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.)

Ms. GIM 383, ff. 148v–149r

А

Лексїконъ треязычный.

сирѣчь

реченїй славенскихъ, еллиногреческихъ и латинскихъ

сокровище

изъ различныхъ древнихъ и новыхъ книгъ

собраное

и по славенскому алфавиту въ чинъ

разположеное.

ΛΕΞΙΚΟΝ ΤΡΙΓΛΩΤΤΟΝ

ἤτοι

Λέξεων σλαβονικῶν, ἑλληνικῶν τὲ καὶ λατινικῶν

θησαυρὸς

Ἐκ διαφόρων παλαιῶν τὲ καὶ νέων βιβλίων

συλλεχθεὶς

καὶ κατὰ τὸ σλαβονικὸν ἀλφαβητάριον εἰς τάξιν

διατεθεὶς

DICTIONARIUM TRILINGUE

hoc est

Dictionum Slauonicarum Graecarum & Latinarum

thesaurus

Ex varys antiquis ac recentioribus libris

collectus

Et iuxta Slauonicum alphabetum in ordinem

dispositus

А

Лексїконъ

славеногреколатинскїй, въ пользу мудролюбивому учащихся вразумленїю.

А

Азъ, двое знаменуетъ: въ азбуцѣ убо славенской первому писмени имя азъ: ἄλφα. в языцѣ же славенскомъ мѣстоименїе перваго лица, азъ, простѣ глаголемо, а, ἐγὼ, ἔγωγε, ego. А, союзъ раздѣлительный, разумъ дѣлящь, яко рещи: азъ читаю, а ты спиши. азъ глаголю, а ты пиши. Полагается же всегда напреди реченїй, ибо не можемъ рещи, ты а: онъ а: δὲ, γὲ, verò, quidem, certe, autem. Есть иногда и гласъ посланнина ѿкликающагося, а. И знакъ числа перваго славянскимъ. А, двойственнаго числа, третїяго лица, вмѣсто рещи двухъ ихъ.

А со Б

Абїе, скоро, въ тотъ часъ, αὐτίκα, εὐθὺς, εὐθέως, παραχρῆμα, παραυτίκα, statim, confestim, velocitèr, citò, mox, illico.

А со В

Авва, ὁ ἀββᾶς, ὁ ἡγούμενος, ὁ καθηγούμενος, abbas.

Авраамъ, ὁ ἀβραάμ, abraham.

