Database of Slovak Verbs

Radovan Garabík

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

2011-11-24

Introduction

- complex verb morphology
- several existing approaches & databases
- one database rules them all
- reuse existing data (if possible)

Verb Morphology

- 3 persons
- singular, plural
- infinitive, indicative, L-participle, imperative, negation, active participle, passive participle, deverbative
- gender (only sometimes)
- (zombie category: past active participle)

Anomalies

- impersonal verbs
- greetings: verbs only in the imperative – *ahoj, ahojte, čau, čaute, vitaj, vitajte ...*
- verbs without an infinitive – *pošiel*
- verbs without negation – *nenávidieť*
- negation of *byť* – *nie je*

Beyond the Morphology

- conditional
- past conditional
- reflexive pairs
- aspect pairs
- tense: plusquamperfect/past/present
- ... or plusquamperfect/past/future

repetitive/habitual

- semi-morphology change (new lexeme)

repetitive/habitual

- semi-morphology change (new lexeme)
- zamiluvávavava (zamilovať→zamilovávať→zamilovávavať→zamilovávavavať)
- pretrvávavavaní (pretrvať→pretrvávať→pretrvávavať→pretrvávavavať)

Tagset

- Slovak National Corpus tagset

Levenshtein edit operations

- character insertion, deletion or substitution
- Levenshtein distance: $\rho(s_1, s_2)$ – minimal number of Levenshtein edit operations needed to transform $s_1 \rightarrow s_2$
- lemma → tag, wordform
- tag: formal description of grammar categories
- sequence of Levenshtein edit operations applied to lemma: →wordform
- applied to another lemma – the same paradigm

- technical trick: go backwards
- *abdikovať abeced-ovať abon-ovať abricht-ovať absent-ovať absolutiz-ovať absolv-ovať absorb-ovať abstin-ovať abstrah-ovať*
- another technical trick: use NFKD Unicode normalization
- *chodiť* → *chodím*, *volať* → *volám*

- technical trick: go backwards
- *abdikovať abeced-ovať abon-ovať abricht-ovať absent-ovať absolutiz-ovať absolv-ovať absorb-ovať abstin-ovať abstrah-ovať*
- another technical trick: use NFKD Unicode normalization
- *chodiť* → *chodím, volať* → *volám*
-
  - *chodiť, volať*

- technical trick: go backwards
- *abdikovať abeced-ovať abon-ovať abricht-ovať absent-ovať absolutiz-ovať absolv-ovať absorb-ovať abstin-ovať abstrah-ovať*
- another technical trick: use NFKD Unicode normalization
- *chodiť* → *chodím*, *volať* → *volám*
-
    - *chodi~~ť~~*, *vola~~ť~~*
    - *chodi*, *vola*

- technical trick: go backwards
- *abdikovať abeced-ovať abon-ovať abricht-ovať absent-ovať absolutiz-ovať absolv-ovať absorb-ovať abstin-ovať abstrah-ovať*
- another technical trick: use NFKD Unicode normalization
- *chodiť* → *chodím*, *volať* → *volám*
-
    - *chodiť*, *volať*
    - *chodi*, *vola*
    - *chodi***'**, *vola***'** ≡ *chodí*, *volá*

- technical trick: go backwards
- *abdikovať abeced-ovať abon-ovať abricht-ovať absent-ovať absolutiz-ovať absolv-ovať absorb-ovať abstin-ovať abstrah-ovať*
- another technical trick: use NFKD Unicode normalization
- *chodiť* → *chodím*, *volať* → *volám*
-
    - *chodi~~ť~~*, *vola~~ť~~*
    - *chodi*, *vola*
    - *chodi***'**, *vola***'** ≡ *chodí*, *volá*
    - *chodí*+**m**, *volá*+**m**

- technical trick: go backwards
- *abdikovať abeced-ovať abon-ovať abricht-ovať absent-ovať absolutiz-ovať absolv-ovať absorb-ovať abstin-ovať abstrah-ovať*
- another technical trick: use NFKD Unicode normalization
- *chodiť* → *chodím*, *volať* → *volám*
-
    - *chodi~~ť~~*, *vola~~ť~~*
    - *chodi*, *vola*
    - *chodi***’**, *vola***’** ≡ *chodí*, *volá*
    - *chodí*+**m**, *volá*+**m**
    - *chodím*, *volám*

Slovak Morphology database

- paradigm classes based on Levenshtein edit operations
- 76885 lemmas, 923441 unique wordforms, 2472721 entries
- approach successful, but linguistically opaque

