A Comprehensive Comparison of Open-Source Libraries for Handwritten Text Recognition in Norwegian

Martin Maarand, Yngvil Beyer, Andre Kåsen, Knut T. Fosseide and Christopher Kermorvant

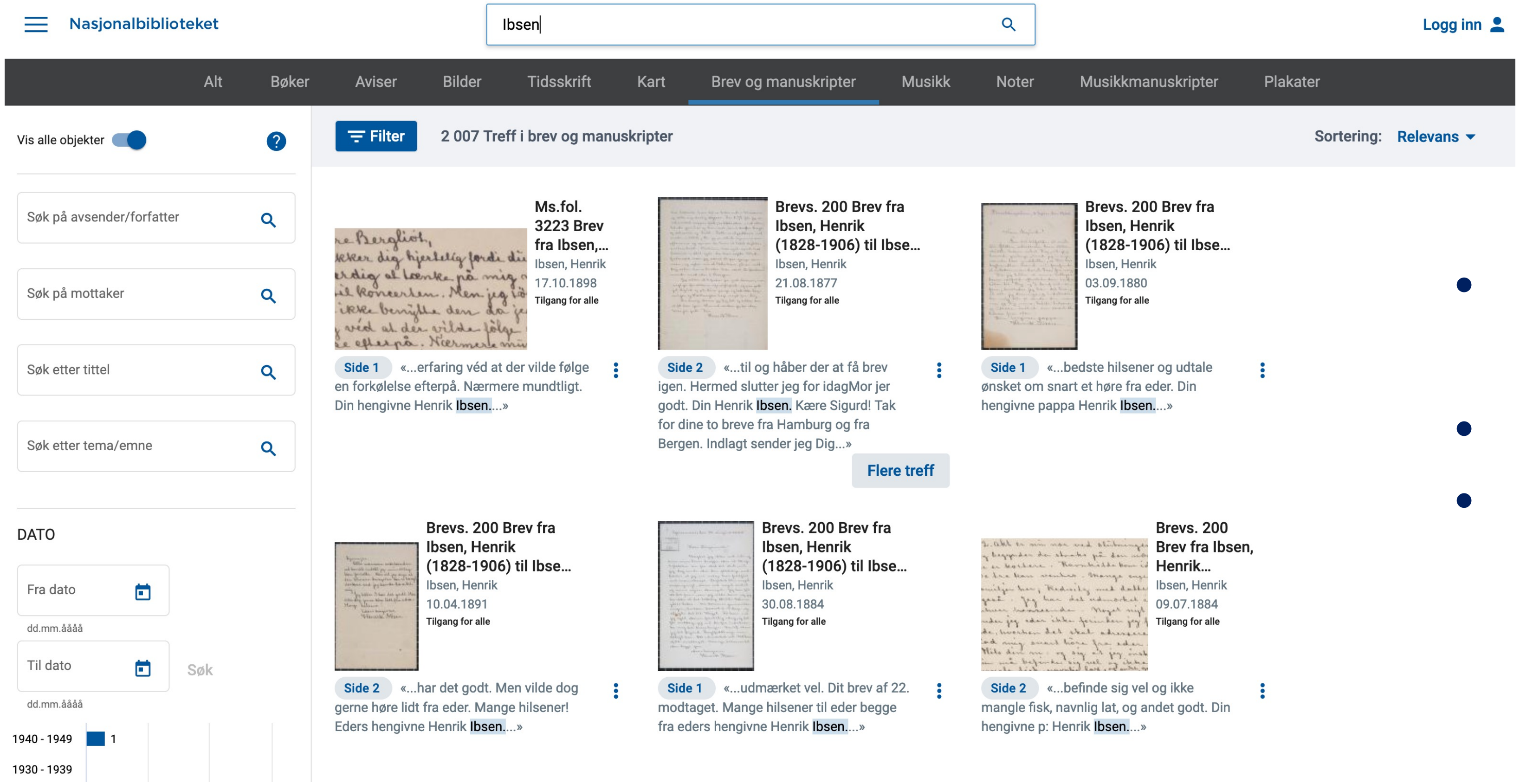
DAS 2022 – La Rochelle

TEKLIA

Nasjonalbiblioteket

LUMEX

Searching the manuscripts at Nasjonalbiblioteket



Ibsen

2 007 Treff i brev og manuskripter

Ms.fol. 3223 Brev fra Ibsen,...
Ibsen, Henrik
17.10.1898
Tilgang for alle
Side 1 «...erfaring véd at der vilde følge en forkølelse efterpå. Nærmere mundtligt. Din hengivne Henrik Ibsen....»

Brevs. 200 Brev fra Ibsen, Henrik (1828-1906) til Ibse...
Ibsen, Henrik
21.08.1877
Tilgang for alle
Side 2 «...til og håber der at få brev igen. Hermed slutter jeg for idagMor jer godt. Din Henrik Ibsen. Kære Sigurd! Tak for dine to breve fra Hamburg og fra Bergen. Indlagt sender jeg Dig...»

Brevs. 200 Brev fra Ibsen, Henrik (1828-1906) til Ibse...
Ibsen, Henrik
03.09.1880
Tilgang for alle
Side 1 «...bedste hilsener og udtale ønsket om snart et høre fra eder. Din hengivne pappa Henrik Ibsen....»

