Universal Dependency Treebank for Latvian: a Pilot

Lauma Pretkalniņa, Laura Rituma and Baiba Saulīte

University of Latvia, Institute of Mathematics and Computer Science

Universal Dependencies

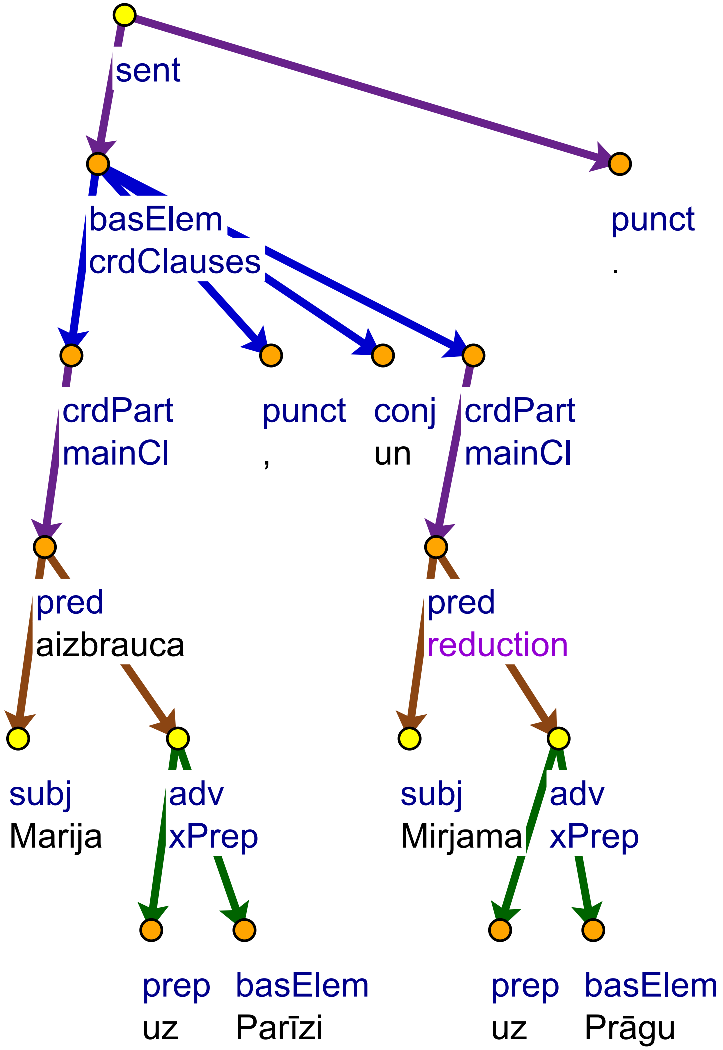
- Cross-lingual initiative
- Unified annotation guidelines
- Emphasis on similar annotations for similar phenomena across different languages
- More than 40 languages
- Latvian included since v1.3.

Latvian UD Treebank

- Size: 20K tokens, 1.1K sentences
- Genre: newswire
- Source: Latvian Treebank
- Conversion procedure: automatic

Latvian Treebank

- In development since 2010
- 3,9K sentences
- Various text genres
- Hybrid annotation model:
  - dependency relations form tree’s backbone
  - each dependency node can be either word or phrase



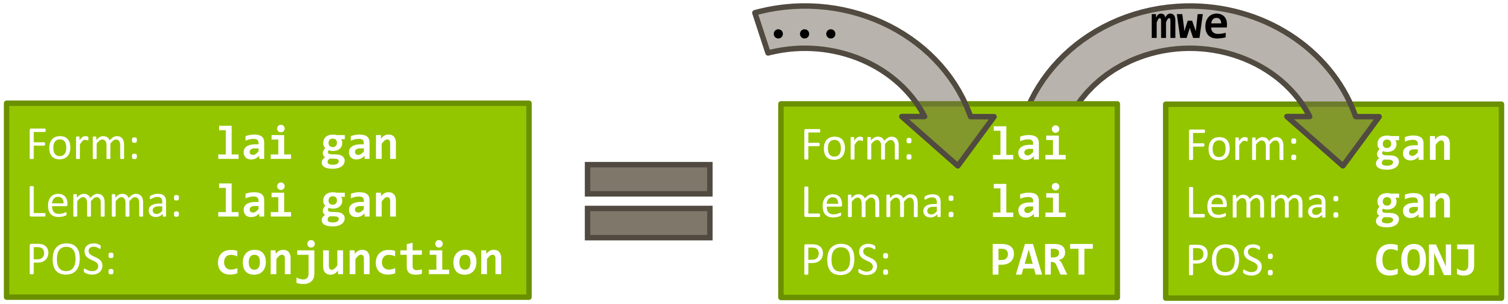
```
Marija aizbrauca uz Parīzi, un  Mirjama uz Prāgu .
Marie  went      to Paris   and Miriam  to Prague.
```

