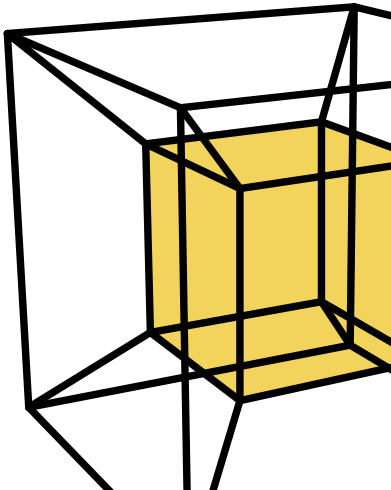
# When Tesseract Does It Alone

## Optical Character Recognition of Medieval Texts

**Vítek Novotný**
**witiko@mail.muni.cz**

Faculty of Informatics, Masaryk University

December 8, 2020



# Introduction

- In the AHISTO project, the Department of Auxiliary Historical Sciences and Archive Studies (DAHSAS) and the NLP Centre will develop a database of Medieval texts.
- Scanned regests (19th–20th cent.) of Hussite texts (1419–1436) from CMS online.
- To enable full-text search, the scanned texts must be optically recognized (OCR).
- We have preprocessed the scanned texts and evaluated three OCR engines:
  1. Google Cloud Vision AI
  2. Tesseract 3, 4, and 3 + 4
  3. OCR-D

# Related Work I

## Google Cloud Vision AI

- Paid proprietary OCR service ($1.50 / 1,000 images) made available by Google in 2019.
- Used by DAHSAS at CMS online.
- Uses non-standard JSON output.
- Detects language at character level.
- Our ground truth for OCR evaluation.

## Tesseract [1, 2]

- Developed in 1980s by Hewlett-Packard, made free open-source in 2005.
- Funded by Google since 2006 and likely used inside Google Cloud Vision AI.
- Since version 2, supports Western languages.
- Since version 3:
  1. supports ideographic and RTL languages,
  2. supports standard hOCR output, and
  3. detects language at paragraph level.
- Since version 4, uses non-accelerated LSTM engine for line OCR.

# Related Work II

## OCR-D [5]

- A free open-source German Research Foundation (DFG)-funded joint project of:
  1. the Berlin-Brand. Academy of Sciences,
  2. the Herzog August Bibliothek,
  3. the Berlin State Library, and
  4. the Karlsruhe Institute of Technology.
- Developed to digitize the German cultural heritage of the 16th–19th century in connection with the VD16, VD17, and VD18 cataloging and archival projects.
- Since February 2020, OCR-D is being deployed to intent organizations.
- Supports configurable heterogeneous OCR workflows:
  1. Enhancement
  2. Binarization
  3. Cropping
  4. Binarization
  5. Denoising
  6. Deskewing
  7. Dewarping
  8. Segmentation
  9. Clipping
  10. Text recognition
  11. Text alignment
  12. Post-correction
- Supports GPU-accelerated Calamari engine [3, 4] (SOTA on Fraktur) for line OCR.
- Supports standard hOCR output, but does not detect language with best OCR engine.
- Supports the OCR4all web front-end for semi-automatic OCR.

# Methods

## Data Preprocessing

- We received 302,909 low-res and 168,113 hi-res scanned images from DAHSAS.
- Low-res images have estimated 145 DPI and are linked to ground truth texts.
- Hi-res images have estimated 414 DPI and are suitable for OCR.
- Less than 187,267 (62%) low-res scanned images come from same books as hi-res.
- To produce our test dataset, we linked low-res and hi-res images as follows:
  1. All scanned images were pre-processed: rescaled to 512 × 512 px and binarized.
  2. Pre-processed hi-res images were indexed in a vector DB by Hamming distance.
  3. For each of the 187,267 pre-processed low-res images:
     1. We retrieved its ground truth (GT) text.
     2. We retrieved 100 nearest hi-res images.
  4. For each of the 100 nearest hi-res images:
     1. We OCRd hi-res image by Tesseract 4.
     2. We reranked by TF-IDF cossim(OCR, GT).
  5. If 1NN before and after reranking match, we link the low-res and 1NN hi-res image.
  6. If all low-res scanned images of a book were linked, we call the book linked.
- We linked 108 (19–29%) books containing 65,348 (39%) hi-res scanned images.

# Quantitative Evaluation

## Speed

- To evaluate speed, we report the wall clock time in days on a single CPU or GPU.
- For Tesseract, we used apollo, asteria04, mir, hypnos1, turnus01, nymfe{23..74}: 492 CPUs.
- For OCR-D, we used epimetheus{1..4} for CPU and apollo, turnus03 for GPU.
- For OCR-D, we also measured each workflow step separately.