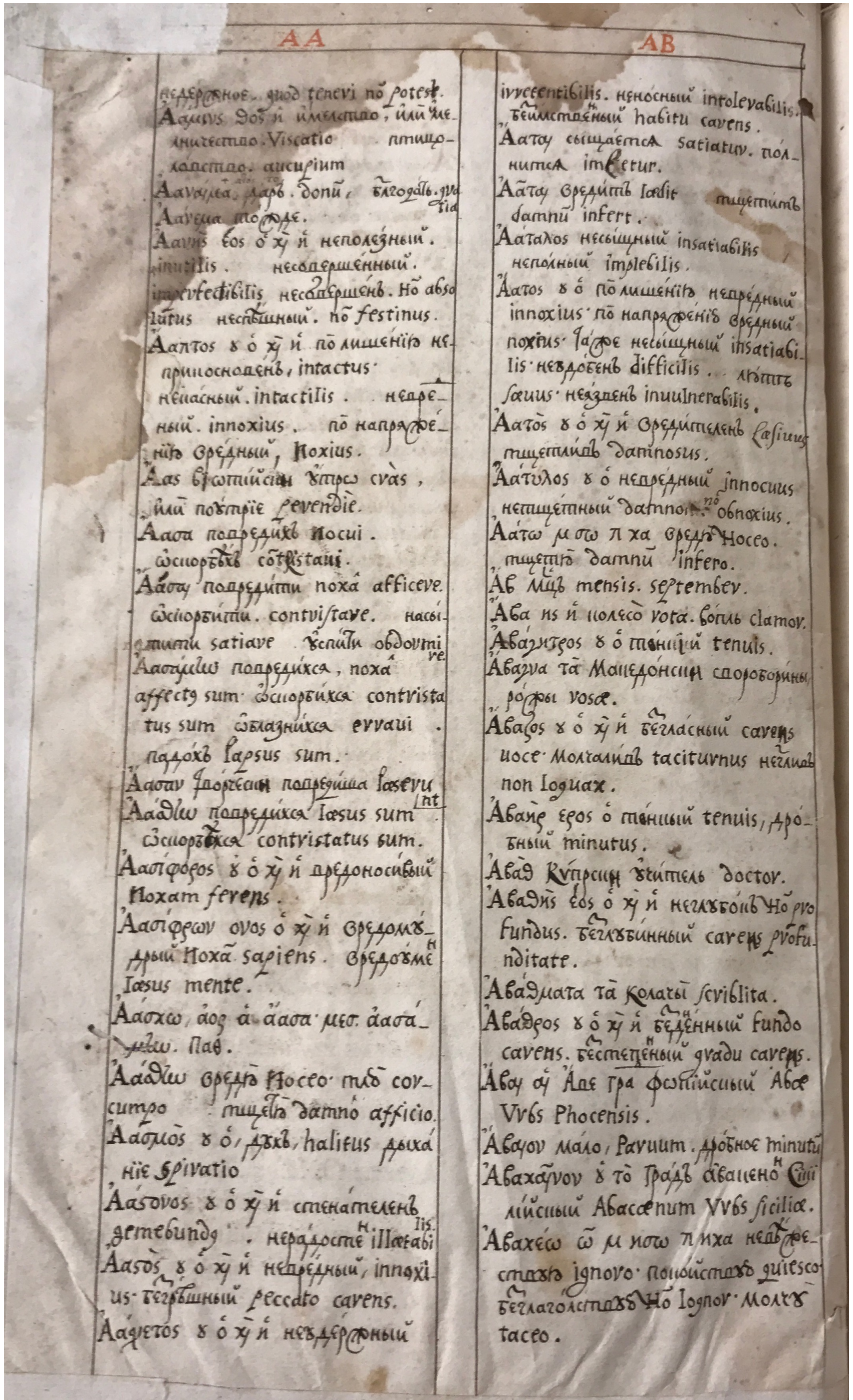
А со Г

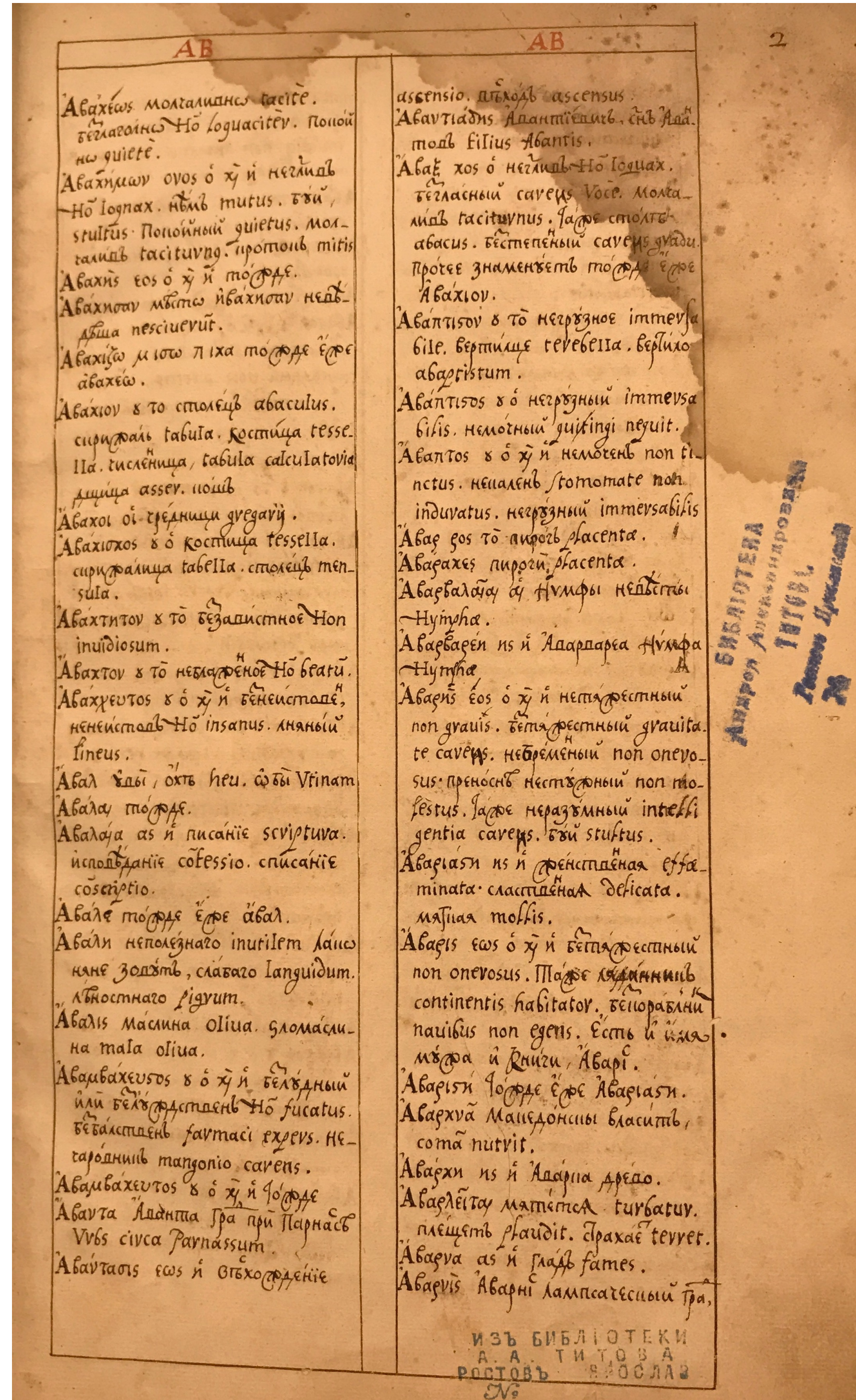
Агъ, гласъ жалѣнїя и сѣтованїя покомълнво, αἲ αἲ, ἰοῦ, οἴμοι, ah, heu, hei.

