Možnosti počítačové analýzy staroslověnských a církevněslovanských textů

Miroslav Vepřek (FF UP) – Radek Čech (FF MU)

Obsah

1. Digitalizace textů a formální úprava pro počítačové analýzy
2. Stylometrická analýza csl. textů českého původu
3. Automatická analýza prostřednictvím Universal Dependencies
4. Geoinformatické zobrazení

Digitalizace a formální úprava

- ruční přepis textů vs. OCR
- font Bukyvede
- sjednocení punktace
- rozepsání zkratek
- speciální znak pro číselnou platnost liter
- sjednocení variantních grafémů: и (и, й, ı, ї, ι), оу (оу, у, ѹ) atp.

Stylometrická analýza

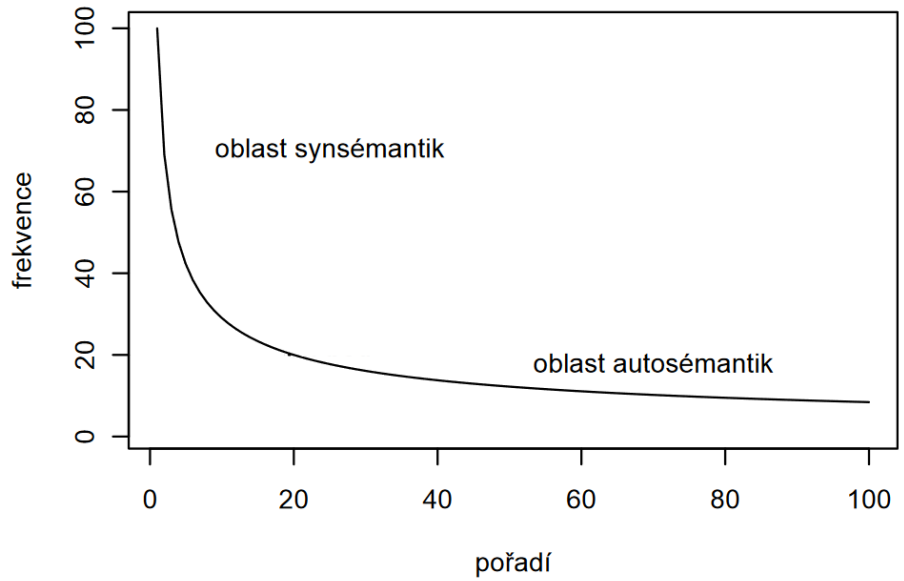
- ověření možnosti aplikace stylometrických metod na csl. texty
- heuristika & hypotézy
- blízkost/vzdálenost památek
  - autorský okruh
  - žánrová blízkost
  - geografické faktory
  - …

Stylometrická analýza - metody

- lexikální diverzita / slovní bohatství
  - klouzavý průměr poměru typů a tokenů
  - proporce hapax legomenon
  - …
- délka jazykových jednotek
  - slovo
  - věta
  - fráze
  - …
- analýza nejfrekventovanějších slov
  - porovnávání proporcí
  - většinou jde o synsémantika – eliminace vlivu tématu

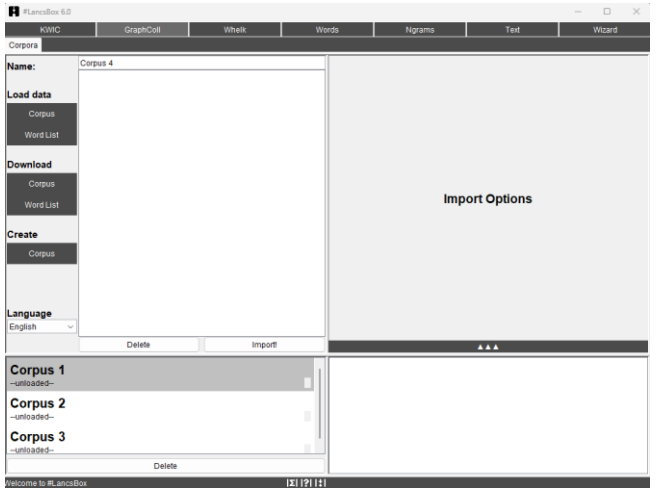
Stylometrická analýza - metody

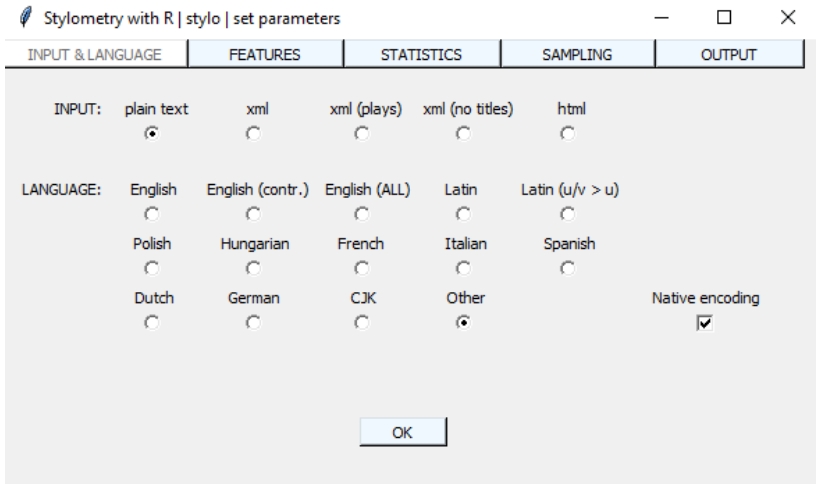
- lexikální diverzita / slovní bohatství
  - klouzavý průměr poměru typů a tokenů
  - proporce hapax legomenon
  - …
- délka jazykových jednotek
  - slovo
  - věta
  - fráze
  - …
- analýza nejfrekventovanějších slov
  - porovnávání proporcí
  - většinou jde o synsémantika – eliminace vlivu tématu



