



**UNIVERSITÄT
HEIDELBERG**
ZUKUNFT
SEIT 1386

Texterkennungstechnologien im Vergleich: Transkribus vs. eScriptorium/*kraken*

Vortrag am Heidelberg Forum Digital Humanities, 9.2.2022

Forschungsgegenstand

- Mehrsprachige Wörterbücher des 17. Jh.
 - Griechisch-Slavisch-Lateinisches Wörterbuch des Epifanij Slavineckij (GSL), verfasst in Moskau in den 1660er Jahren, Handschrift (ca. 1500 Seiten)
 - ca. 1642 Calepino-Übersetzung ins Ostslavische/Kirchenslavische
 - Slavisch-Griechisch-Lateinisches Wörterbuch von Fedor Polikarpov, Moskau 1704, Altdruck (812 Seiten)
 - Überarbeitete und russifizierte Version des GSL
- Hauptziel: Entwicklung und Transformation des Wortschatzes vom Ruthenischen und Kirchenslavischen zum Russischen
- Weitere Quelle: Moskauer Abschrift des GSL (Ende des 17. – Anfang des 18. Jh.)

ЛЕЗІКОНЪ ТРЕАЗЫЧНЫИ.

с н р т ч ъ
реченій славѣнскѣхъ , ѣллиногрѣческѣхъ и латинскѣхъ
сокровище
изъ различныхъ древнихъ и новыхъ книгъ
собраное
и по славѣнскому алфавиту въ чинъ
разположеное

ΛΕΞΙΚΟΝ ΤΡΙΓΛΩΤΤΟΝ

ήτοι
Λέξων σλαβονικῶν , ἑλληνικῶν τε καὶ λατινικῶν
θισαβρὸς
Ἐκ διαφόρων παλαιῶν τε καὶ νέων βιβλίων
συλλεχθεῖς
καὶ κατὰ τὸ σλαβονικὸν ἀλφαιβητάριον εἰς τάξιν
διατεθεῖς

DICTIONARIUM TRILINGUE

hoc est
Dictionum Slauonicarum Græcarum & Latinarum
thesaurus
Ex varis antiquis ac recentioribus libris
collectus
Et iuxta Slauonicum alphabetum in ordinem
dispositus



Л Е З І К О Н З

СЛАВЕНОГРЕКОЛАТИНСКІИ , ВЪ ПОЛЬЗУ
МЪДРОЛЮБИВОМУ ОУЧАЩИХСЯ ВРАЗУМЛЕНІЮ .

Азъ . двои знаменуетъ : въ азъцѣ оубо славѣнской первомъ
писмени има азъ : алфа . въ азъцѣ же славѣнскомъ мѣстоименіе
перваго лица , азъ , прѣтъ глаголемо , а , еуѡ , еуѡуе , ego .
А , соузъ распрѣгательный , разумъ дѣлающе , гакъ рещи :
азъ читаю , а ты спиши . азъ глаголю , а ты поещи .
Полагаетса же всегда напредн реченіи , ибо не можемъ рещи ,
ты а : онъ а : дѣ , уѣ , verò , quidem , certe , autem . естъ
иногда и гласъ посланина ѡканіканцагоуа , а . и знакъ
числа перваго славѣншма . А , двоиственнаго числа ,
трѣтјаго лица , вмѣстѣ рещи двѣхъ и хъ .

А ѡ Б
Аѡе , икоуа , въ тоуъ часѣ , ѡ каѡиуѡиѡе , abbas .
аутиѡа , еуѡѡс , еуѡѡѡс , пара- Аѡраѡиѡа , ѡ ѡзраѡиѡа , abraham .
хѡиѡа , параутиѡа , statim , А ѡ Г
confestim , velociter , cito , Аѡѡ , гласъ жалѣніѡа и ѡѡто-
мох , illico . ваніѡа покѡмѡанѡе , аѡ аѡ ,
А ѡ Б іѡѡ , ѡ іѡѡа , ah , heu , hei .
Аѡѡѡ , ѡ ѡѡѡѡѡс , ѡ іѡѡиѡе , Аѡѡѡѡ , ѡ ѡѡѡѡѡѡс , angelus .

Feodor Polikarpov, Dictionarium trilingue (Moskau, 1704)

AA	AB
Веладное. quod tenere non potest.	irresistibilis. немощный insolevabilis.
Aaquis dos и илматтао, или ие ицесттао. Viscatio пищю- ласттао. ацирпнт	бѣлматтаый habitu cavens.
Аавица. даръ. допи. блягоръ. пи Аавица пофде.	Аагаъ същаетса satiativ. пол- нитса imetur.
Aavis eos o x и неполезный.	Аагаъ бредитъ tabis пищиттаъ damnu inferit.
inimilis. несодерженный.	Аататос несъщный insatiabilis исполный implebilis.
imperfecibilis несодерженъ. Но абво итус несъщный. но festinus.	Аатосъ о о по лишению непердний innoxius: по направленио бредний похисъ. Јаже несъщный insatiabi- lis: невдовенъ difficilis. facius: невдвенъ ininvenerabilis.
Aaplos o x и по лишению не- приснокаченъ, intactus:	Аатосъ о о x и бредиттаченъ Casinus пищиттаъ dampnosus.
неисакный. intactilis. неаре- ный. innoxius. по направле- нио бредний, noxius.	Аатолосъ о о неардний innoxius несъщитный dampno. obnoxius.
Aas бротицтца ипреъ стас, или подитре реventie.	Аатомъ м ств лха бредитъ носсо. пищитъ dampni infero.
Aasa пааредитъ носси.	Ав мѣъ mensis. september.
асиордѣиъ consistant.	Аба ns и носсо vota. болъ clamor.
Аасту пааредити поха afficere.	Абаритосъ о о паннѣи tenuis.
асиоритти. contrivitate. насѣ- титти satiare. бѣлѣи обдоити	Абарта та македонска саорозитни рофри rosa.
Аастѣлѣ пааредитса, поха affectus sum: асиоритца contrivista- tus sum: асиоритца ervavi.	Абасъ о о x и бѣгласный cavens носсо Молтаинѣ taciturnus негмѣъ non loquax.
паарохъ lapsus sum.	Абачеъ еросъ о о паннѣи tenuis, аро- бный minutus.
Аастан Јаорѣити пааредитса lassari	Абадъ Куприциъ дѣителъ doctor.
Аадѣлѣ пааредитса lassus sum	Абадисъ еосъ о о x и негавтѣоубъ Но ро fundus. бѣглвинный cavens рѣоги- nditate.
асиоритца contrivistatus sum.	Абадиата та кѣлати scriblita.
Ааотрофосъ о о x и ардоноситный	Абадѣосъ о о x и бѣдѣннѣи fundo cavens. бѣстепенный гради cavens.
похам ferens.	Авоу оу Аасъ гра фѣлѣиттисий Абѣ Vvbs Phocensis.
Аастрофанъ ovos о о x и бредомъ- дый поха sapiens. бредомѣ Iasus mente.	Авоуонъ мѣло, Равитум. дробное minutu
Аастроу, аогъ аааста мѣо. ааста- мѣо. Пав.	Авахарѣонъ о то фразъ аваченѣоъ Сити мѣиттисий Abacenum Vvbs sicilia.
Аадѣлѣ бредитъ носсо: пѣдъ соу- ситро пищитъ dampno afficio.	Авахѣоо ѿ м ств лха негавтѣ- стасъ ignovo: поноситтаъ quiesco бѣглоглѣсттадъ Но loquor. молѣъ taceo.
Ааотисъ о о, дѣлѣ, halitus дѣлѣ нѣи sprivatio	
Аасовосъ о о x и стѣнателенъ demebund неардѣостѣ illatibi	
Аасосъ о о x и неарднѣи, innox- ius: бѣглвинный recedo cavens.	
Аачетосъ о о x и неардѣитный	