Агглъ, ὁ ἄγγελος, angelus.

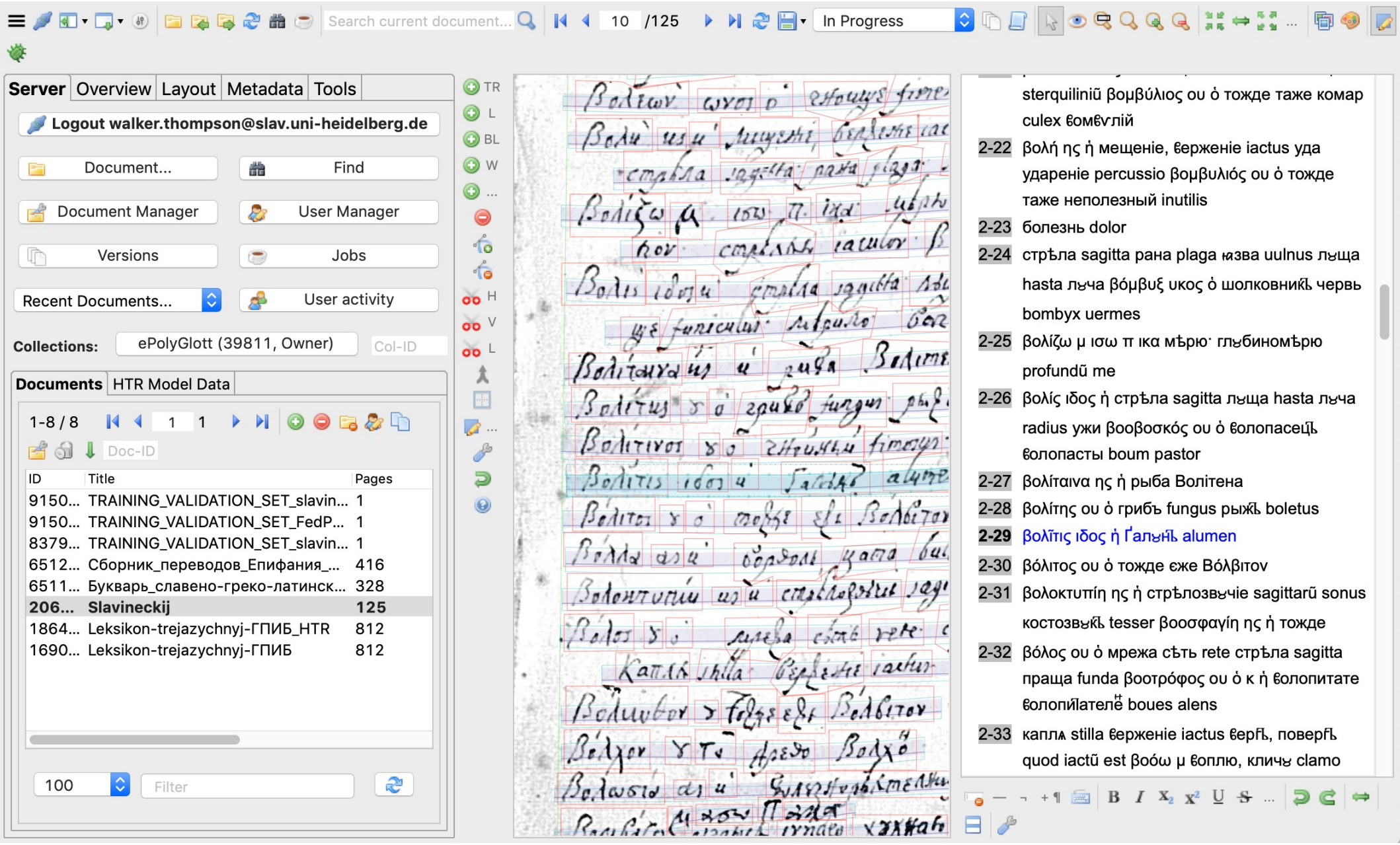
А

Feodor Polikarpov, *Dictionarium trilingue* (Moskau, 1704)



