Open Source Handwritten Text Recognition on Medieval Manuscripts using Mixed Models and Document-Specific Finetuning

Christian Reul[1], Stefan Tomasek[1], Florian Langhanki[1], Uwe Springmann[2]

[1]University of Würzburg
[2]LMU Munich

25.05.2022

Introduction – Model Types and Where to Use Them

- Different approaches and models for different use cases
- Model types
  - Mixed models (MM): trained on a variety of books / manuscripts / printing types / hands / …
  - Document-specific models: trained specifically for the recognition of a single book / …
- More specialized models yield better results but their creation requires more effort
- Example use-cases
  - Mass text recognition, e.g. in libraries or archives
    - Manual correction not feasible
    - Apply MM and hope for the best
  - Preparation of a critical edition
    - 100% error-free text required → high amount of manual correction inevitable
    - Recognition → correction → training → recognition → … until everything is transcribed

  → Gray area in-between

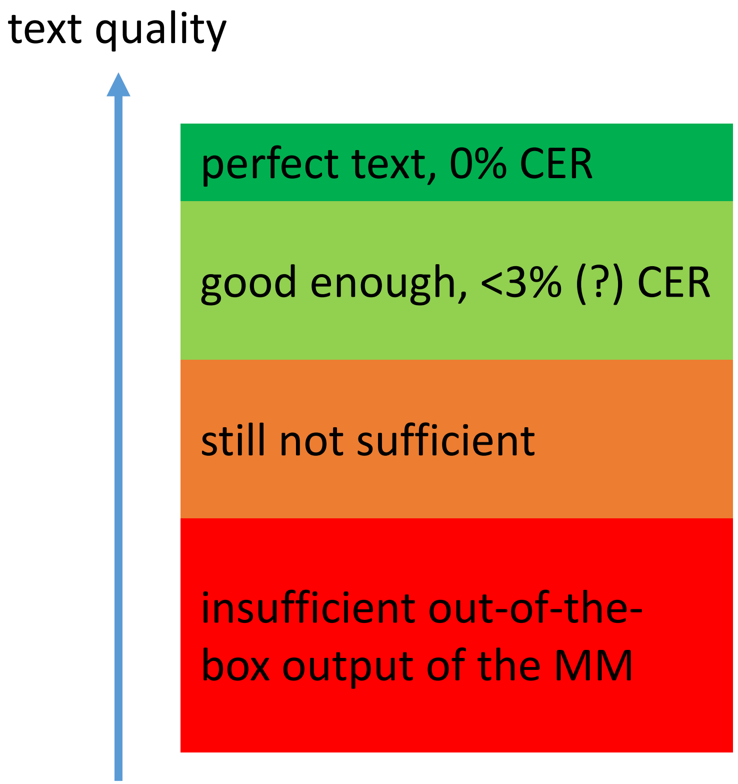
Use-Case – Overview

Digital edition project „Die Kindheit Jesu“ (Childhood of Jesus Christ) at the University of Würzburg

- Material: Medieval manuscripts written …
  - between 1250 and 1350 in Middle High German
  - using two very common German medieval writing styles: Gothic and Bastarda
- Goal: Comparing text variants using a configurable filter system based on different distance measures
- Some distance measures, e.g. Word Mover‘s Distances, can and have to be trained
- Training data needed: Non-normalized text in …
  - (very) good but not perfect quality
  - large quantities → efficiency is key → combining mixed models and document-specific finetuning

Use-Case – Task and Workflow

Task: Get from the out-of-the-box output of the mixed model to a good enough CER in a very efficient manner