AB	AB
Абахѣиосъ молтаинѣиъ creite. бѣглоглѣитно Но loquaciter. поносѣ но quiete.	ascensio. пѣлѣлѣ ascensus.
Абахѣиѣонъ ovos о о x и негмѣъ Но loquax. негмѣъ mutus. бѣи, stultus: поносѣиный quietus. мол- таинѣъ taciturnus. прѣоитѣи мѣлѣи Abavis eos o x и пофде.	Абанѣтиадѣиъ Абанѣиттаинѣъ. сѣбъ Ааб. мѣоъ filius Abantis.
Абахѣиѣонъ мѣлѣи ивахѣиѣонъ негав- тѣиъ rescuunt.	Абанѣ хосъ о о негмѣъ. Но loquax. бѣгласный cavens voce. молта- инѣъ taciturnus. Јаже стѣоитѣо- abacus. бѣстепенный cavens гради прѣоѣтѣи знаменѣтѣиъ пофдеъ ерѣе Абахѣиѣо.
Абахѣиѣо мѣлѣи лѣха пофдеъ ерѣе абахѣиѣо.	Абанѣтисѣиъ о о неардѣиъ immer- sabilis. немѣиттисий distingui nequit.
Абахѣиѣоъ о то стѣоитѣиъ abaculus. сирѣиттѣиъ tabula. кѣстѣица tessē. Иа. исѣлѣица, tabula calculatovia мѣица assen. носѣиъ	Абанѣтосъ о о x и негавтѣиъ non ti- nctus. неисакѣиъ stomomate non indivatus. неардѣиъ immersabilis
Абахѣиѣи ои бредиттисий dregari.	Абанѣъ еосъ то аирѣѣиъ ascensu.
Абахѣиѣохосъ о о кѣстѣица tessella. сирѣиттѣица tabella. стѣоитѣиъ men- sula.	Абарѣалоѣаъ оу тѣиѣиѣиъ негавтѣиъ Нумѣиѣа.
Абахѣитѣиѣоъ о то бѣгалиттѣиѣоъ Non invidiosum.	Абарѣарѣи ns и Аарѣарѣа Нумѣиѣа Нумѣиѣа
Абахѣитѣиѣоъ о то негавтѣиѣоъ Но бѣати.	Абарѣиъ еосъ о о x и негавтѣиѣиъ non gravis. бѣглѣиттисий gravita- te cavens. неардѣиѣиъ non onero- sus: прѣносѣиъ нестѣиѣоиттисий non mo- lestus. Јаже неардѣиѣиъ intelli- gentia cavens. бѣи stultus.
Абахѣитѣиѣоъ о о x и бѣглѣиттѣиѣоъ, неисакѣиѣоъ. Но insanus. инанѣиѣиъ lincus.	Абарѣиѣастѣи ns и бѣглѣиттѣиѣа effa- minata: сластѣиѣа delicata. мѣлѣиъ mollis.
Абалъ ѣлѣи, охъ heu. ѿ бѣи Vinam	Абарѣиъ еосъ о о x и бѣглѣиттѣиѣиъ non onerosus. Маѣѣе ѣлѣиттисий continentis habitator. бѣиоравѣиѣиъ navibus non gens. Естѣиъ и ѣлѣиѣа. мѣлѣиѣа и Рѣиѣиѣиъ Абарѣиѣи.
Авала пофде.	Абарѣитѣиѣоъ Јаже ерѣе Абарѣиѣастѣи.
Авалѣоаъ as и писанѣиъ scriptura. испаѣлѣиттѣиѣа confessio. спѣсакѣиѣа cooperatio.	Абарѣитѣиѣа македонскаъ бласитѣиъ, comā nutrit.
Авалѣе пофдеъ ерѣе абал.	Абарѣитѣи ns и Аабѣиѣа дрѣао.
Авали неардѣиѣоъ inutilem laicu нанѣе зѣлѣитѣиъ, славаго languidum. лѣиттисий rigitum.	Абарѣлѣитѣиѣа мѣиттѣиѣа. турѣиттисий. мѣищитѣиъ plaudit. Трахаѣе tenet.
Авалис маѣиѣица oliua. гломѣиѣи- на mala oliua.	Абарѣиѣа as и мѣлѣиъ fames.
Авалѣихѣиттосъ о о x и бѣлѣднѣиъ или бѣлѣиттѣиѣа. Но ficatus. бѣглѣиттѣиѣа factaci exrens. не- тарѣиттисий marginio cavens.	Абарѣитѣиъ Абарѣиѣи лампѣиттисий Тра.
Авалѣихѣиттѣиѣоъ о о x и Јаже Авалѣиѣа Аадѣитѣиѣа Тра при Парнаѣсѣ Vvbs circa Parnassum.	
Авалѣиттѣиѣоъ еосъ и стѣохѣоѣдѣиѣе	

