Transkribus,

a research platform for the mass digitisation, transcription, recognition and searching of historical documents

Günter Mühlberger

University of Innsbruck,

Digitisation and Digital Preservation Group

READ

Agenda

- Why this meeting?
- Transkribus - Future
- Some updates on technology

Why this meeting?

There is a great chance…

- Excellent prerequisites
- Digitisation by archives and libraries
- Digital Humanities projects
- Long term tradition in (digital) editing
- New technology, but already mature enough for broad usage

…for a Dutch Model

- A large dataset of transcripts available to everyone who wants to train machines to read historical Dutch documents
    - Archives, libraries
    - Scholars
    - Computer scientists
- One or more neural networks (“models”) capable to read any kind of Dutch handwriting of the last 300-400 years with reasonable results
    - Reasonable results would be something around 10% Character Error Rate without further training or adaption
    - Based on Keyword Spotting this will already allow searching with high accuracy
- Data curation as the main task
    - Its not so much an IT task but the task to build and maintain a national dataset of transcriptions

Purpose of the meeting..

- Someone needs to take the lead…
- Spotting and collecting existing digital (and printed) editions
- Negotiating with data providers
- Transforming editions into machine readable data
  - Partly with automated Text2Image matching
  - Partly with support of service providers or crowd
- Transkribus is the logical place for this dataset, but the dataset can of course be used independently from Transkribus

Future of Transkribus

Tuc
2018

Transkribus future

- Projects ends on 30th June 2019
- However, there is a strong demand for Transkribus services so that maintenance of the platform is already safeguarded until 2021
  - EU Project NewsEye (2018-2021)
  - German Science Funds project (2019-2020)
  - Project with National Archive Finland (2019)
  - Project with National Archive Netherlands (2019-2020)
  - National project with cadastre documents from Tyrol (2019-2020)
  - Project with Trinity College Dublin (2019-2021)
  - Project with State Archive Zurich (2019-2020)
  - More to come and under negotiation!

READ-COOP

- European Cooperative Society (SCE) as the legal framework
- Run and further develop the Transkribus platform
- Collaboration of independent entities
- Democratic constitution – members have the final say in the general meeting
- Customers become owners, owners become customers
- Direct benefit of members is the main goal
- No shareholder value
- Open to natural persons as well

Current state of affairs

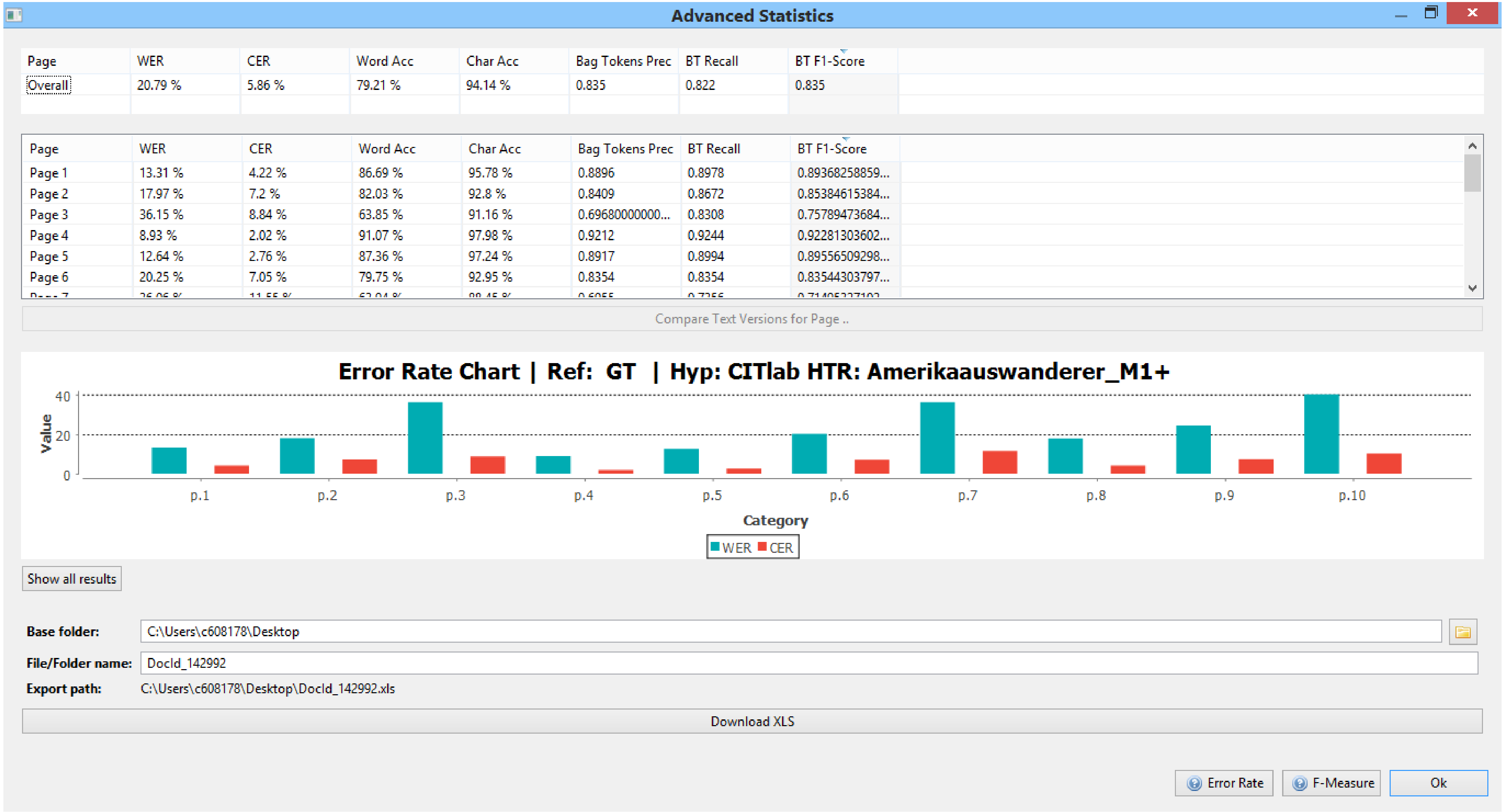
- Statutes in pre-final version
  - Monistic system: Board of directors – general meeting
- Founding members
  - University Innsbruck, University Greifswald, Technical University Valencia, National Archive Finland, British Library, University Library Belgrade, Diocesan Archive Passau, University Rostock, ZAMG Vienna, Picturae, Geneanet France, etc..
  - Everyone invited to join!
- Membership shares
  - From 1000 (minimum) to 5000 EUR, 250 EUR for natural persons
- First board of directors will be formed in the next weeks
- Founding act shall take place before summer holidays
- Official start of business on 1st of July (keep fingers crossed!)

