From Multilingual Dictionary to Lithuanian WordNet

Slovko 2013, Bratislava

Radovan Garabík Indrė Pileckytė

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava

2013-11-13

WordNet

- Princeton WordNet
- BalkaNet, EuroWordNet, WordNet Grid, BabelNet
- Slovak Online – small dictionary (sk, en, de, pl, lt)

Automatic (sk) Synset Generation

- translate synsets, hypernyms and hyponyms
- $\bigcap_{*nym}$ translations[1]
- proofread, fill in the gaps
- $\forall sk : \exists hypernym$

[1] collaboration with Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice

Automatic (lt) Synset Generation

- get rough sk-lt dictionary (eo-sk $\cup$ eo-lt)[2]
- eo-lt: 11 529 entries, 16 268 words
- eo-sk: 7 116 entries, 8 130 words
- result: 3 977 entries, 10 048 Lithuanian words
- proofread the dictionary (‘precision’)
- substitute the literals
- proofread resulting database...

[2] `http://lernu.net/`

Automatic[?] (lt) Synset Generation

- get rough sk-lt dictionary (eo-sk $\cup$ eo-lt)[2]
- eo-lt: 11 529 entries, 16 268 words
- eo-sk: 7 116 entries, 8 130 words
- result: 3 977 entries, 10 048 Lithuanian words
- proofread the dictionary ('precision')
- substitute the literals
- proofread resulting database...

[2] `http://lernu.net/`

Database Structure

- one entry, one synset
- (optional) definition
- (not optional) link to en synsets
- (optional) link to sk synsets: L: M: N