Brevs. 200 Brev fra Ibsen, Henrik (1828-1906) til Ibse...
Ibsen, Henrik
10.04.1891
Tilgang for alle
Side 2 «...har det godt. Men vilde dog gerne høre lidt fra eder. Mange hilsener! Eders hengivne Henrik Ibsen....»

Brevs. 200 Brev fra Ibsen, Henrik (1828-1906) til Ibse...
Ibsen, Henrik
30.08.1884
Tilgang for alle
Side 1 «...udmærket vel. Dit brev af 22. modtaget. Mange hilsener til eder begge fra eders hengivne Henrik Ibsen....»

Brevs. 200 Brev fra Ibsen, Henrik...
Ibsen, Henrik
09.07.1884
Tilgang for alle
Side 2 «...befinde sig vel og ikke mangle fisk, navnlig lat, og andet godt. Din hengivne p: Henrik Ibsen....»

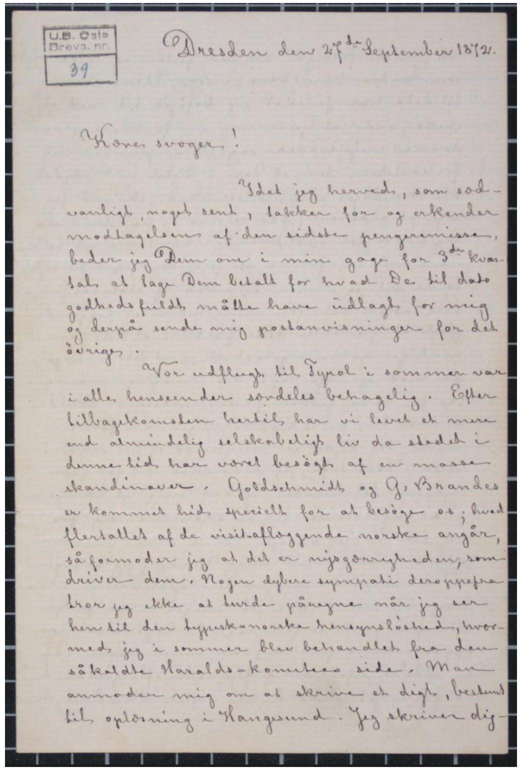
- Searching in meta-data and full text
- Provide faceted search
- Index persons, places, time, etc

HTR at the Nasjonalbiblioteket

Objectives:

- Include handwriting recognition in the standard digitization process
- Use open-source software for document processing
- Produce resources for HTR in Norwegian
- Develop and formalize best practices for HTR

The NorHand Dataset

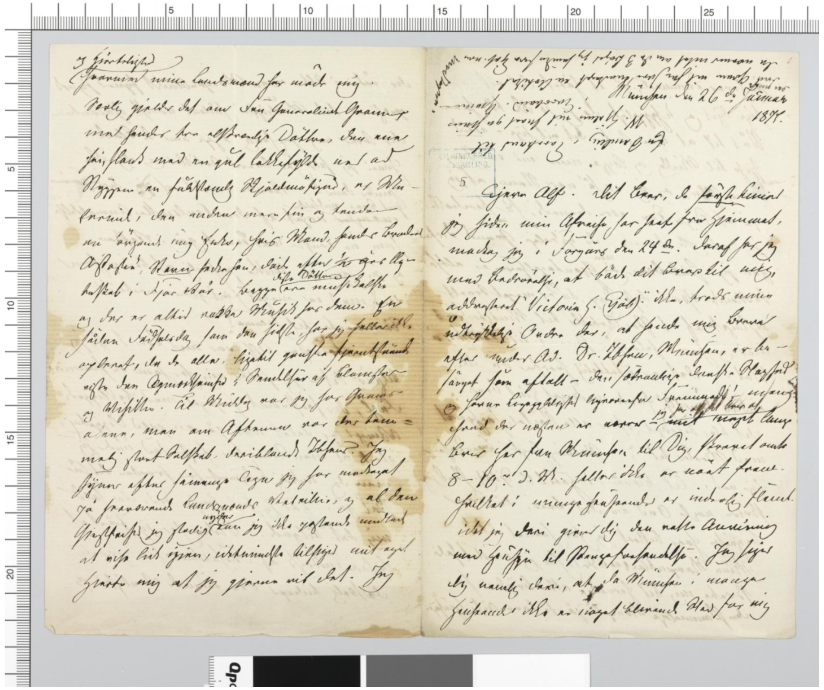


Dresden den 27de September 1872.

Kære svoger!

Idet jeg herved, som sædvanligt, noget sent, takker for og erkender modtagelsen af den sidste pengeremisse, beder jeg Dem om i min gage for 3die kvartal at tage Dem betalt for hvad De til dato godhedsfuldt måtte have udlagt for mig og derpå sende mig postanvisninger for det övrige.