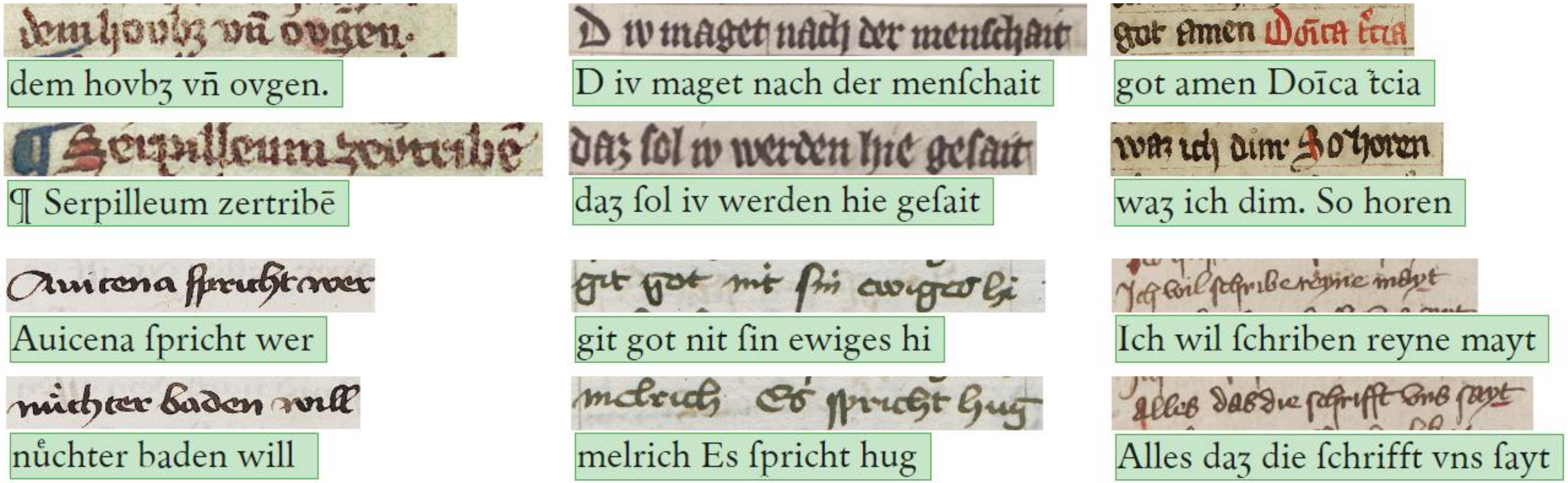


Workflow:

- Train suitable mixed model(s)
- For each manuscript:
  1. Apply the most suitable MM
  2. Manually correct some pages
  3. Train a document-specific model using the newly gained GT
  4. Apply the new model
  5. Repeat steps 2-4, if needed (Iterative Training Approach)
  6. Apply the document-specific model to the remaining pages