01984902
sit down,
sit

01984902
sit down, sit

947
sadnúť si; usadiť sa; posadiť sa

948
sadnúť; usadiť

3708
sadať; usádzať

6423
sadať si

01984902
sit down, sit

947
sadnúť si; usadiť sa; posadiť sa

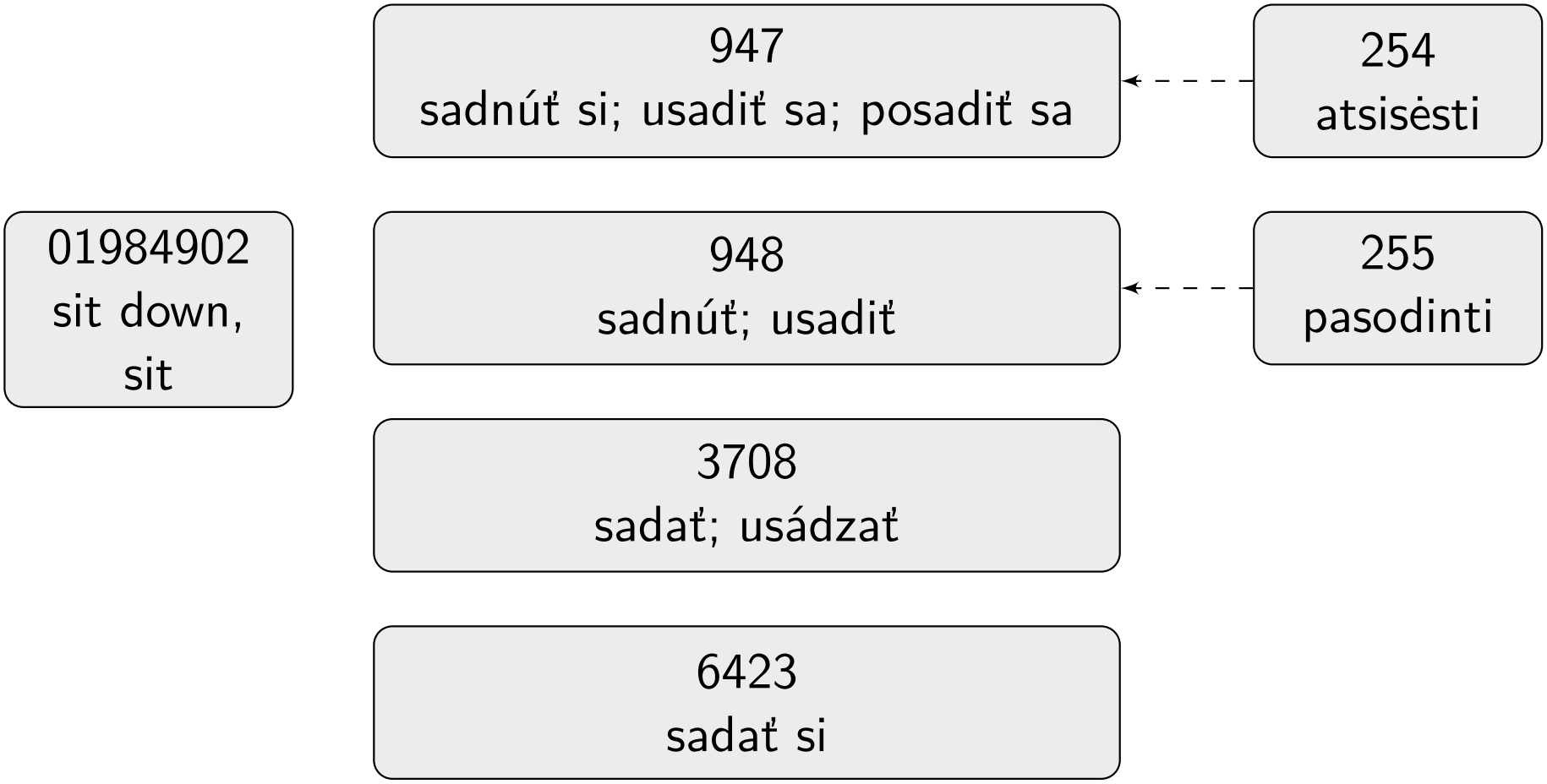
948
sadnúť; usadiť