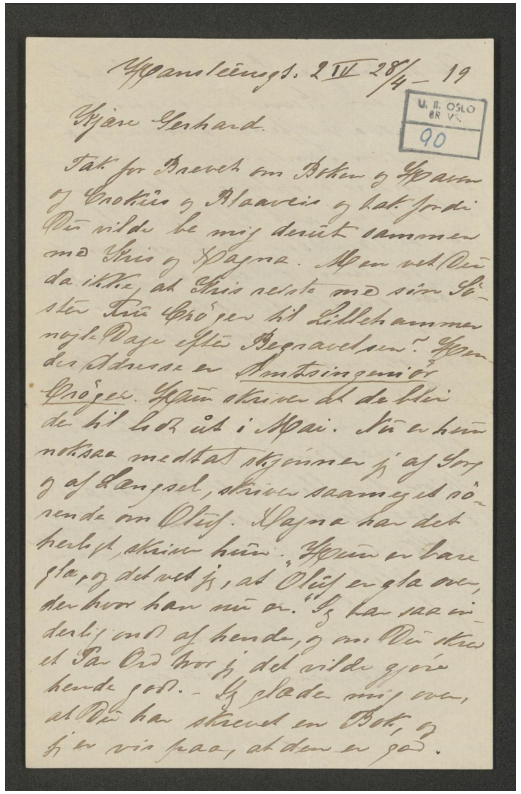
Vor udflugt til Tyrol i sommer var i alle henseender sördeles behagelig. Efter tilbagekomsten hertil har vi levet et mere end almindelig selskabeligt liv da stedet i denne tid har været besögt af en masse skandinaver. Goldschmidt og G. Brandes er kommet hid specielt for at besöge os; hvad flertallet af de visit-afleggende norske angår, så formoder jeg at det er nysgjerrigheden, som driver dem. Nogen dybere sympati deroppefra tror jeg ikke at turde påregne når jeg ser hen til den typisk=norske hensynslöshed, hvormed jeg i sommer blev behandlet fra den såkaldte Haralds-komitees side. Man anmoder mig om at skrive et digt, bestemt til oplösning i Haugesund. Jeg skriver dig-



I Hjertelighed
Sammen min Landsmand har mødt mig.
Især gjælder det om den Generalinde [illegible] med hendes tvo elskværdige Døttre, den ene [illegible] med en gul [illegible] [illegible] en fuldkommen [illegible], et [illegible], den andre munter fin og [illegible] en [illegible] ung Frue, hvis Mand, hendes [illegible], [illegible], død efter et [illegible] i Fjor [illegible]. Begge disse Døttre ere musikalske og der er altid nokke Musik hos dem. En [illegible] som den sidste, har jeg [illegible] oplevet, der de alle, tilsidst ganske [illegible] efter den [illegible] i Familien af [illegible] og [illegible]. Til Middag var jeg hos [illegible] alene, men om Aftenen var der kun = meget stort Selskab. Iblandt Ibsens. Jeg [illegible] jeg har [illegible] [illegible] Landsmands [illegible], og at den [illegible] jeg [illegible] ægte kan jeg ikke [illegible] at vise lidt igjen, [illegible] [illegible] mig at jeg gjerne vil det. Jeg

München den 26de Januar 1877.

Kjære Alf. Dit Brev, der første [illegible] jeg siden min Afreise har faaet fra Hjemmet, modtog jeg i Forgaars den 24de. Deraf ser jeg med Bedrøvelse, at baade dit Brev til mig, adresseret "Victoria H. (Hôtel)" ikke, trods mine udtrykkelige Ordre der, at sende mig Breve efter under Ad. Dr. Ibsen, München, er besørget som aftalt – den [illegible] [illegible] som [illegible] [illegible] [illegible] der [illegible] – [illegible] meget lange Brev har den München til Dig, [illegible] 8–10de d. M. [illegible] ikke er naaet frem. [illegible] i mange [illegible] er [illegible] idet jeg dari giver dig den [illegible] Anvisning med Hensyn til Pengeforsendelse. Jeg siger dig nemlig dari, at da München i mange Henseender ikke er noget [illegible] Sted for mig



