Ensemble Tagging
Slovak Web Data

Vladimír Benko[1,2] & Radovan Garabík[1]

{vladimir.benko,radovan.garabik}@juls.savba.sk

[1] Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
[2] Comenius University, UNESCO Chair in Plurilingual and Multicultural Communication

SlaviCorp 2018
Praha | 24–26 September 2018

Aranea

Aranea: A Family of Gigaword Web Corpora

- Started in 2013
- **Slovak-Centric** (languages spoken and/or taught in **Slovakia** and its neighbouring countries)
- Crawled and pre-processed by at (approximately) **the same time**
- Language-independent processing by **the same tools**
- Language-dependent processing using **the same methodology**
- **The same size** (2 “comparable” versions),
- “Language-neutral” (Latin) names

Aranea

- **Name Araneum** (*pl.* **Aranea**, *n.*)

araneum, aranei

noun
declension: 2nd declension
gender: neuter

Definitions:

1. mass of threads resembling a spider web
2. spider web, cobweb

Age: In use throughout the ages/unknown
Area: Agriculture, Flora, Fauna, Land, Equipment, Rural
Geography: All or none
Frequency: For Dictionary, in top 20,000 words
Source: "Oxford Latin Dictionary", 1982 (OLD)

Aranea

Present state of the Project

- **20 languages** (plus varieties for 3 of them)
- **5 Slavic languages**

  Slovacum, Bohemicum, Russicum, Polonicum, Bulgaricum
- **3 more languages in preparation**

  including *Ucrainicum*
- **5 languages with more than 1 Gtoken**

  including *Slovacum* (3 G), *Bohemicum* (5 G) & *Russicum* (13 G)

Ensemble

Michelle, ma belle

These are words that go together well

My Michelle

Michelle, ma belle

Sont des mots qui vont très bien ensemble

Très bien ensemble

[McCartney & Lennon, 1966]

Ensemble

Ensemble methods

- several tools used to perform (independently) the same task
- aggregating their output
- hoping for improvement of the overall success rate

Ensemble

Ensemble methods

- several tools used to perform (independently) the same task
- aggregating their output
- hoping for improvement of the overall success rate

Ensemble tagging

- if more than one tagger (and/or language model) is available for the respective language
- usually true for “large” languages

Araneum Slovacum

Resources for tagging Slovak (corpus) data

- 1,2 Mtoken manually disambiguated corpus
- Slovak Morphological Database
  100 K basic forms
  approx. 3 M inflected forms

Araneum Slovacum

Resources for tagging Slovak (corpus) data

Language Models for

- **TnT** (Hajič tagset, Garabík)
- **Morče** (SNK tagset, Garabík)
- **MorphoDita** (SNK taget, Garabík)
- **TreeTagger** (several models: simplified SNK tagset, Schmid;
  full SNK tagset, Schmid/Benko;
  ASCII-only model, Benko)
- **RFTagger** (ajka tagset; Medveď)

Araneum Slovacum

Resources for tagging Slovak (corpus) data

Language Models for

- **TnT** (Hajič tagset, Garabík)
- **Morče** (SNK tagset, Garabík)
- **MorphoDita** (**SNK taget, Garabík**)
- **TreeTagger** (several models: simplified SNK tagset, Schmid; **full SNK tagset**, Schmid/**Benko**; ASCII-only model, Benko)
- **RFTagger** (ajka tagset; Medveď)

Morphosyntactic Annotation

<s>O pár minút lob toho istého hráča znova zastavilo brvno.</s>

Morphosyntactic Annotation

<s>O pár minút lob toho istého hráča znova zastavilo brvno.</s>

Morphosyntactic Annotation

<s>
O
pár
minút
lob
toho
istého
hráča
znova
zastavilo
brvno
<g/>
.
</s>

Morphosyntactic Annotation

<s>		
O	o	Eu4
pár	pár	NUns4
minút	minúta	SSfp2
lob	lob	SSip7
toho	to	PFns2
istého	istý	AAms2x
hráča	hráč	SSms4
znova	znova	Dx
zastavilo	zastaviť	VLdscn+
brvno	brvno	SSns4
<g/>		
.	.	Z.
</s>		

