Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

STAATSBIBLIOTHEK ZU BERLIN – PK
Osteuropa-Abteilung

Online-Workshop:

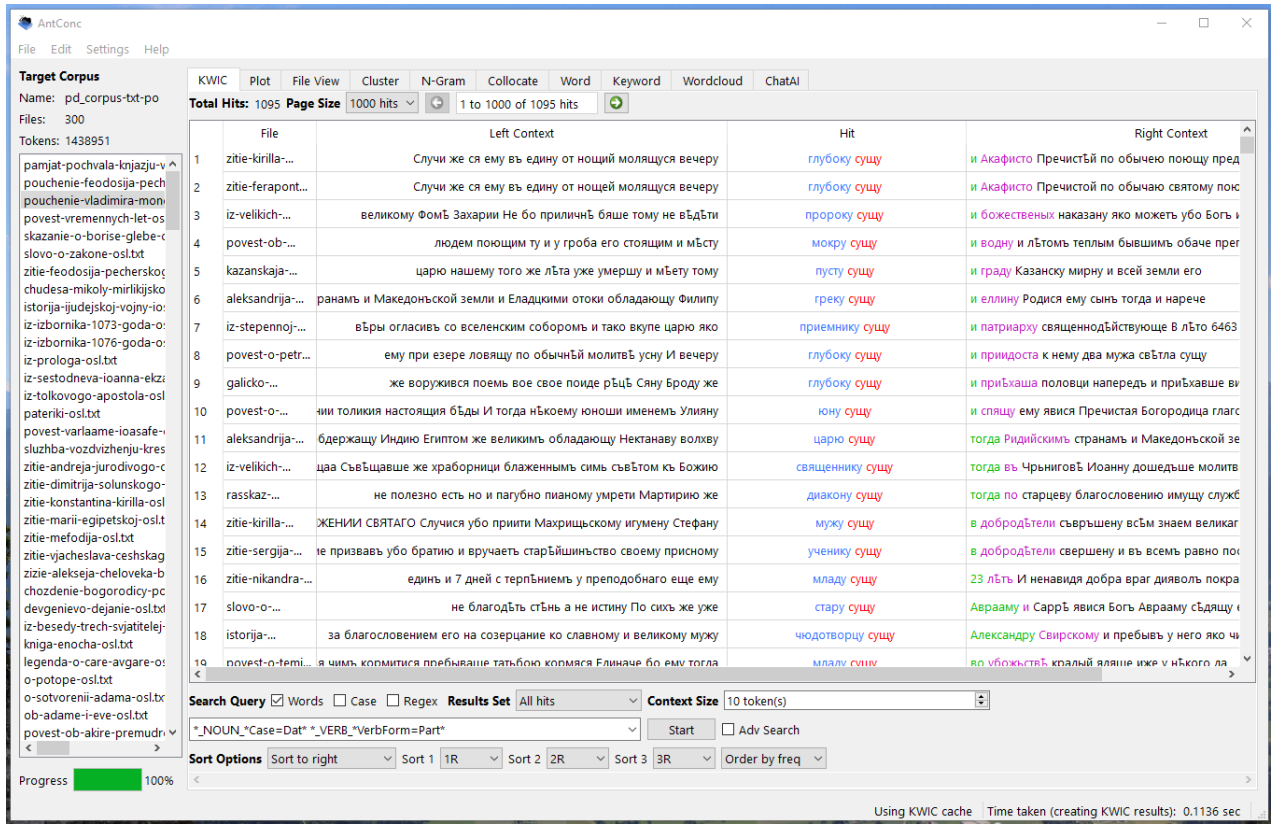
Quantitative Analyse slawischer historischer Texte

– Korpuslinguistische Methoden und Suchtechniken mit AntConc
(Kollokationen, Keywords, N-Grams, annotierte Daten)