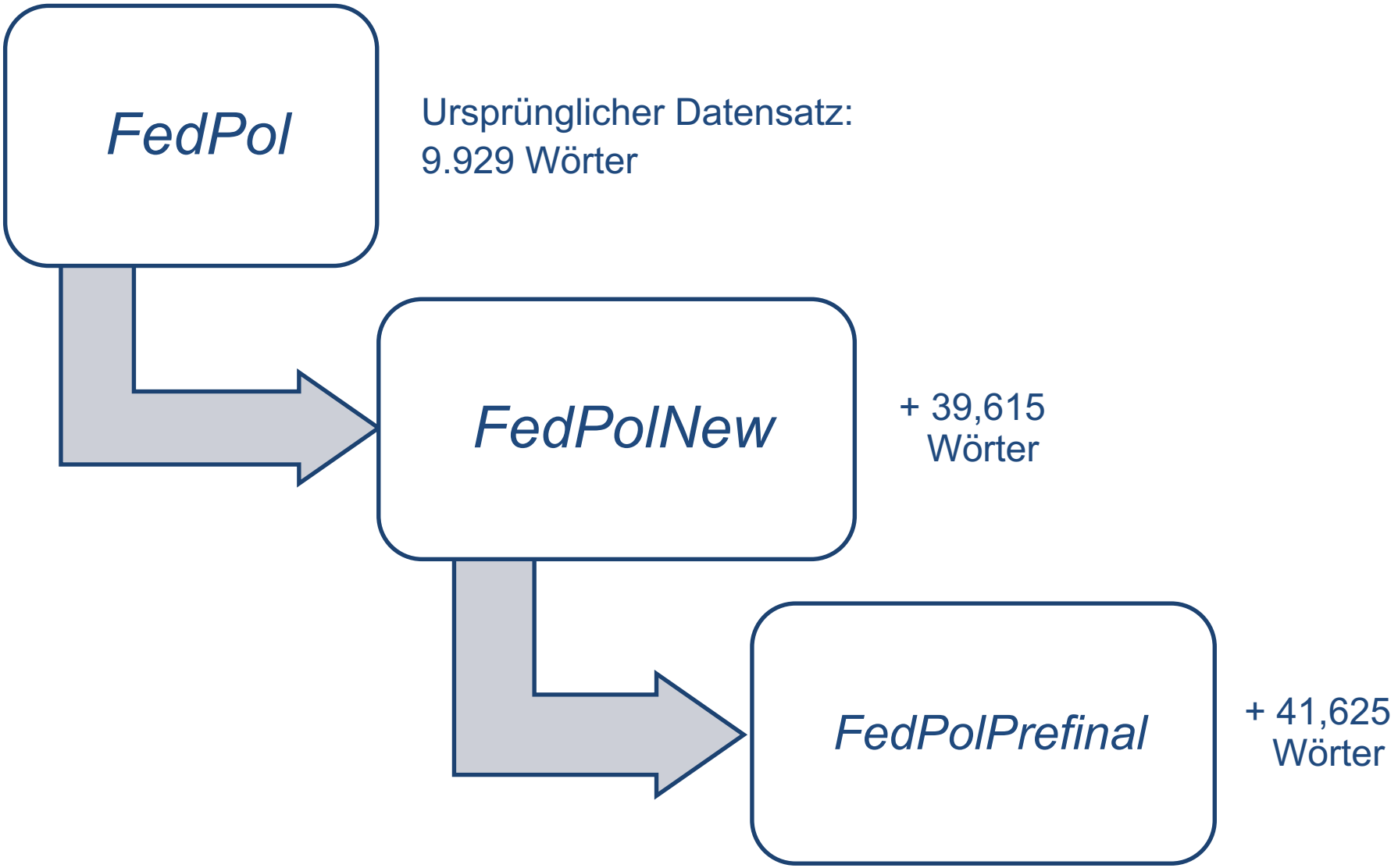


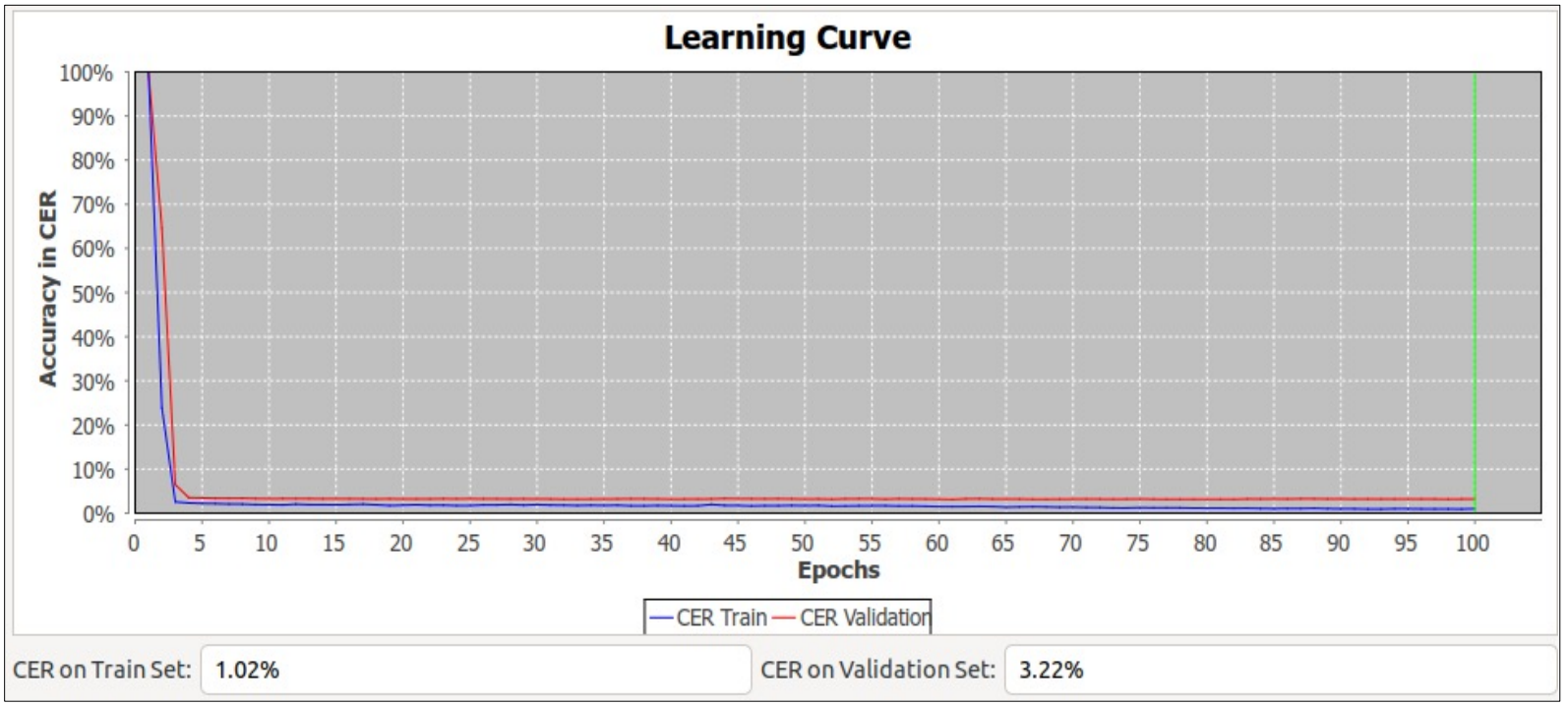
Ms. Titov 67, ff. 1v–2r



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

Trainingverfahren (Transkribus HTR+)





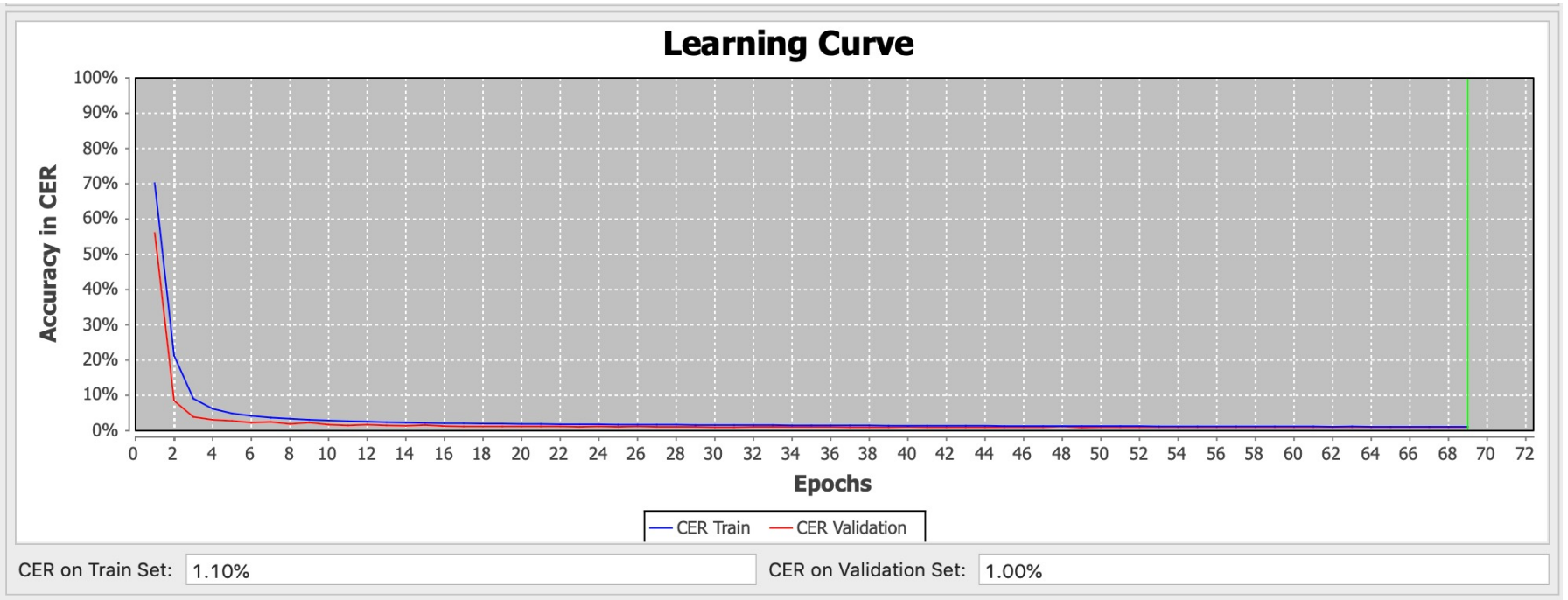
Lernkurve für das Modell *FedPolPrefinal*

- 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 eigegebenen 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