Approach to verbs

- *(ne-)(prefix-)root-suffix*
- *root* can change, too: **pr**ať → **per**iem
- paradigm class: she same set of suffixes, “the same” way of changing the root, the same aspect
- formalize “the same” – the same sequence of Levenshtein edit operations
- split the lemma
- keep the prefix
- add the suffix according to grammar categories
- apply Levenshtein edit operations to the root
- … and we obtain the inflected word form
- negation is handled separately

Database

- MoinMoin wiki engine
- written in Python
- user permissions
- ACL
- web interface
- keep track of changes
- versioning
- custom parser

Database Structure

- two kinds of pages:
  1. paradigm class description

Database Structure

- two kinds of pages:
    1. paradigm class description
    2. verb data

Database Structure

- two kinds of pages:
  1. paradigm class description
  2. verb data
  3. system pages

Database Structure

- two kinds of pages:
  1. paradigm class description
  2. verb data
  3. system pages

Paradigm Class Description

Just a stupid list of *tag: word form*

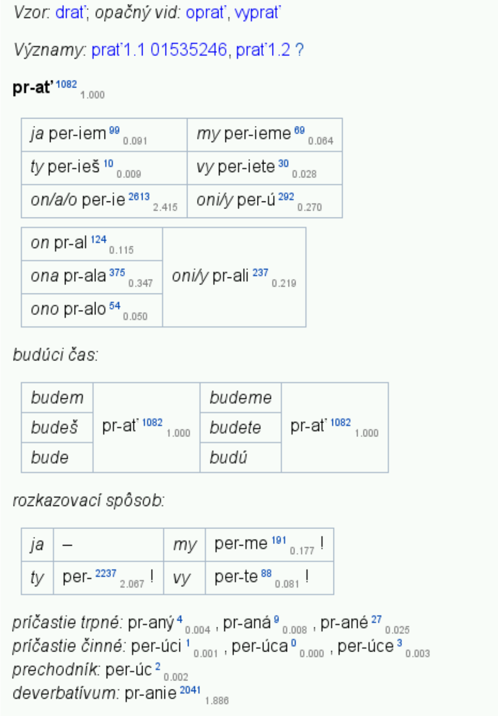
```
VIe+    : dr-ať
VKesa+ : der-iem
...
VLesan+: dr-alo
VLepah+: dr-ali

Gtms1x : dr-aný
Gtfs1x : dr-aná
Gtns1x : dr-ané
Gkms1x : der-úci
Gkfs1x : der-úca
Gkns1x : der-úce
SSns1   : dr-anie
```

- not only tags describing verbs, but also participles (active, passive) and a noun
- active *past* participle only for perfective aspect
- active *present* participle only for imperfective aspect
- … so we reuse the MSD tags for active/passive

Verb Data

```
lema: pr-ať
vzor: drať
vid: oprať, vyprať
synset: prať1.1 01535246, prať1.2 ?
```



Vzor: drať; *opačný vid:* oprať, vyprať

Významy: prať1.1 01535246, prať1.2 ?

pr-ať $^{1082}_{1.000}$

ja per-iem $^{99}_{0.091}$	*my* per-ieme $^{69}_{0.064}$
ty per-ieš $^{10}_{0.009}$	*vy* per-iete $^{30}_{0.028}$
on/a/o per-ie $^{2613}_{2.415}$	*oni/y* per-ú $^{292}_{0.270}$

on pr-al $^{124}_{0.115}$	*oni/y* pr-ali $^{237}_{0.219}$
ona pr-ala $^{375}_{0.347}$	
ono pr-alo $^{54}_{0.050}$	

budúci čas:

budem	pr-ať $^{1082}_{1.000}$	*budeme*	pr-ať $^{1082}_{1.000}$
budeš		*budete*	
bude		*budú*	

rozkazovací spôsob:

ja	–	*my*	per-me $^{191}_{0.177}$!
ty	per- $^{2237}_{2.067}$!	*vy*	per-te $^{88}_{0.081}$!

príčastie trpné: pr-aný $^{4}_{0.004}$, pr-aná $^{9}_{0.008}$, pr-ané $^{27}_{0.025}$
príčastie činné: per-úci $^{1}_{0.001}$, per-úca $^{0}_{0.000}$, per-úce $^{3}_{0.003}$
prechodník: per-úc $^{2}_{0.002}$
deverbatívum: pr-anie $^{2041}_{1.886}$

Bootstraping the Database

- prepare list of paradigm classes (331)
- for each class, create list of Levenshtein edit operations converting the lemma into the word form
- … repeat for each verb
- if the list of edit operations is the same, assign the verb to the given class
- cross check against the corpus (but not the past active participle)

Goals

- 12 thousand verbs (very thorough coverage)
- complete aspect pairs
- hypernyms (according to English WordNet)
- most frequent senses (go just by linguistic intuition, really)

Thank you for the attention