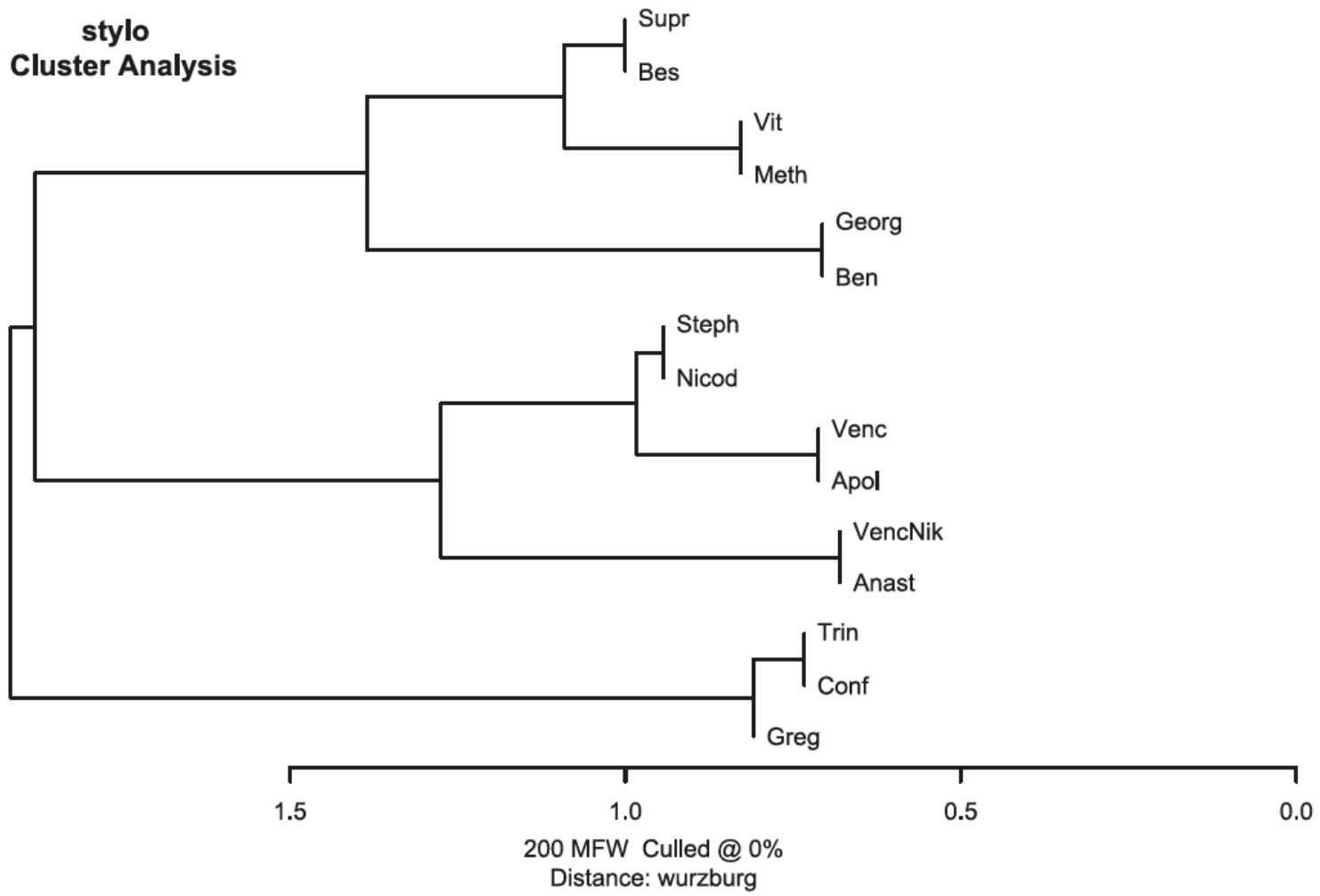
Stylometrická analýza - nástroje

- QuitaUp
- Stylo
- LancBox
- …

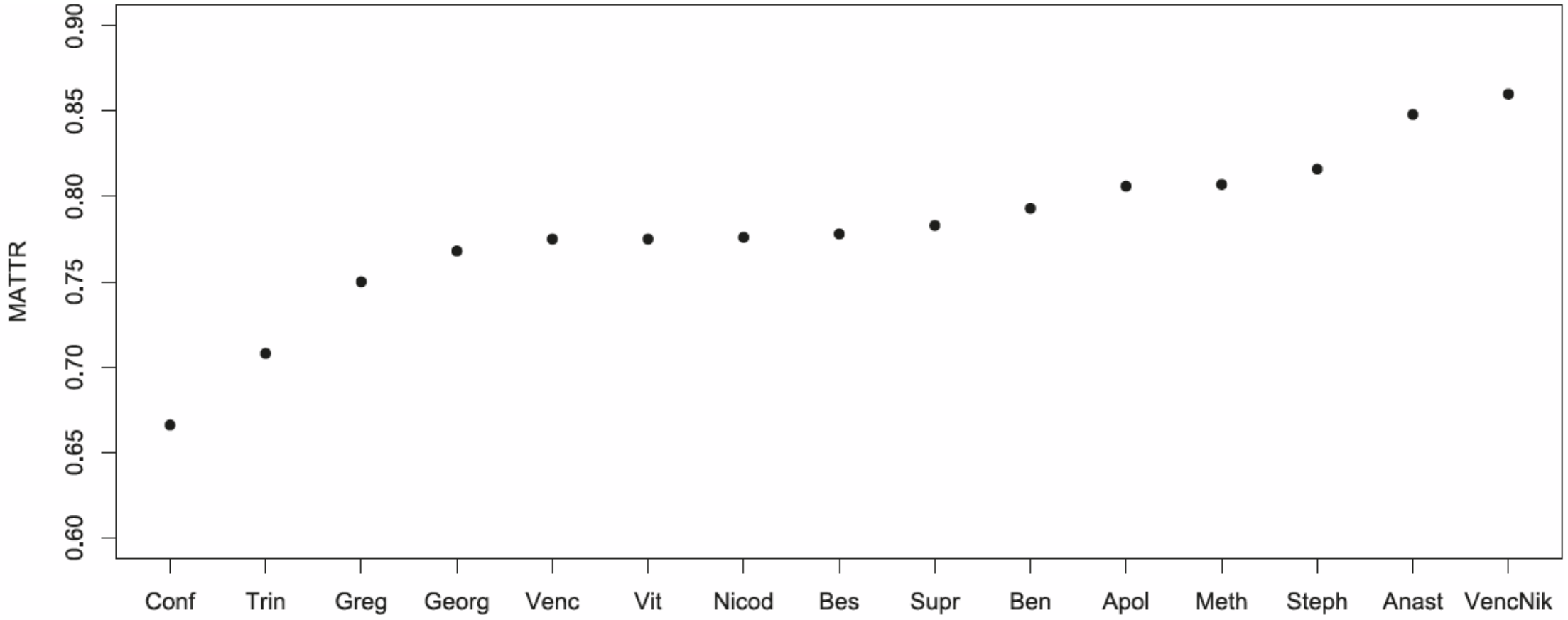




Stylometrická analýza – dílčí výsledky



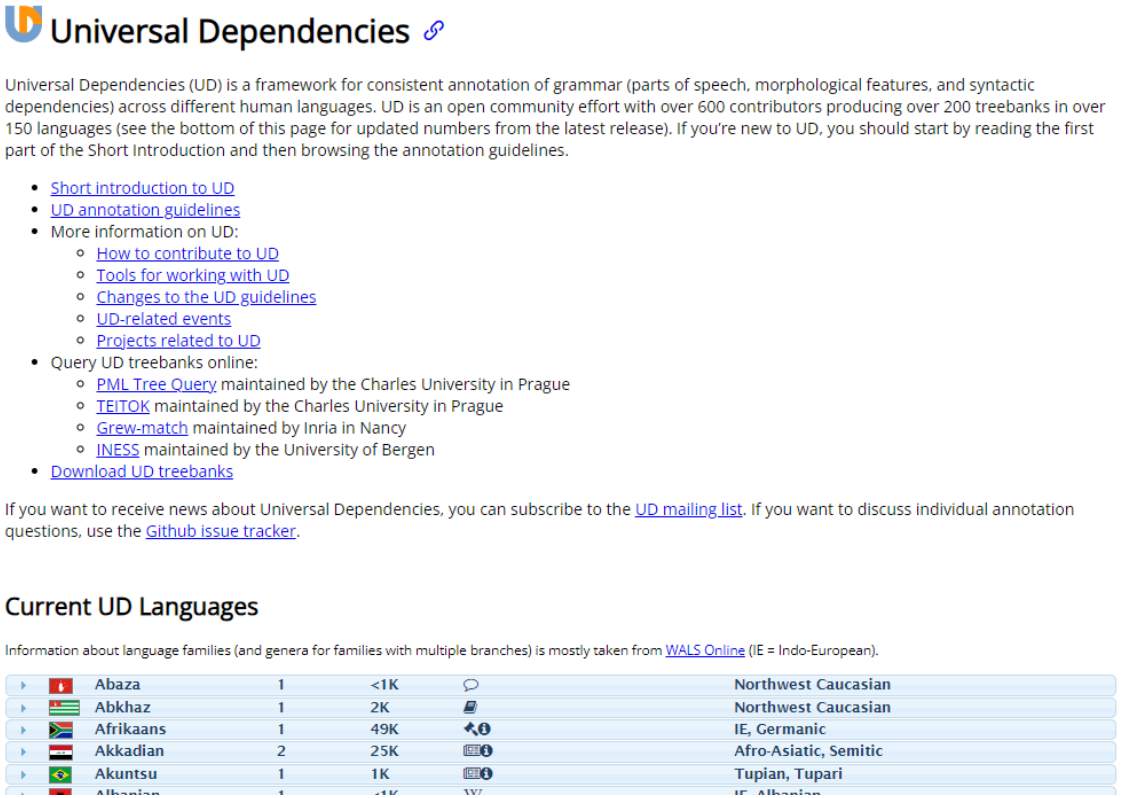
Stylometrická analýza – dílčí výsledky



Universal Dependecies

- jednotný anotační systém
  - lemmatizace
  - morfologické značkování
  - syntax
    - syntaktické funkce
    - dependenční stromy

- https://universaldependencies.org/



Universal Dependencies

Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with over 600 contributors producing over 200 treebanks in over 150 languages (see the bottom of this page for updated numbers from the latest release). If you're new to UD, you should start by reading the first part of the Short Introduction and then browsing the annotation guidelines.