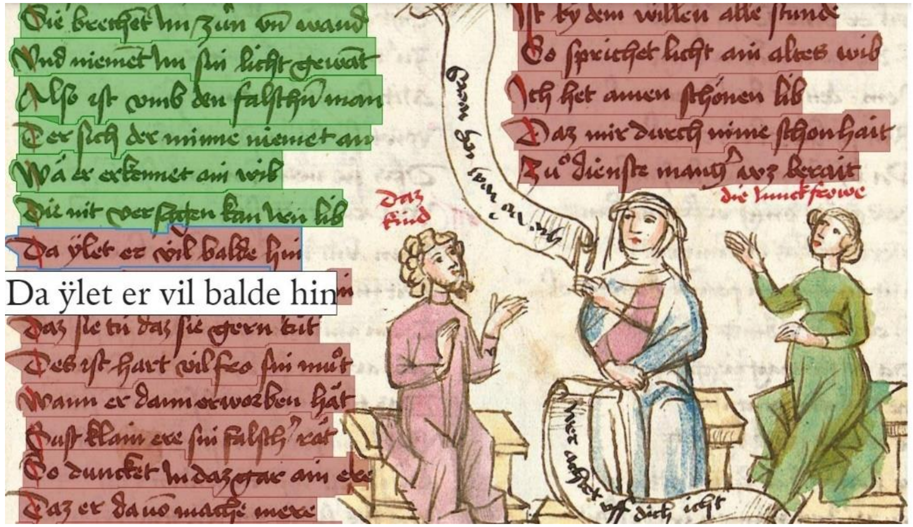
Data – Mixed Model Training

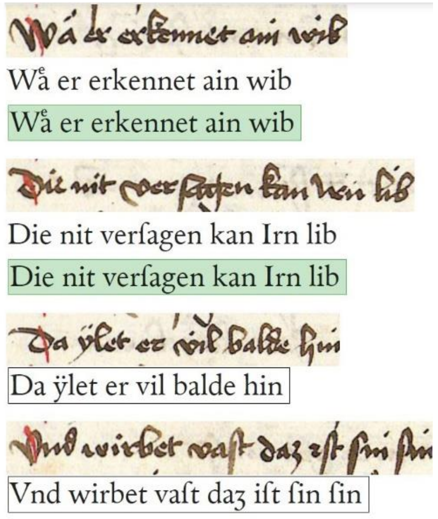
- Suitable manuscripts from ca. 1200 to 1500: Kindheit Jesu, Marienleben, Parzival, …
- Self-created or compiled from freely available sources
- Final corpus: 35 works (at least 30 different hands), close to 300 pages, ca. 12.5k lines



Data – Preparation

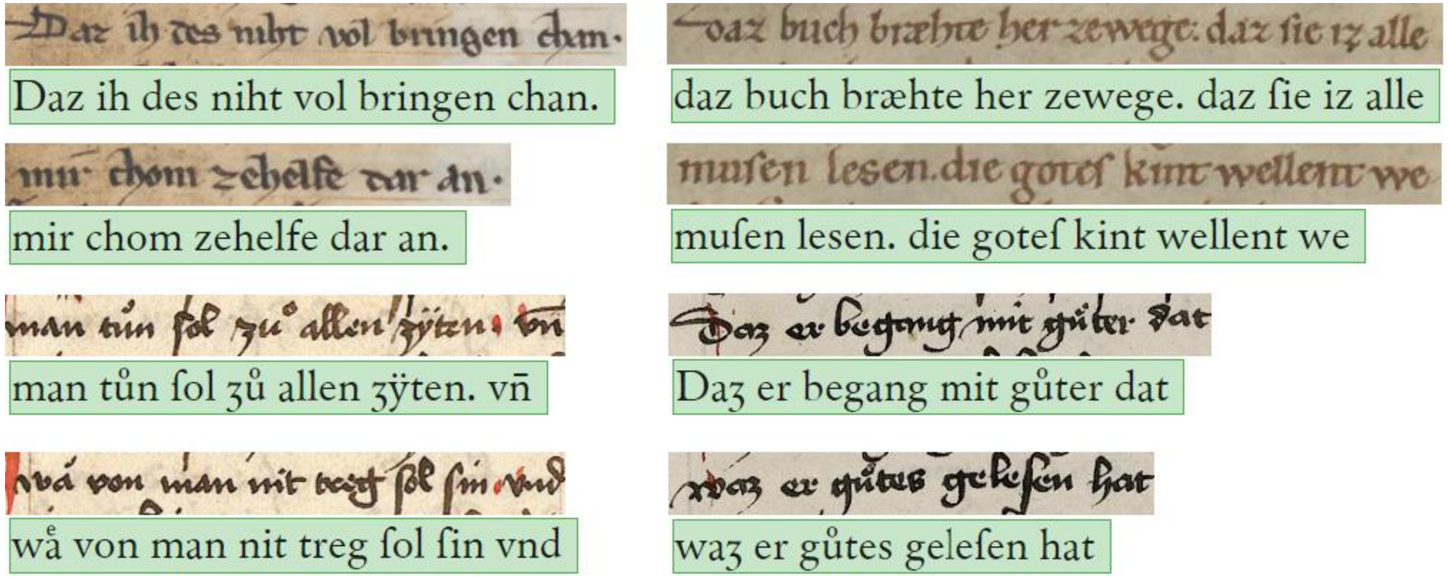
- Fully open source workflow!
- Preprocessing via OCR-D (https://github.com/OCR-D) processors
- OCR4all and LAREX
  - Segmentation and ground truth production
  - https://github.com/OCR4all
- Calamari
  - Recognition, training, evaluation
  - https://github.com/Calamari-OCR