## Accuracy

- To evaluate accuracy, we report Character Error Rate (CER) and Word Error Rate (WER): [6]

$$\text{ErrorRate}(A, B) = \frac{\text{EditDistance}(A, B)}{\text{Maximum}(|A|, |B|)} \times 100\%$$

- For WER, we lower-cased and deaccented the texts to model our full-text search use case.
- Although CER and WER are correlated, we are mainly interested in WER, which is harder.

# Qualitative Evaluation

- To better understand the strengths and weaknesses of the individual OCR engines, we compare OCR outputs side-by-side on different scanned images:
  1. A random page from the test dataset.
  2. The page with the worst WER on the best engine.
  3. The page with the best WER on the worst engine.
- To inspect ground truth, we compare the ratio of improvement to deterioration.

# Results

## Quantitative Evaluation I

| | **Tesseract 3** | Tesseract 4 | OCR-D (GPU) | Tesseract 3 + 4 | OCR-D (CPU) |
|---|---|---|---|---|---|
| Time (days) | **61.12** | 127.69 | 140.39 | 172.11 | 174.07 |

Table: The speed of optical character recognition on the test dataset for different OCR engines.

| | **Tesseract 4** | Tesseract 3 | Tesseract 3 + 4 | OCR-D |
|---|---|---|---|---|
| Character Error Rate (CER, %) | **9.81** | 10.60 | 12.00 | 19.88 |
| Word Error Rate (WER, %) | **13.77** | 16.04 | 18.70 | 30.94 |

Table: The accuracy of optical character recognition on the test dataset for different OCR engines.

# Results

## Quantitative Evaluation II

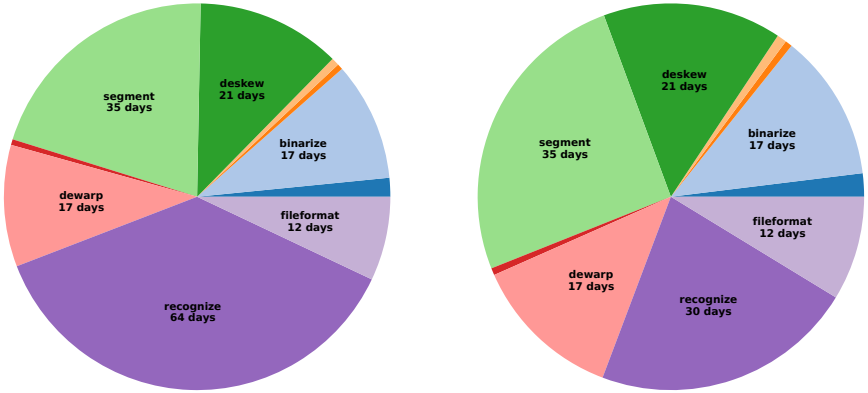
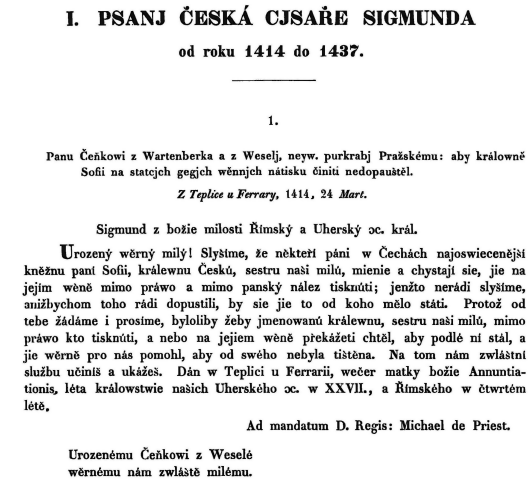



Figure: The speed of individual workflow steps of OCR-D on CPU (left) and GPU (right).

# Qualitative Evaluation I

## Random Page (book 792, page 5)



I. PSANJ ČESKÁ CJSAŘE SIGMUNDA

od roku 1414 do 1437.

1.

Panu Čeňkowi z Wartenberka a z Weselj, neyw. purkrabj Pražskému: aby králownĕ Sofii na statcjch gegjch wĕnnjch nátisku činiti nedopauštĕl.

*Z Teplice u Ferrary,* 1414, 24 *Mart.*

Sigmund z božie milosti Římský a Uherský oc. král.