Business

- Transkribus platform
  - Software as a service: We expect that users will work with the platform and connected services in a rather independent way (as it is already the case)
  - However, specific support and development tasks can also be offered for special projects
- Subscription fee for the platform itself
  - Around 3000-4000 EUR per year
- Page based prices for services
  - Between 10 and 16 Cent per page for HTR processing
  - Large quantities can be negotiated
- Discounts for coop members

Recent work and updates

Error rate tool – Advanced mode and sample mode



Advanced Statistics

Page	WER	CER	Word Acc	Char Acc	Bag Tokens Prec	BT Recall	BT F1-Score
Overall	20.79 %	5.86 %	79.21 %	94.14 %	0.835	0.822	0.835

Page	WER	CER	Word Acc	Char Acc	Bag Tokens Prec	BT Recall	BT F1-Score
Page 1	13.31 %	4.22 %	86.69 %	95.78 %	0.8896	0.8978	0.89368258859...
Page 2	17.97 %	7.2 %	82.03 %	92.8 %	0.8409	0.8672	0.85384615384...
Page 3	36.15 %	8.84 %	63.85 %	91.16 %	0.69680000000...	0.8308	0.75789473684...
Page 4	8.93 %	2.02 %	91.07 %	97.98 %	0.9212	0.9244	0.92281303602...
Page 5	12.64 %	2.76 %	87.36 %	97.24 %	0.8917	0.8994	0.89556509298...
Page 6	20.25 %	7.05 %	79.75 %	92.95 %	0.8354	0.8354	0.83544303797...
Page 7	36.96 %	11.55 %	63.04 %	88.45 %	0.6955	0.7356	0.71405337102...

Compare Text Versions for Page ..

Error Rate Chart | Ref: GT | Hyp: CITlab HTR: Amerikaauswanderer_M1+

Show all results

Base folder: C:\Users\c608178\Desktop

File/Folder name: DocId_142992

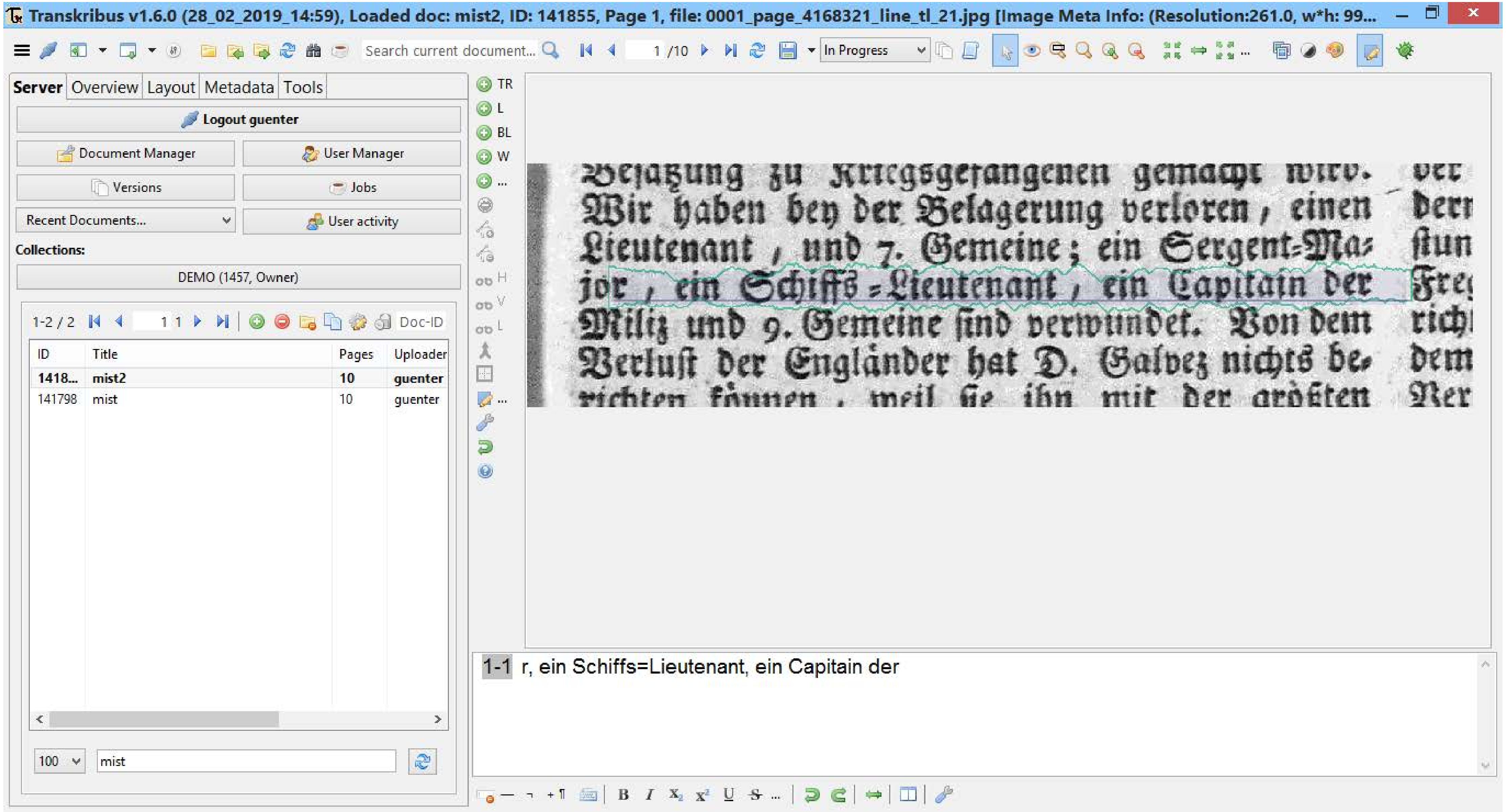
Export path: C:\Users\c608178\Desktop\DocId_142992.xls