БИБЛИОТЕКА
 Андрей Аввакумъ
 1870
 Андрей Аввакумъ

ИЗЪ БИБЛИОТЕКИ
 А. А. ТИТОВА
 РОСТОВЪ 1908
 №

Ms. Titov 67, ff. 1v-2r

Server Overview Layout Metadata Tools

Logout walker.thompson@slav.uni-heidelberg.de

Document... Find

Document Manager User Manager

Versions Jobs

Recent Documents... User activity

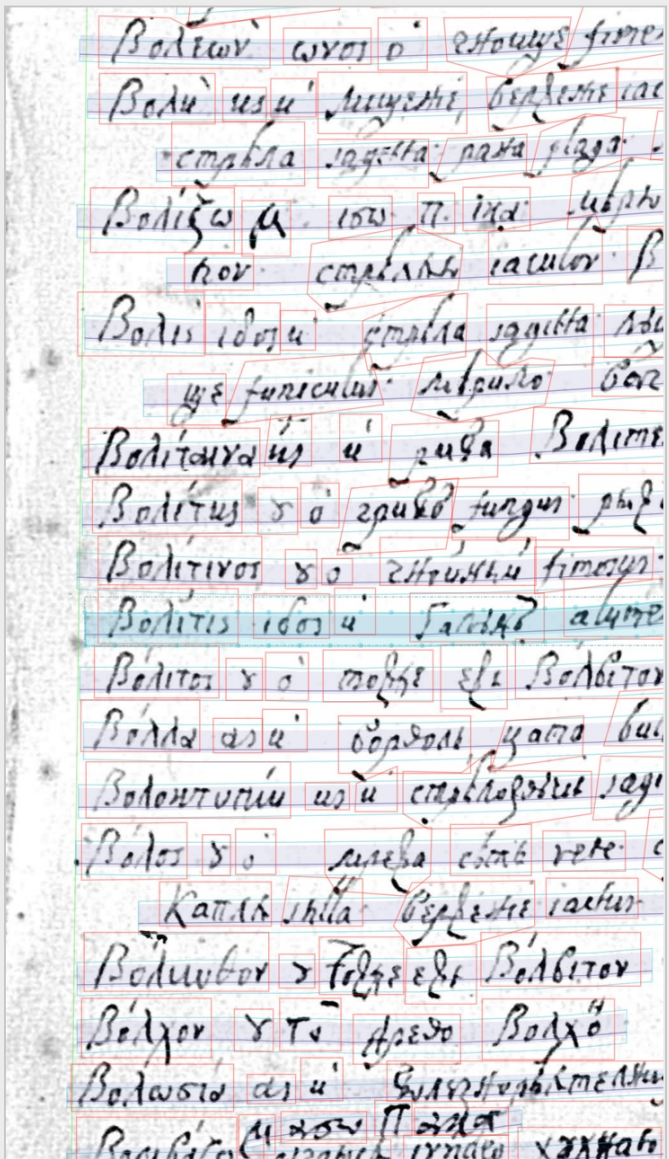
Collections: ePolyGlott (39811, Owner) Col-ID

Documents HTR Model Data

1-8 / 8

ID	Title	Pages
9150...	TRAINING_VALIDATION_SET_slavin...	1
9150...	TRAINING_VALIDATION_SET_FedP...	1
8379...	TRAINING_VALIDATION_SET_slavin...	1
6512...	Сборник_переводов_Епифания_...	416
6511...	Букварь_славено-греко-латинск...	328
206...	Slavineckij	125
1864...	Leksikon-trejazychnyj-ГПИБ_HTR	812
1690...	Leksikon-trejazychnyj-ГПИБ	812

100 Filter



sterquiliniū βομβύλιος ου ὃ τοжде таже комар
 culex βομῆνλιῳ

2-22 βολή ης ἢ μεσσηνίε, ὀρθετε ἰακίον
 удареніе percussio βομβυλιός ου ὃ τοжде
 таже бесполезный inutilis

2-23 болезнь dolor

2-24 στρῆλα sagitta рана plaga κῆρυα uulnus лѣща
 hasta лѣща бѣмбυξ υκος ὃ шолковникъ червь
 bombyx uermes

2-25 βολίζω μ ἰσω π ἰκα μῆρυῳ γλῆβινομῆρυῳ
 profundū me

2-26 βολίς ἰδος ἢ στρῆλα sagitta лѣща hasta лѣща
 radius ужи βοοβοσκος ου ὃ βολοпασεῖς
 βολοпасты boum pastor

2-27 βολίταινα ης ἢ ρυθα Βολίταινα

2-28 βολίτης ου ὃ γρῖβος fungus ρυθῆ boletus

2-29 βολίτις ἰδος ἢ Γαληνῆ alumen

2-30 βόλιτος ου ὃ τοжде εἰ Βόλιτων

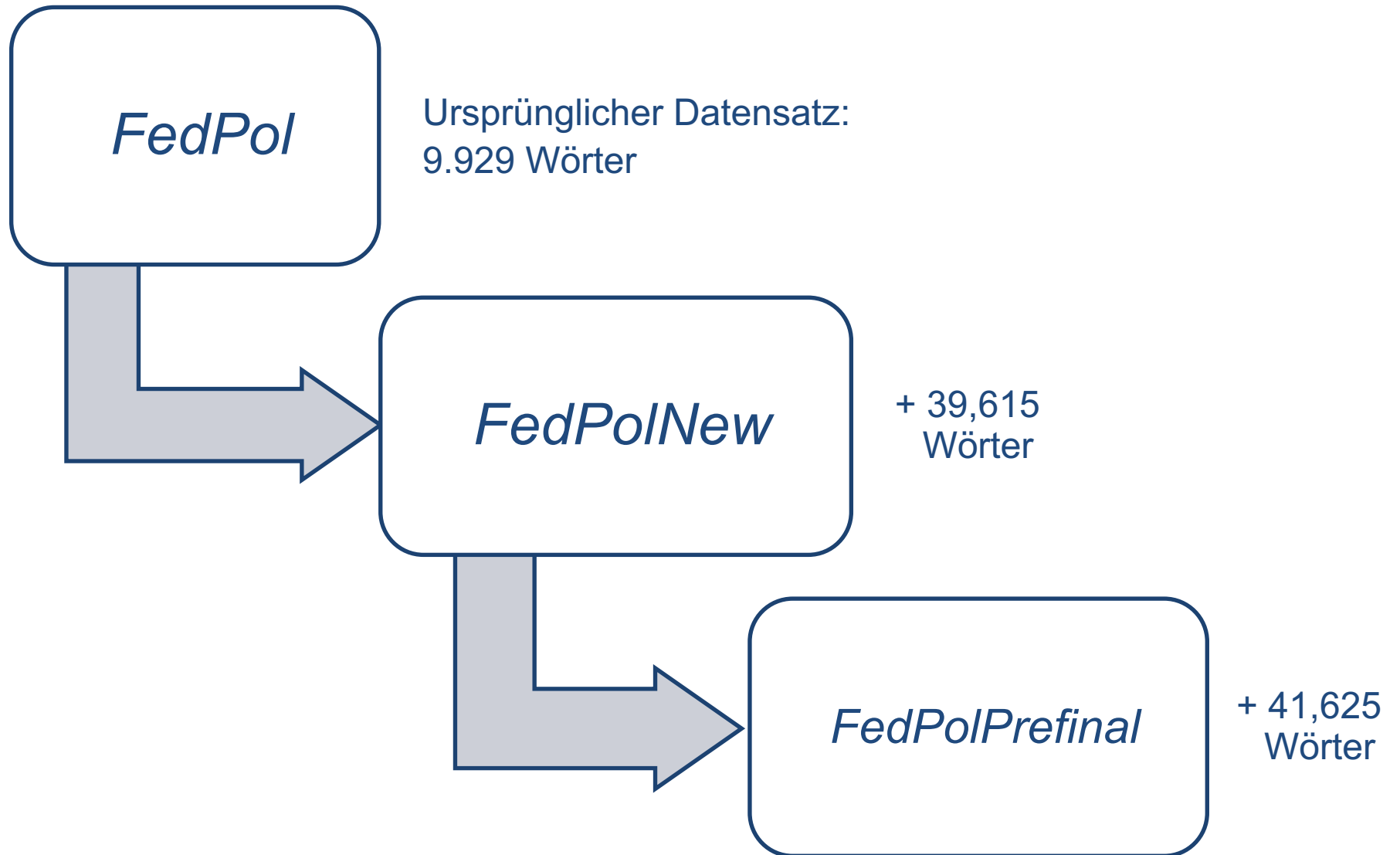
2-31 βολοκτυπίη ης ἢ στρῆλοζωχίε sagittarū sonus
 κοστοζωχίε tesser βοοσφάγη ης ἢ τοжде