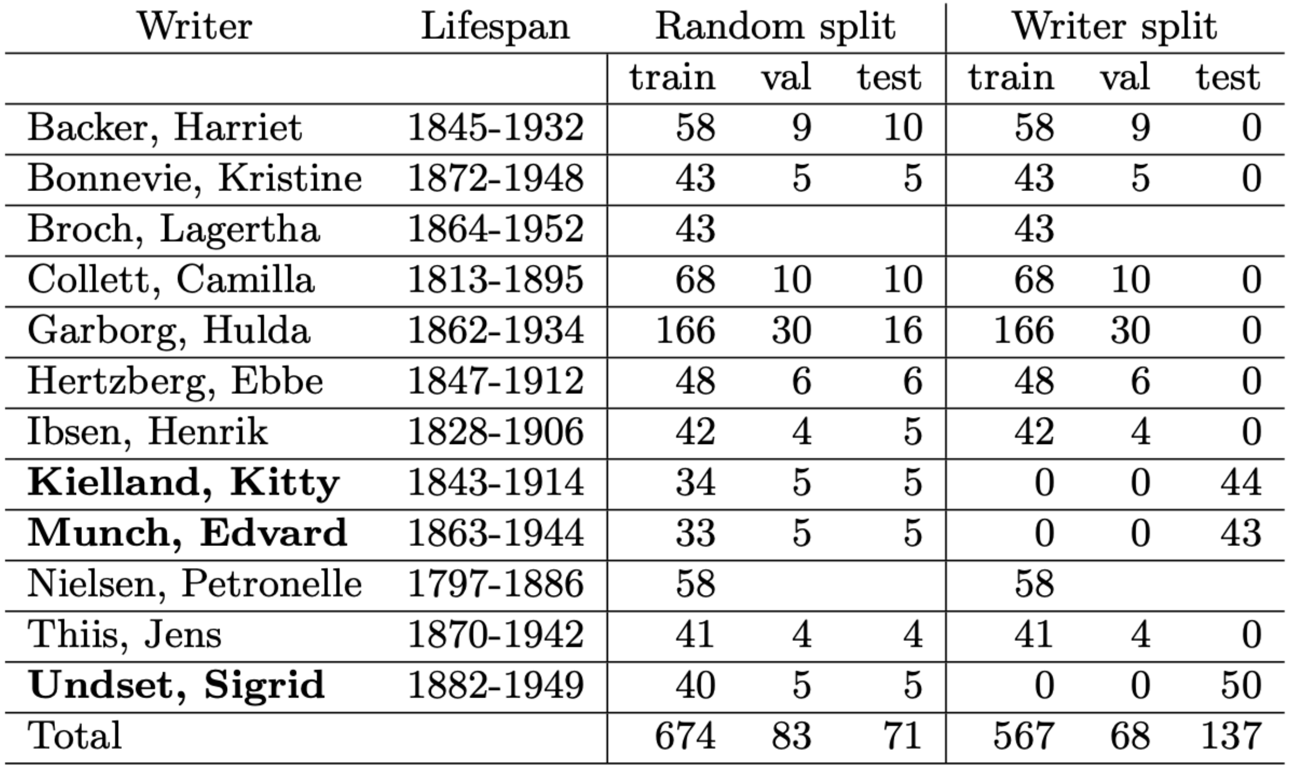
Hansteensgt. 2 IV 28/4 – 19

Kjære Gerhard.

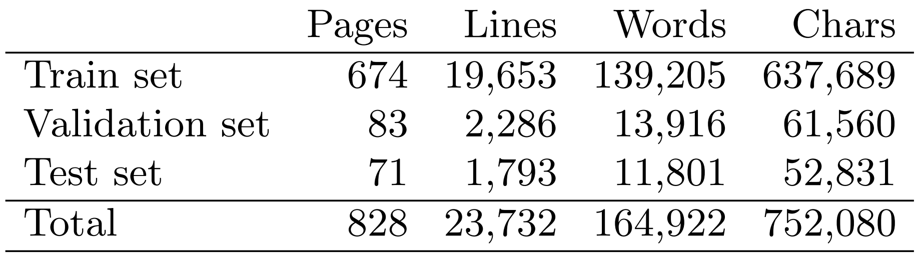
Tak for Brevet om Boken og Haven og Crokus og Blaaveis og tak fordi Du vilde be mig derut sammen me Kris og Dagna. Men vet Du da ikke, at Kris reiste me sin Söster Fru Kröger til Lillehammer nogle Dage efter Begravelsen? Hendes Adresse er Amtsingeniör Kröger. Hun skriver at de blir der til Midt ut i Mai. Nu er hun noksaa medtat skjönner jeg af Sorg og af Længsel, skriver saa meget rörende om Oluf. Dagna har det herligt, skriver hun; "Hun er bare glad og det vet jeg, at Oluf er gla over der hvor han nu er." Jeg har saa inderlig ondt af hende, og om Du skrev et Par Ord tror jeg det vilde gjöre hende godt. – Jeg glæder mig over, at Du har skrevet en Bok, og jeg er vis paa, at den er god.

Letters from Henrik Ibsen (1872), Camilla Collett (1877) and Harriet Backer (1919).

The NorHand Dataset



Writer	Lifespan	Random split			Writer split		
		train	val	test	train	val	test
Backer, Harriet	1845-1932	58	9	10	58	9	0
Bonnevie, Kristine	1872-1948	43	5	5	43	5	0
Broch, Lagertha	1864-1952	43			43		
Collett, Camilla	1813-1895	68	10	10	68	10	0
Garborg, Hulda	1862-1934	166	30	16	166	30	0
Hertzberg, Ebbe	1847-1912	48	6	6	48	6	0
Ibsen, Henrik	1828-1906	42	4	5	42	4	0
Kielland, Kitty	1843-1914	34	5	5	0	0	44
Munch, Edvard	1863-1944	33	5	5	0	0	43
Nielsen, Petronelle	1797-1886	58			58		
Thiis, Jens	1870-1942	41	4	4	41	4	0
Undset, Sigrid	1882-1949	40	5	5	0	0	50
Total		674	83	71	567	68	137



	Pages	Lines	Words	Chars
Train set	674	19,653	139,205	637,689
Validation set	83	2,286	13,916	61,560
Test set	71	1,793	11,801	52,831
Total	828	23,732	164,922	752,080

- Manual transcription at line level
- Available in Page XML format
- Official splits provided
- Version 1 (more to come)