Vladimir Neumann
23.1.2025

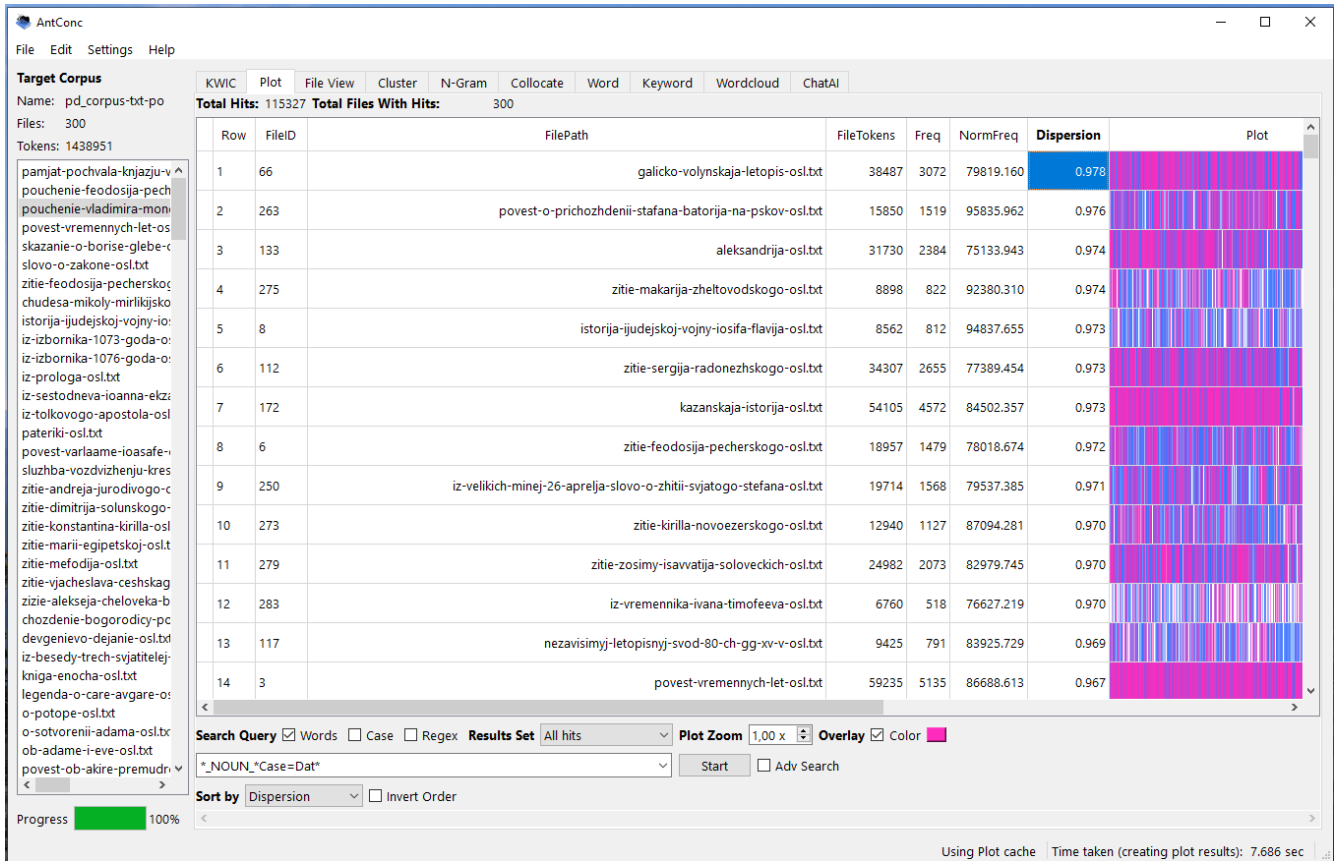
Überblick

- Grundlagen der quantitativen Textanalyse
  - Untersuchung linguistischer Muster in Texten
- Warum AntConc?
  - Vorstellung von „AntConc“, einem kostenfreien und plattformunabhängigen Tool zur Korpusanalyse
- Slawische historische Texte und Korpora
  - Typische Merkmale, Datenmodellierung, Erstellung eigener Korpora
- Praxis mit AntConc:
  - KWIC, Cluster & N-Gram, Collocate Tool, Keywords
- Arbeiten mit annotierten Daten



Grundlagen der quantitativen Textanalyse

- Untersuchung sprachlicher Muster in Texten
- Nutzung moderner korpuslinguistischer Techniken
- Statistische Auswertung von Textdaten für historische und linguistische Fragestellungen
- Datengetriebenes Lernen und Forschungsansätze
  - Starten mit AntConc, dann Wechseln zur fortgeschritteneren Applikationen oder Methodiken (Python, Natural Language Processing, Maschine Learnning)



Warum AntConc?