Urozený wĕrný milý! Slyšíme, že někteří páni w Čechách najoswiecenĕjší kněžnu pant Sofii, králewnu Českú, sestru naši milú, mienie a chystají sie, jie na jejím wĕně mimo práwo a mimo panský nález tisknúti; jenžto nerádi slyšíme, anižbychom toho rádi dopustili, by sie jie to od koho mĕlo státi. Protož od tebe žádáme i prosíme, byloliby žeby jmenowanú králewnu, sestru naši milú, mimo práwo kto tisknúti, a nebo na jejiem wĕně překážeti chtĕl, aby podlé ní stál, a jie wĕrně pro nás pomohl, aby od swého nebyla tištĕna. Na tom nám zwláštní službu učiníš a ukážeš. Dán w Teplici u Ferrarii, wečer matky božie Annuntiationis, léta králowstwie našich Uherského oc. w XXVII., a Římského w čtwrtém létĕ,

Ad mandatum D. Regis: Michael de Priest.

Urozenému Čeňkowi z Weselé
wĕrnému nám zwláštĕ milému.

## Google Cloud Vision AI (0% CER, 0% WER)

I. PSANJ ČESKÁ CJSARE SIGMUNDA
od roku 1414 do 1437.

Panu Čeńkowi z Wartenberka a z Weselj, neyw. purkrabj Pražskému: aby králownĕ Sofii na statejch gegjch wĕnnjch nátisku činiti nedopauštĕl.

Z Teplice u Ferrary, 1414, 24 Mart.

Sigmund z božie milosti Rimský a Uherský oc. král.

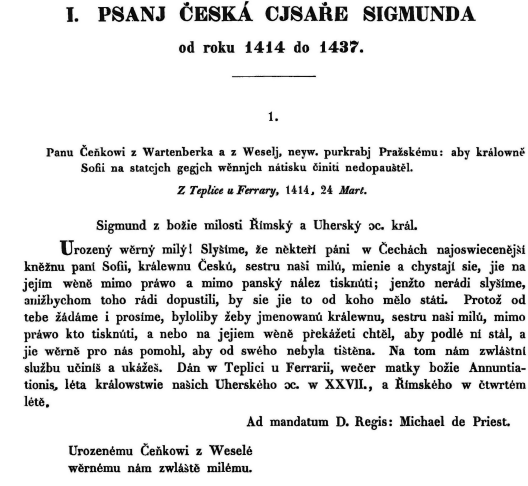
Urozený wĕrný mily! Slyšíme, že někteří páni w Čechách najoswiecenĕjší kněžnu pant Sofii, králewnu Českú, sestru naši milú, mienie a chystaji sie, jie na jejím wĕně mimo práwo a mimo panský nález tisknúti; jenžto nerádi slyšíme, anižbychom toho rádi dopustili, by sie jie to od koho mĕlo státi. Protož od tebe žádáme i prosime, byloliby žeby jmenowanú králewnu, sestru naši milú, mimo prawo kto tisknúti, a nebo na jejiem wĕně překážeti chtĕl, aby podle ní stál, a jie wĕrně pro nás pomohl, aby od swého nebyla tištĕna. Na tom nám zwláštní službu učiniš a ukážeš. Dán w Teplici u Ferrarii, weder matky božie Annuntia-tionis, léta králowstwie našich Uherského uc. w XXVII., a Římskéhow čtwrtém létĕ,

Ad mandatum D. Regis: Michael de Priest.

Urozenému Čeńkowi z Wesele
wĕrnému nám zwláštĕ milému.

# Qualitative Evaluation II

## Random Page (book 792, page 5)



I. PSANJ ČESKÁ CJSAŘE SIGMUNDA

od roku 1414 do 1437.

1.

Panu Čeňkowi z Wartenberka a z Weselj, neyw. purkrabj Pražskému: aby králownĕ Sofii na statcjch gegjch wĕnnjch nátisku činiti nedopauštĕl.

*Z Teplice u Ferrary*, 1414, 24 *Mart.*

Sigmund z božie milosti Římský a Uherský oc. král.

Urozený wĕrný milý! Slyšíme, že někteří páni w Čechách najoswiecenĕjší kněžnu paní Sofii, králewnu Českú, sestru naši milú, mienie a chystají sie, jie na jejím wĕně mimo práwo a mimo panský nález tisknúti; jenžto nerádi slyšíme, anižbychom toho rádi dopustili, by sie jie to od koho mělo státi. Protož od tebe žádáme i prosíme, bylo liby žeby jmenowanú králewnu, sestru naši milú, mimo práwo kto tisknúti, a nebo na jejiem wĕně překážeti chtĕl, aby podlé ní stál, a jie wĕrně pro nás pomohl, aby od swého nebyla tištĕna. Na tom nám zwlášťní službu učiníš & ukážeš. Dán w Teplici u Ferrarii, wečer matky božie Annuntiationis, léta králowstwie našich Uherského oc. w XXVII., a Římského w čtwrtém létĕ.

Ad mandatum D. Regis: Michael de Priest.

Urozenému Čeňkowi z Weselé
wĕrnému nám zwláště milému.