Download: https://zenodo.org/record/6542056

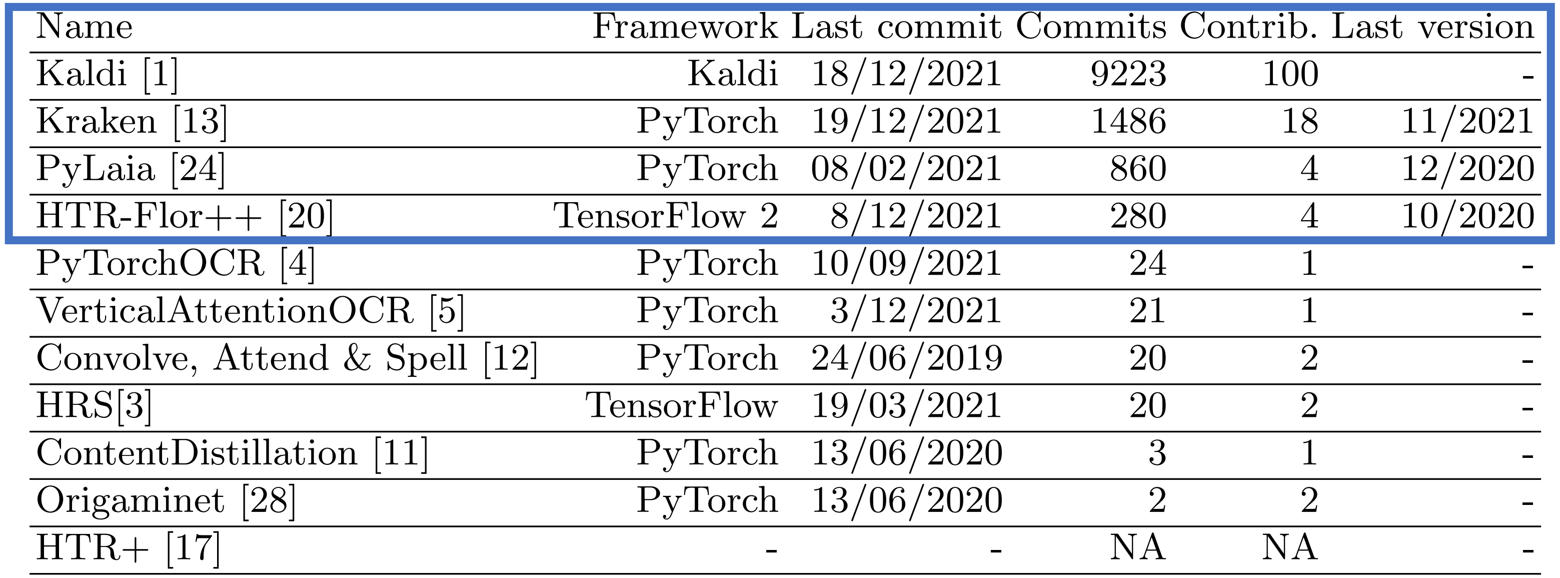
Survey of recent open source HTR libraries

- Survey of HTR libraries used in IJDAR, ICDAR, ICFHR, DAS, ICPR papers
- Between 2019 and 2021
- Open source
- Compared to state-of-the-art systems on publicly available databases of handwritten documents in European languages

10 libraries + HTR+ from Transkribus

Selection of open source HTR libraries

Selected



Name	Framework	Last commit	Commits	Contrib.	Last version
Kaldi [1]	Kaldi	18/12/2021	9223	100	-
Kraken [13]	PyTorch	19/12/2021	1486	18	11/2021
PyLaia [24]	PyTorch	08/02/2021	860	4	12/2020
HTR-Flor++ [20]	TensorFlow 2	8/12/2021	280	4	10/2020
PyTorchOCR [4]	PyTorch	10/09/2021	24	1	-
VerticalAttentionOCR [5]	PyTorch	3/12/2021	21	1	-
Convolve, Attend & Spell [12]	PyTorch	24/06/2019	20	2	-
HRS[3]	TensorFlow	19/03/2021	20	2	-
ContentDistillation [11]	PyTorch	13/06/2020	3	1	-
Origaminet [28]	PyTorch	13/06/2020	2	2	-
HTR+ [17]	-	-	NA	NA	-