- „AntConc“
  - Tool zur Korpusanalyse: kostenfrei und plattformunabhängig
  - leichtfüßig, aber funktionsstark
- Kernfunktionen:
  - KWIC (Key-Word-In-Context)
  - Plot-Analyse
  - Wortlisten
  - Cluster und N-Gram Tools
- Unterstützung für:
  - Komplexe Suchanfragen und
  - Suche in annotierten Daten

- AntConc-Download
- Kernfunktionen

 · CC BY 4.0

Slawische historische Texte und Korpora

- Slawische historische Texte und Korpora
  - typische Merkmale, große Korpora
- Eigene Sammlungen von Texten / Korpora
  - Grundlagen der Datenmodellierung
  - Erstellung von eigener Korpora
- Herausforderungen:
  - Sammeln von Daten
  - Uneinheitliche Orthographie
  - Sprachliche Variation

- Anthologie-Charakter: früher gedruckt, heute digital
- Korpora in Slawistik: CorpSlav-Collection, National-Korpora, Spezial-Korpora, RRuDi, PolDi, VMČ, Drevnerusskij Korpus, Cyrillomethodiana (BG), Kirchenslavica-Digital
- Probleme: Datenformate sind oft proprietär (ANNIS), Systeme nicht völlig offen (SketchEngine, NoSketchEngine), Copyright / Urheberrecht
- Digital Humanities: Daten und Datenmodellierung
- Daten-Beschaffung (Internet, Intranet), -Homogenisierung und –Normalisierung
- Beispiele:
  - Bibliotheka Literatury Drevnej Rusi
  - Kirchenslavica-Digital (Barlaam)
  - Syntacticus-Treebank (Github)
  - lib.pravmir.ru

 · CC BY 4.0

Praktische Anwendungen

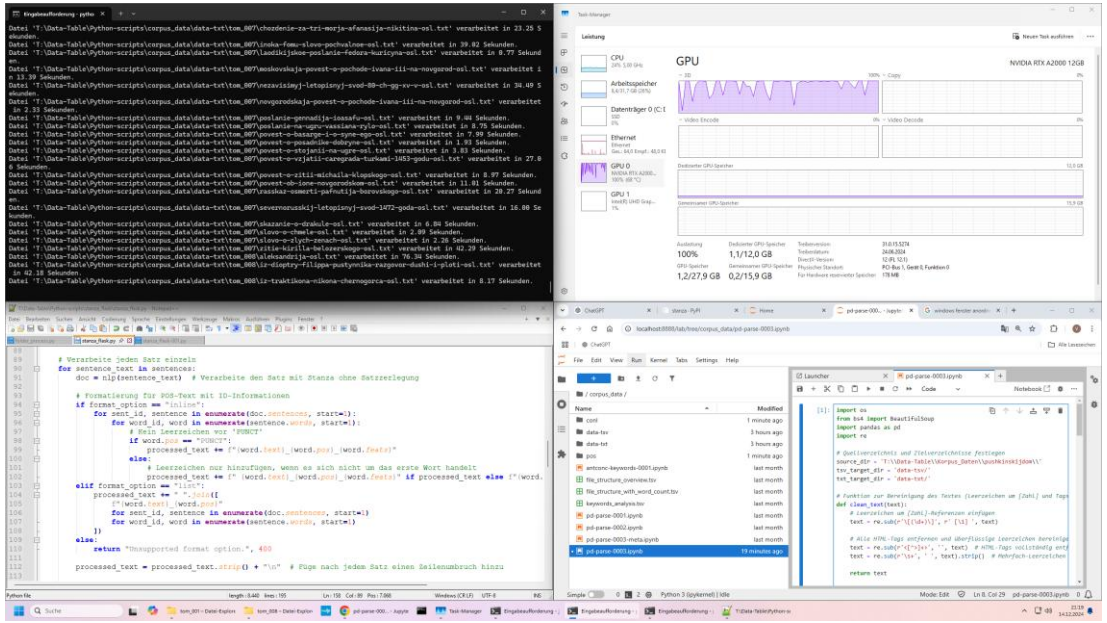
- KWIC: Analyse der syntaktischen Umgebung und semantischen Variationen relevanter Begriffe
- Cluster & N-Gram: Identifikation wiederkehrender Phrasen und stilistischer Muster
- Collocate Tool: Untersuchung typischer Wortverbindungen und semantischer Felder
- Keyword: Analyse durch Vergleich: Target- und Referenzkorpora