- Short introduction to UD
- UD annotation guidelines
- More information on UD:
  - How to contribute to UD
  - Tools for working with UD
  - Changes to the UD guidelines
  - UD-related events
  - Projects related to UD
- Query UD treebanks online:
  - PML Tree Query maintained by the Charles University in Prague
  - TEITOK maintained by the Charles University in Prague
  - Grew-match maintained by Inria in Nancy
  - INESS maintained by the University of Bergen
- Download UD treebanks

If you want to receive news about Universal Dependencies, you can subscribe to the UD mailing list. If you want to discuss individual annotation questions, use the Github issue tracker.

Current UD Languages

Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

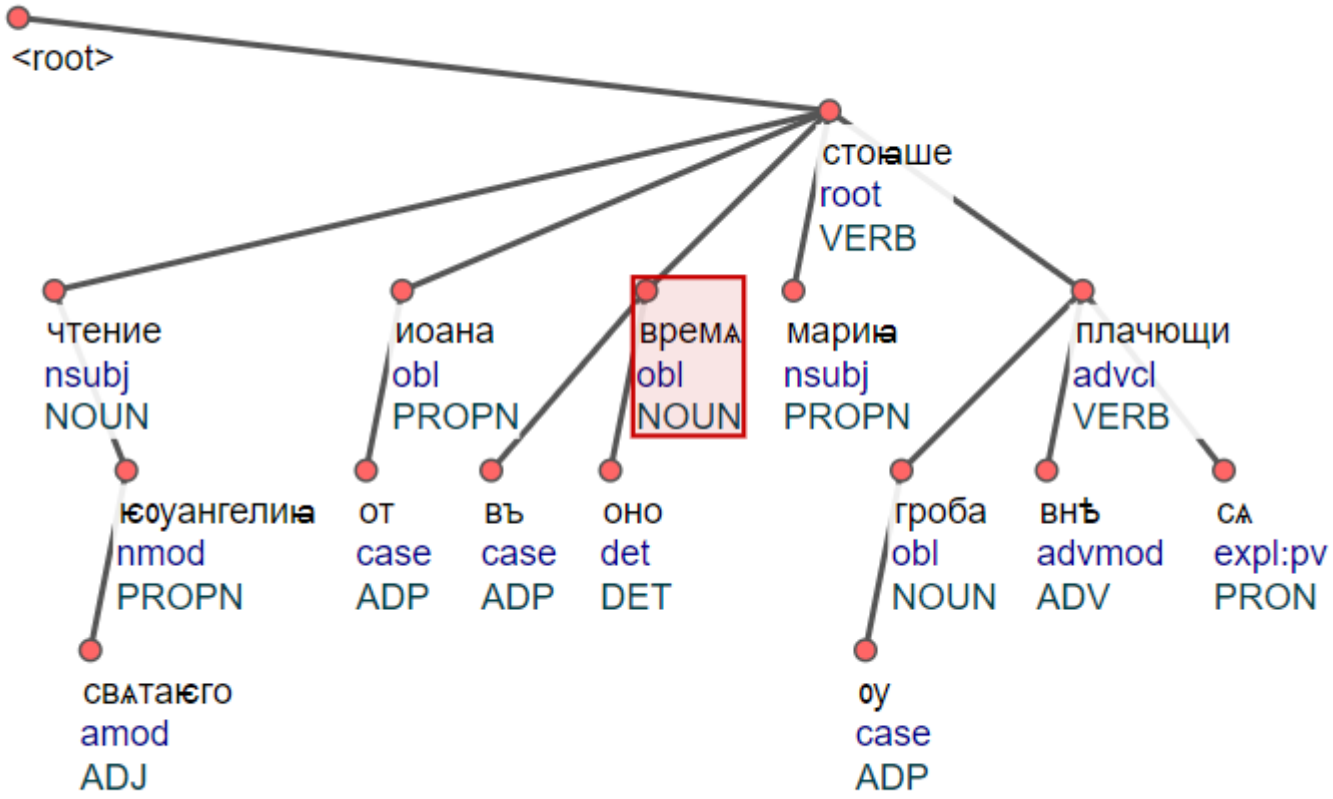
Abaza	1	<1K		Northwest Caucasian
Abkhaz	1	2K		Northwest Caucasian
Afrikaans	1	49K		IE, Germanic
Akkadian	2	25K		Afro-Asiatic, Semitic
Akuntsu	1	1K		Tupian, Tupari

Universal Dependecies

чтение сѧтаѥго ѥѳуангелиꙗ от иоана въ оно времѧ мариꙗ стоꙗше ѳу гроба внѣ плачющи сѧ

Hide empty attributes

deprel	obl
feats	Case=Acc|Gender=Neut|Number=Sing
form	времѧ
head	10
id	8
lemma	времѧ
misc	SpacesAfter=\s\s|TokenRange=46:51
upostag	NOUN
xpostag	Nb