Download XLS

Error Rate

F-Measure

Ok

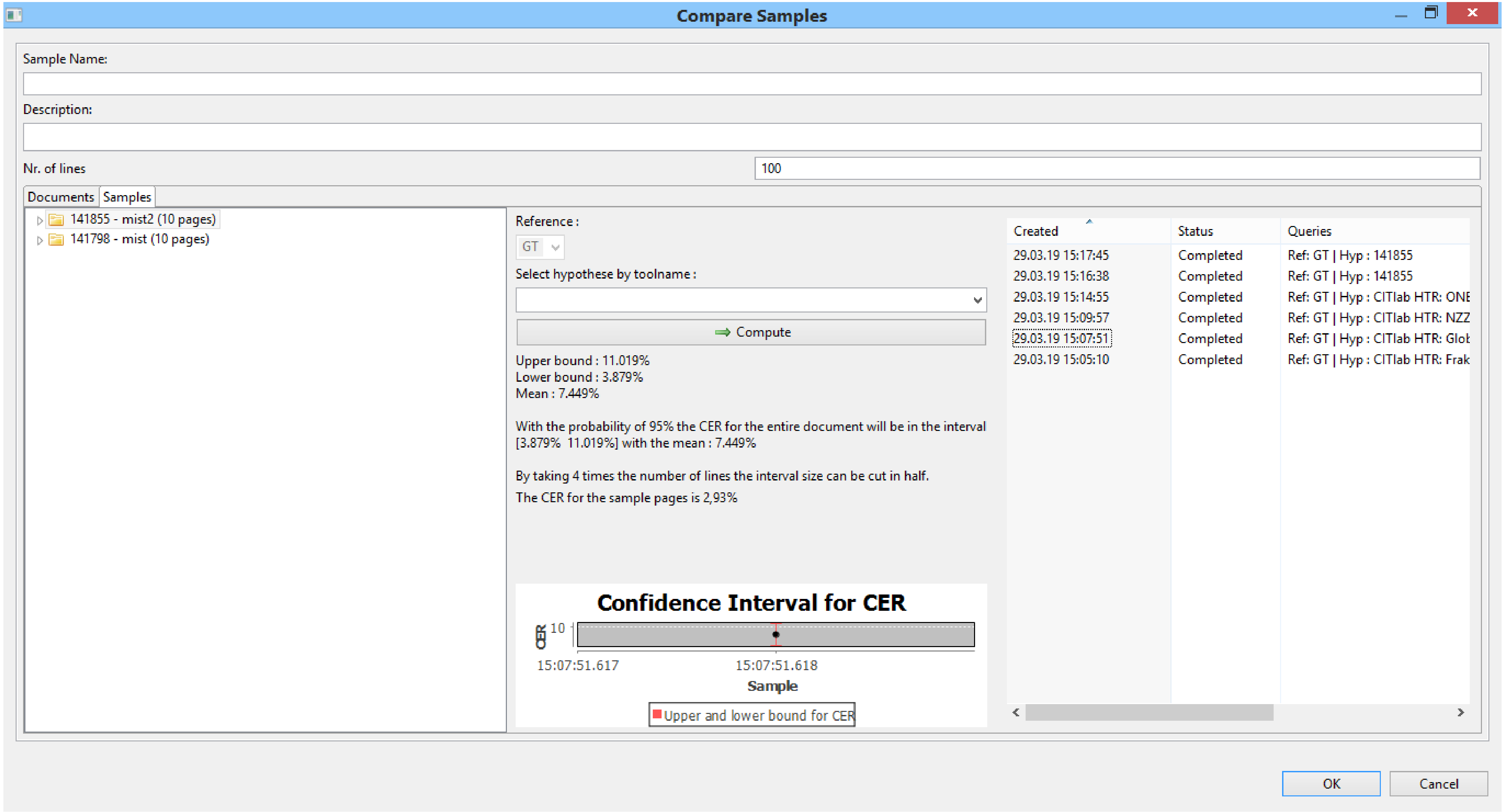


Transkribus v1.6.0 (28_02_2019_14:59), Loaded doc: mist2, ID: 141855, Page 1, file: 0001_page_4168321_line_tl_21.jpg [Image Meta Info: (Resolution:261.0, w*h: 99...