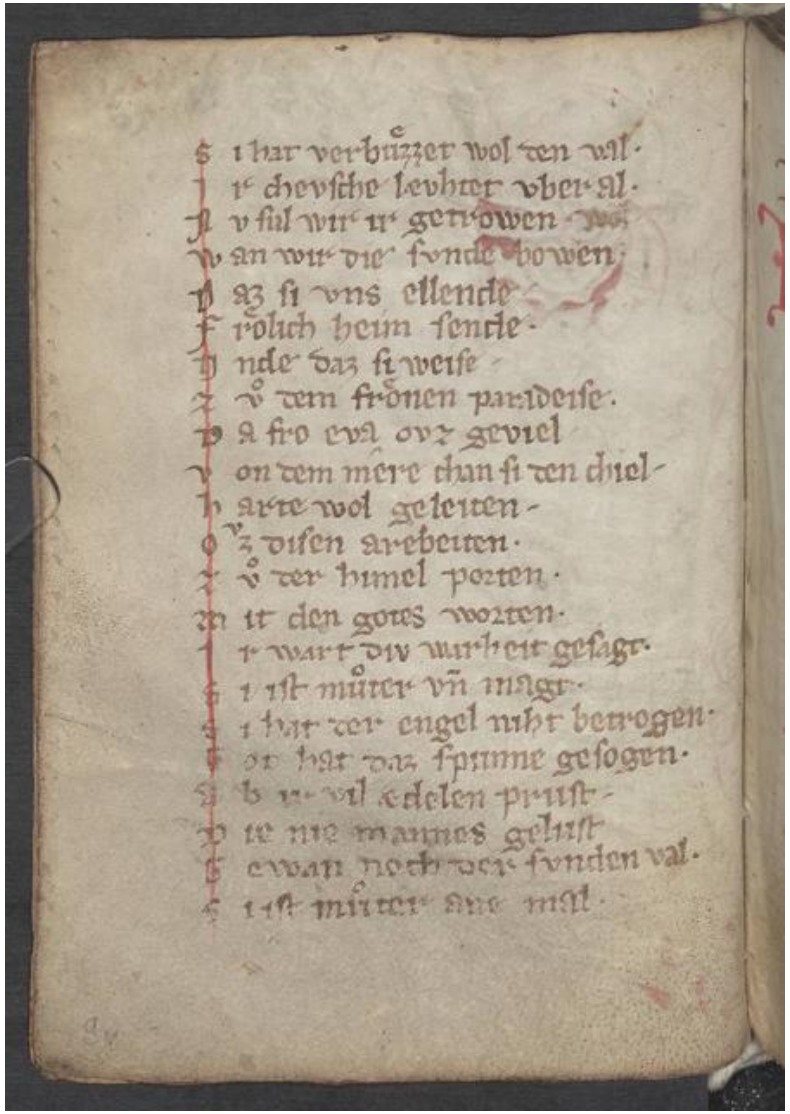


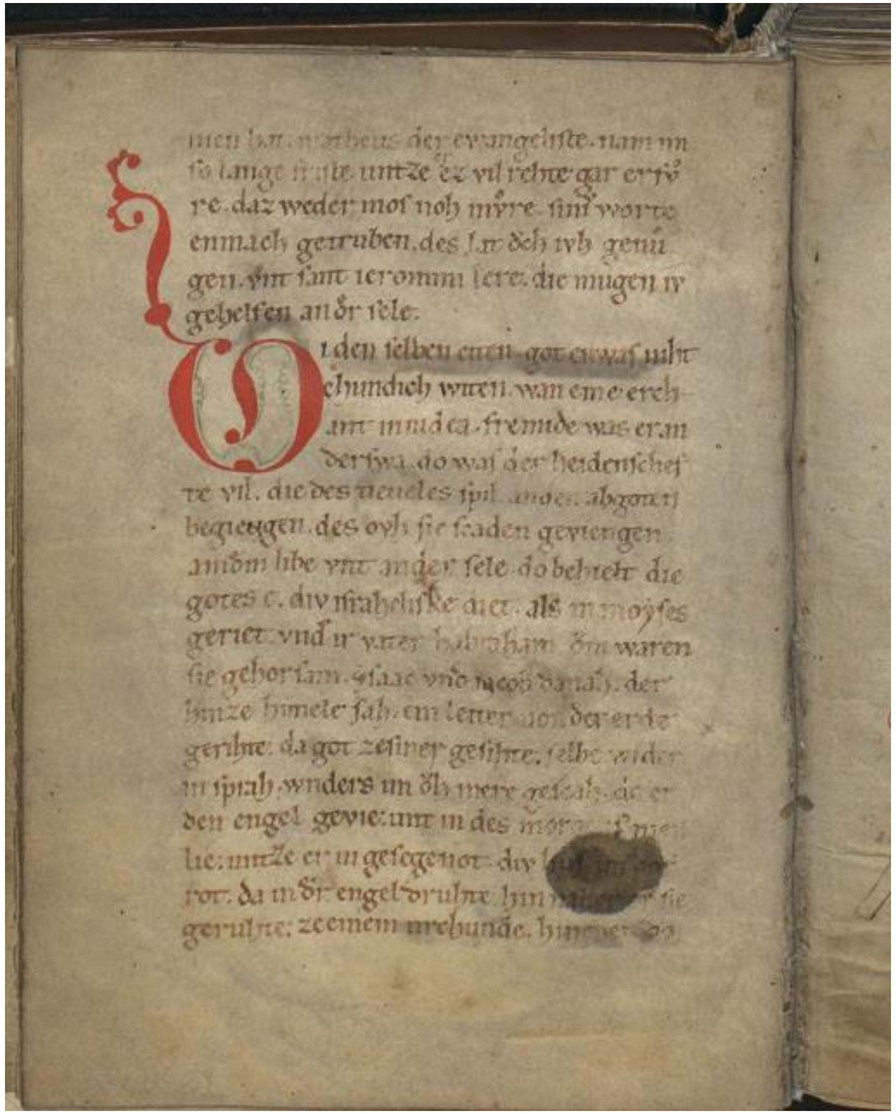
Data – Evaluation

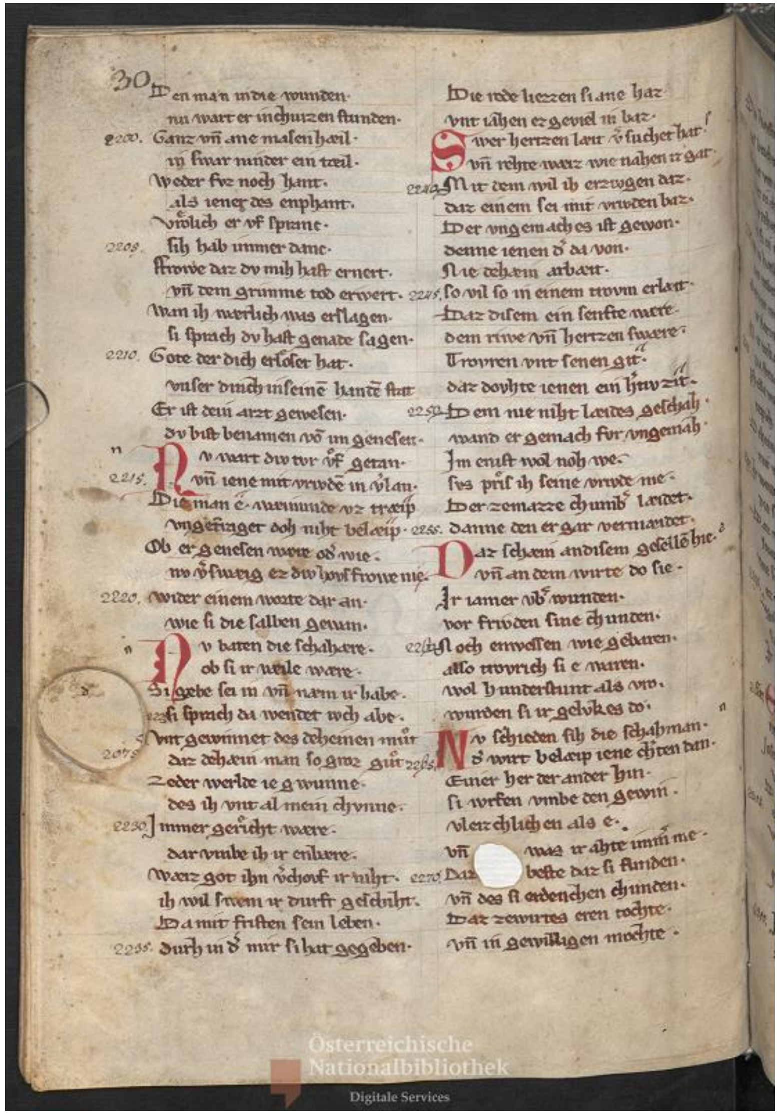
- Four manuscripts (2x Gothic, 2x Bastarda) written by previously unseen hands
- Kindheit Jesu, Marienleben, Der Welsche Gast (x2)
- Each manuscript further divided into train (32 pages) and eval (8-18 pages) subsets