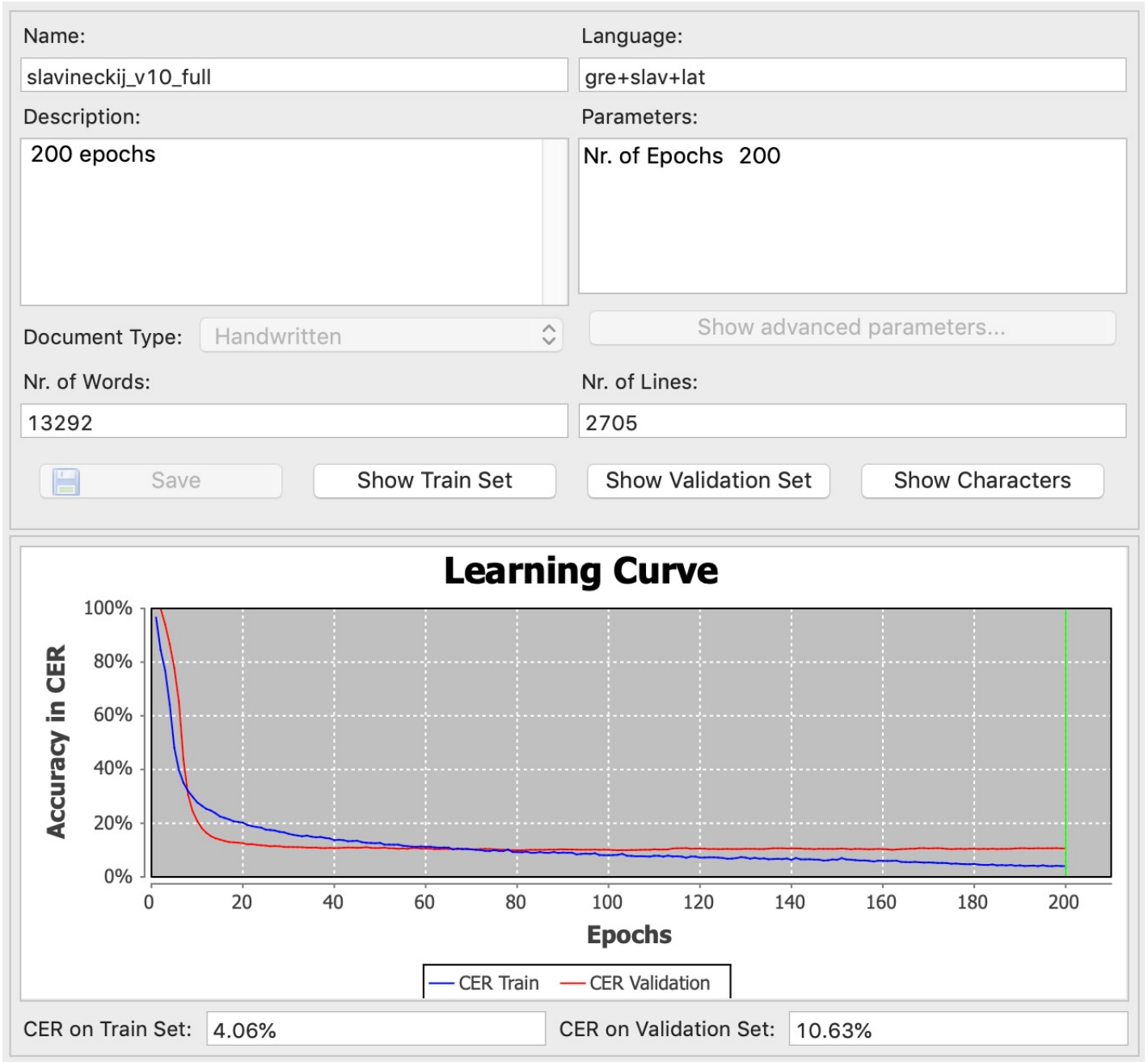
Slavineckij: aktueller Stand

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

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

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

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



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

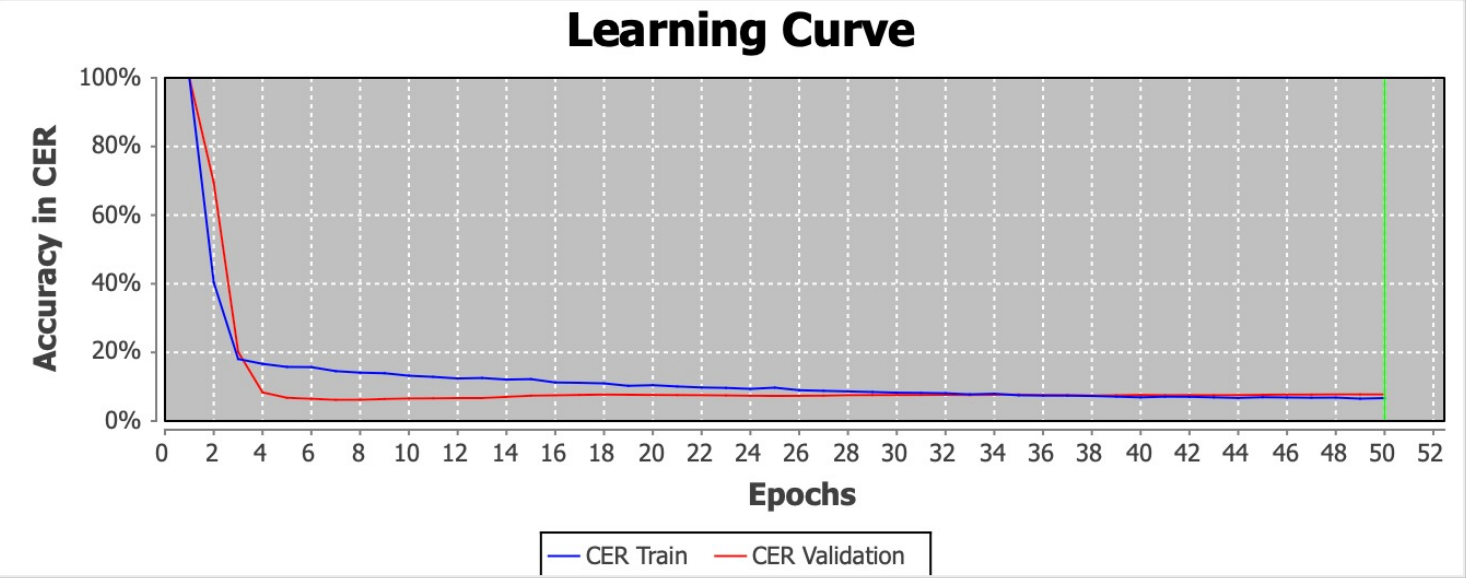
Document Type: Handwritten

Show advanced parameters...

Nr. of Words: 11836

Nr. of Lines: 2489

Save | Show Train Set | Show Validation Set | Show Characters



CER on Train Set: 6.69%

CER on Validation Set: 7.74%

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

Escriptorium - Homepag
Escriptorium - Update a
localhost:8000

eScriptorium Home Contact My Documents Hello admin

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.