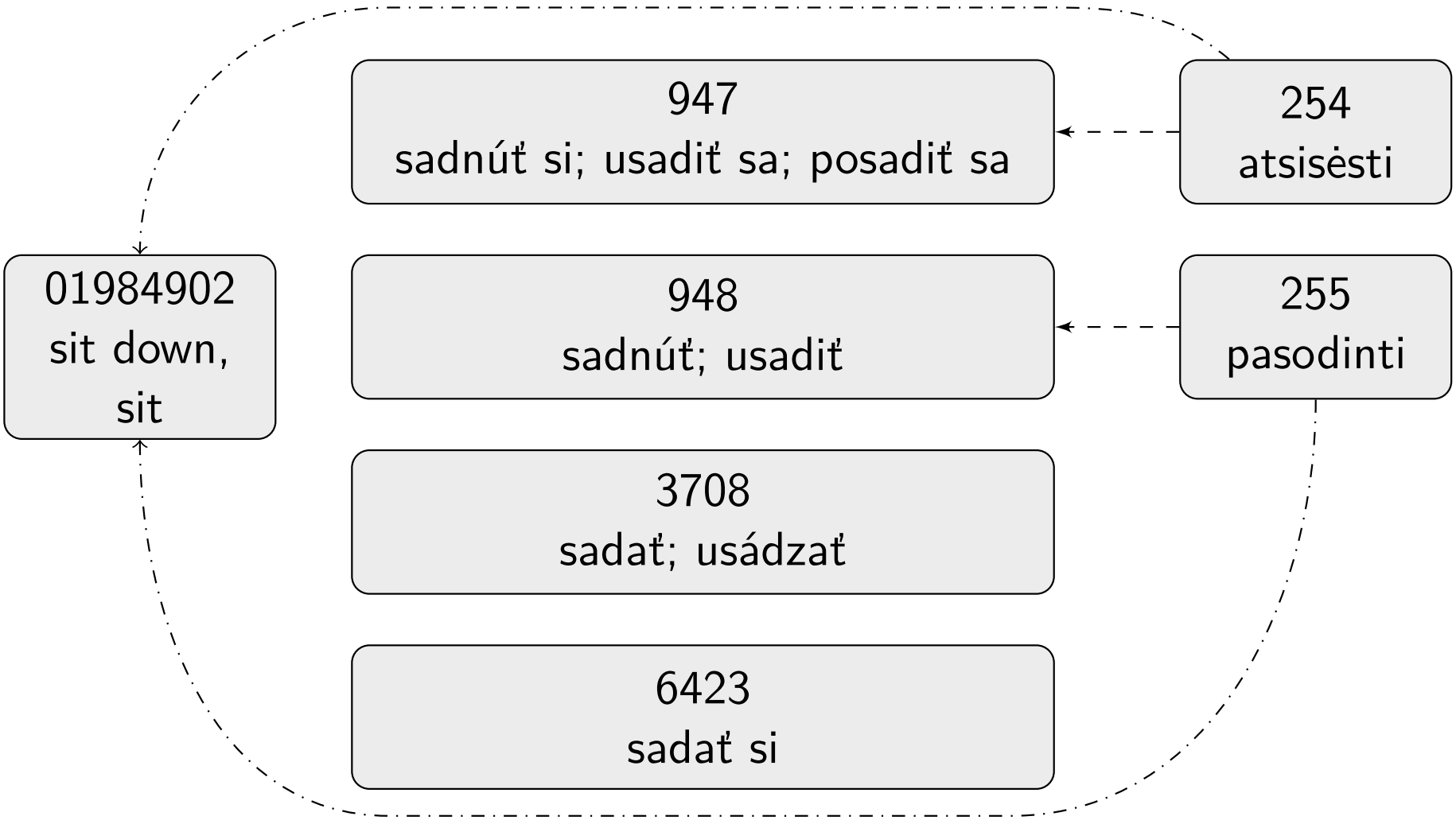
3708
sadať; usádzať

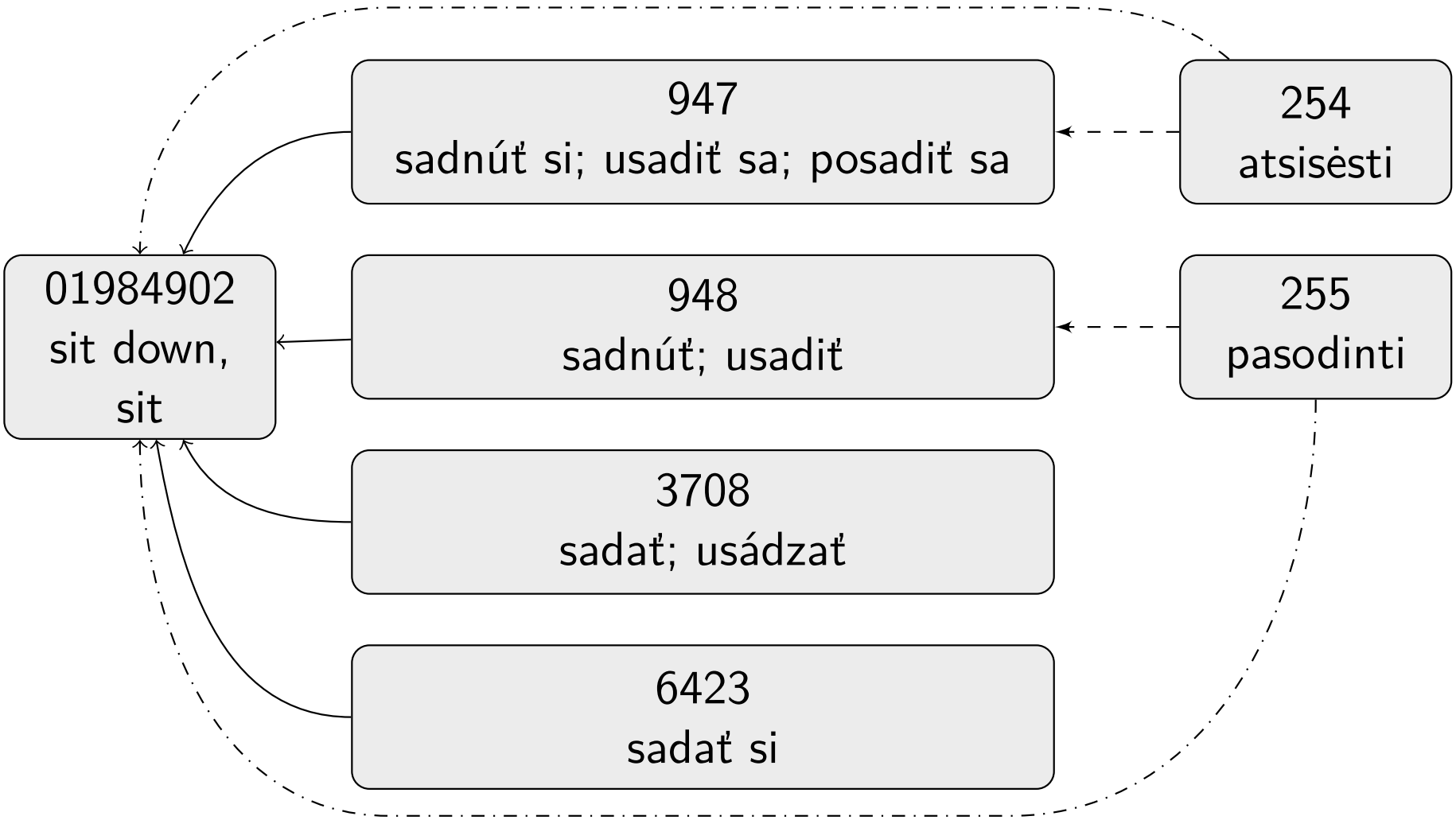
6423
sadať si

254
atsisėsti

255
pasodinti


947
sadnúť si; usadiť sa; posadiť sa
254
atsisėsti
01984902
sit down,
sit
948
sadnúť; usadiť
255
pasodinti
3708
sadať; usádzať
6423
sadať si


