

TILR Group 2: Lexicon schemas and related data models

Trippel, Maxwell, Corbett, Prince, Manning,
Grimes, Moran

Overview

- Lexicon schemas:
 - comparison
 - used cases
 - transformation
- De Facto standards
 - what are we looking for
 - which one would we adopt

Use cases: What is a lexicon?

- DVD based lexicons
- printouts
- machine learning NLP
- lexicon management tool
- webbased
- merging lexicons
- print dictionaries
- reusable for electronic purposes
- DATR
- LMF
- LIFT
- MILE
- Toolbox/ Shoebox/
MDF/ FLEX

Purpose

- Writing a lexicon
- interchange format

DATR

- works!
 - application
 - programming language describing lexicons
 - defining theories
 - testing hypothesis based on the compiler
 - steep learning curve
- sdatr --> statistical datr
 - not maintained???
 - as interchange format? Yes, but no inheritance, why?
 - nested structures when interchanging
 - no syntactic checking
 - creating flat dumps

MILE

- inheritance model
- enforces checking
- steep learning curve

LIFT

- original problem:
 - shoebox: non-Roman font issue
 - Standard Data Format (SDF) implies some structure not explicitly stated, given in order of datcats
- supports ontologies
- tree structures but also back references

Use cases

- (morphological) lexicon --> reuse in parser
(morphological/syntactic) parser
- inducing sound change patterns from a lexicon based
- based on lexicons including meaning

Issues

- Data category selection
 - but interchange of storage
 - datcats need to be listed in interchange file
 - Implicit: application in Unicode
 - representation of signs of signlanguages?
 - legacy data
 - dependencies management (e.g. fonts)
 - lossless conversion (roundtrip)
 - commonality: the core of the intersection
 - flat structure for interchange
 - references to material out there
- ambiguity
 - implication of hierarchies to be avoided
 - extensible list of datcats
 - reuse of standards for example Feature Structure Representation
 - it has to work --> lexicon structure
 - lexicons imply a lot of grammar
 - paradigm classes
 - interchange format
 - multiple writing systems
 - multiple glossing systems
 - implementable

Problem

- non monotonic inheritance
- full expandable
 - full form lexicon in the exchange
 - non expanded form
- verbose data format
- easy to understand, iconicity -->long time archiving
- relation to ontologies and datcat repositories

Working plan

- design of testsets
- implementation of testsets according to our understanding
- emailing the standards authors on help
 - where we got stuck
 - evaluating our understanding

Test case requirements

- prose description
- attempt of encode in one of the "standard" formats
- examples in Unicode
- electronically available for cut and paste
- comma separated items

Test cases

- Russian (Greville Corvett)
- complex morphology (Mike Maxwell)
- semantic example, systematic similarity / polysemy (Steve Moran)
- multiple gloss/multiple writing example (Cambell Price)
- nested feature structure (Mike Maxwell)

Russian lexicon entry: DATR

- 274 komnata

<> == N

<gloss> == 'room'

<stem> == komnat

<mor class> == II.

- i.e. komnata, N, II, fem, 'room'.