Universal Dependecies

sent_id = 1

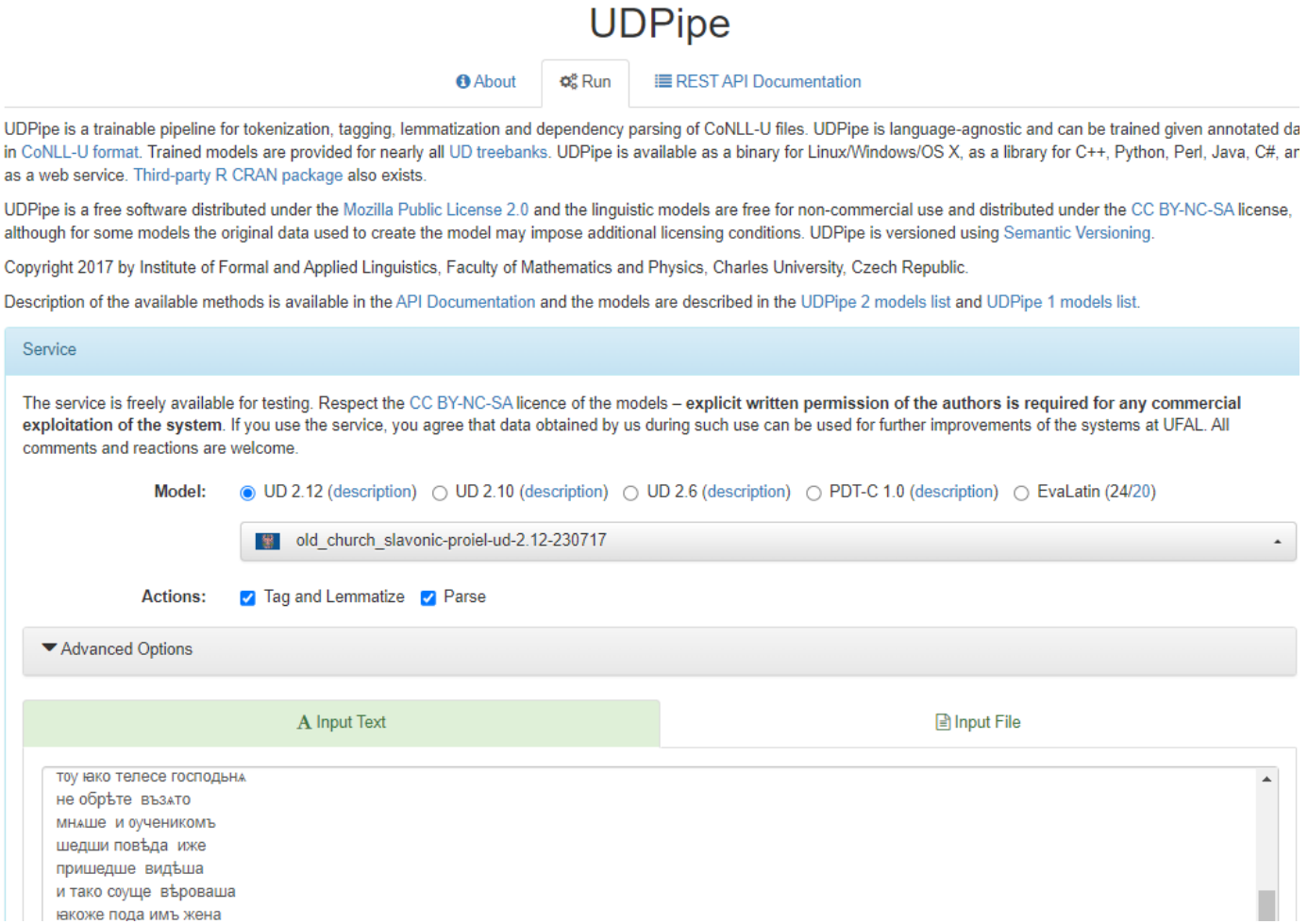
text = чтение сѫтаѥго ѭоуангелиꙗ от иоана въ оно времѧ мариꙗ стоꙗше оу гроба внѣ плачющи сѧ

1	чтение	чтение	NOUN	Nb	Case=Nom|Gender=Neut|Number=Sing	10	nsubj	_	TokenRange=0:6
2	сѫтаѥго	сѫтъ	ADJ	A-	Case=Gen|Degree=Pos|Gender=Masc|Number=Sing	3	amod	_	TokenRange=7:15
3	ѭоуангелиꙗ	ѭоуангелии	PROPN	Ne	Case=Gen|Gender=Masc|Number=Sing	1	nmod	_	SpacesAfter=\s\s|TokenRange=16:25
4	от	отъ	ADP	R-	_	5	case	_	TokenRange=27:29
5	иоана	иоанъ	PROPN	Ne	Case=Gen|Gender=Masc|Number=Sing	10	obl	_	SpacesAfter=\s\r\n\s|TokenRange=30:35
6	въ	въ	ADP	R-	_	8	case	_	TokenRange=39:41
7	оно	онъ	DET	Pd	Case=Acc|Gender=Neut|Number=Sing	8	det	_	TokenRange=42:45
8	времѧ	времѧ	NOUN	Nb	Case=Acc|Gender=Neut|Number=Sing	10	obl	_	SpacesAfter=\s\s|TokenRange=46:51
9	мариꙗ	мариꙗ	PROPN	Ne	Case=Nom|Gender=Fem|Number=Sing	10	nsubj	_	SpacesAfter=\s\r\n|TokenRange=53:58
10	стоꙗше	стоꙗти	VERB	V-	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act	0	root	_	TokenRange=61:67
11	оу	оу	ADP	R-	_	12	case	_	TokenRange=68:69
12	гроба	гробъ	NOUN	Nb	Case=Gen|Gender=Masc|Number=Sing	14	obl	_	SpacesAfter=\s\s|TokenRange=70:75
13	внѣ	внѣ	ADV	Df	_	14	advmod	_	SpacesAfter=\r\n|TokenRange=77:80
14	плачющи	плакати	VERB	V-	Tense=Pres|VerbForm=Inf|Voice=Act	10	advcl	_	TokenRange=82:89
15	сѧ	себе	PRON	Pk	Case=Acc|Number=Sing|Person=3|PronType=Prs|Reflex=Yes	14	expl:pv	_	SpacesAfter=\s\s|TokenRange=90:92

UDPipe

- nástroj pro automatickou analýzu vlastních textů

https://lindat.mff.cuni.cz/services/udpipe/



UDPipe - staroslověnština

- nástroj trénován na *The Old Church Slavonic UD treebank*
  - *Codex Marianus New Testament* translation (parts of the Gospels)
  - *Codex Suprasliensis* (lives of saints and homilies),
  - *Kiev Missal* (liturgy)
  - *Psalterium Sinaiticum*
  - extracts from the *Codex Zographensis New Testament manuscript*

- více viz:

https://github.com/UniversalDependencies/UD_Old_Church_Slavonic-PROIEL/blob/master/README.md

UDPipe - staroslověnština

UD for Old Church Slavonic

Tokenization and Word Segmentation

- In general, words are delimited by whitespace characters.
- Punctuation such as commas and periods is not included in the data. Occasionally a punctuation symbol (typically a hyphen) is part of a word as in *вы-инѣ*.
- There are no multi-word tokens.
- There are no words with spaces.

