>
<td>‘She was going to cook it, but didn’t.’</td>
<td>‘She always used to cook it.’</td>
</tr>
<tr>
<td>a...ki-</td>
<td></td>
<td>d₂kaʔarasiki</td>
<td>d₂kaʔarásis</td>
<td>d₂kaʔarásisarís</td>
<td>d₂kaʔarasīkiss</td>
</tr>
<tr>
<td>aorist</td>
<td></td>
<td>‘I heard that she cooked it.’</td>
<td>‘I heard she was cooking it.’</td>
<td>‘I heard she was going to cook it.’</td>
<td>‘I heard she always used to cook it.’</td>
</tr>
<tr>
<td>quotative</td>
<td></td>
<td>d₂kaʔarasiki</td>
<td>d₂kaʔarásis</td>
<td>d₂kaʔarásisarís</td>
<td>d₂kaʔarasīkiss</td>
</tr>
<tr>
<td>aʔa...kt-</td>
<td></td>
<td>d₂kaʔarasiki</td>
<td>d₂kaʔarásis</td>
<td>d₂kaʔarásisarís</td>
<td>d₂kaʔarasīkiss</td>
</tr>
<tr>
<td>future</td>
<td></td>
<td>‘She will cook it.’</td>
<td>‘She’ll be cooking it.’</td>
<td></td>
<td>‘It will be her job to cook it every time.’</td>
</tr>
<tr>
<td>keʔe-</td>
<td></td>
<td>keʔarasiki</td>
<td>keʔarásis</td>
<td></td>
<td>keʔarasíkiss</td>
</tr>
<tr>
<td>future</td>
<td></td>
<td>‘I heard she’ll cook it.’</td>
<td>‘I heard she’ll be cooking it.’</td>
<td></td>
<td>‘I heard it will be her job to cook it every time.’</td>
</tr>
<tr>
<td>quotative</td>
<td></td>
<td>ehèʔarasiki</td>
<td>ehèʔarásis</td>
<td></td>
<td>ehèʔarasíkiss</td>
</tr>
</tbody>
</table>
A “Linguist’s” XML Wichita Dictionary

Wichita Database (Rood) → XML (2006)
Wichita Challenge 1: Headwords?

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

Root: /tarʔa:ti/ ‘cure, doctor’

ti:ckiciyé:sʔastarʔa:c
“he is doctoring some dogs” (source 1973-20)

ta- i- uc- kiciye:- s- ?ak- tarʔa:ti -s
pres-pfocus-prev.dat-dog-inc-patns-doctor-impf
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) ... 
   \[=> \text{then *all* verb entries start with } /t/!\]

2. Use a nominalized form (participle) ... 
   \[\Rightarrow \text{again, *all* verb entries start with } /n/!\]

3. Use another tense/mode form (other complexities ...)

4. Decide on a verb-by-verb basis (?).
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:rʔicʔi-hrih
imper-2.sub-loc.on-sit  ppl-1.sub-shoulder-be-loc
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
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
From Wichita XML to LMF

Wichita XML → LMF
Wichita XML → LMF, points of discussion

1. Example is under Sense, but is all of it sense?
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.
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
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
Relation between headmorph and examples
Relation between and example and lemmas (headmorph)
Relation between and headmorphs (lemmas)
Points of discussion
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!
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