Conversion procedure

1. Retokenize
2. Work out morphology
    1. Determine UPOS
    2. Add as much FEATS as possible
3. Work out syntax
    1. Determine dependency role
    2. Adjust tree structure

Tokenization

- What we did?
  - Got rid of “words with spaces”



Form: lai gan
Lemma: lai gan
POS: conjunction

=

...

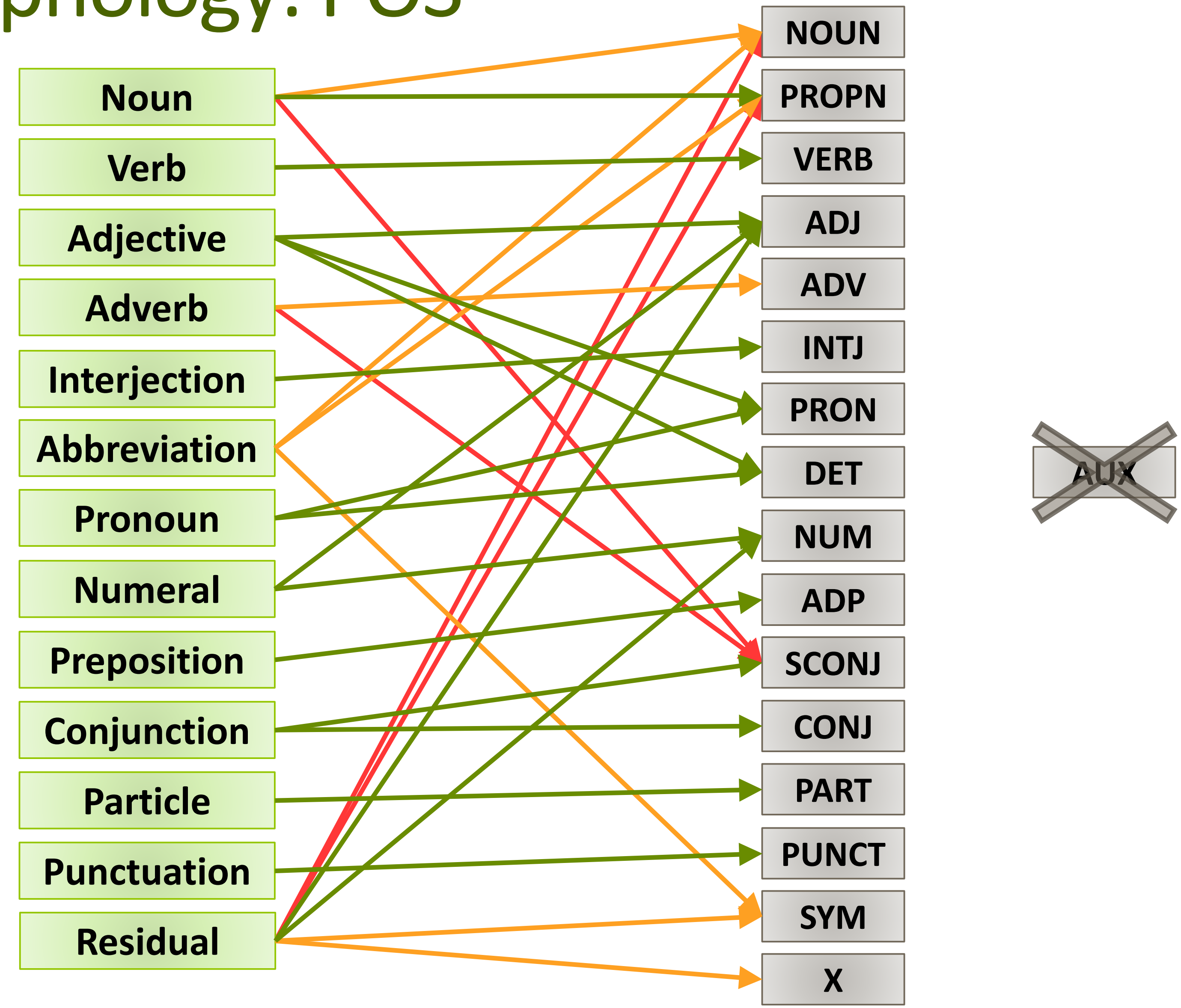
mwe

Form: lai
Lemma: lai
POS: PART

Form: gan
Lemma: gan
POS: CONJ

- What is still missing?
  - Reflexive verb = direct verb + reflexive pronoun

Morphology: POS



Noun
Verb
Adjective
Adverb
Interjection
Abbreviation
Pronoun
Numeral
Preposition
Conjunction
Particle
Punctuation
Residual

NOUN
PROPN
VERB
ADJ
ADV
INTJ
PRON
DET
NUM
ADP
SCONJ
CONJ
PART
PUNCT
SYM
X

~~AUX~~

Morphology: lexico-grammatical features

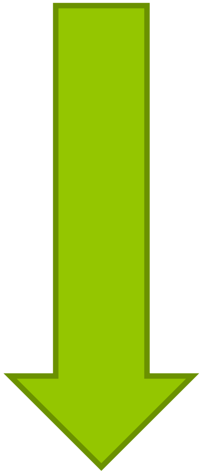
- ✓ `Gender`, `Number`, `Case`, `Definite`, `Degree`
- ✓ `VerbForm`, `Mood`, `Tense`, `Voice`, `Person`, `Aspect` (participles only), `Negative` (non-participle verbs only)
- ✓ `PronType`, `NumType`, `Poss`, `Reflex` (pronouns and verbs)

Sometimes we miss:

- ✘ `VerbForm=Part`, `Voice` (adjectives like *vienota* ‘unified’)
- ✘ `VerbForm=Trans` (adverbs like *salīdzinoši* ‘comparatively’)
- ✘ `Negative` (any nouns, adjectives, e.g., *neapzināts* ‘unconscious’)
- ✘ `NumType` (nouns like *miljons* ‘million’, *puse* ‘half’, some adverbs like *divpadsmitreiz* ‘twelfth time’)

Syntax: overview

- Latvian Treebank = dependencies + phrases + ellipses



1. Remove childless ellipsis nodes
2. Determine UD role for each node
3. Rework tree structure:
   - transform phrases to dependency subtrees
   - remove remaining ellipses