ID	Title	Pages	Uploader
1418...	mist2	10	guenter
141798	mist	10	guenter

Beſatzung zu Kriegsgefangenen gemacht wird.
Wir haben bey der Belagerung verloren, einen
Lieutenant, und 7. Gemeine; ein Sergent=Ma=
jor, ein Schiffs=Lieutenant, ein Capitain der
Miliz und 9. Gemeine ſind verwundet. Von dem
Verluſt der Engländer hat D. Galvez nichts be=
richten können, weil ſie ihn mit der größten

1-1 r, ein Schiffs=Lieutenant, ein Capitain der



Compare Samples

Sample Name:

Description:

Nr. of lines: 100

Documents | Samples

- 141855 - mist2 (10 pages)
- 141798 - mist (10 pages)

Reference : GT

Select hypothesis by toolname :

Upper bound : 11.019%
Lower bound : 3.879%
Mean : 7.449%

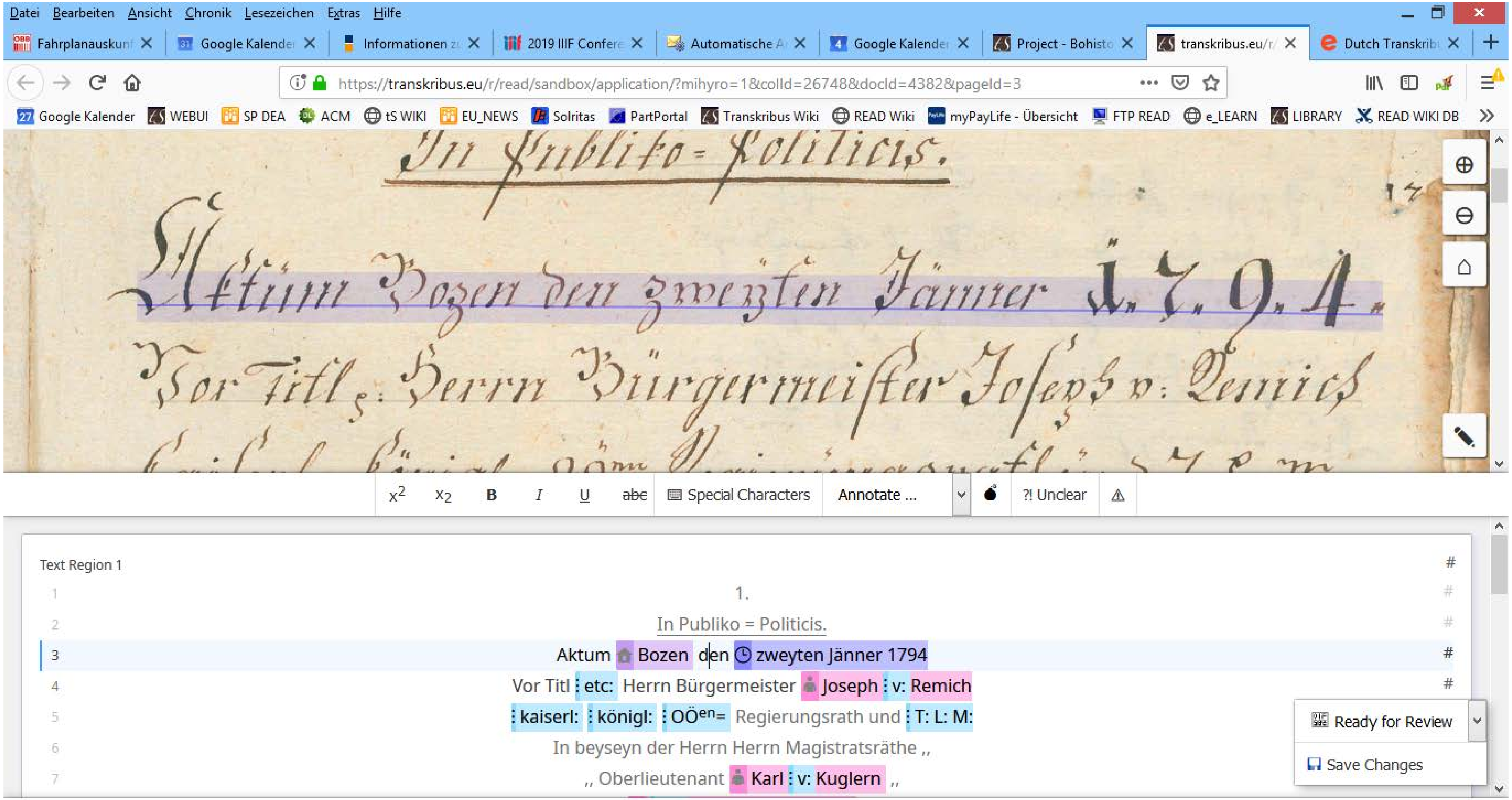
With the probability of 95% the CER for the entire document will be in the interval [3.879% 11.019%] with the mean : 7.449%

By taking 4 times the number of lines the interval size can be cut in half.

The CER for the sample pages is 2,93%

Created	Status	Queries
29.03.19 15:17:45	Completed	Ref: GT | Hyp : 141855
29.03.19 15:16:38	Completed	Ref: GT | Hyp : 141855
29.03.19 15:14:55	Completed	Ref: GT | Hyp : CITlab HTR: ONE
29.03.19 15:09:57	Completed	Ref: GT | Hyp : CITlab HTR: NZZ
29.03.19 15:07:51	Completed	Ref: GT | Hyp : CITlab HTR: Glok
29.03.19 15:05:10	Completed	Ref: GT | Hyp : CITlab HTR: Frak

Browser based transcription interface (including tagging and baseline correction)



Text Region 1

1.

In Publiko = Politicis.

Aktum Bozen den zweyten Jänner 1794

Vor Titl etc: Herrn Bürgermeister Joseph v: Remich

kaiserl: königl: OÖ^{en}= Regierungsrath und T: L: M:

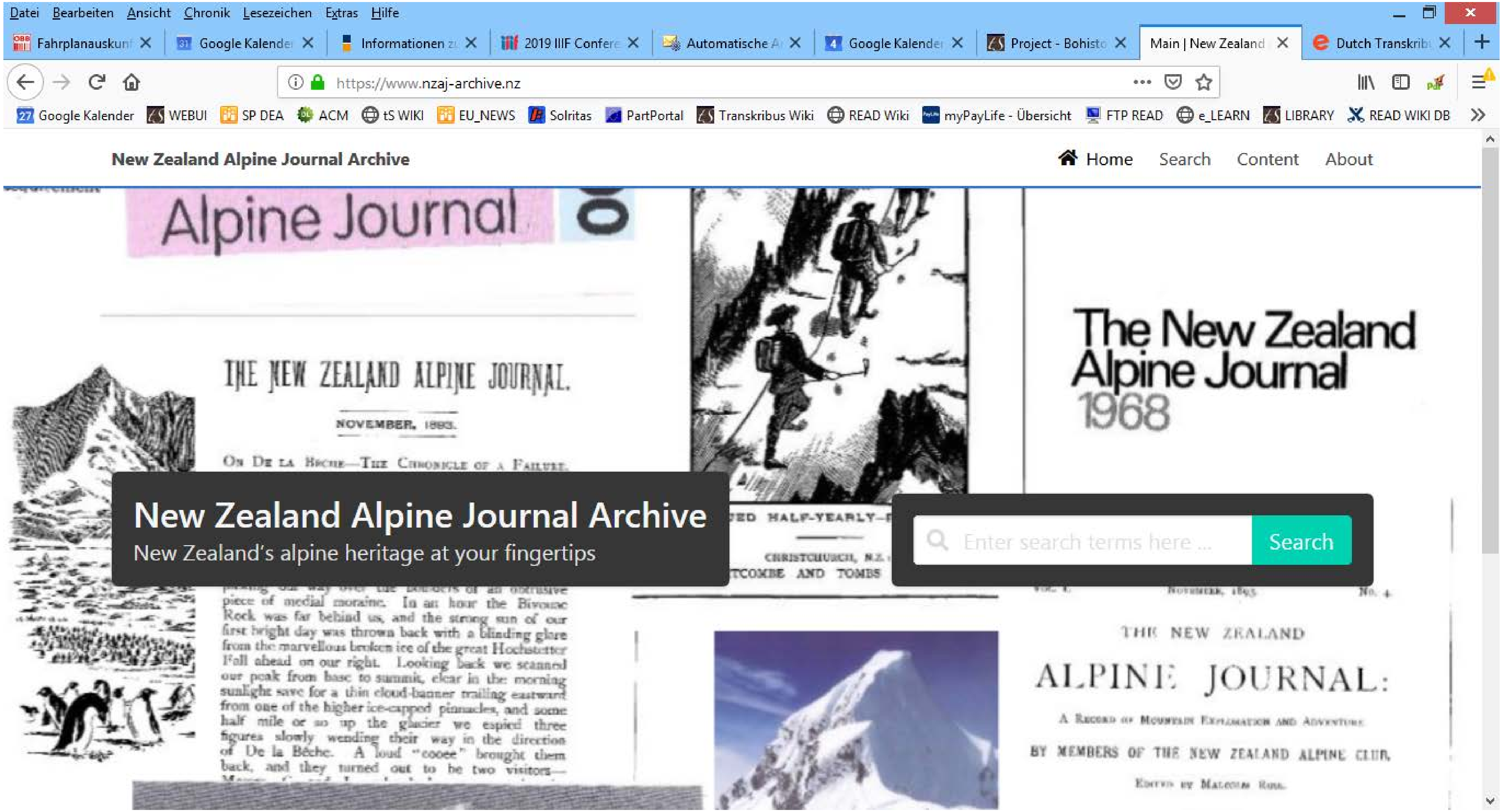
In beyseyn der Herrn Herrn Magistratsräthe ,,