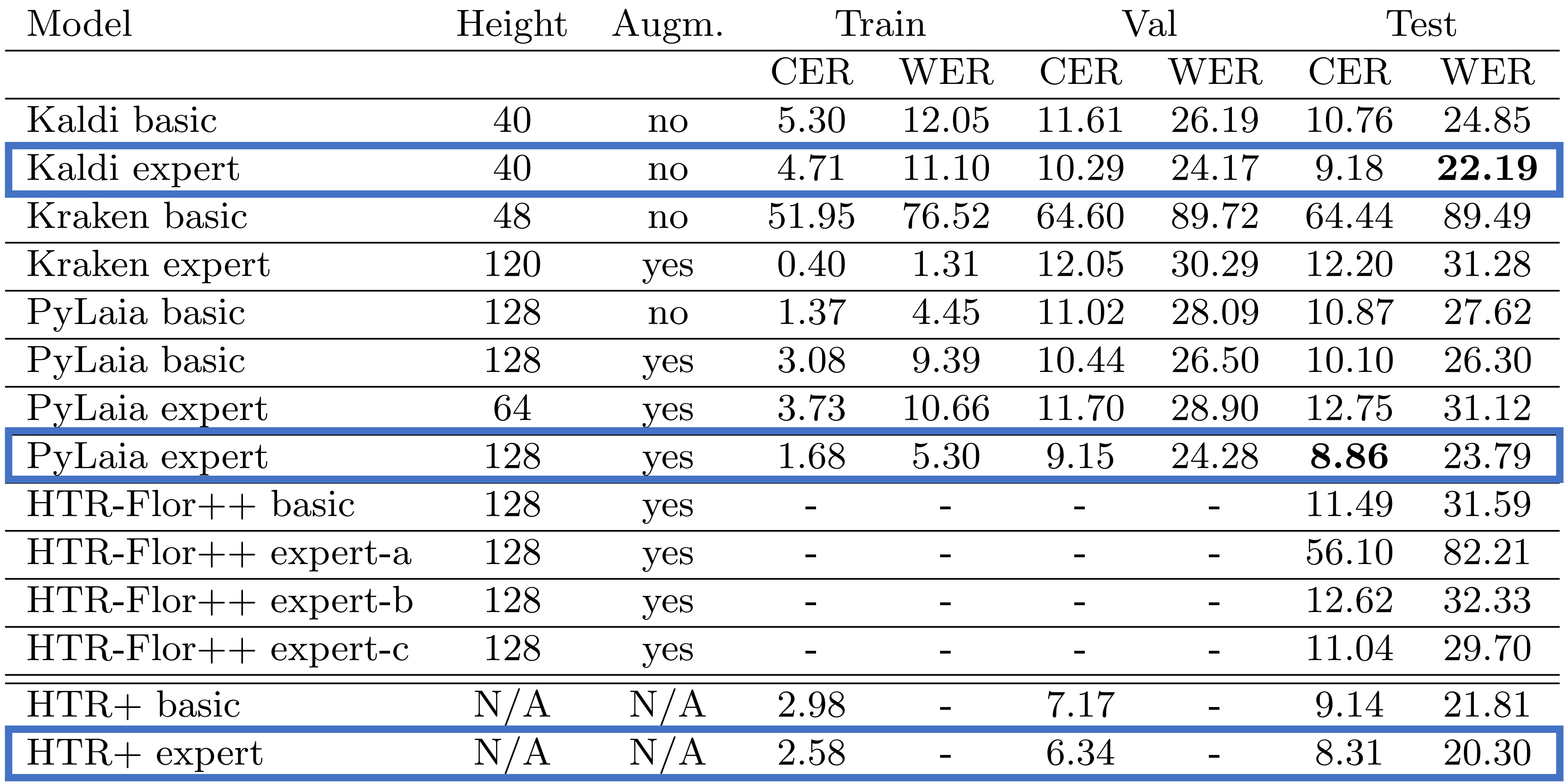
and HTR+

- Number of commits: active development
- Number of contributors: future maintenance
- Date of last commit: recently updated
- Date of last version/package: best practice of software development

Training of HTR models

- We trained the models from bounding boxes and manual transcriptions
- For each library, 2 setups:
  - Basic model: from the documentation (non-expert)
  - Expert model: with the support of the creators of the libraries
- Vertical lines are ignored
- Training with random split and writer split

Recognition results (random split)