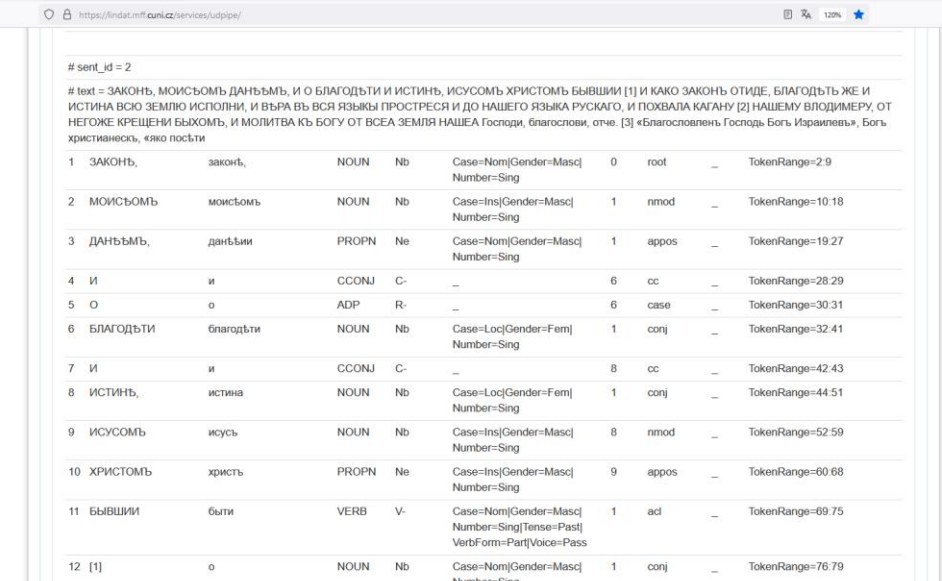
- Präpositionen, Adjektive, Partizipien, Pronomena / Trunkierung, Reguläre Ausdrücke
- гд[еѣ] vs. кд[еѣ]
- zeigt benachbarte Wortgruppen auf der Grundlage der Suchbedingung an
- durchsucht das gesamte Korpus nach allen Clustern der Größe 'N'
- zeigt Wörter an mit einem bestimmten Abstand zum Suchbegriff
- zeigt Wörter an, die im Zielkorpus im Vergleich zu den Wörtern im Referenzkorpus auf der Grundlage eines statistischen Maßes (d. h. „Schlüsselwörter“) ungewöhnlich häufig vorkommen

 · CC BY 4.0

Arbeiten mit annotierten Daten

- Einsatz von Tags wie POS (Parts of Speech) und Lemmata
- Pos-„Syntax“ im Antconc: Type_POS(UD)
- Nutzung von annotierten Daten für zielgerichtete linguistische Analysen
- Einführung in Regular Expressions (Regex) zur Suchoptimierung.





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

sent_id = 2

text = ЗАКОНѢ, МОИСѢОМѢ ДАНѢѢМѢ, И О БЛАГОДѢТИ И ИСТИНѢ, ИСУСОМѢ ХРИСТОМѢ БЫВШИИ [1] И КАКО ЗАКОНѢ ОТИДЕ, БЛАГОДѢТѢ ЖЕ И ИСТИНА ВСЮ ЗЕМЛЮ ИСПОЛНИ, И ВѢРА ВѢ ВСЯ ЯЗЫКЫ ПРОСТРЕСЯ И ДО НАШЕГО ЯЗЫКА РУСКАГО, И ПОХВАЛА КАГАНУ [2] НАШЕМУ ВЛОДИМЕРУ, ОТ НЕГОЖЕ КРЕЩЕНИ БЫХОМѢ, И МОЛИТВА КѢ БОГУ ОТ ВСЕА ЗЕМЛЯ НАШЕА Господи, благослови, отче. [3] «Благословленѣ Господь Богъ Израилевѣ», Богъ христианескъ, «яко посѣти