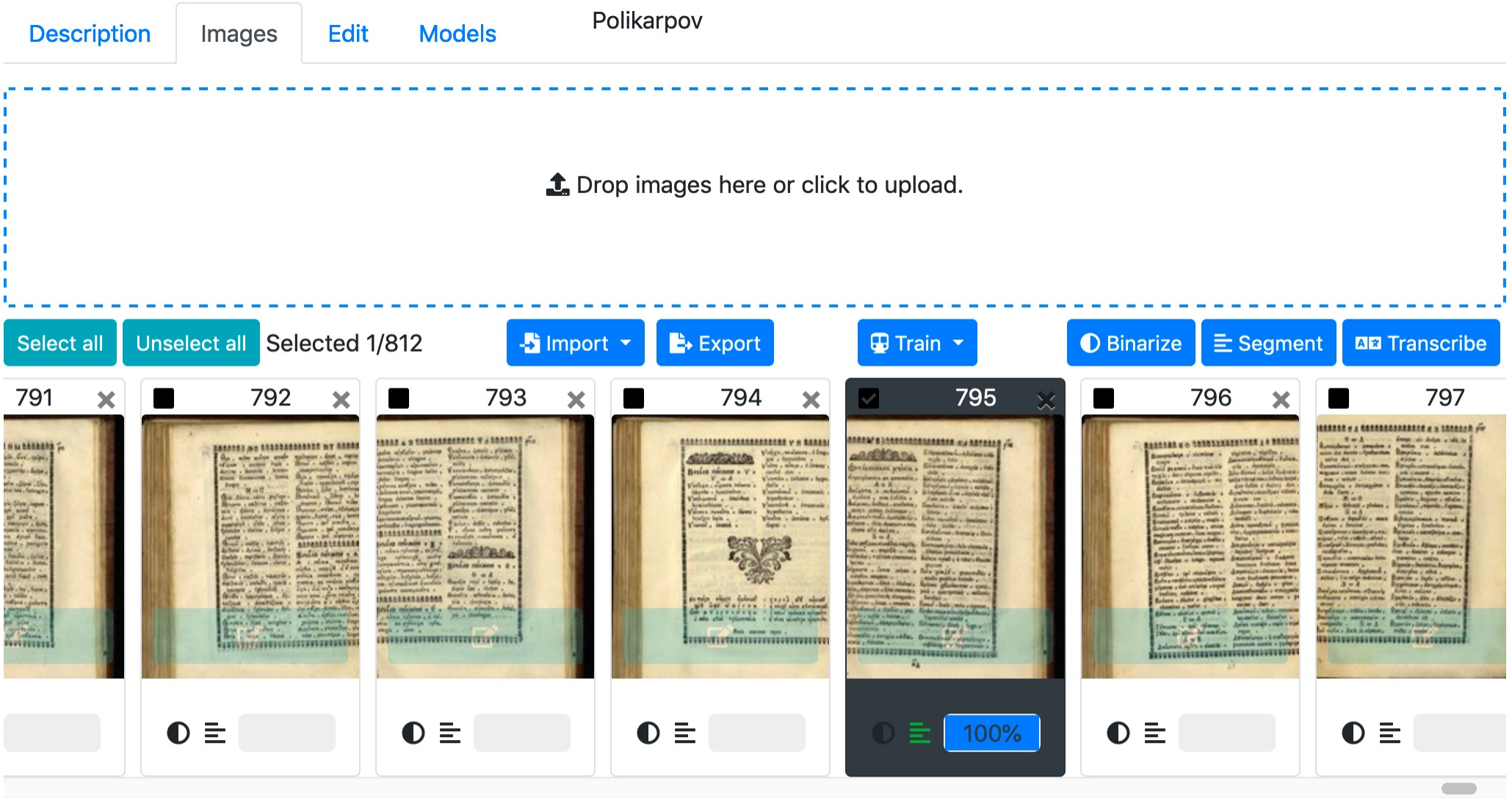
Manual Edition

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.

Die Willkommensseite für eScriptorium im Browser

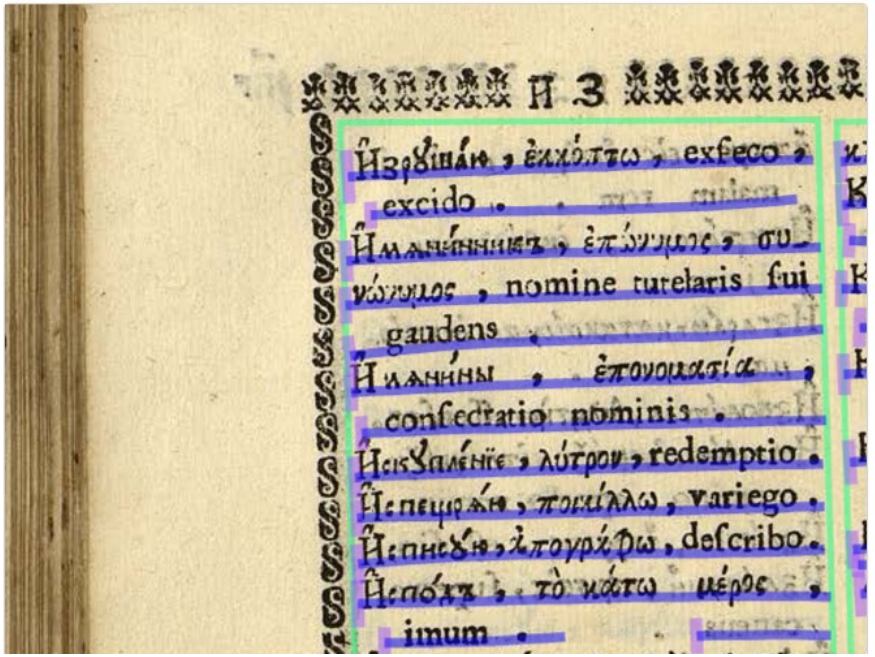


Die grafische Oberfläche von eScriptorium (aktuelle Version)

Description Images Edit Models Polikarpov Appendix Test

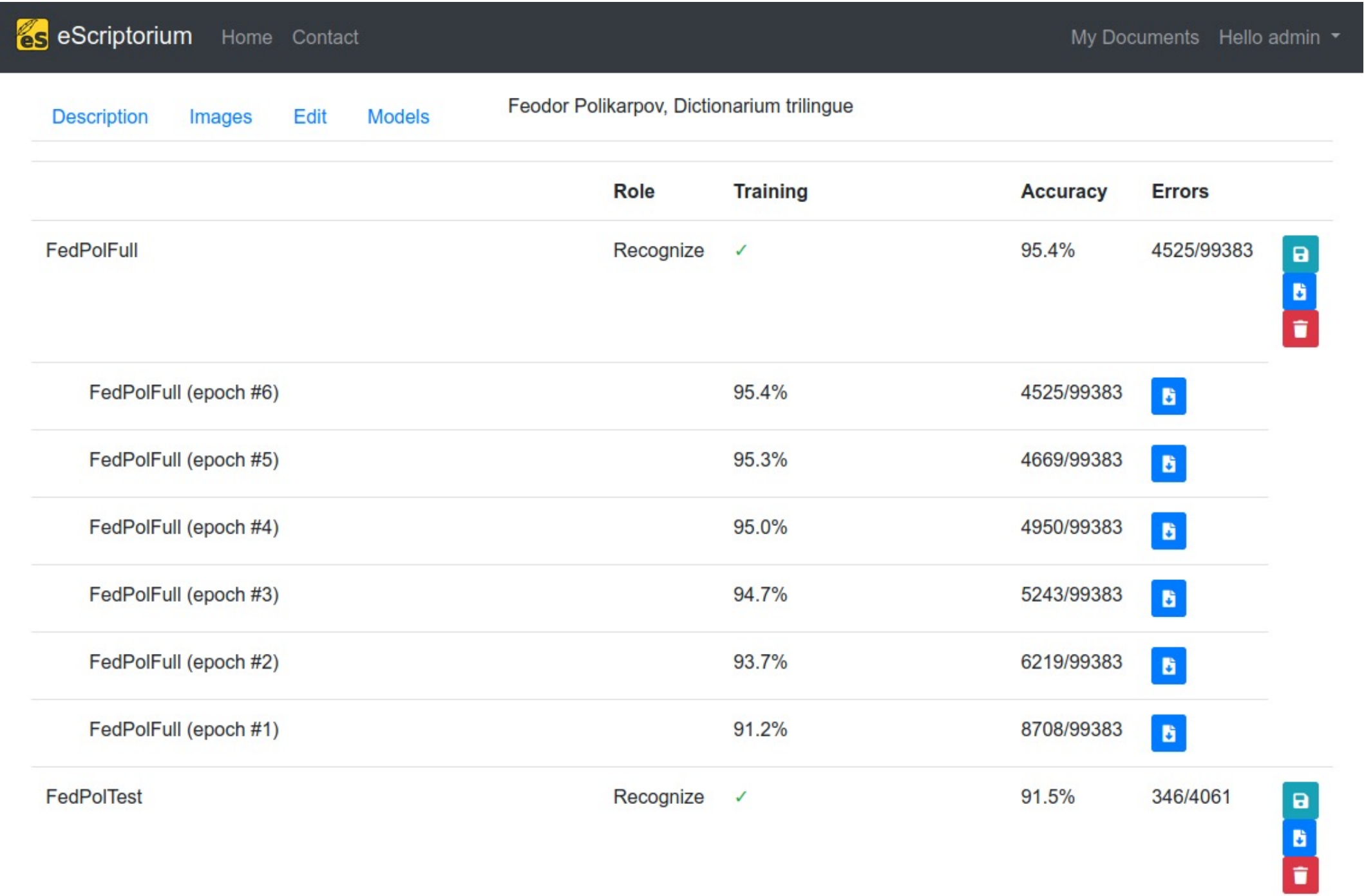
Element 4 - Polikarpov_798.png - (2256x2642)