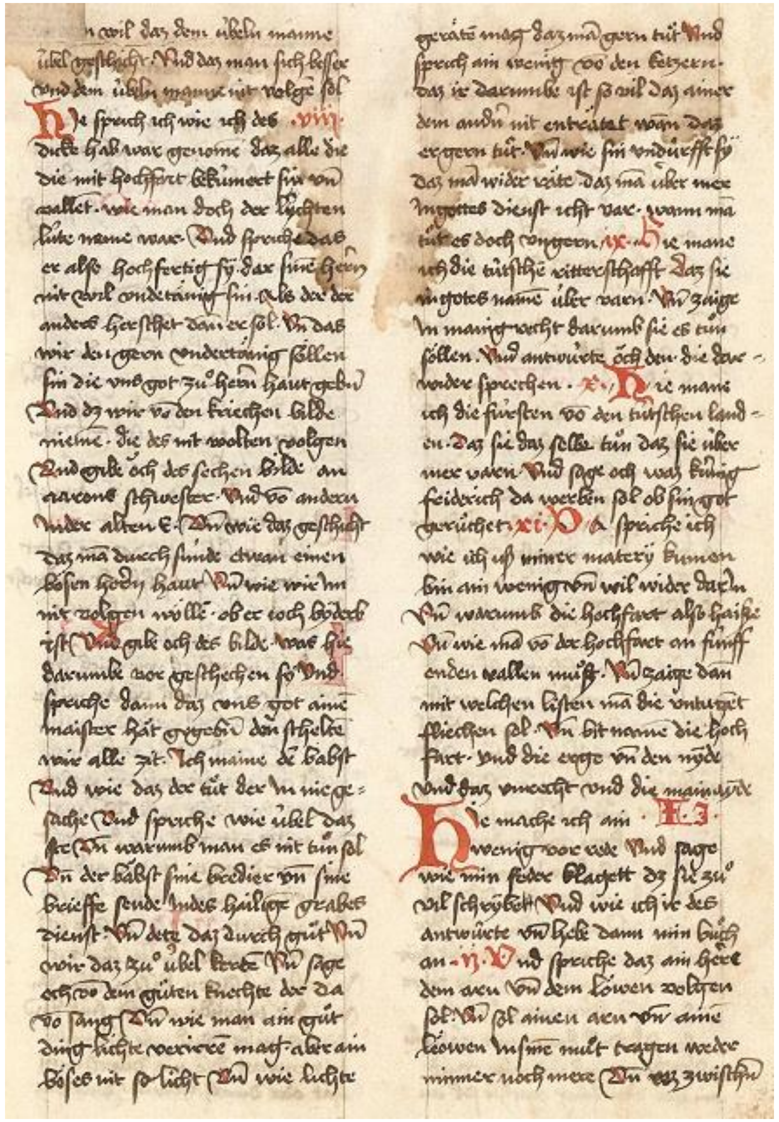
Gothic

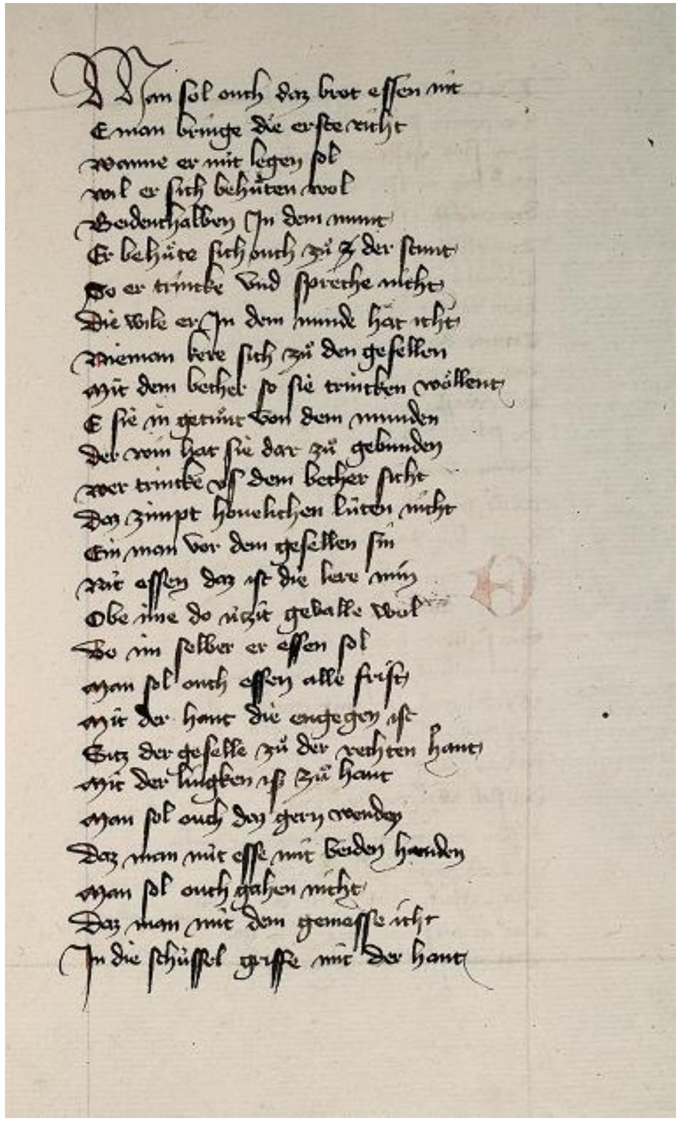






Bastarda





Experiments – General Methodology

- Using Calamari version 2.1 and mainly its default parameters / best practices
  - Five models per voting ensemble
  - Five augmentations per original line
  - Early stopping after five epochs without improvement
  - …
- Evaluation
  - Apply models to evaluation data
  - Standardize the output analogously to the training data
  - Calculate the CER using the Calamari evaluation script

Experiments – Training Pipeline and Preliminary Experiments

- Training pipeline to successively refine the models

  0: Starting point: LSH-4 (strong MM for material printed between 1450 and 1900)