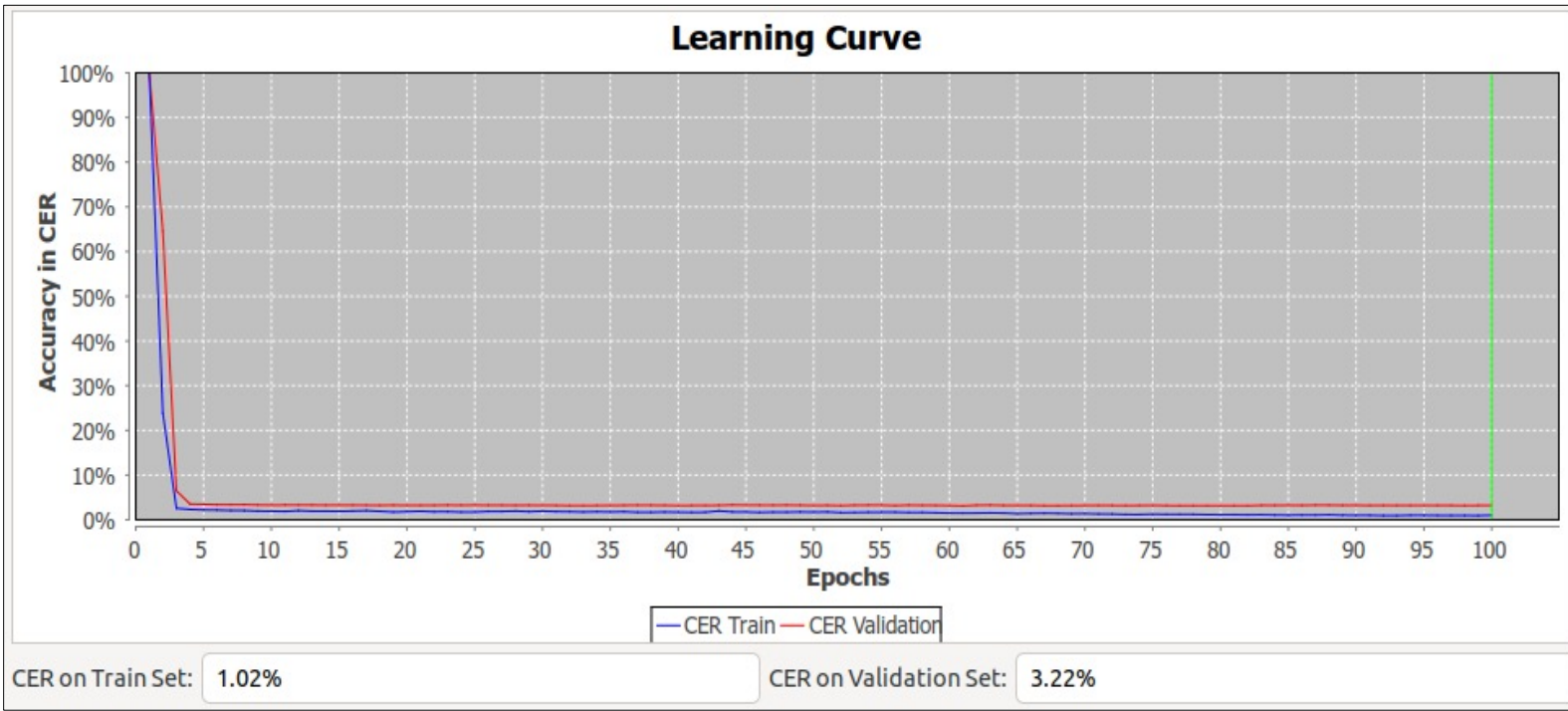
2-32 βόλος ου ὃ μρεβα εἰσὺς rete στρῆλα sagitta
 праща funda βοοτρόφος ου ὃ κ ἢ βολοпитате
 βολοпῖлателῆ boues alens

2-33 καπλά stilla ὀρθετε ἰακίον πορθε, πορθε
 quod iactū est βοῶ μ βοπλῳ, κληχρ clamor

Die grafische Benutzeroberfläche von Transkribus (Expertentient, v1.16.1)

Trainingverfahren (Transkribus HTR+)