kraken:fedpolfull



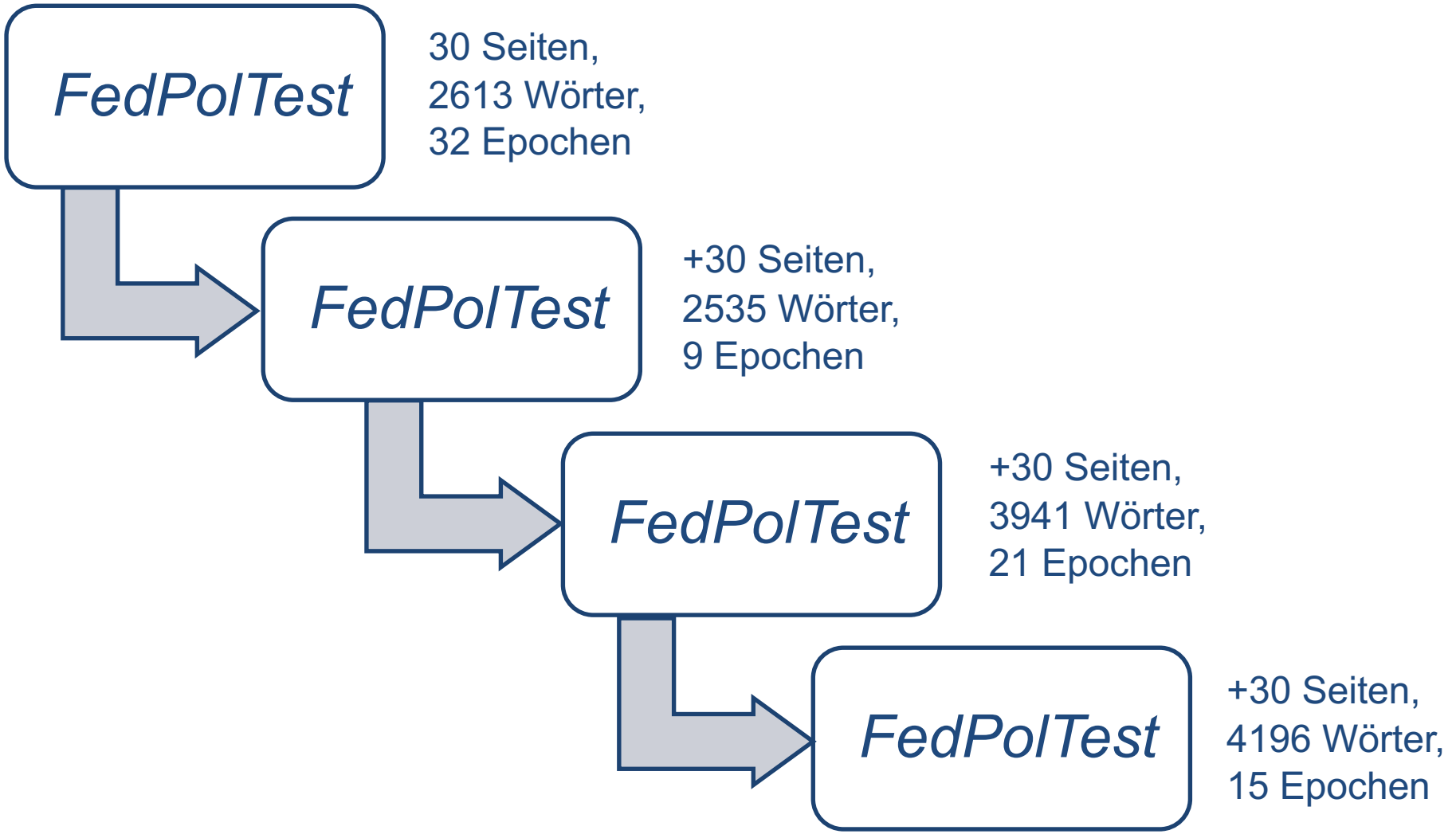
1 Изрѣшáю, ἐκάτεω, exfeco,
2 excido.
3 Именúнникъ, ἐπώνυμος, συ-
4 νώνυμος, nomine tutelaris fui
5 gaudens.
6 Именúны, ἐπονομασία,
7 confecratio nominis.
8 Искупленïе, λύτρον, redemptio.
9 Испещрáю, ποικίλλω, variego.
10 Описáю, ἀ πογράφω, deſcribo.
11 Испóлъ, τὸ κάτω μέρος,
12 imum.