  1: Utilize the entire training set (Gothic + Bastarda)

  2: Train two writing-style-specific models by exclusively using Gothic/Bastarda data

- Summary of preliminary experiments
  - Application of in-domain MMs (Gothic → Gothic and Bastarda → Bastarda) yielded best results
    - Out-of-the-box application: very clearly
    - Starting point for finetuning: in-domain > combined > out-of-domain >>> starting from scratch

    → used in-domain MMs for all experiments
  - Calamari predefined network structure *deep3* performed best → used for all experiments
    (cnn=40:3x3,pool=2x2,cnn=60:3x3,pool=2x2,cnn=120:3x3,lstm=200,lstm=200,lstm=200,dropout=0.5)

Experiments – Simulating an Iterative Training Approach

- Iteratively doubling the amount of training data
  - 2, 4, 8, 16, 32 pages (4 subsumes 2, …)
  - Starting point: the respective most suitable MM
  - ootb: out-of-the-box result of the MM
- Evaluating on a fixed evaluation set and calculating the CER; averaged over all manuscripts
- Decent ootb CER considering the material and MMs
- Training on just two pages cuts the CER in half
- CER of 2-3% already completely sufficient for our use-case, many downstream tasks, and efficient further transcription

# pages	CER in %	Impr. in %
ootb	6.22	-
2	3.27	48
4	2.58	21
8	2.17	16
16	1.94	11
32	1.65	15

Example Results After Finetuning on Just Two Pages

uon ſo heiliclichem ſite. gab im got ſinen ſe

uon ſo heiliclichem ſite. gab im got ſinen ſe

gen. himeltov und regen. daz ez michel ge-

gen. himeltov und regen. daz ez michel ge-

nuht gewan. ſwaz er buwen began. an ſwiv

nuht gewan. ſwaz er buwen began. an ſwiv

ſich der herre uerlie. mit heil ez fur ſih gie.

ſich der herre uerlie. mit heil ez fur ſih gie

v̄ ergab im ſælicliche. er was alſo vihe riche.

v̄ ergab im ſælicliche. er was alſo vihe riche.

Er erwirbet amit rům minur daȝ

Er erwirbet mit rům nur daȝ

Daȝ ain ieglich man daȝ gicht

Daȝ ain ieglich man daȝ gicht

Daȝ ſie es tet durch ain bo̊ſwicht

Daȝ ſie es tet durch ain bo̊ſwicht

Des iſt ſie wol ain bo̊ſes wib

Des iſt ſie wol ain bo̊ſes wib

Daȝ ſie geleſtert hab Imrn lib

Daȝ ſie geleſtert hab Irn lib

line image
diff view
ground truth

diff view shows deletions and insertions necessary to turn the prediction into the GT

Example:
model predicted „m“,
correct would have been „rn“

Experiments – Effect of Mixed Models

- Using MMs as starting points compared to training from scratch
- Unexpectedly vast improvements …
  - when using small amounts of GT (85% for 2 pages)
  - but also for large training sets (39% for 32 pages)
- Suitable MM not always available or not strong enough but …
- … even starting from an out-of-domain MM yields large improvements (>50%) over starting from scratch

# pages	without MM CER in %	with MM CER in %	Impr. in %
2	21.12	3.27	85
4	10.74	2.58	76
8	5.82	2.17	63
16	3.78	1.94	49
32	2.72	1.65	39

Conclusion

- Using a fully open source workflow we …
  - created a training corpus covering two widely used German medieval writing styles
  - trained two performant and freely available MMs
- Document-specific finetuning quickly led to very low error rates
  - 2-4 pages of training material already enough to reach our goal of <3% CER in most cases
  - Use of suitable MM as a starting point improved effectiveness and efficiency considerably … but even a less suited MM is better than nothing
- Future Work
  - Improving our MM as a side-effect of our data collection
  - Opening up new domains by applying the same procedure
  - Automatic quality estimation using Calamari’s intrinsic confidence values