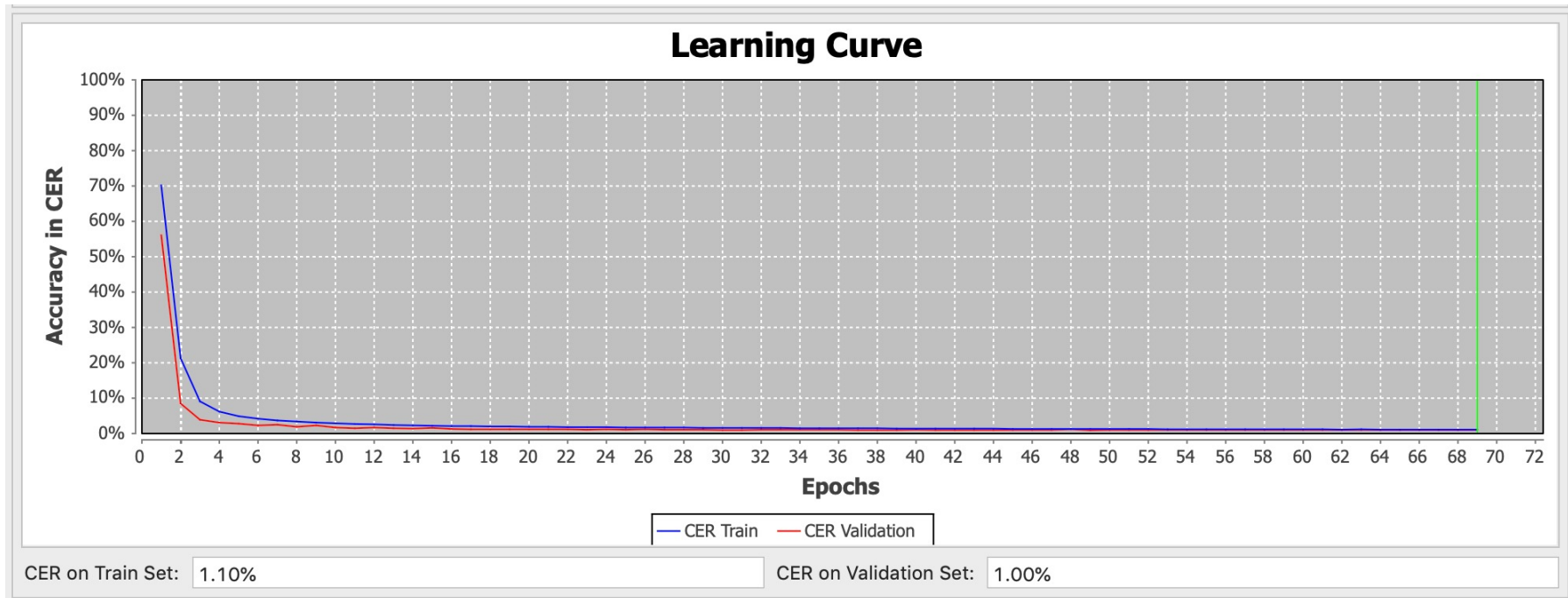
Lernkurve für das Modell *FedPolPrefinal*

- Qualitativ hochwertige Ergebnisse, allerdings vermehrte Transkriptionsfehler bei Großbuchstaben
- Erfolgreiche Versuche mit anderen Altdrucken (bspw. dreisprachiges Fibelbuch von Polikarpov, 1701)

Trainingverfahren (Transkribus PyLaia)

FedPolPyLaia

alle 744 bisher schon eingegebenen
korrigierten Seiten, insg. 113.834 Wörter



Lernkurve für das Modell *FedPolPyLaia*

- Sehr genaue Anpassung an die Trainingsdaten aus dem Polikarpov-Wörterbuch
- Niedrige Qualität der Transkriptionen bei Versuchen mit anderen Altdrucken

Slavineckij: aktueller Stand

Transkriptionen vorhanden: 144ff. (↑**21**)

Segmentierung in Transkribus fertig: 125ff. (↑**26**)

Eingabe in Transkribus fertig: 51ff. (↑**23**)

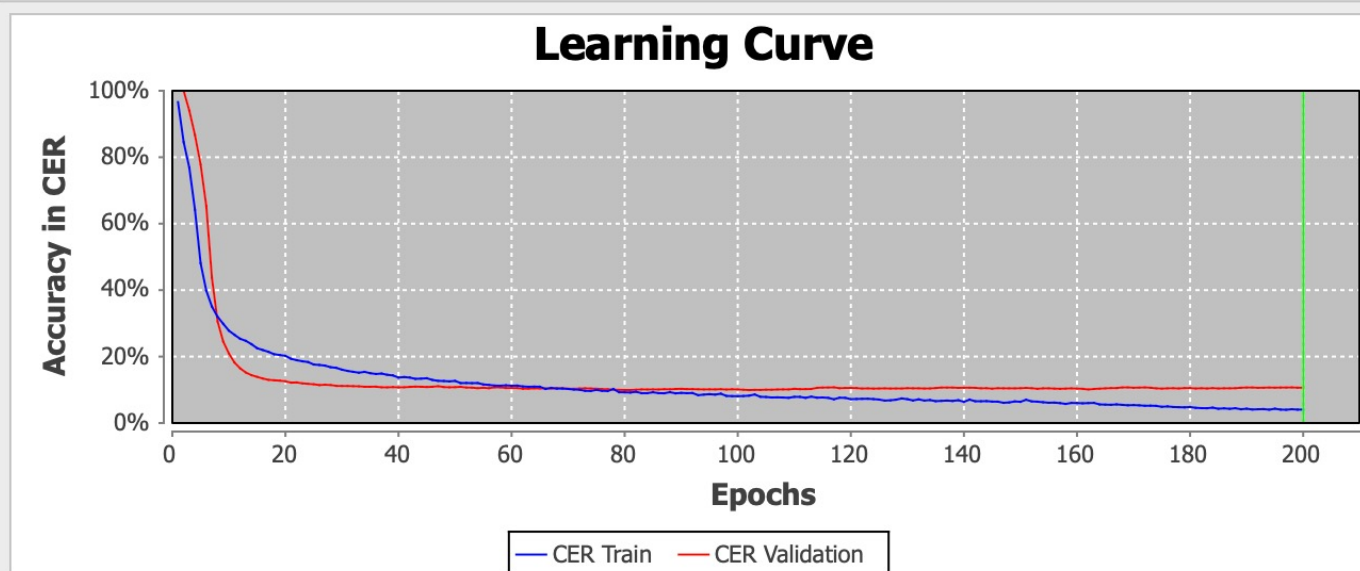
Training abgeschlossen (slavineckij_v12): 51ff. (↑**39**)

Name: Language:

Description: Parameters:

Document Type:

Nr. of Words: Nr. of Lines:



CER on Train Set: CER on Validation Set:

Das letzte Modell *slavineckij_v10* (aus dem Projekt ePolyGlott)