1	ЗАКОНѢ,	законѣ,	NOUN	Nb	Case=Nom|Gender=Masc|Number=Sing	0	root	_	TokenRange=2:9
2	МОИСѢОМѢ	моисѣомъ	NOUN	Nb	Case=Ins|Gender=Masc|Number=Sing	1	nmod	_	TokenRange=10:18
3	ДАНѢѢМѢ,	данѣѣии	PROPN	Ne	Case=Nom|Gender=Masc|Number=Sing	1	appos	_	TokenRange=19:27
4	И	и	CCONJ	C-	_	6	cc	_	TokenRange=28:29
5	О	о	ADP	R-	_	6	case	_	TokenRange=30:31
6	БЛАГОДѢТИ	благодѣти	NOUN	Nb	Case=Loc|Gender=Fem|Number=Sing	1	conj	_	TokenRange=32:41
7	И	и	CCONJ	C-	_	8	cc	_	TokenRange=42:43
8	ИСТИНѢ,	истина	NOUN	Nb	Case=Loc|Gender=Fem|Number=Sing	1	conj	_	TokenRange=44:51
9	ИСУСОМѢ	исусъ	NOUN	Nb	Case=Ins|Gender=Masc|Number=Sing	8	nmod	_	TokenRange=52:59
10	ХРИСТОМѢ	христъ	PROPN	Ne	Case=Ins|Gender=Masc|Number=Sing	9	appos	_	TokenRange=60:68
11	БЫВШИИ	быти	VERB	V-	Case=Nom|Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part|Voice=Pass	1	acl	_	TokenRange=69:75
12	[1]	о	NOUN	Nb	Case=Nom|Gender=Masc|	1	conj	_	TokenRange=76:79

```
O_ADP_None ЗАКОНѢ_PROPN_Case=Loc|Gender=Masc|Number=Sing,_PUNCT_None МОИСѢОМѢ_PROPN_Case=Nom|Gender=Masc|Number=Sing
ДАНѢѢМѢ_PROPN_Case=Nom|Gender=Masc|Number=Sing,_PUNCT_None И_CCONJ_None O_ADP_None
БЛАГОДѢТИ_PROPN_Case=Loc|Gender=Masc|Number=Sing И_CCONJ_None ИСТИНѢ_PROPN_Case=Loc|Gender=Masc|Number=Sing,_PUNCT_None
ИСУСОМѢ_PROPN_Case=Loc|Gender=Masc|Number=Sing ХРИСТОМѢ_PROPN_Case=Loc|Gender=Masc|Number=Sing
БЫВШИИ_PROPN_Case=Loc|Gender=Masc|Number=Sing И_CCONJ_None КАКО_ADV_Degree=Pos ЗАКОНѢ_PROPN_Case=Gen|Gender=Masc|Number=Sing
ОТИДЕ_PROPN_Case=Loc|Gender=Masc|Number=Sing,_PUNCT_None БЛАГОДѢТѢ_PROPN_Case=Loc|Gender=Masc|Number=Sing ЖЕ_PART_None
И_CCONJ_None ИСТИНА_PROPN_Case=Loc|Gender=Masc|Number=Sing ВСЮ_ADP_None ЗЕМЛЮ_PROPN_Case=Loc|Gender=Masc|Number=Sing
ИСПОЛНИ_PROPN_Case=Loc|Gender=Masc|Number=Sing,_PUNCT_None И_CCONJ_None ВѢРА_PROPN_Case=Nom|Gender=Masc|Number=Sing ВѢ_ADP_None
ВСЯ_DET_Case=Loc|Gender=Masc|Number=Sing|PronType=Dem ЯЗЫКЫ_PROPN_Case=Loc|Gender=Masc|Number=Sing
ПРОСТРЕСЯ_PROPN_Case=Gen|Gender=Masc|Number=Sing И_CCONJ_None ДО_ADP_None НАШЕГО_PROPN_Case=Gen|Gender=Masc|Number=Sing
ЯЗЫКА_PROPN_Case=Gen|Gender=Masc|Number=Sing РУСКАГО_PROPN_Case=Gen|Gender=Masc|Number=Sing,_PUNCT_None И_CCONJ_None
ПОХВАЛА_NOUN_Case=Gen|Gender=Fem|Number=Sing КАГАНУ_NOUN_Case=Gen|Gender=Fem|Number=Sing
НАШЕМУ_PROPN_Case=Gen|Gender=Masc|Number=Sing ВЛОДИМЕРУ_NOUN_Case=Dat|Gender=Masc|Number=Sing,_PUNCT_None ОТ_ADP_None