,, Oberlieutenant Karl v: Kuglern ,,

Technical updates

Simple search interface for New Zealand Alpine Journal

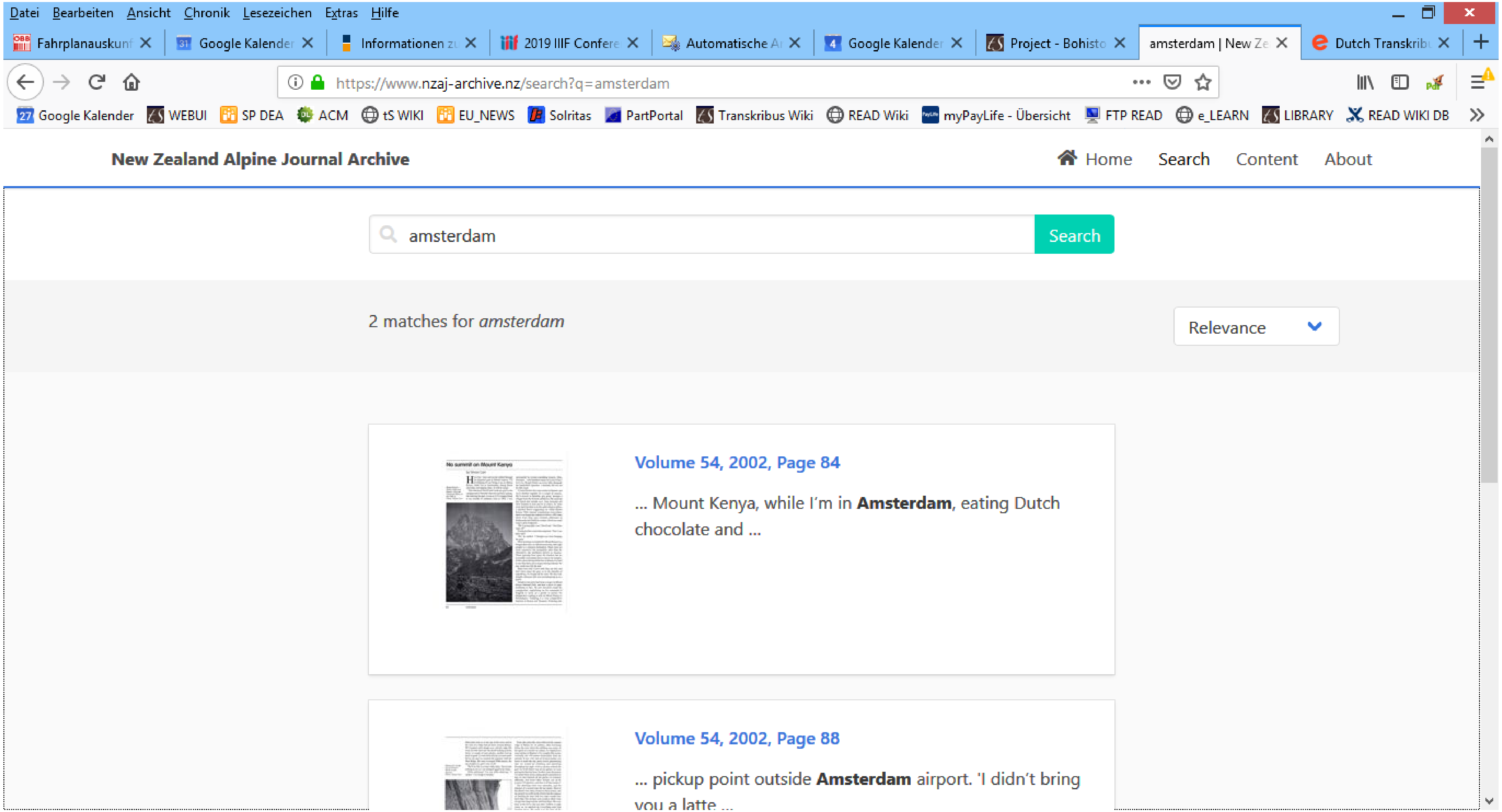
http://nzaj-archive.nz/



New Zealand Alpine Journal Archive

New Zealand Alpine Journal Archive

New Zealand's alpine heritage at your fingertips



amsterdam

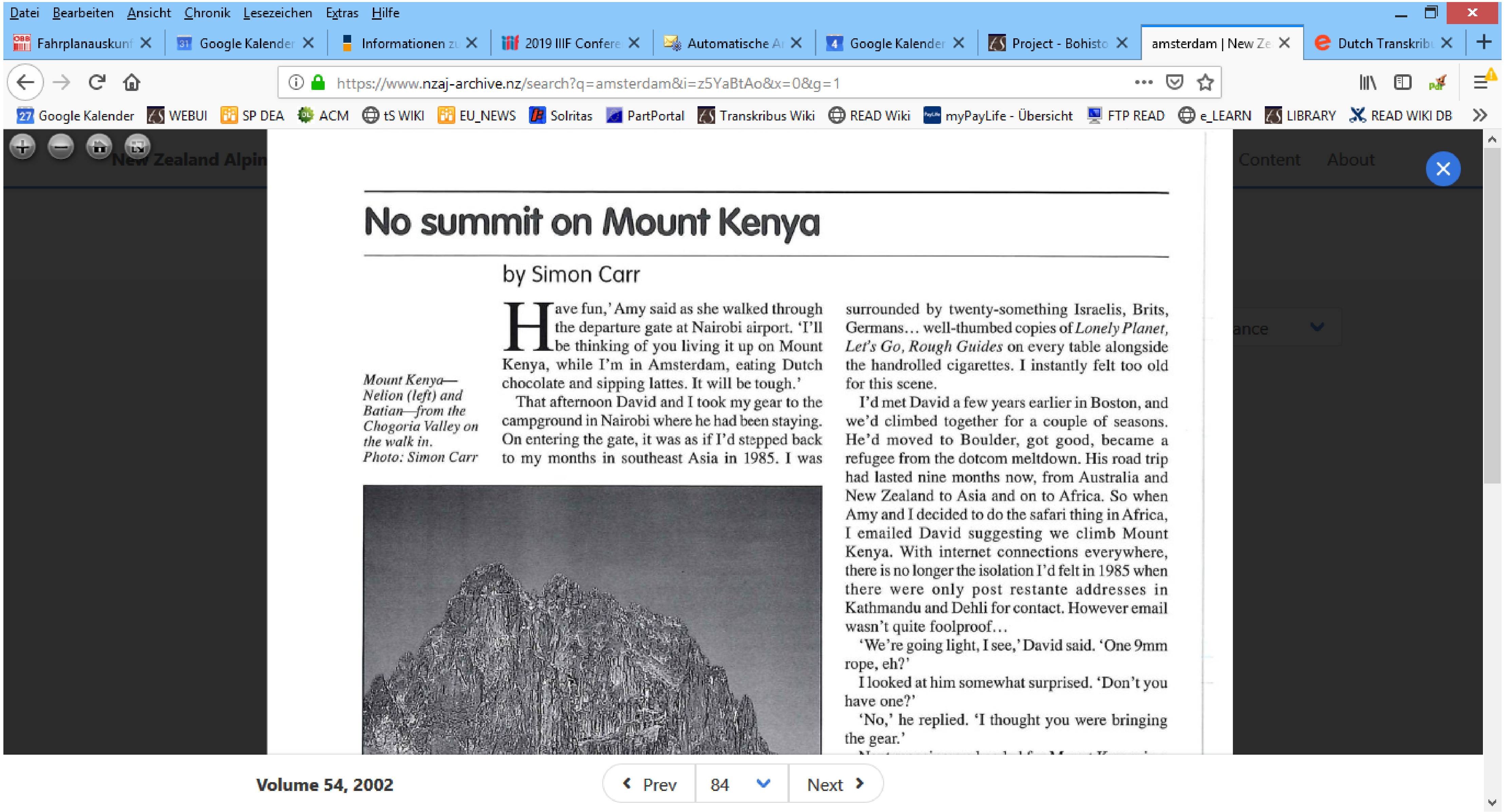
2 matches for *amsterdam*

Volume 54, 2002, Page 84

... Mount Kenya, while I'm in **Amsterdam**, eating Dutch chocolate and ...

Volume 54, 2002, Page 88

... pickup point outside **Amsterdam** airport. 'I didn't bring you a latte ...