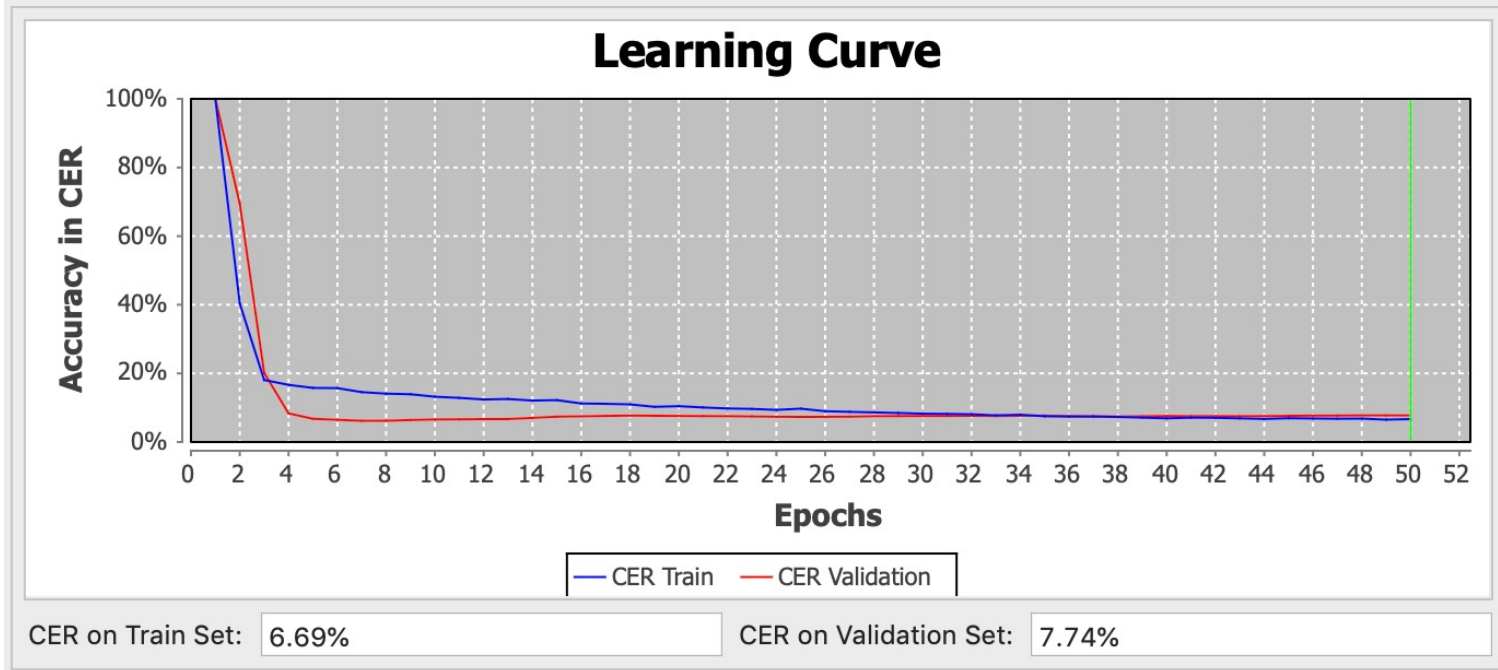
Name: slavineckij_v12 Language: Greek, (Russian) Church Slavonic, Latin

Description: With new training data from Q42021 and Q12022 Parameters:
 Nr. of Epochs 50
 HTR Base Model ID 38290
 HTR Base Model Name slavineckij_v11

Document Type: Handwritten Show advanced parameters...

Nr. of Words: 11836 Nr. of Lines: 2489

Save Show Train Set Show Validation Set Show Characters



Verbesserung von fast 3% gegenüber *slavineckij_v10* + Vermeidung von Überanpassung

eScriptorium

A project providing digital recognition of handwritten documents using machine learning techniques.



Data Interchange

Import or Export transcriptions with Alto or Page XML, Import images as zip or IIIF. Access data from any application through a full Rest API.



Manual Edition

Make use of an ergonomic user interface leveraging modern browser technology to edit segmentations and transcriptions.



Automatic Transcription

Train and apply new neural networks to vastly speed up the transcription process of large corpora.

Description

Images

Edit

Models

Polikarpov

Drop images here or click to upload.

Select all

Unselect all

Selected 1/812

Import

Export

Train

Binarize

Segment

Transcribe

791 X

792 X

793 X

794 X

795 X 100%

796 X

797 X

Die grafische Oberfläche von eScriptorium (aktuelle Version)

Description

Images

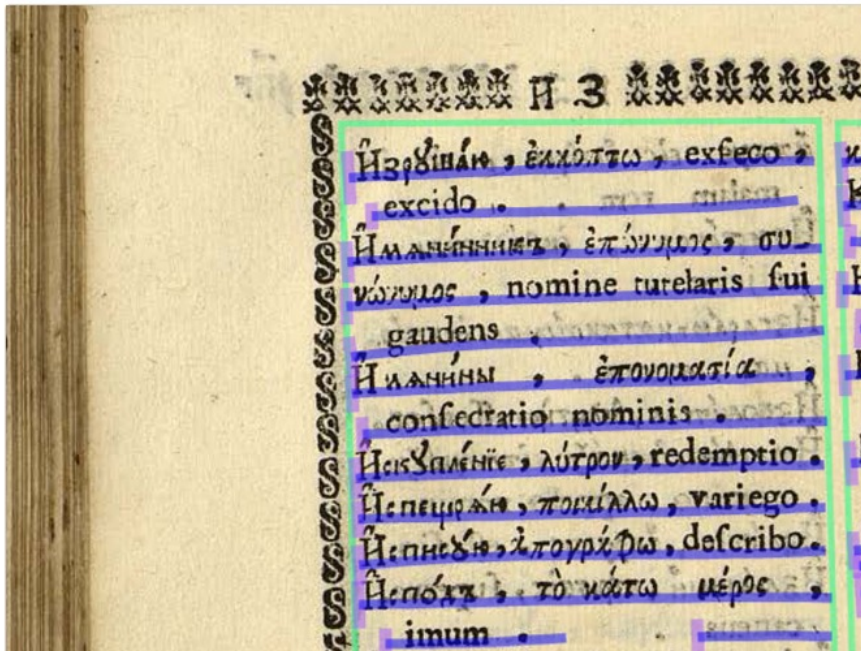
Edit

Models

Polikarpov Appendix Test

Element 4 - Polikarpov_798.png - (2256x2642)

kraken:fedpolfull















- 1 Ιζρδшаю, ἐκάτεω, exfeco,
- 2 excido.
- 3 Иманѣнникъ, ἐπώνυμος, συ-
- 4 νώνυμος, nomine tutelaris fui
- 5 gaudens.
- 6 Иманѣны, ἐπονομασία,
- 7 confecratio nominis.
- 8 Искѣплѣнїе, λύτρον, redemptio.
- 9 Испещрѣю, ποικίλλω, variego.
- 10 Исписѣю, ἀπογράφω, describo.
- 11 Исπόлъ, τὸ κάτω μέρος,
- 12 imum.