Morphology

Tags

- Old Church Slavonic uses 14 universal POS categories. There are no particles, punctuation and other symbols in the data.
- The only auxiliary verb (AUX) in Old Church Slavonic is *быти* "to be". It is used as copula (*съ ними* ***есть*** *женихъ* "the bridegroom is with them").
- There are five main (de)verbal forms, distinguished by the value of the VerbForm feature:
  - Infinitive `Inf`, tagged VERB or AUX.
  - Finite verb `Fin`, tagged VERB or AUX.
  - Participle `Part`, tagged VERB or AUX.
  - Resultative participle `PartRes`, tagged VERB or AUX.
  - Supine `Sup`, tagged VERB or AUX.

Nominal Features

- Nominal words (NOUN, PROPN and PRON) have an inherent Gender feature with one of three values: `Masc`, `Fem` or `Neut`.
  - The following parts of speech inflect for `Gender` because they must agree with nouns: ADJ, DET, NUM, VERB, AUX. For verbs (includin auxiliaries), only participles can inflect for gender. Finite verbs don't.
- The three values of the Number feature are `Sing`, `Dual`, and `Plur`. The following parts of speech inflect for number: NOUN, PROPN, PRON, ADJ, DET, NUM, VERB, AUX (finite and participles).
- Case has 7 possible values: `Nom`, `Gen`, `Dat`, `Acc`, `Ins`, `Loc`, `Voc`. It occurs with the nominal words, i.e., NOUN, PROPN, PRON, ADJ, DET, NUM. For verbs (VERB) and auxiliaries (AUX) it occurs with participles (`VerbForm=Part`).

https://universaldependencies.org/cu/index.html

UDPipe – výsledky

ł = 3					
Се нънѣ сбыстъ сѧ пророческое слово еже глаголааше господь нашь исѳусъ христосъ					
1	Се	се	INTJ	I-	_
2	нънѣ	нънѣ	ADV	Df	_
3	сбыстъ	сбыти	VERB	V-	Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act
4	сѧ	себе	PRON	Pk	Case=Acc|Number=Sing|Person=3|PronType=Prs|Reflex=Yes
5	пророческое	пророческъ	ADJ	A-	Case=Nom|Degree=Pos|Gender=Neut|Number=Sing
6	слово	слово	NOUN	Nb	Case=Nom|Gender=Neut|Number=Sing
7	еже	иже	PRON	Pr	Case=Acc|Gender=Neut|Number=Sing|PronType=Rel
8	глаголааше	глаголати	VERB	V-	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act
9	господь	господь	NOUN	Nb	Case=Nom|Gender=Masc|Number=Sing
10	нашь	нашь	DET	Ps	Case=Nom|Gender=Masc|Number=Sing|Person=1|Poss=Yes
11	исѳусъ	исѳусъ	PROPN	Ne	Case=Nom|Gender=Masc|Number=Sing
12	христосъ	христ(ос)ъ	PROPN	Ne	Case=Nom|Gender=Masc|Number=Sing

UDPipe – výsledky

id = 22					
тыı сѣдиши о десноую бога отьца владъıи въ вѣкъı азъ тѧ молю отпоусти грѣхъı моѧ					
1	тыı	тъı	PRON	Pp	Case=Nom|Gender=Masc|Number=Sing|Person=2|PronType=Prs
2	сѣдиши	сѣдѣти	VERB	V-	Mood=Ind|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin|Voice=Act
3	о	о	ADP	R-	_
4	десноую	деснъ	ADJ	A-	Case=Acc|Degree=Pos|Gender=Masc|Number=Plur
5	бога	богъ	NOUN	Nb	Case=Gen|Gender=Masc|Number=Sing
6	отьца	отьць	NOUN	Nb	Case=Gen|Gender=Masc|Number=Sing
7	владъıи	владъ	VERB	V-	Case=Nom|Gender=Masc|Number=Sing|Tense=Pres|VerbForm=Part|Voice=Act
8	въ	въ	ADP	R-	_
9	вѣкъı	вѣкъ	NOUN	Nb	Case=Acc|Gender=Masc|Number=Plur
10	азъ	азъ	PRON	Pp	Case=Nom|Number=Sing|Person=1|PronType=Prs
11	тѧ	тъı	PRON	Pp	Case=Acc|Number=Sing|Person=2|PronType=Prs
12	молю	молѥ	VERB	V-	_
13	отпоусти	отпоустити	VERB	V-	Mood=Imp|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin|Voice=Act
14	грѣхъı	грѣхъ	NOUN	Nb	Case=Acc|Gender=Masc|Number=Plur
15	моѧ	мои	DET	Ps	Case=Acc|Gender=Masc|Number=Plur|Person=1|Poss=Yes