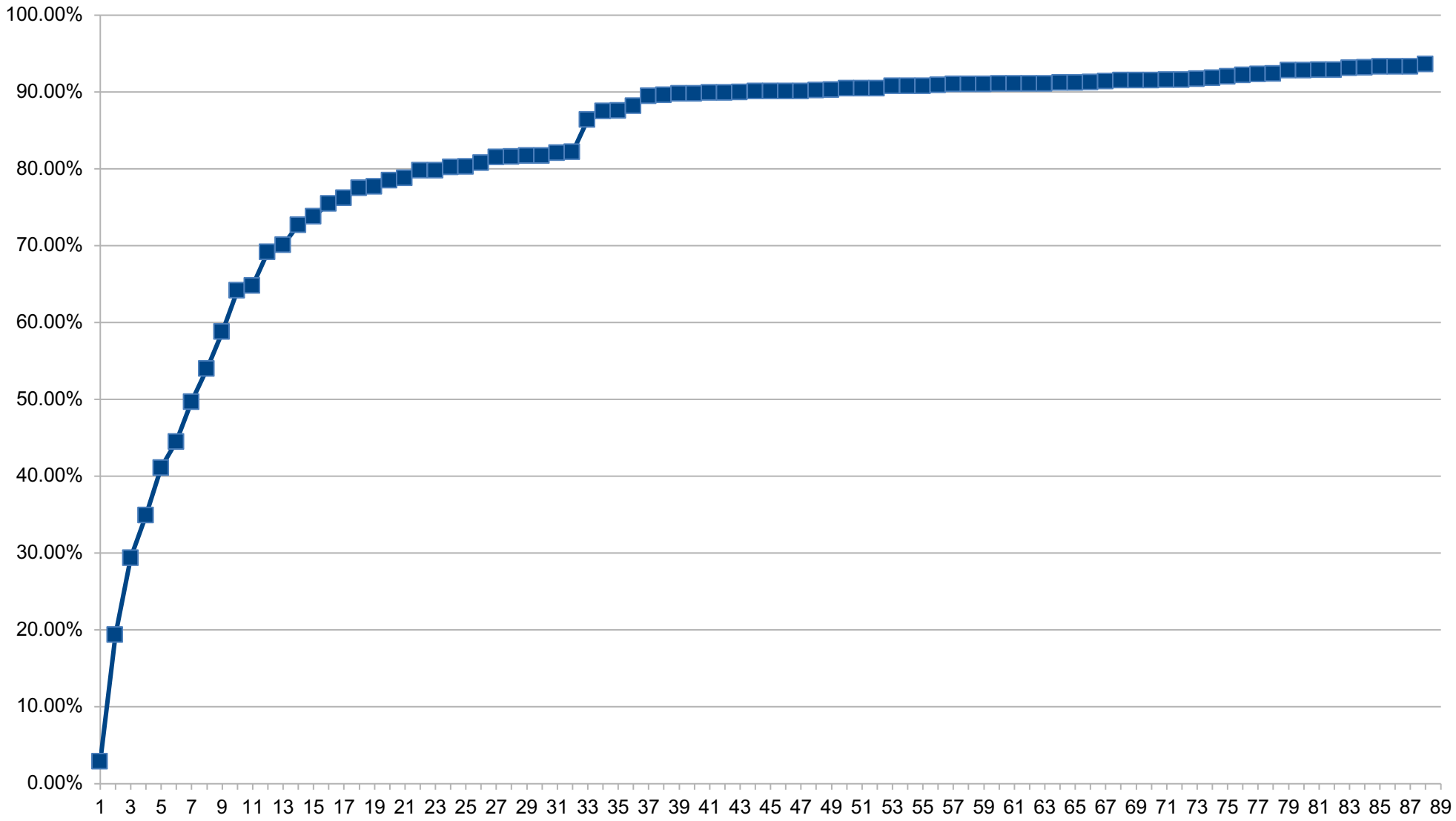
Segmentierung und Texteingabe



Modelldateien speichern und herunterladen

Trainingverfahren (kraken, 1. Versuch)



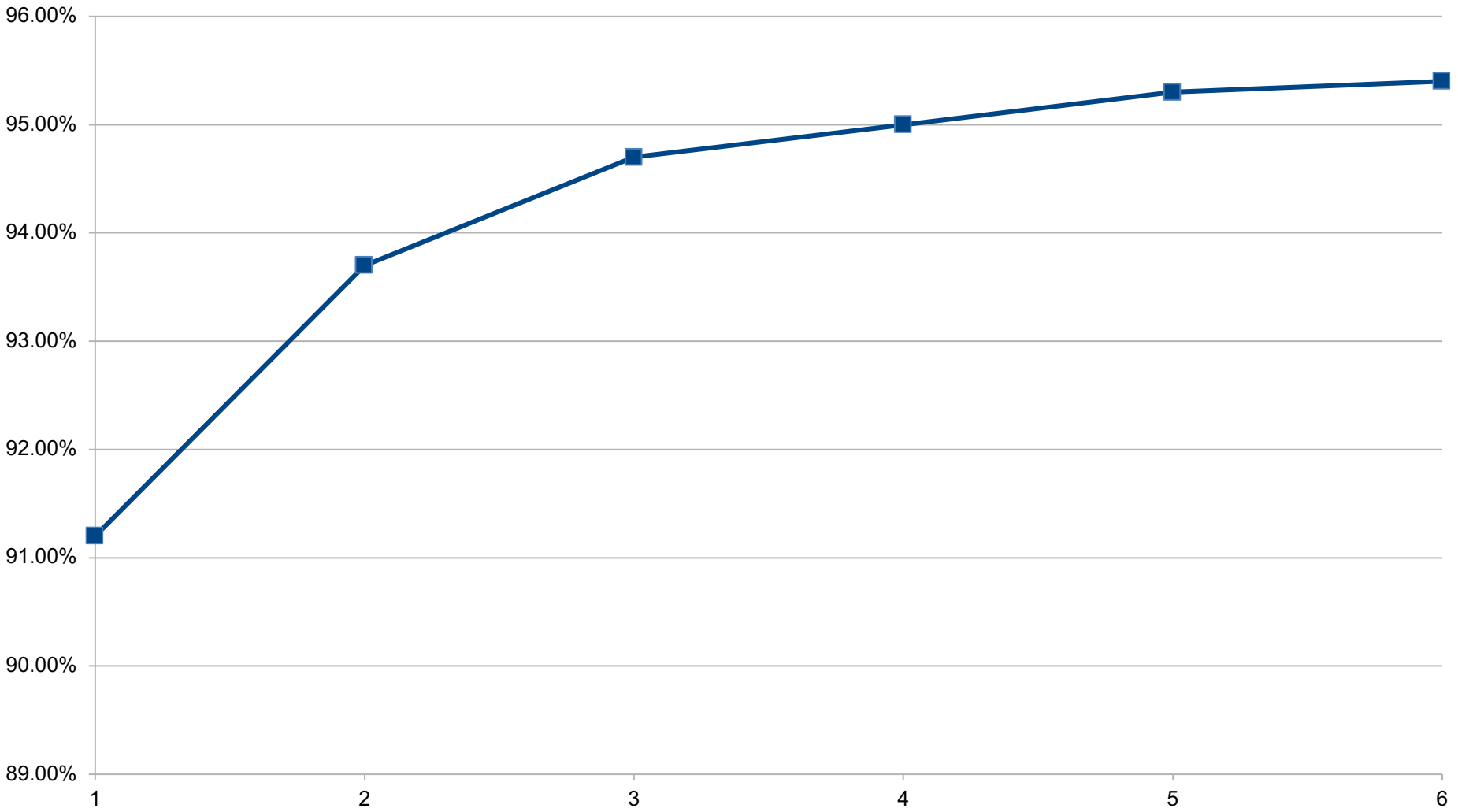


Lernkurve für das Modell *FedPolTest*

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 t*ranskribiert 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!