Segmentierung und Texteingabe

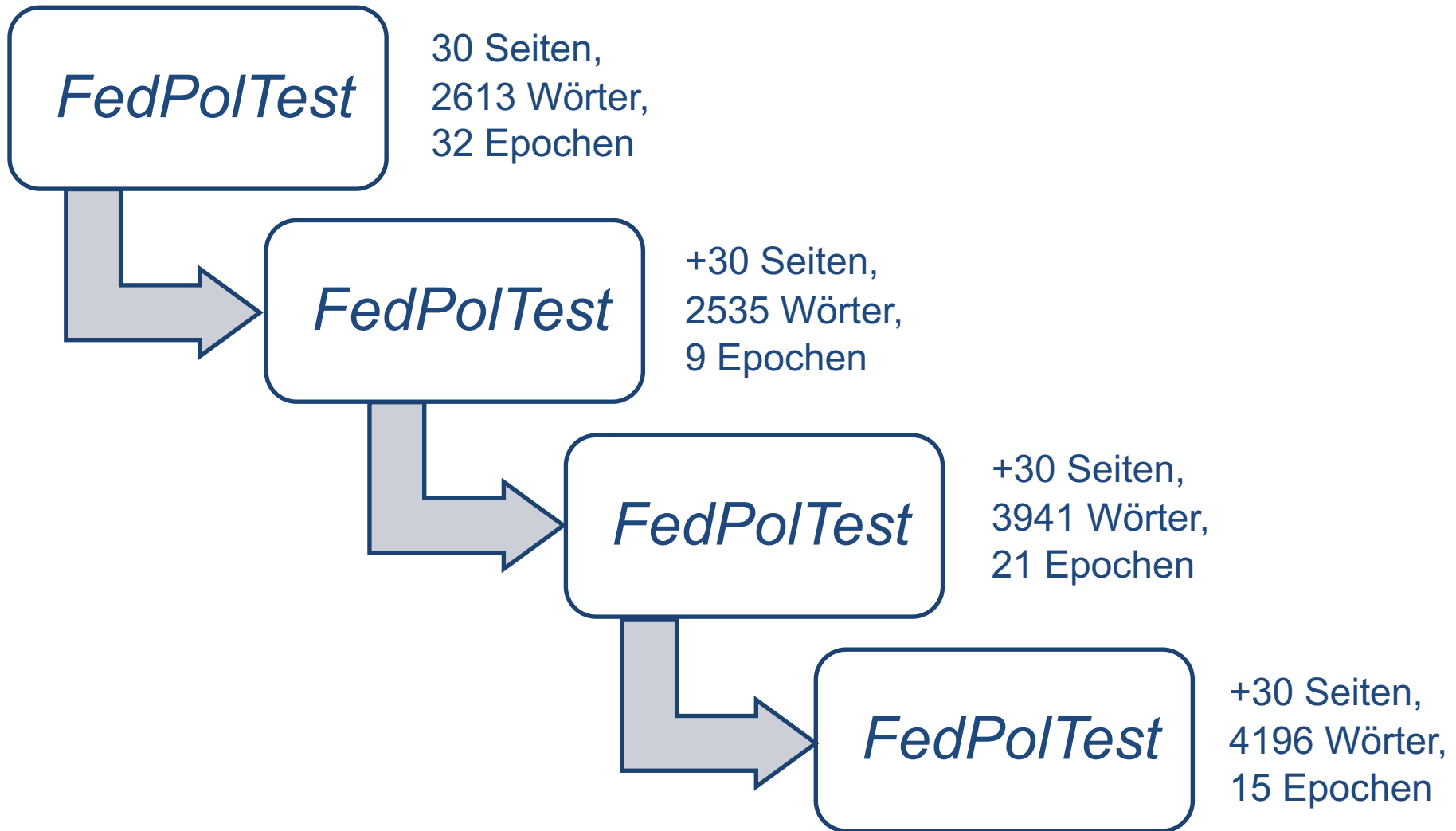
[Description](#)
[Images](#)
[Edit](#)
[Models](#)

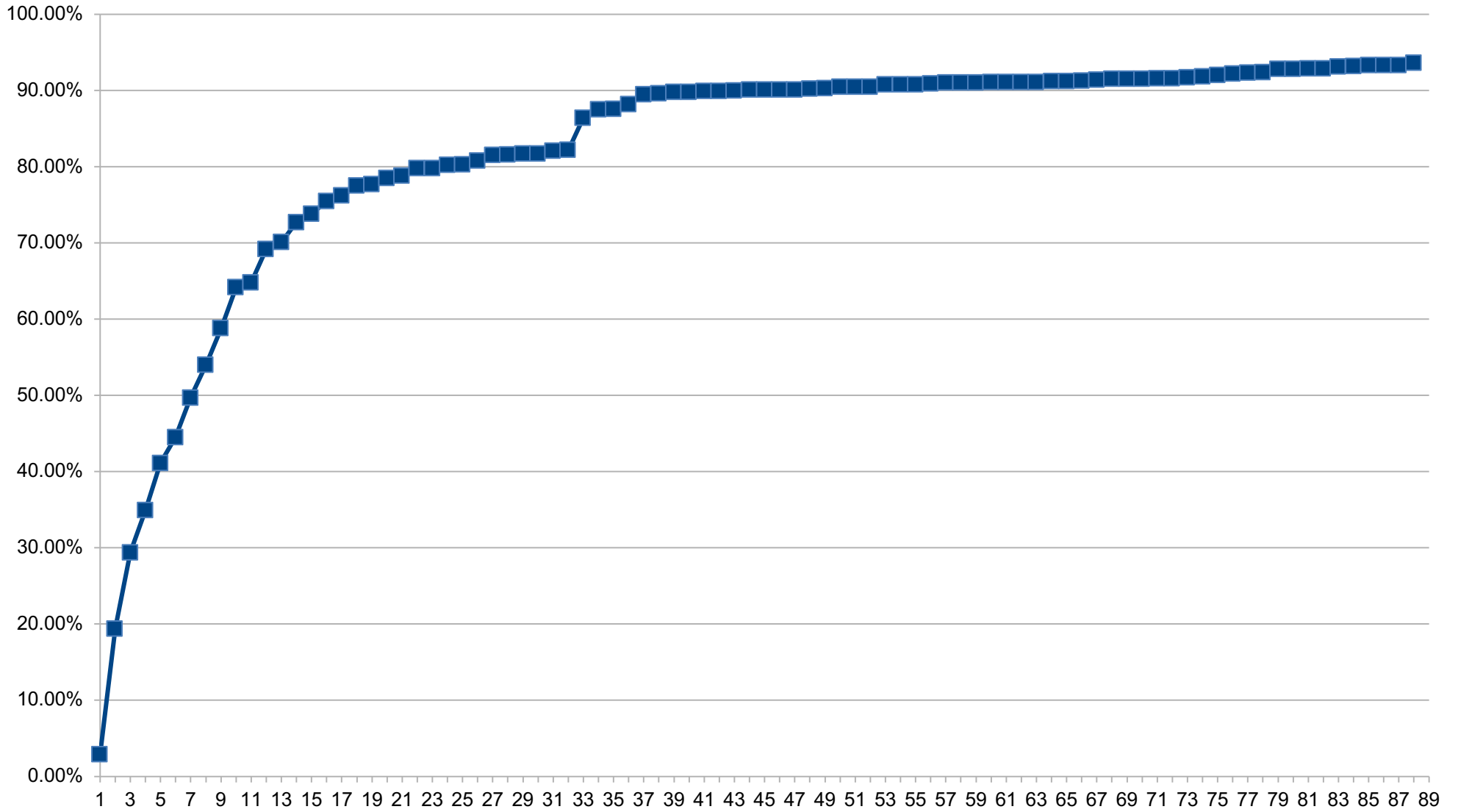
Feodor Polikarpov, Dictionarium trilingue