## Tesseract 3 (3.96% CER, 6.36% WER)

]. PSANJ ČESKÁ CJSAŘE SIGMUNDA
od roku 1414 do 1437.

Panu Čeňkowi z Wartenberka a z Weselj, neyw. purkrabj Pražskému: aby králownĕ Sofii na. statcjch gegjch wénnjch nátisku činiti nedopauštĕl.

Z Teplice uFermry, 1414, 24 Mart.

Sigmund z božie milosti Římský a Uherský oc. král.

Urozený wĕrný milý! Slyšíme, že někteří páni w Čechách najoswiecenĕjší kněžnu paní Sofii, králewnu Českú, sestru nasi milú, mienie a chystají sie, jie na jejím wĕně mimo práwo a mimo panský nález tisknúti, jenžto nerádi slyšíme, anižbychom toho rádi dopustili, by sie jie to od koho mělo státi. Protož od tebe žádáme i prosíme, bylo liby žeby jmenowanú králewnu, sestru naši milú, mimo práwo kto tisknúti, a nebo na jejiem wĕně překážeti chtĕl, aby podlé ní stál, a jie wĕrně pro nás pomohl, aby od swého nebyla tištĕna. Na tom nám zwlástní službu učiníš & ukážeš. Dán W Teplici u Ferrarii, wečer matky božie Annuntia-tionis, létá královstwie našich Uherského oc. w XXVII., & Římského w čtwrtém létĕ.

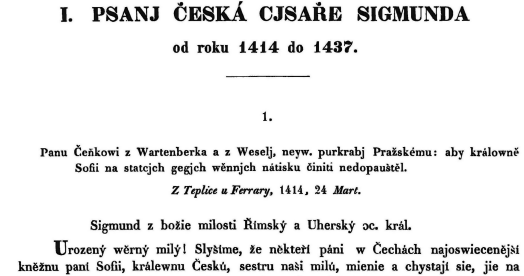
Ad mandatum D. Regis: Michael de Priest.

Urozenému Čeňkowi z Weselé
Wĕrnému nám zwláště milému.

42% of changes improved the ground truth.

# Qualitative Evaluation III

## Random Page (book 792, page 5)



I. PSANJ ČESKÁ CJSAŘE SIGMUNDA

od roku 1414 do 1437.

1.

Panu Čeňkowi z Wartenberka a z Weselj, neyw. purkrabj Pražskému: aby královně Sofii na statcjch gegjch wěnnjch nátisku činiti nedopauštěl.

*Z Teplice u Ferrary,* 1414, 24 *Mart.*

Sigmund z božie milosti Římský a Uherský oc. král.

Urozený wěrný milý! Slyšíme, že někteří páni w Čechách najoswiecenější kněžnu paní Sofii, králewnu Česků, sestru naši milú, mienie a chystají sie, jie na jejím wěně mimo práwo a mimo panský nález tisknúti; jenžto nerádi slyšíme, anižbychom toho rádi dopustili, by sie jie to od koho mělo státi. Protož od tebe žádáme i prosíme, bylo liby žeby jmenowanú králewnu, sestru naši milú, mimo práwo kto tisknúti, a nebo na jejiem wěně překážeti chtěl, aby podlé ní stál, a jie wěrně pro nás pomohl, aby od swého nebyla tištěna. Na tom nám zwláštní službu učiníš a ukážeš. Dán w Teplici u Ferrarii, wečer matky božie Annuntiationis, léta králowstwie našich Uherského oc. w XXVII., a Římského w čtwrtém létě,

Ad mandatum D. Regis: Michael de Priest.

Urozenému Čeňkowi z Weselé
wěrnému nám zwláště milému.

## Tesseract 4 (4.95% CER, 8.57% WER)

I. PSANJ ČESKÁ CJSAŘE SIGMUNDA
od roku 1414 do 1437.

Panu Čeňkowi z Wartenberka a z Weselj, neyw. purkrabj Pražskému: aby královně Sofii na statcjch gegjch wénnjch nátisku činiti nedopauštěl.

Z Teplice u Ferrary, 1414, 24 Mart.

Sigmund z bozie milosti Římský a Uherský oc. král.