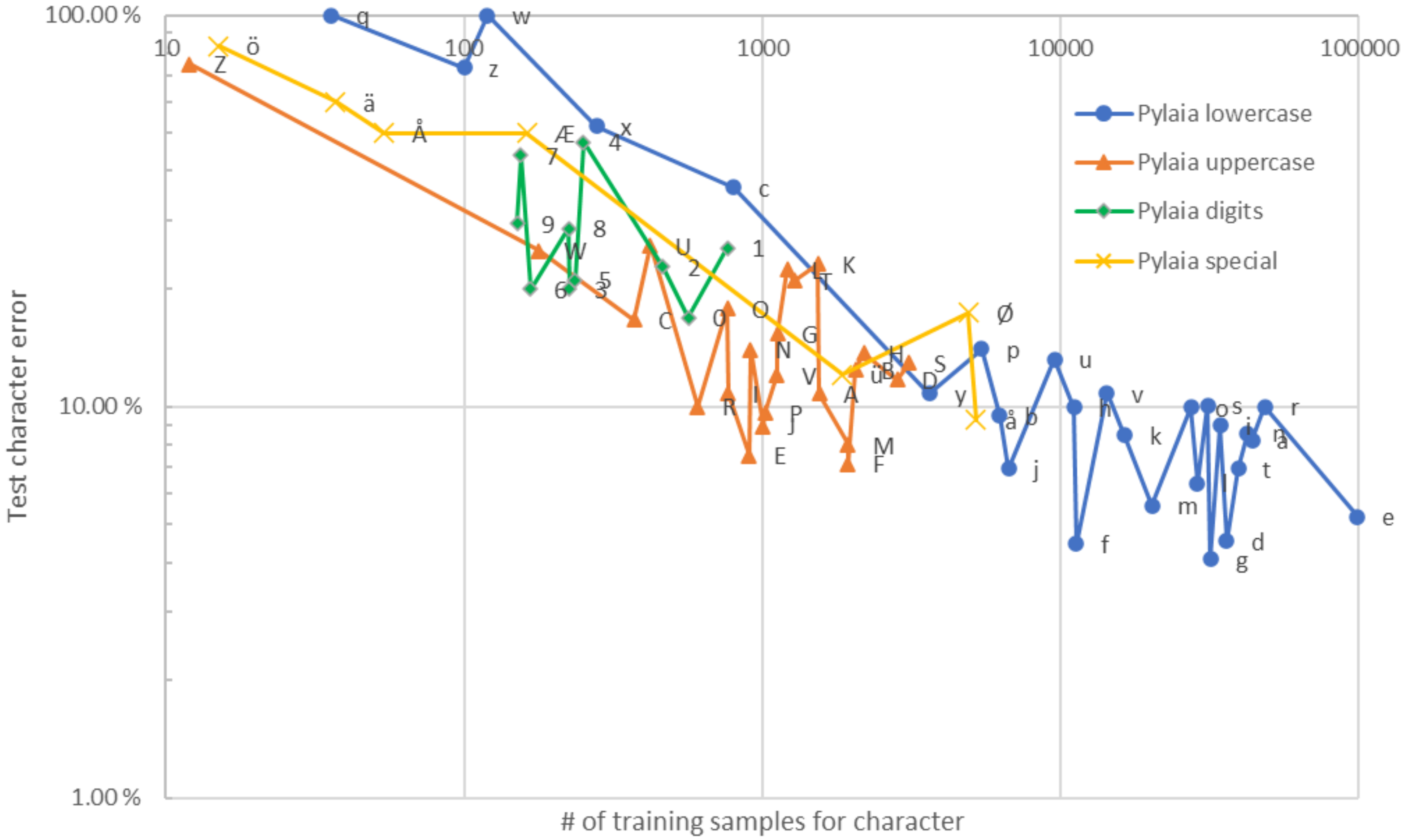


Model	Height	Augm.	Train		Val		Test	
			CER	WER	CER	WER	CER	WER
Kaldi basic	40	no	5.30	12.05	11.61	26.19	10.76	24.85
Kaldi expert	40	no	4.71	11.10	10.29	24.17	9.18	**22.19**
Kraken basic	48	no	51.95	76.52	64.60	89.72	64.44	89.49
Kraken expert	120	yes	0.40	1.31	12.05	30.29	12.20	31.28
PyLaia basic	128	no	1.37	4.45	11.02	28.09	10.87	27.62
PyLaia basic	128	yes	3.08	9.39	10.44	26.50	10.10	26.30
PyLaia expert	64	yes	3.73	10.66	11.70	28.90	12.75	31.12
PyLaia expert	128	yes	1.68	5.30	9.15	24.28	**8.86**	23.79
HTR-Flor++ basic	128	yes	-	-	-	-	11.49	31.59
HTR-Flor++ expert-a	128	yes	-	-	-	-	56.10	82.21
HTR-Flor++ expert-b	128	yes	-	-	-	-	12.62	32.33
HTR-Flor++ expert-c	128	yes	-	-	-	-	11.04	29.70
HTR+ basic	N/A	N/A	2.98	-	7.17	-	9.14	21.81
HTR+ expert	N/A	N/A	2.58	-	6.34	-	8.31	20.30