	Role	Training	Accuracy	Errors	
FedPolFull	Recognize	✓	95.4%	4525/99383	  
FedPolFull (epoch #6)		95.4%	4525/99383		
FedPolFull (epoch #5)		95.3%	4669/99383		
FedPolFull (epoch #4)		95.0%	4950/99383		
FedPolFull (epoch #3)		94.7%	5243/99383		
FedPolFull (epoch #2)		93.7%	6219/99383		
FedPolFull (epoch #1)		91.2%	8708/99383		
FedPolTest	Recognize	✓	91.5%	346/4061	  

Modelldateien speichern und herunterladen

Trainingverfahren (kraken, 1. Versuch)



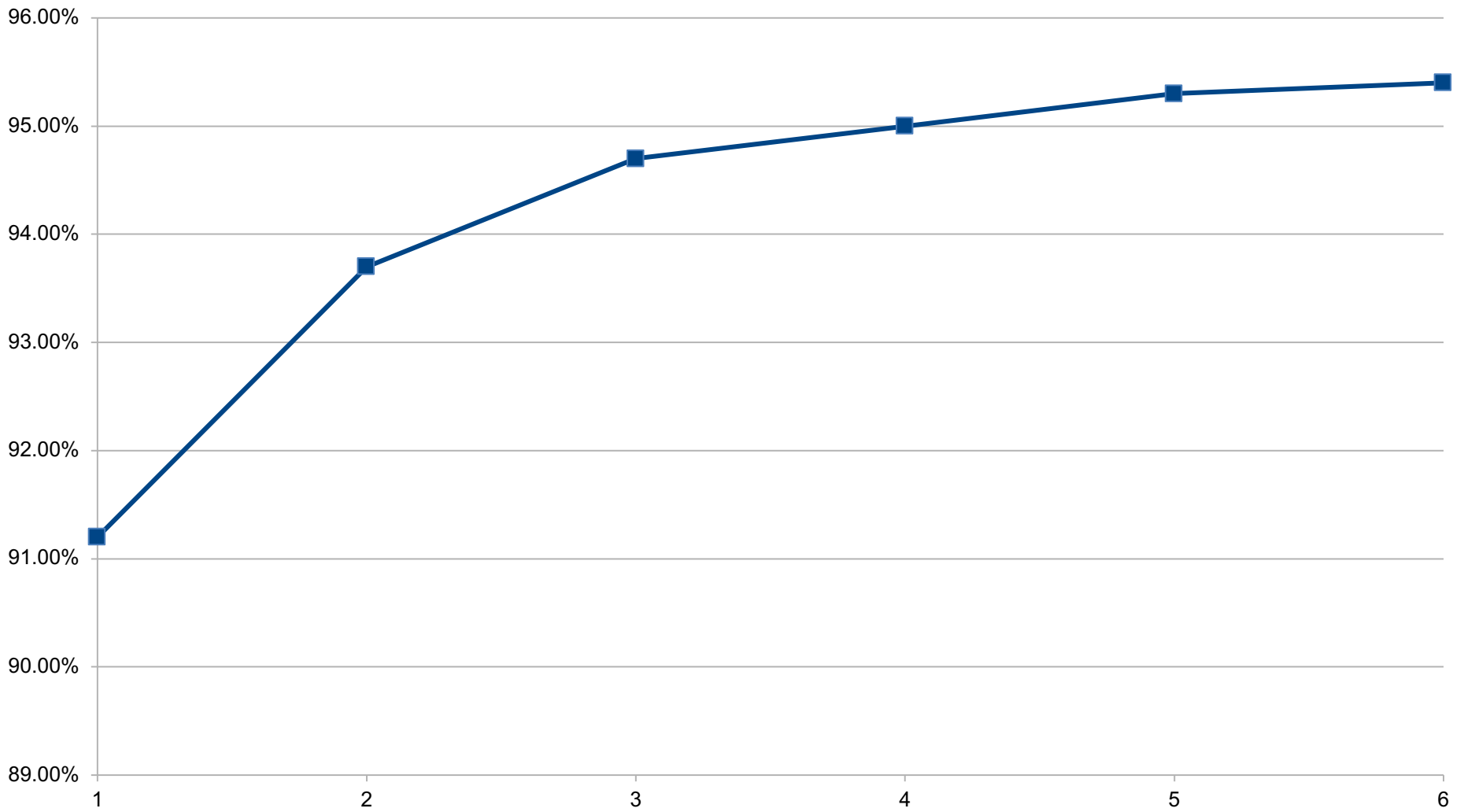


Lernkurve für das Modell *FedPolTest*

Trainingverfahren (kraken, 2. Versuch)

FedPolFull

744 Seiten,
insg. 99383 Wörter,
6 Epochen



Lernkurve für das Modell *FedPolFull*

Vorteile von eScriptorium gegenüber Transkribus

- Konzeptionell kostenfrei und quelloffen
- OCR-Engine *kraken* transkribiert Diakritika genauer und konsequenter als Transkribus/HTR+
- Komplexe Layoutanalyse mit *kraken* gelingt besser, Segmentierung kann auch leicht trainiert werden
- Lokale Installation mit browserbasierter grafische Oberfläche, freier Zugriff auf alle Dateien, lässt sich ohne Weiteres als Server im lokalen Netz einrichten
- Import von Daten aus Transkribus möglich dank offenen Standards (PageXML, Alto)

Nachteile von eScriptorium gegenüber Transkribus

- Wie PyLaia, gut anpassbar auf einzelne Dokumente aber möglicherweise schlechter generalisierbar (muss noch versucht werden)
- Manuelle Segmentierung noch nicht ergonomisch
- Vorladung und GPU-Beschleunigung werden automatisch ausgeschaltet bei größeren Datenvolumen
- Modell-Training mit CUDA/cuDNN nur auf Linux
- Technisch anspruchsvoll, erstmaliges Einrichten recht kompliziert, erfordert Kenntnisse von Git, Python/Django, SQL, Redis... (als Alternative: Docker, aber Containerisierung nicht optimal wegen erschwertem Zugriff auf Dateien und Verzeichnisse)

Weitere technische Ergebnisse

- Aufbau einer technischen Infrastruktur für eScriptorium auf der Plattform bwCloud (Hosting der Browser-Oberfläche und Backend)
- Aneignung des Umgangs mit dem OCR-Engine *kraken* per Kommandozeile, erfolgreiche Probeversuche mit Modelltraining
- Entwicklung von eigenen Tools (Skripts) für unixoide Systeme zur automatischen Umkodierung und Transliteration altkyrillischer Texte
- Veröffentlichung einer Projektseite

Danke für Ihre Aufmerksamkeit!