The Lithuanian Treebank ALKSNIS

Agnė Bielinskienė
Loïc Boizou
Jolanta Kovalevskaitė
Erika Rimkutė

Vytautas Magnus university
Centre of Computational Linguistics

Baltic HLT 2016, Riga

Introduction

- There were several attempts to design Lithuanian treebanks, but none were sufficiently developed and publicly available.
- The current ALKSNIS treebank is funded for the period 2015-2016 by the project *Lithuanian Membership in International Scientific Research Infrastructure – Common Language Resources and Technology Infrastructure Consortium (CLARIN ERIC)*.
- Basic resource presentation.

Past attempts

- 2007-2008: project “Internet resources: Annotated Corpus of the Lithuanian Language and Tools of Annotation (ALKA 2)” (VMU-CCL):
  - 1566 sentences.
- 2013: VILSINTEKS (Institute of Lithuanian Language):
  - 3 sentences.

Structure

- 4 parts (full texts):
  - newspapers (699 sentences);
  - magazines (690 sentences);
  - modern prose (716 sentences);
  - legal texts (250 sentences).
- Total:
  - 2355 sentences;
  - About 32 000 tokens.

Realization

- Automatic segmentation, tagging and parsing;
  - Initial parsing results differ significantly from the actual structure.
- Correction by linguists.
  - vizualization with TrED.

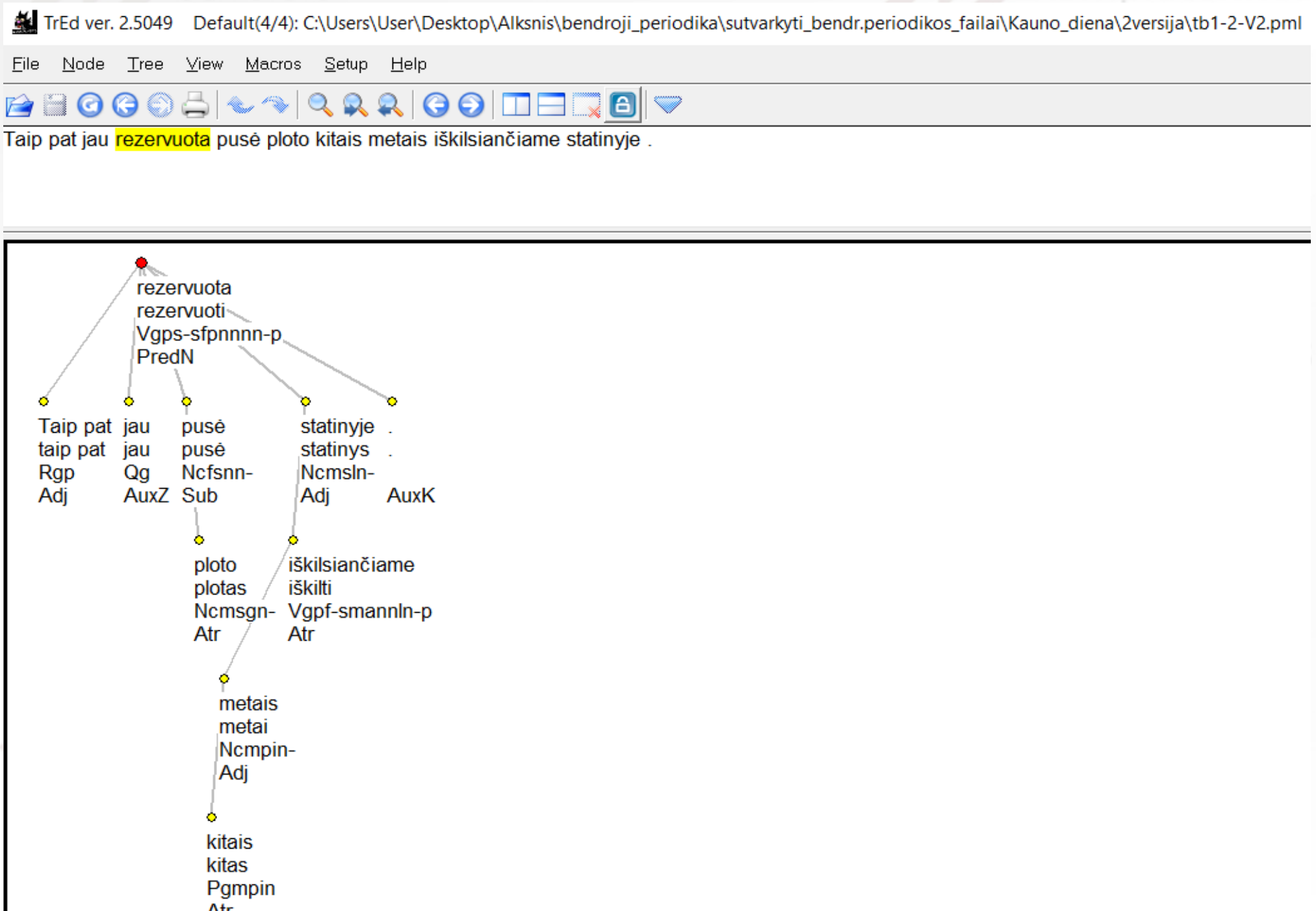
Information provided

- Word form, lemma, morphological information.
- Dependency connections and node syntactic functions.