НЕГОЖЕ_PROPN_Case=Gen|Gender=Masc|Number=Sing КРЕЩЕНИ_NOUN_Case=Gen|Gender=Masc|Number=Sing
БЫХОМѢ_PROPN_Case=Gen|Gender=Masc|Number=Sing,_PUNCT_None И_CCONJ_None МОЛИТВА_PROPN_Case=Nom|Gender=Masc|Number=Sing
КѢ_ADP_None БОГУ_NOUN_Case=Dat|Gender=Masc|Number=Sing ОТ_ADP_None ВСЕА_DET_Case=Gen|Gender=Fem|Number=Sing|PronType=Tot
ЗЕМЛЯ_PROPN_Case=Gen|Gender=Masc|Number=Sing НАШЕА_PROPN_Case=Gen|Gender=Masc|Number=Sing
Господи_NOUN_Case=Nom|Gender=Masc|Number=Sing,_PUNCT_None
благослови_VERB_Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act,_PUNCT_None
отче_ADV_Degree=Pos._PUNCT_None«_PUNCT_None
Благословленѣ_VERB_Aspect=Perf|Case=Nom|Gender=Masc|Number=Sing|Tense=Past|Variant=Short|VerbForm=Part|Voice=Pass
Господь_NOUN_Case=Nom|Gender=Masc|Number=Sing Богъ_NOUN_Case=Nom|Gender=Masc|Number=Sing
Израилевѣ_PROPN_Case=Nom|Gender=Masc|Number=Sing»_PUNCT_None,_PUNCT_None Богъ_NOUN_Case=Nom|Gender=Masc|Number=Sing
христианескъ_ADJ_Case=Nom|Degree=Pos|Gender=Masc|Number=Sing|Variant=Short,_PUNCT_None «яко_SCONJ_None
посѣти_VERB_Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act и_CCONJ_None
сѣтвори_VERB_Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act
избавление_NOUN_Case=Acc|Gender=Neut|Number=Sing людемь_NOUN_Case=Dat|Gender=Masc|Number=Plur
своимъ_DET_Case=Dat|Gender=Masc|Number=Plur|Poss=Yes|PronType=Prs|Reflex=Yes»_PUNCT_None,_PUNCT_None яко_SCONJ_None
не_PART_Polarity=Neg презрѣ_VERB_Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act до_ADP_None
конца_NOUN_Case=Gen|Gender=Masc|Number=Sing тварі_NOUN_Case=Gen|Gender=Masc|Number=Sing
своеа_DET_Case=Gen|Gender=Masc|Number=Sing|Poss=Yes|PronType=Prs|Reflex=Yes
идольскыимъ_ADJ_Case=Ins|Degree=Pos|Gender=Masc|Number=Sing мракомъ_NOUN_Case=Ins|Gender=Masc|Number=Sing
одержимѣ_ADJ_Case=Ins|Degree=Pos|Gender=Masc|Number=Sing быти_VERB_Aspect=Perf|VerbForm=Inf|Voice=Act и_CCONJ_None
бѣсовьскыимъ_ADJ_Case=Ins|Degree=Pos|Gender=Neut|Number=Sing служеваниемь_NOUN_Case=Ins|Gender=Neut|Number=Sing
гыбнути_VERB_Aspect=Imp|VerbForm=Inf|Voice=Act._PUNCT_None
 нъ_CCONJ_None оправдѣ_VERB_Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act прежде_ADV_Degree=Pos
племя_NOUN_Case=Nom|Gender=Neut|Number=Sing Авраамле_PROPN_Case=Gen|Gender=Masc|Number=Sing
```