Morphosyntactic Annotation

<s>			
O	o	Eu4	1
pár	pár	NUns4	1
minút	minúta	SSfp2	1
lob	lob	SSip7	0
toho	to	PFns2	1
istého	istý	AAms2x	1
hráča	hráč	SSms4	1
znova	znova	Dx	1
zastavilo	zastaviť	VLdscn+	1
brvno	brvno	SSns4	1
<g/>			
.	.	Z.	1
</s>			

Morphosyntactic Annotation

<s>			
O	o	Eu4	1
pár	pár	NUns4	1
minút	minúta	SSfp2	1
lob	lob	SSip7	0
toho	to	PFns2	1
istého	istý	AAms2x	1
hráča	hráč	SSms4	1
znova	znova	Dx	1
zastavilo	zastaviť	VLdscn+	1
brvno	brvno	SSns4	1
<g/>			
.	.	Z.	1
</s>			

Morphosyntactic Annotation

<s>			
O	o	Eu4	1
pár	pár	NUns4	1
minút	minúta	SSfp2	1
lob	lob	SSip7	0
toho	to	PFns2	1
istého	istý	AAms2x	1
hráča	hráč	SSms4	1
znova	znova	Dx	1
zastavilo	zastaviť	VLdscn+	1
brvno	brvno	SSns4	1
<g/>			
.	.	Z.	1
</s>			

Morphosyntactic Annotation

<s>				
O	o	Pp	Eu4	1
pár	pár	Nm	NUns4	1
minút	minúta	Nn	SSfp2	1
lob	lob	Nn	SSip7	0
toho	to	Pn	PFns2	1
istého	istý	Aj	AAms2x	1
hráča	hráč	Nn	SSms4	1
znova	znova	Av	Dx	1
zastavilo	zastaviť	Vb	VLdscn+	1
brvno	brvno	Nn	SSns4	1
<g/>				
.	.	Zz	Z.	1
</s>				

Morphosyntactic Annotation

word	lemma	atag	tag	ztag
<s>				
O	o	Pp	Eu4	1
pár	pár	Nm	NUns4	1
minút	minúta	Nn	SSfp2	1
lob	lob	Nn	SSip7	0
toho	to	Pn	PFns2	1
istého	istý	Aj	AAms2x	1
hráča	hráč	Nn	SSms4	1
znova	znova	Av	Dx	1
zastavilo	zastaviť	Vb	VLdscn+	1
brvno	brvno	Nn	SSns4	1
<g/>				
.	.	Zz	Z.	1
</s>				

Morphosyntactic Annotation

word	lemma	atag	tag	ztag	lemma	tag	prec
<s>							
O	o	Pp	Eu4	1	o	Eu4	19
pár	pár	Nm	NUns4	1	pár	NUns4	35
minút	minúta	Nn	SSfp2	1	minúta	SSfp2	01
lob	lob	Nn	SSip7	0	lobiť	VMdsb+	XX
toho	to	Pn	PFns2	1	ten	PFms4	04
istého	istý	Aj	AAms2x	1	istý	AAms4x	08
hráča	hráč	Nn	SSms4	1	hráč	SSms4	02
znova	znova	Av	Dx	1	znova	Dx	01
zastavilo	zastaviť	Vb	VLdscn+	1	zastaviť	VLdscn+	03
brvno	brvno	Nn	SSns4	1	brvno	SSns1	03
<g/>							
.	.	Zz	Z.	1	Z	Z	01
</s>							

Morphosyntactic Annotation

word	lemma	atag	tag	ztag	lemma	tag	prec
<s>							
O	o	Pp	Eu4	1	o	Eu4	19
pár	pár	Nm	NUns4	1	pár	NUns4	35
minút	minúta	Nn	SSfp2	1	minúta	SSfp2	01
lob	lob	Nn	SSip7	0	lobiť	VMdsb+	XX
toho	to	Pn	PFns2	1	ten	PFms4	04
istého	istý	Aj	AAms2x	1	istý	AAms4x	08
hráča	hráč	Nn	SSms4	1	hráč	SSms4	02
znova	znova	Av	Dx	1	znova	Dx	01
zastavilo	zastaviť	Vb	VLdscn+	1	zastaviť	VLdscn+	03
brvno	brvno	Nn	SSns4	1	brvno	SSns1	03
<g/>							
.	.	Zz	Z.	1	Z	Z	01
</s>							