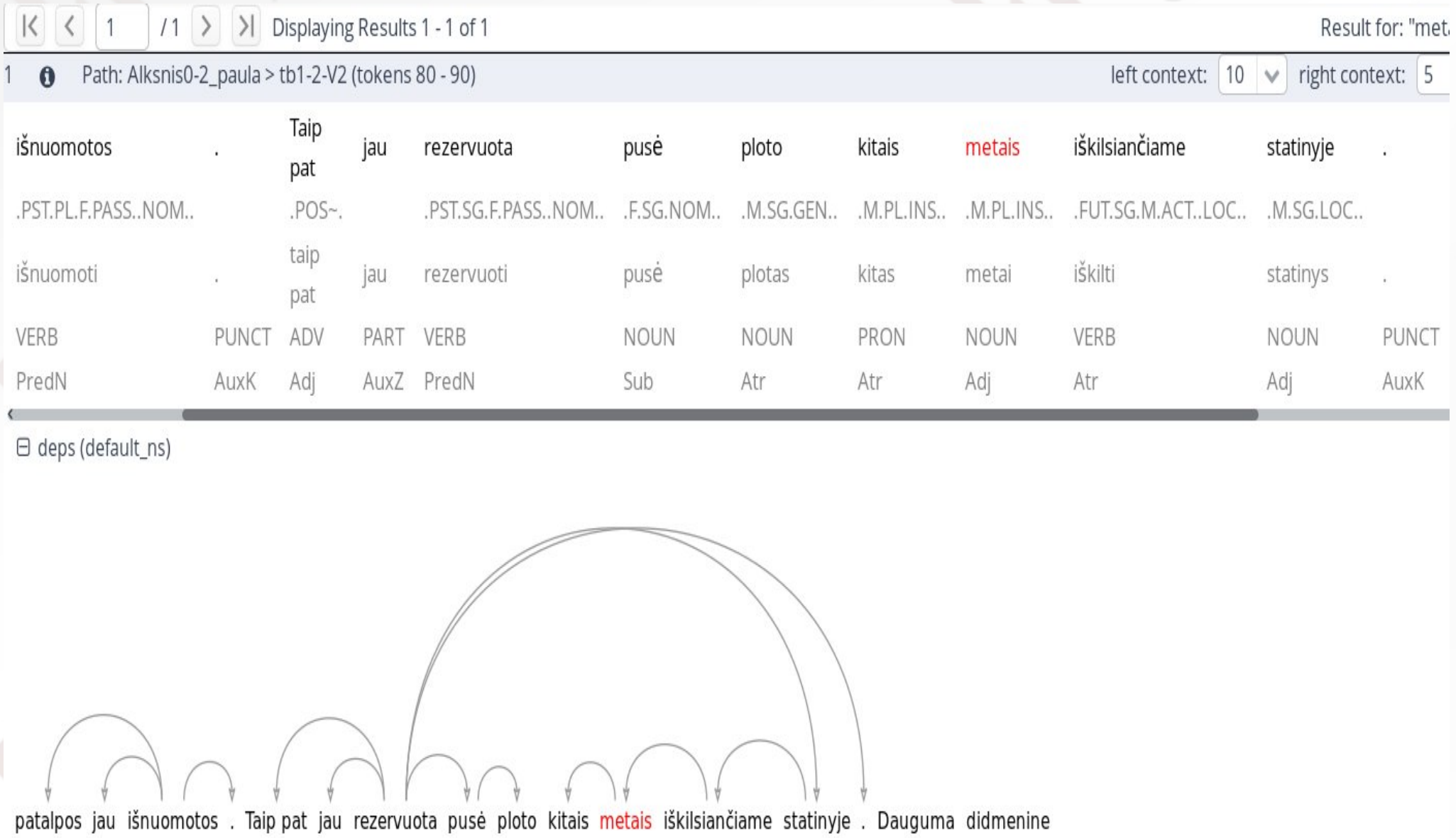
Formats

- 2 formats:
  - PML (for TrED);
  - Paula (for generating Annis).
- Different POS and morphological information encodings:
  - PML: inspired by MULTEXT-East (for shortness);
  - Annis: UD POS + Leipzig Glossing abbreviations (with required additions).

Taip pat jau rezervuota pusė ploto kitais metais iškilsiančiame statinyje (PML)



Taip pat jau rezervuota pusė ploto kitais metais iškilsiančiame statinyje (PAULA)



Syntactic tags

Pred (PredN, PredV)

Sub

Obj

Atr

Adj

Aux (AuxK, AuxC, AuxZ, AuxP, AuxL)

Coord

Par

ExD

_Co (Atr_Co)

Pred_ (Pred_Obj)

Structural choices

- PDT-inspired…
  - Dependency model;
  - Many common tags and decisions;
- … but:
  - Fewer auxiliary categories;
  - No ATV category (adverbial or integrated in a complex predicate);
  - No Apos category.

Alksnis development

- For now:
  - Still few corrections needed;
  - More or less stable: recent discussion about of one structural choice;
  - Availability.
- For tomorrow (depending on funding…):
  - Enlarge;
  - Enrich;
  - UD conversion;
  - New parser (statistical).

Links

- Online initial draft

  http://clarin-lt.lt/?p=205

- Annis
  - Only on internal server

Thank you for your attention!