- Inherits from mor class II the inflections

– N komnata

– A komnatu

– G komnaty

– D komnate

– I komnatoj

– L komnate

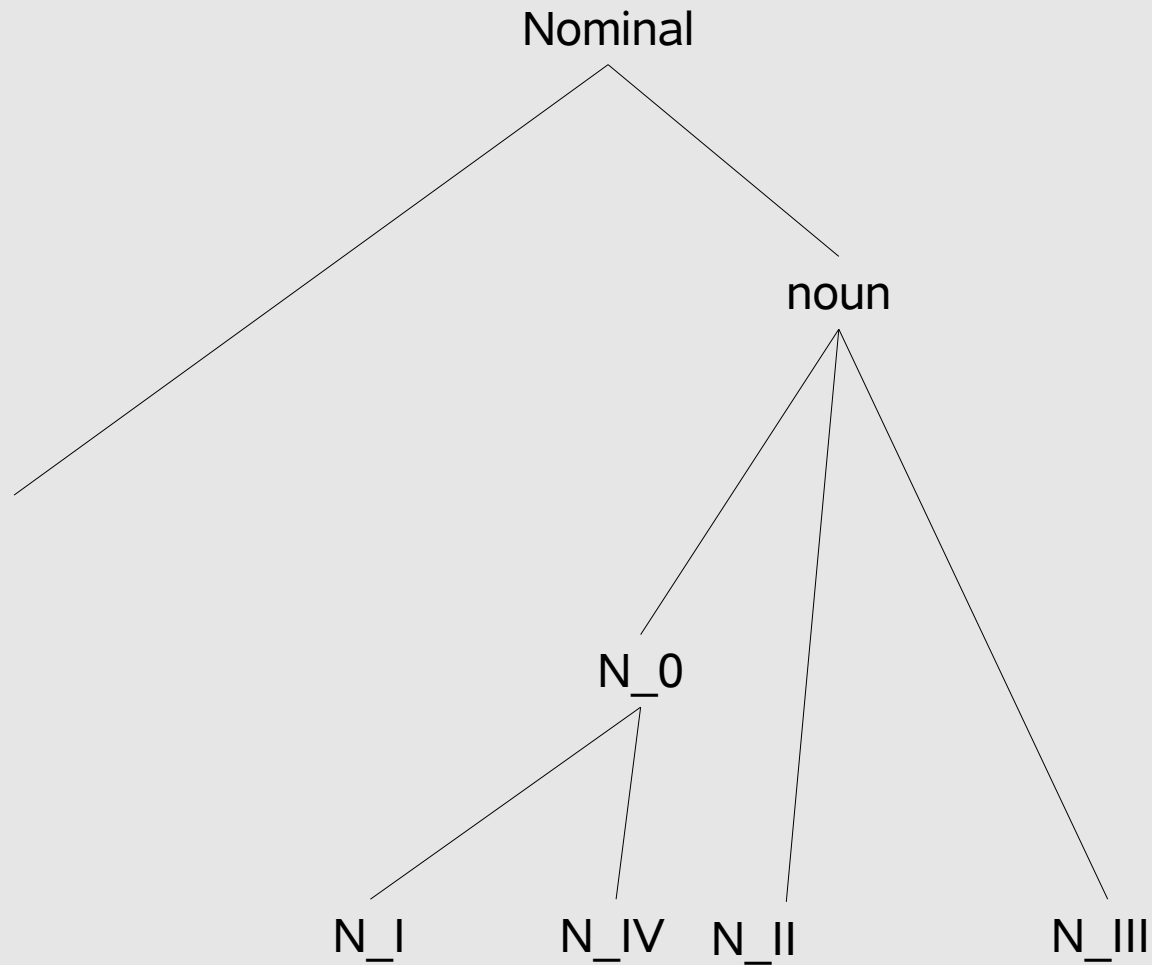
- Inherits feminine noun

...

Russian Example: LIFT

```
<?xml version="1.0" encoding="UTF-8"?>
<?oxygen RNGSchema="file:/home/ttrippel/emeld/tilr/lift.rng" type="xml"?>
<lift>
  <entry id="komnata274">
    <lexical-unit>
      <form lang="latin">
        <text>komnata</text>
      </form>
      <grammatical-info type="N" stem="kamnat"
gender="fem"></grammatical-info>
      <gloss lang="en"><text>room</text></gloss>
      <trait name="mor class" value="2"></trait>
      <reference type="m" ref="3">
        <range name="mor class"></range>
      </lexical-unit>
    </entry>
  </lift>
```

Inheritance Model



Complex morphology

- to be provided by Mike Maxwell

Semantic example: polysemy

- Cherry:

<> == NOUN

<mor root> == cherry

<sem gloss i> == sweet red berry with pip

<sem gloss 2> == tree bearing <sem gloss i>

<sem gloss 3> == wood from <sem gloss 2>.

Multiple gloss/multiple writing example

- to be provided by Cambell Price

Nested Feature Structure

$$\begin{bmatrix} \text{ERG} \\ \text{ABS} \end{bmatrix} \begin{bmatrix} \begin{bmatrix} \text{PERS} & 1 \\ \text{NUM} & \text{SG} \end{bmatrix} \\ \begin{bmatrix} \text{PERS} & 2 \\ \text{NUM} & \text{PL} \end{bmatrix} \end{bmatrix}$$

Implementation of testsets and communication with "standard authors"

- LMF --> Thorsten Trippel
- LIFT --> Cambell Prince
- OLIF --> Mike Maxwell
- DATR --> ?

Time plan

- finishing of test set within 10 days or earlier
- implementing and developer contacts within 10 days after
- report due in three weeks (!)

Which standard would we adopt?

- One that
 - can represent our lexicons
 - inheritance
 - ambiguity
 - writing systems
 - is usable
 - easy enough
 - illustrated by good examples
 - has implementation(s)
- One that has all the issues solved.....

Bottom up lexicon standard wanted!