Urozený wěrný milý! Slyšííme, že někteří páni w Čechách najoswiecenější kněžnu paní Sofii, králewnu Česků, sestru naši milů, mienie a chystají sie, jie na jejím wěeně mimo práwo a mimo panský nález tisknůti; jenžto nerádi slyšíme, anižbychom toho rádi dopustili; by sie jie to od koho mělo státi. ProtoZ od tebe žádáme i prosime, bylo liby Zeby jmenowanü králewnu, sestru naài milü, mimo práwo kto tisknüti, a nebo na jejiem wéné prekázeti chtél, aby podlé ni stál, a jie wěrně pro nás pomohl, aby od swého nebyla tištěna. Na tom nám zwláštní sluZbu ucini$ a ukážeš. Dán w Teplici u Ferrarii, wečer matky boZie Annuntia-tionis, léta králowstwie našich Uherského sc. w XXVIL, a Římského w čtwrtém létě,

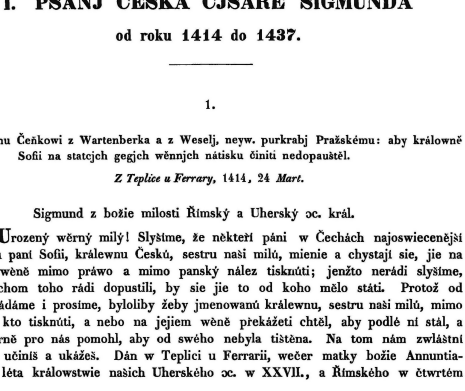
Ad mandatum D. Regis: Michael de Priest.

Urozenému Ceünkowi z Weselé
wérnému náàm zwlásté milému.

28% of changes improved the ground truth.

# Qualitative Evaluation IV

## Random Page (book 792, page 5)



I. PSANJ ČESKÁ CJSAŘE SIGMUNDA

od roku 1414 do 1437.

1.

Panu Čeňkowi z Wartenberka a z Weselj, neyw. purkrabj Pražskému: aby královně Sofii na statcjch gegjch wěnnjch nátisku činiti nedopauštěl.

*Z Teplice u Ferrary,* 1414, 24 *Mart.*

Sigmund z božie milosti Římský a Uherský oc. král.

Urozený wěrný milý! Slyšíme, že někteří páni w Čechách najoswiecenější kněžnu paní Sofii, králewnu Českú, sestru naši milú, mienie a chystají sie, jie na jejím wěně mimo práwo a mimo panský nález tisknúti; jenžto nerádi slyšíme, anižbychom toho rádi dopustili, by sie jie to od koho mělo státi. Protož od tebe žádáme i prosíme, bylo liby žeby jmenowanú králewnu, sestru naši milú, mimo práwo kto tisknúti, a nebo na jejiem wěně překážeti chtěl, aby podlé ní stál, a jie wěrně pro nás pomohl, aby od swého nebyla tištěna. Na tom nám zwláštní službu učiníš a ukážeš. Dán w Teplici u Ferrarii, wečer matky božie Annuntiationis, léta královstwie našich Uherského oc. w XXVII., a Římského w čtwrtém létě.

Ad mandatum D. Regis: Michael de Priest.

Urozenému Čeňkowi z Weselé
wěrnému nám zwláště milému.

I. PSANJ ČESKÁ CJSAŘE SIGMUNDA
od roku 1414 do 1437.

Panu Čeňkowi z Wartenberka a z Weselj, neyw. purkrabj Pražskému: aby královně Sofii na statcjch gegjch wěnnjch nátisku činiti nedopauštěl.

Z Teplice u Ferrary, 1414, 24 Mart.

Sigmund z božie milosti Římský a Uherský oc. král.

Urozeny wörny mily! Sly&tme, ze některí páni w Cechách najoswiecenéjsi kné£nu pani Sofii, králewnu Ceskü, sestru nasi milü, mienie a chystají sie, jie na jejím wéné mimo práwo a mimo pansky nález tisknüti; jenZto nerádi slysíme, aniZbychom toho rádi dopustili, by sie jie to od koho mélo státi. ProtoZ od tebe Zádáme i prosíme, bylo liby Zeby jmenowanü královnu, sestru naài milü, mimo práwo kto tisknüti, a nebo na jejiem wéné prekázeti chtél, aby podlé ni stál, a jie wérné pro nás pomohl, aby od swého nebyla tisténa. Na tom näm zwlästni sluZbu ucini$ a ukàáZes. Dáàn w Teplici u Ferrari, wecer matky bozie Annuntia-tionis, léta královstwie nasich Uherského oc. w XXVIL, a Rimskeho w ótwrtém lété,

Ad mandatum D. Regis: Michael de Priest.

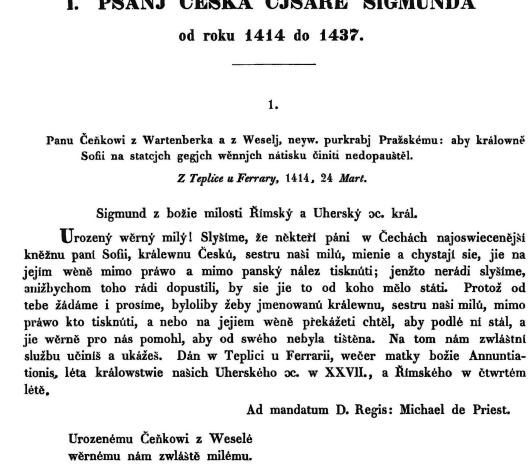
Urozenému Ceünkowi z Weselé
wérnému náàm zwlásté milému.