Rückblick und Ausblick

- Fragen?
- Ausbau der Kenntnisse im zweiten Teil der Schulung am 10.2.2025:
  - Automatische Annotation von Kirchenslavica mit „Stanza“ – Anwendung auf klassische altkirchenslawische Texte und das Korpus der Kirchenslawischen Drucke
- Welche Schulungen im Bereich der Digital Humanities in Slawistik werden noch gewünscht?
  - Scripting mit Python (Daten-Harvesting, Daten Bereinigung, XML-Parsing, Visualisierungstechniken)
  - Arbeiten mit Daten-Indizes: Index-Aufbau, Index-Durchsuchbarkeit

 · CC BY 4.0

Vielen Dank, ich freue mich auf Ihre Fragen!

Dr. Vladimir Neumann
Fachreferat Slawistik (Schwerpunkt Belarus, Polen), Griechenland
Osteuropa-Abteilung, SBB
Vladimir.Neumann@sbb.spk-berlin.de
Tel. +4930 – 266 435640

Morphosyntaktische und lexikalische Besonderheiten

	Lat.	Ruth.	Poln.	
706	VERB.Part.Pres.Act.	Adv.Part.Pres. (-uči)	Adv.Part.Pres. (-ąc)	M+L: Adverbialpartizip des Präsens lat. Gleichzeitigkeit, pol. Gleichzeitigkeit (pol. unvollendete Verben)
88	VERB.Part.Pres.Act.Abl.	Adv.Part.Pres. + Noun.Acc.	Adv.Part.Pres. + Noun.Acc.	
2097	VERB.Part.Past.Pass.	Adv.Part.Pres. (-še/-vše)	Adv.Part.Pres. (-wszy)	M: aktives Adverbialpartizip des Präteritums; lat.+pol. Vozeitigkeit (pol. vollendete Verben)
	VERB.Part.Past.Pass. + esse.Fin.	Adj.Part.Past.Pass (-nyi, -no) + [est'] l-Partizip + sja Past (l-Partizip)	Adj.Part.Past.Pass (-ny, -no) + [iest] Part.Past.Pass (-ny, -no) + [iest] Past (l-Partizip)	L:Passives Präteritalpartizip
	VERB.Part.Past.Pass. + sit	Part.Past.Pass (-nyi, -no) + [est']	Part.Past.Pass (-ny, -no) + [iest]	Konjunktivform nach dem Gebrauch der romanischen Sprachen
168	VERB.Part.Fut.Act.	Pres.	Pres. [+ sie]	Lat. Nachzeitigkeit
	VERB.Part.Fut.Act. + esse.Fin.	Inf. Inf. + byti Präs.	Inf. Inf. + być Inf. + mieć	faciles habiturus es = повóлныхъ бꙋдєшъ мѣ́ть = łatwych będziesz miał
	VERB.Part.Fut.Act. + esse.Inf.	Inf.Act.	Inf.Act.	Futurinfinitiv

Morphosyntaktische und lexikalische Besonderheiten

	Lat.	Ruth.	Poln.	
964	VERB.Inf.Pres.Act.	Inf.Act. (-iti)	Inf.Act. (-ć)	
	VERB.Inf.Pres.Act.Acc.	Inf.Act. + Noun.Dat/Acc.	Inf.Act. + Noun.Acc.	Acl: viam nancisci = найтѝ дорóгꙋ = drogę znaleść
412	VERB.Inf.Pres.Pass.	VSubst (-enja) Inf.Act. (+ sja)	Inf.Act. (-ć) Inf.Act. (+sja)	
99	VERB.Inf.Past.Act.	l-Part.	l-Part.	Aci: coelos ascendisse = на Нѐо взышóлъ = do nieba wstąpił (na + Akk. / do + Lok.)
232	VERB.Gdv.	na + Vsubst Inf. l-Partizip [+ Inf.]	do/dla + VSubst do + VSubst l-Partizip	lat. Verbaladj. (mit passivischer Bdtung, drückt Notwendigkeit aus)
	VERB.Gdv. + esse.Inf.	Inf. + sja + měti.Prät.	Inf. + (-by) + się + mieć.Prät.	appellandum esse = назвáтисѧ мѣ́ло = nazywaćby się miało
71	VERB.Ger.Gen.	l+Part. + Gen. Vsubst. l-Part.	Inf. + Acc. Präs. + się Vsubst.	lat. Verbalsubst.
913	VERB.Sub.Pres.Act.			
	VERB.Sub.Pres.Act. + sit	l-Partizip	l-Partizip	
901	VERB.Sub.Past.Act.	l-Partizip Pres. Pres.Act.	l-Partizip Inf. + mieć.Prät. Pres.Act. [+ się]	voluptates exciperet = пресѣкáетъ = zabierać miała
114	VERB.Sub.Past.Pass.	l-Partizip	l-Partizip	
308	VERB.Sub.Pqp.Act.	l-Partizip	l-Partizip	kein Pqp.Passiv bei Billius