01984902
sit down, sit
947
sadnúť si; usadiť sa; posadiť sa
948
sadnúť; usadiť
3708
sadať; usádzať
6423
sadať si
254
atsisėsti
255
pasodinti


01984902
sit down,
sit
947
sadnúť si; usadiť sa; posadiť sa
948
sadnúť; usadiť
3708
sadať; usádzať
6423
sadať si
254
atsisėsti
255
pasodinti

Synset Microformat

- formalized rules for the synset

```
<synset> ::= [<synset-annot>] <anotated-literal> |
             <annotated-literal> ";" <synset>
<annotated-literal> ::= [<literal-annot>] <literal>
<literal-annot> ::= "+" | ""
<synset-annot> ::= "-" | "?" | ""
```

Nouns, Adjectives and Adverbs

- very straightforward
- mostly 1:1
- gender distinction

10020890 doctor, doc, physician, MD, Dr., medico	
2204 daktaras; gydytojas	4914 daktarė; gydytoja

10020890 doctor, doc, physician, MD, Dr., medico

2204 daktaras; gydytojas	4914 daktarė; gydytoja
	

Verbs

- not very straightforward
- mostly not 1:1
- aspect
- reflexivity

Aspect

- 2 aspects: perfective, imperfective
- Slovak:
  - perfective + *-va-* $\rightarrow$ imperfective
  - *pretrvať* $\rightarrow$ ... $\rightarrow$ *pretrvávavavaní*
  - prefixes + imperfective $\rightarrow$ perfective
  - *robiť* $\rightarrow$ $\{$*u,vy,za,pre,do,na,od*$\}$*robiť*
- keep both forms, if they exist

Aspect

- Lithuanian:
  - verbs without prefix are imperfective
  - motion verbs – imperfective in present tense, perfective in past simple
  - neutral verbs – *mirti*
- prefer imperfective, keep perfective if there is the same meaning in Lithuanian

Reflexivity

- Slovak:
  - separate reflexive pronoun/particle *sa*, *si*
  - treat reflexive verbs as single units
  - cover both reflexive and non-reflexive variants – two synsets
  - reflexivity overlaps with transitivity → the same en synset
- Lithuanian:
  - reflexive affix *-si*, *-s*
  - affix for prefixless verbs: *sukti* → *sukti**s***
  - infix after the prefix morpheme: *nuprausti* → *nu**si**prausti*
  - map the synset to Slovak ones

Proofreading

- Slovak:
  - verify literals in the synset
  - verify synset position in the hierarchy

Proofreading

- Slovak:
  - verify literals in the synset
  - verify synset position in the hierarchy
  - ... twice.

Proofreading

- Slovak:
  - verify literals in the synset
  - verify synset position in the hierarchy
  - ... twice.
- Lithuanian:
  - proofread Lithuanian synset connected to proofread Slovak ones
  - look for their connection to English and Slovak

Zdrojový anglický synset		
Anglické synsety	**Slovenské synsety**	**Litovské synsety**
EN synset: 01984902 { sit down, sit } take a seat +SYN | -SYN	SK synset: 947 { sadnúť si; usadiť sa; posadiť sa } | +EN | -EN | × | ☐ | (i) SK synset: 948 { sadnúť; usadiť } | +EN | -EN | × | ☐ | (i) SK synset: 3708 { sadať; usádzať } | +EN | -EN | × | ☐ | (i) SK synset: 6423 { sadať si } | +EN | -EN | × | ☐ | (i)	LT synset: 254 { atsisėsti } | +EN | -EN | × | ☐ | (i) Prislúchajúci SK synset: 947; Pripoj: 948, 3708, 6423 LT synset: 255 { pasodinti } | +EN | -EN | × | ☐ | (i) Prislúchajúci SK synset: 948; Pripoj: 947, 3708, 6423

Sources Used

- Dabartinis lietuvių kalbos žodynas
- Lietuvos Respublikos terminų bankas
- Valstybinė lietuvių kalbos komisija
- Anglų-lietuvių kalbų kompiuterijos žodynėlis

Sources Used

- Dabartinis lietuvių kalbos žodynas
- Lietuvos Respublikos terminų bankas
- Valstybinė lietuvių kalbos komisija
- Anglų-lietuvių kalbų kompiuterijos žodynėlis
- use only “correct” words at this stage

Current Status and Plans

- 10145 noun synsets
- 2099 adjective synsets
- 683 adverbial synsets
- 533 verbal synsets
- Affero GPL v. 3, CC BY-SA 3.0, ODbL v1.0
- VisDic/DEBVisDic

Thank you for the attention