12% of changes improved the ground truth.

## Tesseract 3 + 4 (8.35% CER, 12.43% WER)

# Qualitative Evaluation V

## Random Page (book 792, page 5)



I. PSANJ ČESKÁ CJSAŘE SIGMUNDA

od roku 1414 do 1437.

1.

Panu Čeňkowi z Wartenberka a z Weselj, neyw. purkrabj Pražskému: aby králownĕ Sofii na statcjch gegjch wĕnnjch nátisku činiti nedopaušt̃l.

*Z Teplice u Ferrary,* 1414, 24 *Mart.*

Sigmund z božie milosti Římský a Uherský oc. král.

Urozený wĕrný milý! Slyšíme, že někteří páni w Čechách najoswiecenĕjší kněžnu paní Sofii, králewnu Českú, sestru naši milú, mienie a chystají sie, jie na jejím wĕně mimo práwo a mimo panský nález tisknúti; jenžto nerádi slyšíme, anižbychom toho rádi dopustili, by sie jie to od koho mělo státi. Protož od tebe žádáme i prosíme, bylo liby žeby jmenowanú králewnu, sestru naši milú, mimo práwo kto tisknúti, a nebo na jejiem wĕně překážeti chtĕl, aby podlé ní stál, a jie wĕrně pro nás pomohl, aby od svého nebyla tištĕna. Na tom nám zwlášt̃ní službu učiníš a ukážeš. Dán w Teplici u Ferrarii, wečer matky božie Annuntiationis, léta králowstwie našich Uherského oc. w XXVII., a Římského w čtwrtém létĕ.

Ad mandatum D. Regis: Michael de Priest.

Urozenému Čeňkowi z Weselé
wĕrnému nám zwlášt̃ě milému.

## OCR-D (13.95% CER, 27.17% WER)

J. PaAN SINA
od roku 1414 do 1437.

Panu Cekowi 2 artenberka a 2 Veselj, neyw. purkrabj Praiskému: aby krâlown Sofii na statejch gegjeh wènnjeh nâtisku öiniti nedopaustẽl.

Teplice u Ferrar, 1414, 24 Mart.

Sigmund z boie milosti Rimsky a Uhersk oc. krâl.

rozeny wörny milyl Slysime, e nekteri pâni w Cechâch najoswiecenèjsi knênu pani Sofii, kralewnu Cesk, sestru nasi milu, mienie a chystaji sie, jie na jejim wenè mimo prâwo a mimo pansky nalez tisliti; jenàto nerâdi slysime, anizbychom toho radi dopustili, by sie jie to od koho mêlo stâti. Proto od tebe adame i prosime, bylоliby eby jmenowanu kralewnu, sestru nasi milu, mimo prâwo kto tiskniti, a nebo na jejiem wènè prekâieti chtél, aby podlé ni stal, a jie wvrnẽ pro nas pomohl, aby od svého nebyla tistêna. Na tom nâm zwlastni sluibu ucinis a ukâes. Dân w Teplici u Ferrarii, wecer matky boie Annuntia-tionis, léta krâlowstwie nasich Uherského oe. w XXVJ., a Rimského w êtwrtém létè.

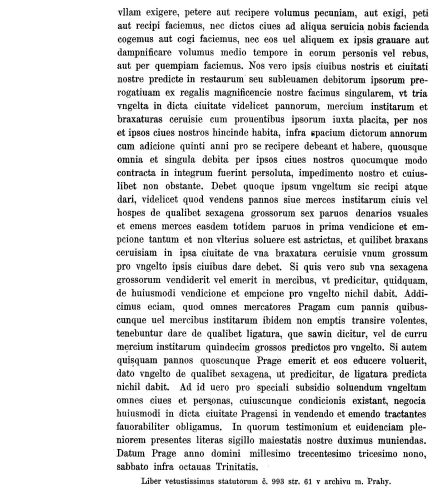
Ad mandatum D. Regis: Michael de Priest.

Urozenému Ceükowi 2 Veselé
wêrnému nam zlastè milému.

5% of changes improved the ground truth.