UDPipe – výsledky

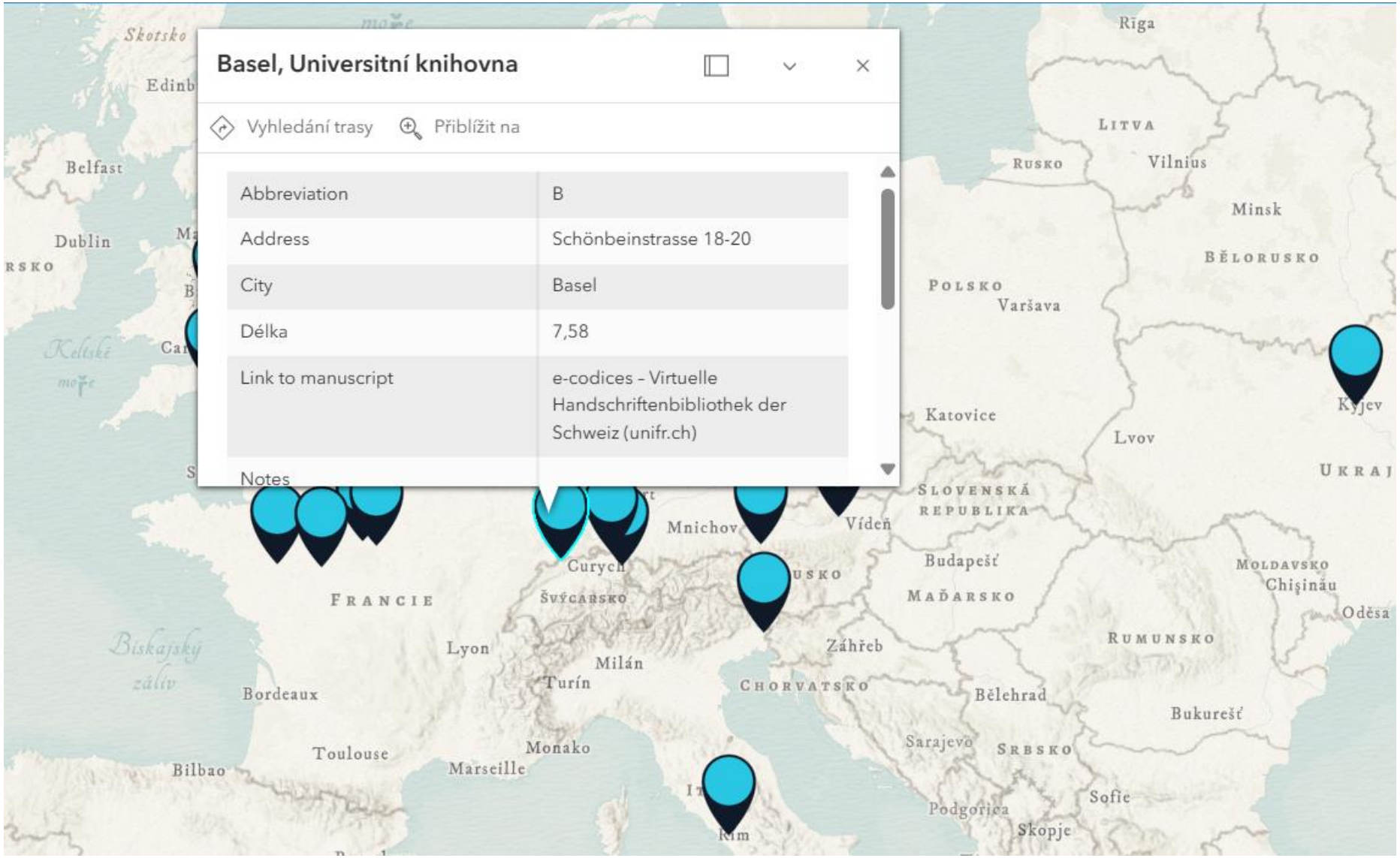
	A	B	C	D	E
1	Bes_30				
2		tokenizace	lemmatizace	POS	tag
3	celkový počet	132	132	132	132
4	chyby	2	10	5	7
5	úspěšnost	0,985	0,924	0,962	0,947
6					
7					
8	vyznani_hrichu				
9	celkový počet	335	335	335	335
10	chyby	4	16	10	18
11	úspěšnost	0,988	0,952	0,970	0,946
12					
13					
14	Venc				
15	celkový počet	183	183	183	183
16	chyby	2	13	10	18
17	úspěšnost	0,989	0,929	0,945	0,902