No summit on Mount Kenya

by Simon Carr

Mount Kenya—Nelion (left) and Batian—from the Chogoria Valley on the walk in.
Photo: Simon Carr

Have fun,' Amy said as she walked through the departure gate at Nairobi airport. 'I'll be thinking of you living it up on Mount Kenya, while I'm in Amsterdam, eating Dutch chocolate and sipping lattes. It will be tough.'

That afternoon David and I took my gear to the campground in Nairobi where he had been staying. On entering the gate, it was as if I'd stepped back to my months in southeast Asia in 1985. I was surrounded by twenty-something Israelis, Brits, Germans… well-thumbed copies of *Lonely Planet, Let's Go, Rough Guides* on every table alongside the handrolled cigarettes. I instantly felt too old for this scene.

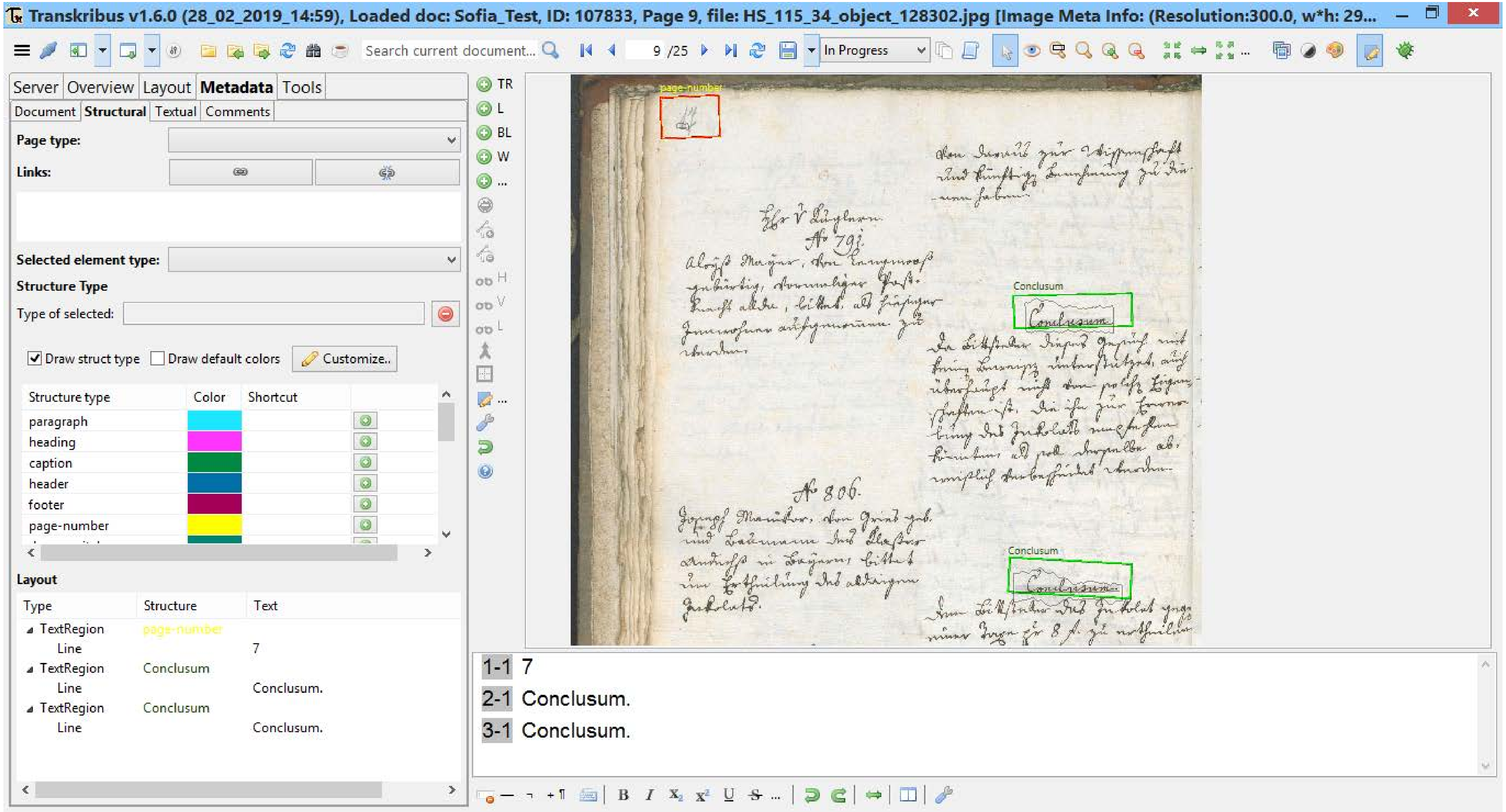
I'd met David a few years earlier in Boston, and we'd climbed together for a couple of seasons. He'd moved to Boulder, got good, became a refugee from the dotcom meltdown. His road trip had lasted nine months now, from Australia and New Zealand to Asia and on to Africa. So when Amy and I decided to do the safari thing in Africa, I emailed David suggesting we climb Mount Kenya. With internet connections everywhere, there is no longer the isolation I'd felt in 1985 when there were only post restante addresses in Kathmandu and Dehli for contact. However email wasn't quite foolproof…

'We're going light, I see,' David said. 'One 9mm rope, eh?'

I looked at him somewhat surprised. 'Don't you have one?'

'No,' he replied. 'I thought you were bringing the gear.'

Trainable layout analysis tool



Transkribus v1.6.0 (28_02_2019_14:59), Loaded doc: Sofia_Test, ID: 107833, Page 9, file: HS_115_34_object_128302.jpg [Image Meta Info: (Resolution:300.0, w*h: 29...

Search current document... 9 /25 In Progress

Server | Overview | Layout | **Metadata** | Tools

Document | **Structural** | Textual | Comments

Page type:

Links:

Selected element type:

Structure Type

Type of selected:

☑ Draw struct type □ Draw default colors Customize..