# Qualitative Evaluation VI

## Best WER w/ OCR-D (book 117, page 230)



vllam exigere, petere aut recipere volumus pecuniam, aut exigi, peti aut recipi faciemus, nec dictos ciues ad aliqua seruicia nobis facienda cogemus aut cogi faciemus, nec eos uel aliquem ex ipsis grauare aut dampnificare volumus medio tempore in eorum personis vel rebus, aut per quempiam faciemus. Nos vero ipsis ciuibus nostris et ciuitati nostre predicte in restaurum seu subleuamen debitorum ipsorum prerogatiuam ex regalis magnificencie nostre facimus singularem, vt tria vngelta in dicta ciuitate videlicet pannorum, mercium institarum et braxaturas ceruisie cum prouentibus ipsorum iuxta placita, per nos et ipsos ciues nostros hincinde habita, infra spacium dictorum annorum cum adicione quinti anni pro se recipere debeant et habere, quousque omnia et singula debita per ipsos ciues nostros quocumque modo contracta in integrum fuerint persoluta, impedimento nostro et cuius-libet non obstante. Debet quoque ipsum vngeltum sic recipi atque dari, videlicet quod vendens pannos siue merces institarum ciuis vel hospes de qualibet sexagena grossorum sex paruos denarios vsuales et emens merces easdem totidem paruos in prima vendicione et empcione tantum et non vlterius soluere est astrictus, et quilibet braxans ceruisiam in ipsa ciuitate de vna braxatura ceruisie vnum grossum pro vngelto ipsis ciuibus dare debet. Si quis vero sub vna sexagena grossorum vendiderit vel emerit in mercibus, vt predicitur, quidquam, de huiusmodi vendicione et empcione pro vngelto nichil dabit. Addicimus eciam, quod omnes mercatores Pragam cum pannis quibuscunque uel mercibus institarum ibidem non emptis transire volentes, tenebuntur dare de qualibet ligatura, que sawin dicitur, vel de curru mercium institarum quindecim grossos predictos pro vngelto. Si autem quisquam pannos quoscunque Prage emerit et eos educere voluerit, dato vngelto de qualibet sexagena, ut predicitur, de ligatura predicta nichil dabit. Ad id uero pro speciali subsidio soluendum vngeltum omnes ciues et personas, cuiuscunque condicionis existant, negocia huiusmodi in dicta ciuitate Pragensi in vendendo et emendo tractantes fauorabiliter obligamus. In quorum testimonium et euidenciam pleniorem presentes literas sigillo maiestatis nostre duximus muniendas. Datum Prage anno domini millesimo trecentesimo tricesimo nono, sabbato infra octauas Trinitatis.

Liber vetustissimus statutorum č. 993 str. 61 v archivu m. Prahy.

vllam exigere, petere aut recipere volumus pecuniam, aut exigi, peti
…
vngelta in dicta ciuitate videlicet pannorum, mercium institarum et braxaturas ceruisie cum prouentibus ipsorum iuxta placita, per nos et ipsos ciues nostros hincinde habita, infra spacium dictorum annorum cum adicione quinti anni pro se recipere debeant et habere, quousque omnia et singula debita per ipsos ciues nostros quocumque modo contracta in integrum fuerint persoluta, impedimento nostro et cuius- libet non obstante. Debet quoque ipsum vngeltum sic recipi atque dari, videlicet quod vendens pannos siue merces institarum ciuis vel hospes de qualibet sexagena grossorum Sex paruos denarios vsuales et emens merces easdem totidem paruos in prima vendicione et em- pcione tantum et non vlterius soluere est astrictus, et quilibet braxans ceruisiam in ipsa ciuitate de vna braxatura ceruisie vnum grossum pro vngelto ipsis ciuibus dare debet. Si quis vero sub vna sexagena grossorum vendiderit vel emerit in mercibus, vt predicitur, quidquam, de huiusmodi vendicione et empcione pro vngelto nichil dabit. Addicimus eciam, quod omnes mercatores Pragam cum pannis quibuscunque uel mercibus institarum ibidem non emptis transire volentes,
…
Liber vetustissimus statutorum c. 993 str. 61 V archivu m. Prahy.

## OCR-D (0.29% CER, 0.00% WER)

0% of changes improved the ground truth.