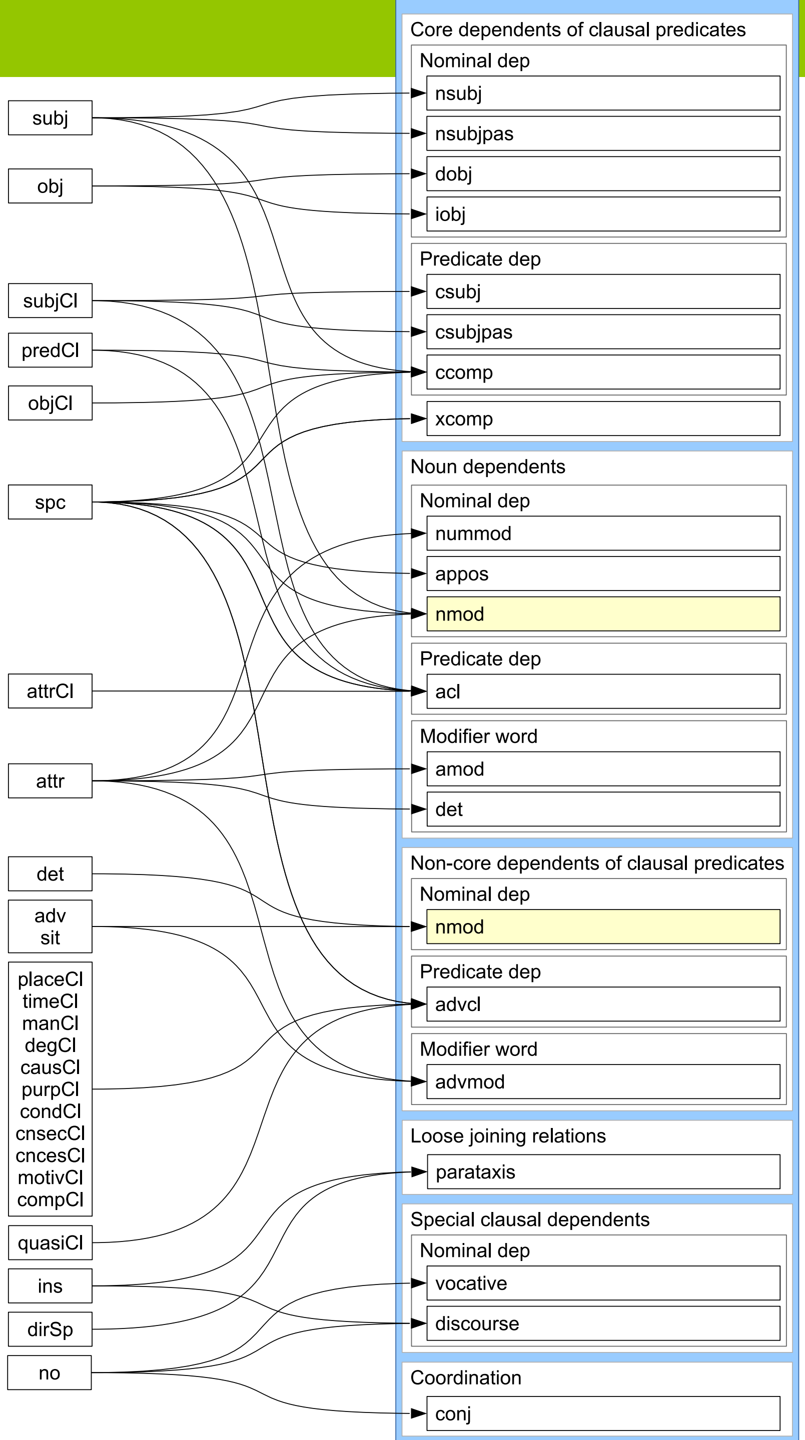
- Latvian UD Treebank = pure dependency trees

Syntax: roles

- Highly asymmetrical relation
- UD roles – POS related
  LVTB roles – more abstract
- Morphotags and structure must be consulted, e.g.,

  $\mathtt{attr}_{\mathtt{pronoun}}$ = $\mathtt{det}$

  $\mathtt{subj}_{\mathtt{pronoun}}$ = $\mathtt{nsubj}$ OR $\mathtt{nsubjpas}$



subj
obj
subjCl
predCl
objCl
spc
attrCl
attr
det
adv
sit
placeCl
timeCl
manCl
degCl
causCl
purpCl
condCl
cnsecCl
cncesCl
motivCl
compCl
quasiCl
ins
dirSp
no

Core dependents of clausal predicates
Nominal dep
nsubj
nsubjpas
dobj
iobj
Predicate dep
csubj
csubjpas
ccomp
xcomp

Noun dependents
Nominal dep
nummod
appos
nmod
Predicate dep
acl
Modifier word
amod
det

Non-core dependents of clausal predicates
Nominal dep
nmod
Predicate dep
advcl
Modifier word
advmod

Loose joining relations
parataxis

Special clausal dependents
Nominal dep
vocative
discourse

Coordination
conj

Syntax: major problems

- Proper distinction between `ccomp` and `xcomp`
  - *viņš mācīja peldēt* ‘he taught [someone] to swim’
  - *viņš iemācījās peldēt* ‘he learned to swim’
- Ellipsis analysis
  - *Marie went to Paris, Miriam — to Prague* is analyzed without `remnants`

Syntax: rare problems

- No explicitly marked `lists`
- Complex predicates with non-neutral word order

kļūt	*izglītots*	*viņš*	*gribēja*
become.INF	educated	he	want.PST.3SG

‘he wanted to become educated’

Future work

- Release better quality corpus with corrected transformation errors
  - Official release UD v1.4
  - Regular updates in GitHub repo `UD_Latvian` dev branch
- Release all Latvian Treebank as UD corpus
  - UD v1.4 or UD v2.0
  - Provide data for Shared Task
- Further…
  - Extend corpus, introduce language specific subroles
  - Make available tokenizing/tagging/parsing tools

Thank you!

Mākslīgā intelekta laboratorija
LU MII

LATVIJAS UNIVERSITĀTE
ANNO 1919
UNIVERSITY OF LATVIA