Morphosyntactic Annotation

word	lemma	atag	tag	ztag	lemma	tag	prec
<s>							
O	o	Pp	Eu4	1	o	Eu4	19
pár	pár	Nm	NUns4	1	pár	NUns4	35
minút	minúta	Nn	SSfp2	1	minúta	SSfp2	01
lob	lob	Nn	SSip7	0	lobiť	VMdsb+	XX
toho	to	Pn	PFns2	1	ten	PFms4	04
istého	istý	Aj	AAms2x	1	istý	AAms4x	08
hráča	hráč	Nn	SSms4	1	hráč	SSms4	02
znova	znova	Av	Dx	1	znova	Dx	01
zastavilo	zastaviť	Vb	VLdscn+	1	zastaviť	VLdscn+	03
brvno	brvno	Nn	SSns4	1	brvno	SSns1	03
<g/>							
.	.	Zz	Z.	1	Z	Z	01
</s>							

Possible Use of Ensemble-Tagged Data

- Provide multiple annotations to users
- If more than 2 annotations available:
  "Voting" on items where taggers disagree
- If only 2 annotations are available:
  - tag items where taggers **agree** (both for OOV and non-OOV lexical units)
  - tag items where taggers **disagree** and

    aggregate by rules

    correct manually

Crowdsourcing

The Task

- Enlarge morphological lexicon by manually checked and corrected most frequent out-of-vocabulary (OOV) word forms derived from a large (3 Gigaword) web corpus

The Crowd

- 46 undergraduate students

The Assigment

- cca. 3,300 lines of OOVs each
- two-fold setup (each line checked by independent annotators)

Data to Annotate

85	zohrávka	zohrávka	Yx	%	Nn	SSfs1
69	zohrávke	zohrávke	Yx	%	Nn	SSfs6
64	zohrávku	zohrávku	Yx	%	Nn	SSfs4
130	zohrávky	zohrávky	Yx	%	Nn	SSfp4
71	zohrávky	zohrávky	Yx	%	Nn	SSfs2
89	zohriata	zohriaty	Aj	Gtfs1x	Aj	AAfs1x
299	zohriata	zohriaty	Aj	Gtfs1x	Aj	Gtfs1x
75	zohriateho	zohriati	Aj	AAis2x	Aj	AAis2x
66	zohriateho	zohriatí	Aj	AAis2x	Aj	AAis2x
64	zohriateho	zohriati	Aj	Gkis2x	Aj	AAis2x
57	zohriatej	zohriaty	Aj	Gtfs6x	Aj	Gtfs6x
110	zohriate	zohriaty	Aj	Gtns1x	Vb	VKdpb-
66	zohriate	zohriaty	Aj	Gtns1x	Vb	VKdpb+

Data to Annotate

	Id	Word	Lemma	Lemmb	bTag	aTag
0	sk_076396	zohrávka	zohrávka	zohrávka		Nn
1	sk_076397	zohrávke	zohrávke	zohrávke		Nn
2	sk_076398	zohrávku	zohrávku	zohrávku		Nn
0	sk_076399	zohrávky	zohrávky	zohrávky		Nn
1	sk_076400	zohriata	zohriaty	zohriaty		Aj
2	sk_076401	zohriateho	zohriati	zohriati		Aj
0	sk_076402	zohriateho	zohriatí	zohriatí		Aj
1	sk_076403	zohriatej	zohriaty	zohriaty		Aj
2	sk_076404	zohriate	zohriaty	zohriaty		Vb

The Results

	Count	%	%
Assigned lines	77,169	100.00	
Lines annotated at least once	76,413	99.02	
Lines annotated twice	60,048	77.81	100.00
Lines agreed on lemma	39,469	51.15	65.73
Lines agreed on lemma and PoS	33,371	43.24	55.57