# Qualitative Evaluation VII

## Worst WER w/ Tesseract 4 (book 118, page 1327), Tesseract 4 (81.56% CER, 100% WER)

1085, 1133, 1149, 1150; purkrabí 1077.
*ze Žebráka* Zajícové: Oldřich 339, Zbyněk 228, 339.
*Žebrání* žáků 1122.
*Želetice* 1127.
*Železo*, ferrum 487, 583, 584, 590—592, 733, 816, 942; bavorské 599, chebské 599, kadaňské 599; ruční, ferrum manuale 18, 68, 79.
*Želivo*, Zelew, klášter 901.
*Žena obecní*, mulier communis, gemaine frau 20.
*Žernoseky*, Zrnoseky, Srnossek 422, 423.
*Žertéři*, truffatores, teuscher 13.
*Žestoky*, Zestok, 1013—1015.
*Žháři*, paliči, incendiarii, incensores, mortprenner 19, 20, 64, 75, 336, 364, 365, 1142, 1144.
*Žhářství*, zapálení, pálení, oheň, incendium, ignis, fewer, brant, mortprant, 19, 20, 66, 77, 181, 182, 209, 210, 688, 692, 753, 754, 1046, 1141, 1143; noční, nocturna incendia 186, 187; úkladné, secretum seu mortiferum incendium 240, 241, 601, 640.
*Žídé* v Čechách, služebníci komory královské, Judei Boemie, regie camere servi, die Juden in dem lande zu Behem gesessen, camerknechte 20, 26, 161, 328, 329, 346—348, 352, 437, 438, 559, 597, 767, 768, 821, 871, 872, 937, 946; v Chebsku, die Juden zu Eger 200, 201, 390, 443, 444, 708, 721, 809, 810, 837, 866.
*Žirotín* hrad, 1003, 1163.
*Žitava*, Sittavia 189, 410, 411, 538.
*Žitavský* Jan, Sittauwer, přísežný na Horách Kutných 443.
*ze Žitavy*, Sittawia, Častalov 6, Chval 41, Jindřich 6, 41.
*Žitenice* u Litoměřic 462.
*Žito*, siligo 216, 305, 307, 792.
*Žitětice*, Zewslicz, ves 939.
*Žlěby* 162, 163, 165, 246—248.
*Žlutice*, oppidum Luticz, Lueticz 487, 615, 616.
*Žofie*, Žofka, královna Česká 928, 930, 972, 1106, 1137, 1138, 1152, 1153, 1182.
*Žoky* chmele, sacci humuli, 484, 485, solní, sacci 483.
*Žoldnéři* městští, stipendiarii 1114, 1116.
*Žumburk*, Sumirburch, Sumerburc 48 49, 107, 108, 1185.

…

ze Žebráka Zajícové: Oldřich 339, Zbyněk 228, 339.

Žebrání žáků 1122.

Želetice 1127.

Železo, ferrum 487, 583, 584, 590—592, 783, 816, 942; bavorské 599, chebské 599, kadaüské 599; ruční, ferrum manuale 18, 68, 79.

…

Židé v Čechách, služebníci komory krá-lovské, Judei Boemie, regie camere servi, die Juden in dem lande zu Behem gesessen, camerknechte 20, 26, 161, 328, 329, 346—348, 852, 437, 438, 559, 597, 167, 768, 821, 871, 872, 987, 946; v Chebsku, die Juden zu Eger 200, 201, 390, 448, 444, 708, 721, 809, 810, 837, 866.

…

Ziutice, oppidum Luticz, Lueticz 487, 615, 616.

Žofie, Žofka, královna Česká 928, 930, 972, 1106, 1137, 1138, 1152, 1153, 1182.

Zoky chmele, sacci humuli, 484, 485, solní, sacci 483.

Zoldnéři městští, stipendiarii 1114, 1116.

Zumburk, Sumirburch, Sumerburc 48 49, 107, 108, 1185.

100% of changes improved the ground truth.

# Future Work and Conclusion

- Qualitative evaluation suggests Google Cloud Vision AI is a dubious ground truth.
- Tesseract 4 is the second fastest, the most accurate, and detects language.
- OCR-D + Calamari is:
  1. comparable to Tesseract 4 in speed,
  2. likely more accurate than Tesseract 4, but
  3. undermined by poor pre-trained models,
  4. without language detection.
- OCR-D can align outputs from OCR engines (Calamari + Tesseract 4): holy grail?

# Bibliography I

[1] Ray Smith. “An overview of the Tesseract OCR engine”. In: *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Vol. 2. IEEE. 2007, pp. 629–633.

[2] Ray Smith, Daria Antonova, and Dar-Shyang Lee. “Adapting the Tesseract open source OCR engine for multilingual OCR”. In: *Proceedings of the International Workshop on Multilingual OCR*. 2009, pp. 1–8.

[3] Christoph Wick, Christian Reul, and Frank Puppe. “Calamari-a high-performance tensorflow-based deep learning package for optical character recognition”. In: *arXiv preprint arXiv:1807.02004* (2018).

[4] Christian Reul et al. “State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines”. In: *arXiv preprint arXiv:1810.03436* (2018).

# Bibliography II

[5] Clemens Neudecker et al. “OCR-D: An end-to-end open source OCR framework for historical printed documents”. In: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*. 2019, pp. 53–58.

[6] R William Soukoreff and I Scott MacKenzie. “Measuring errors in text entry tasks: An application of the Levenshtein string distance statistic”. In: *CHI'01 extended abstracts on Human factors in computing systems*. 2001, pp. 319–320.

MUNI
FACULTY
OF INFORMATICS