Structure type	Color	Shortcut
paragraph	🟦	
heading	🟪	
caption	🟩	
header	🟦	
footer	🟥	
page-number	🟨	

Layout

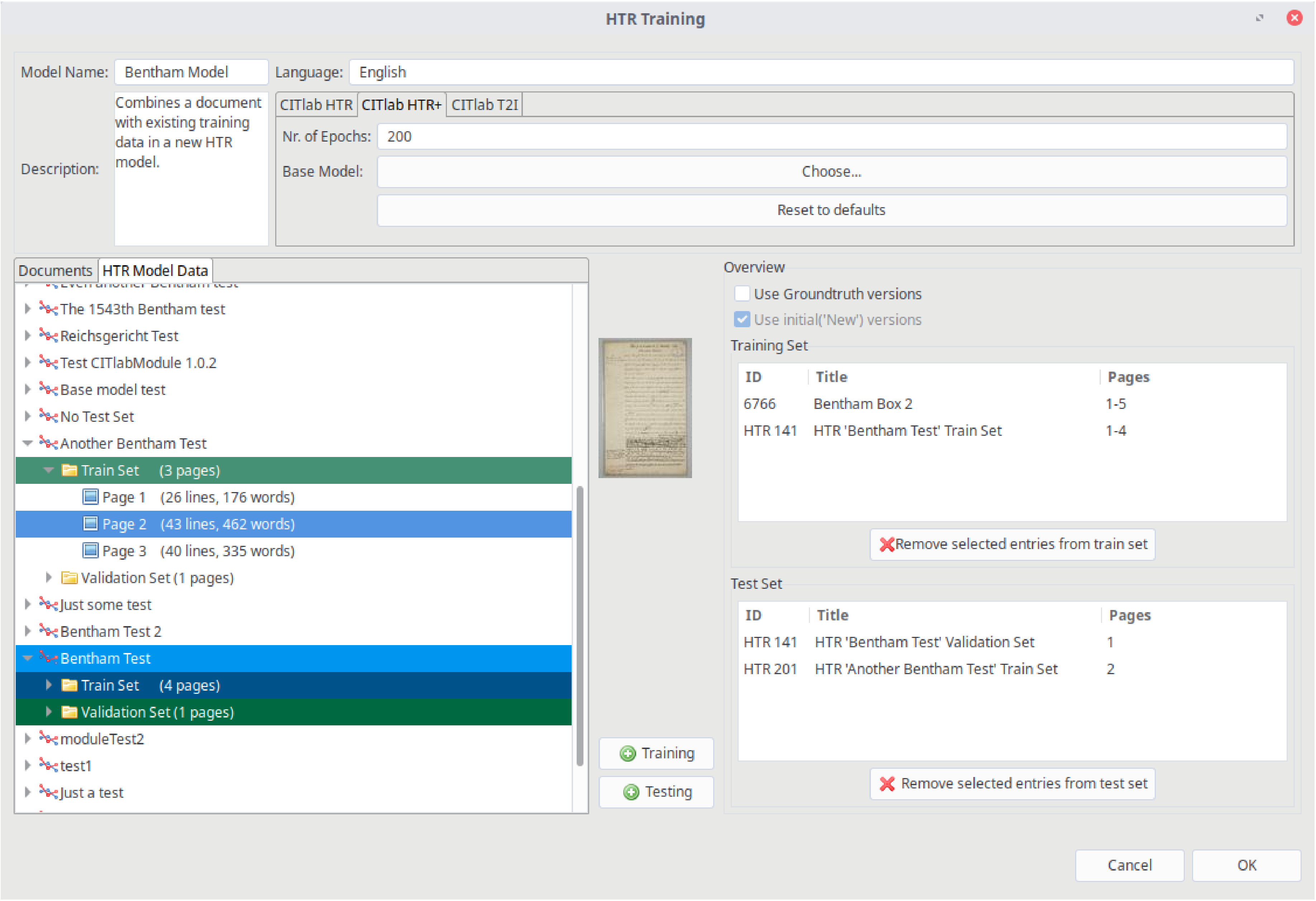
Type	Structure	Text
TextRegion	page-number	
Line		7
TextRegion	Conclusum	
Line		Conclusum.
TextRegion	Conclusum	
Line		Conclusum.

1-1 7

2-1 Conclusum.

3-1 Conclusum.

New training interface for HTR models



HTR Training

Model Name: Bentham Model

Language: English

Description: Combines a document with existing training data in a new HTR model.

Nr. of Epochs: 200

Base Model:

- Even another Bentham test
- The 1543th Bentham test
- Reichsgericht Test
- Test CITlabModule 1.0.2
- Base model test
- No Test Set
- Another Bentham Test
  - Train Set (3 pages)
    - Page 1 (26 lines, 176 words)
    - Page 2 (43 lines, 462 words)
    - Page 3 (40 lines, 335 words)
  - Validation Set (1 pages)
- Just some test
- Bentham Test 2
- Bentham Test
  - Train Set (4 pages)
  - Validation Set (1 pages)
- moduleTest2
- test1
- Just a test

Overview

☐ Use Groundtruth versions

☑ Use initial('New') versions

Training Set

ID	Title	Pages
6766	Bentham Box 2	1-5
HTR 141	HTR 'Bentham Test' Train Set	1-4

Test Set

ID	Title	Pages
HTR 141	HTR 'Bentham Test' Validation Set	1
HTR 201	HTR 'Another Bentham Test' Train Set	2

ScanTent now delivered for testing to interested users all around the world

scantent.eu

Yesterday…

- 19.258 images uploaded by single users
- 82 new users
- 322 unique logins to Transkribus expert client
- 1246 jobs processed
- 5 new HTR models trained by users

Thank you for your attention!

Further information

https://read.transkribus.eu/

https://transkribus.eu/

https://read.transkribus.eu/coop/

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 674943.

European Commission