UDPipe – možnosti využití

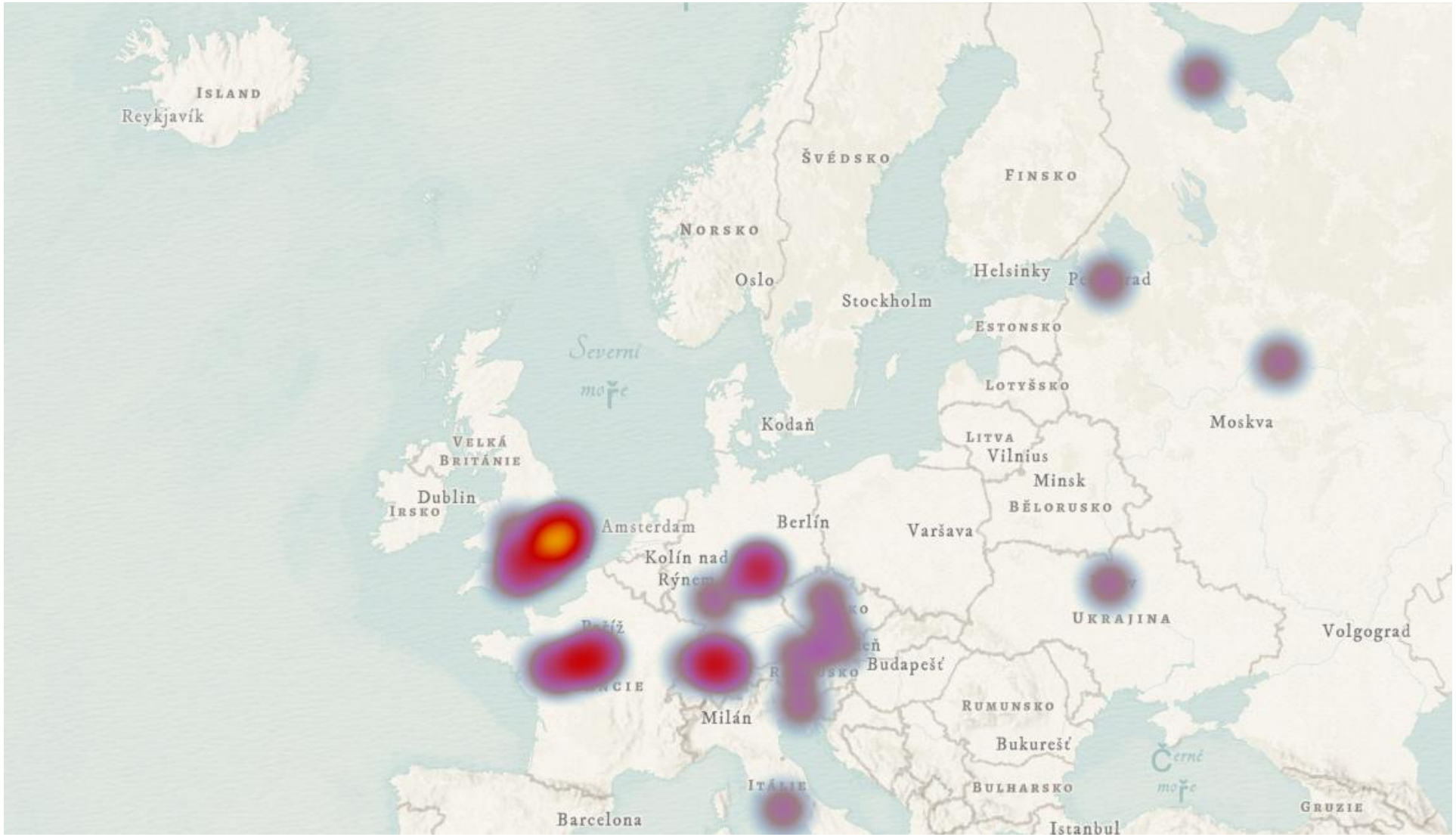
- frekvenční slovníky k památkám
- lemmatizace textu – index verborum
- podklady pro filologickou analýzu
  - slovní druhy
  - morfologické kategorie
  - slovosled
  - syntaktické vztahy
  - …
- kvantitativní stylistika
  - proporce vybraných kategorií

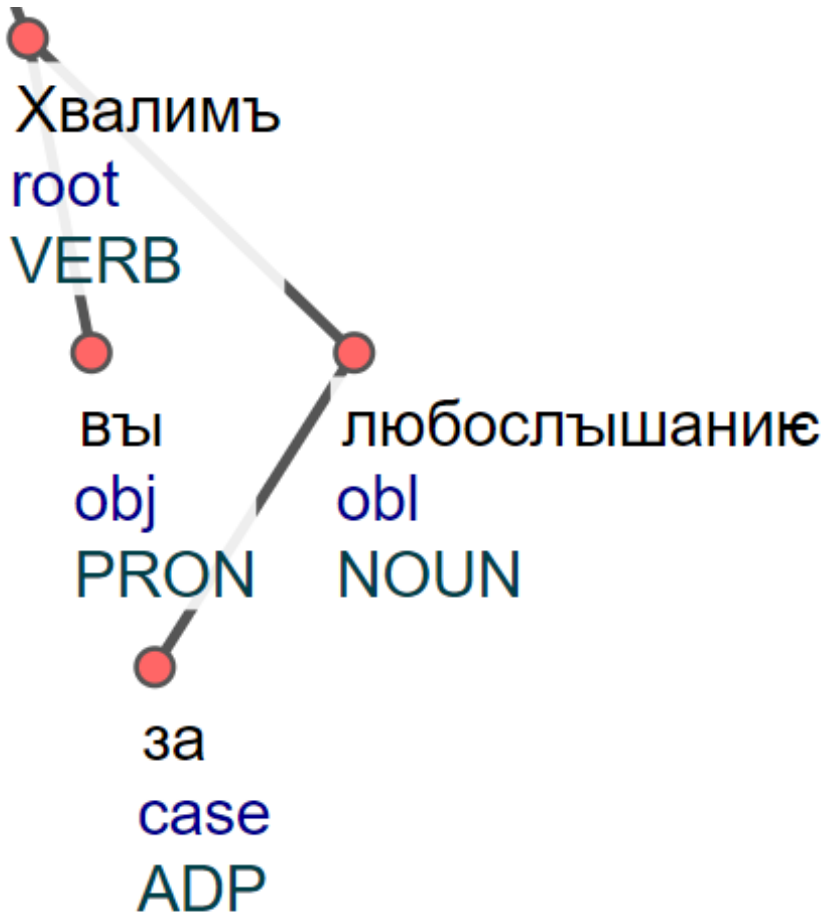
Geoinformatika

Geoinformatika



Geoinformatika




Хвалимъ
root
VERB
въı
obj
PRON
любослъıшанию
obl
NOUN
за
case
ADP