Help of an expert is usefull

Data augmentation improves the model

Detailed CER analysis

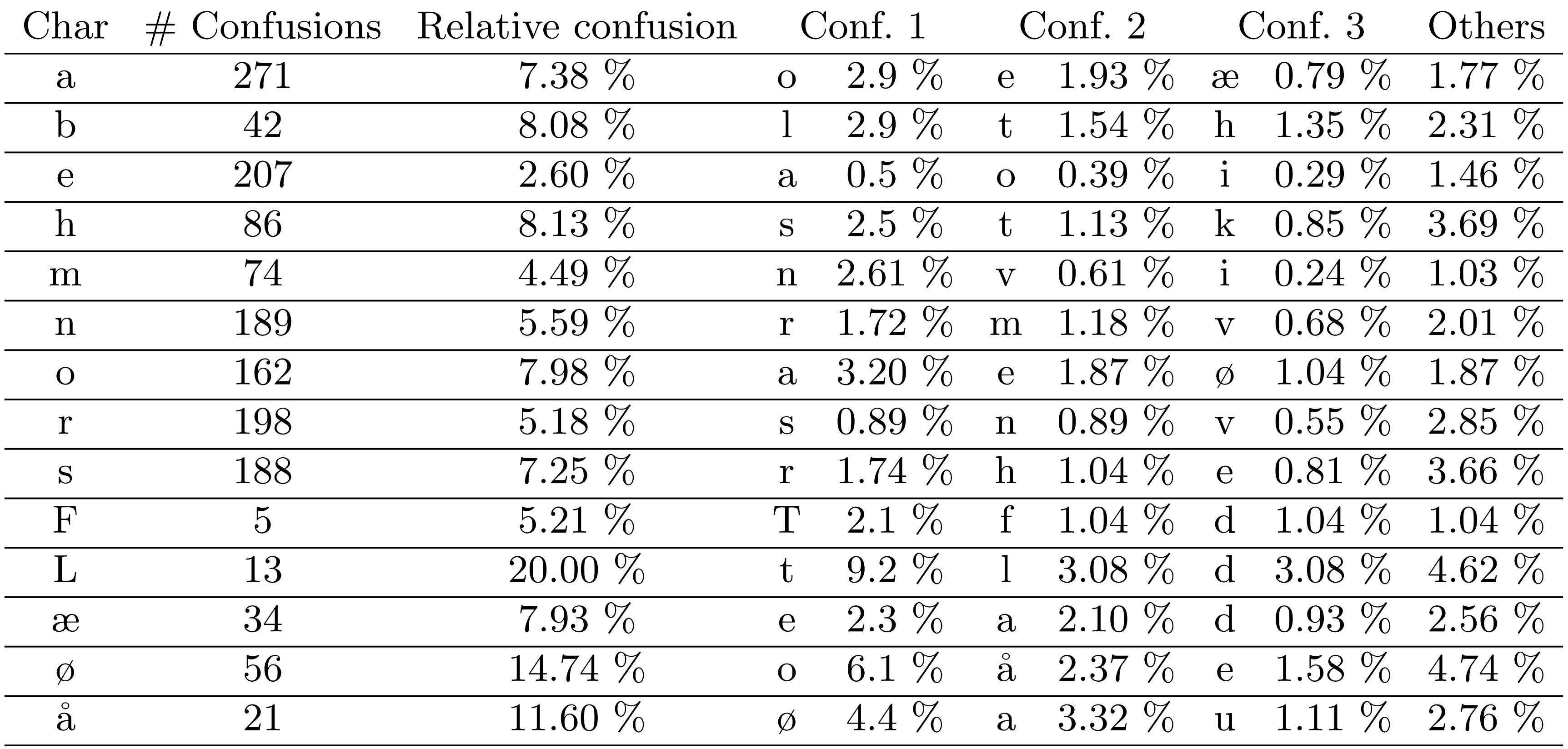


Pylaia Expert model

No language model

Strong correlation between CER and number of training samples

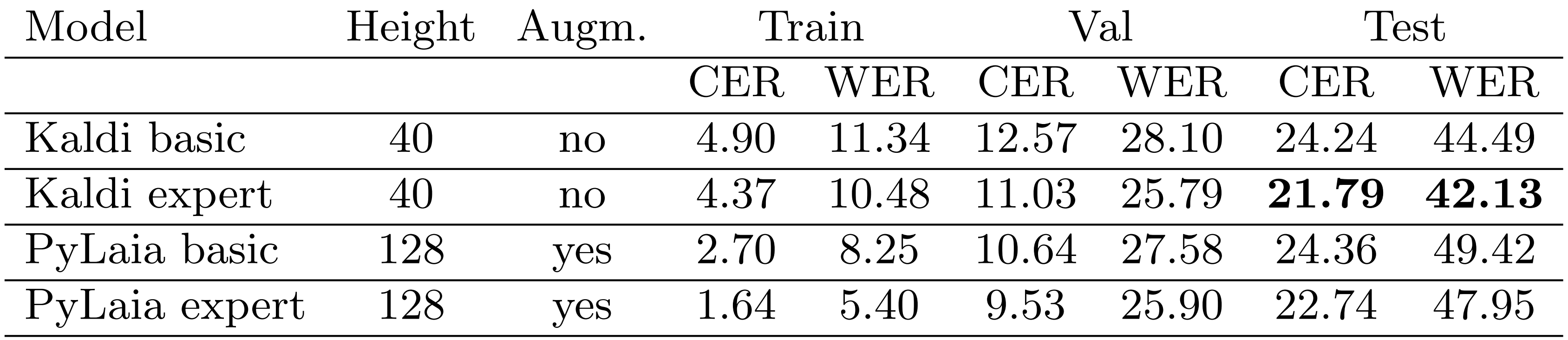
Most common confusion



Char	# Confusions	Relative confusion	Conf. 1		Conf. 2		Conf. 3		Others
a	271	7.38 %	o	2.9 %	e	1.93 %	æ	0.79 %	1.77 %
b	42	8.08 %	l	2.9 %	t	1.54 %	h	1.35 %	2.31 %
e	207	2.60 %	a	0.5 %	o	0.39 %	i	0.29 %	1.46 %
h	86	8.13 %	s	2.5 %	t	1.13 %	k	0.85 %	3.69 %
m	74	4.49 %	n	2.61 %	v	0.61 %	i	0.24 %	1.03 %
n	189	5.59 %	r	1.72 %	m	1.18 %	v	0.68 %	2.01 %
o	162	7.98 %	a	3.20 %	e	1.87 %	ø	1.04 %	1.87 %
r	198	5.18 %	s	0.89 %	n	0.89 %	v	0.55 %	2.85 %
s	188	7.25 %	r	1.74 %	h	1.04 %	e	0.81 %	3.66 %
F	5	5.21 %	T	2.1 %	f	1.04 %	d	1.04 %	1.04 %
L	13	20.00 %	t	9.2 %	l	3.08 %	d	3.08 %	4.62 %
æ	34	7.93 %	e	2.3 %	a	2.10 %	d	0.93 %	2.56 %
ø	56	14.74 %	o	6.1 %	å	2.37 %	e	1.58 %	4.74 %
å	21	11.60 %	ø	4.4 %	a	3.32 %	u	1.11 %	2.76 %

Pylaia expert model

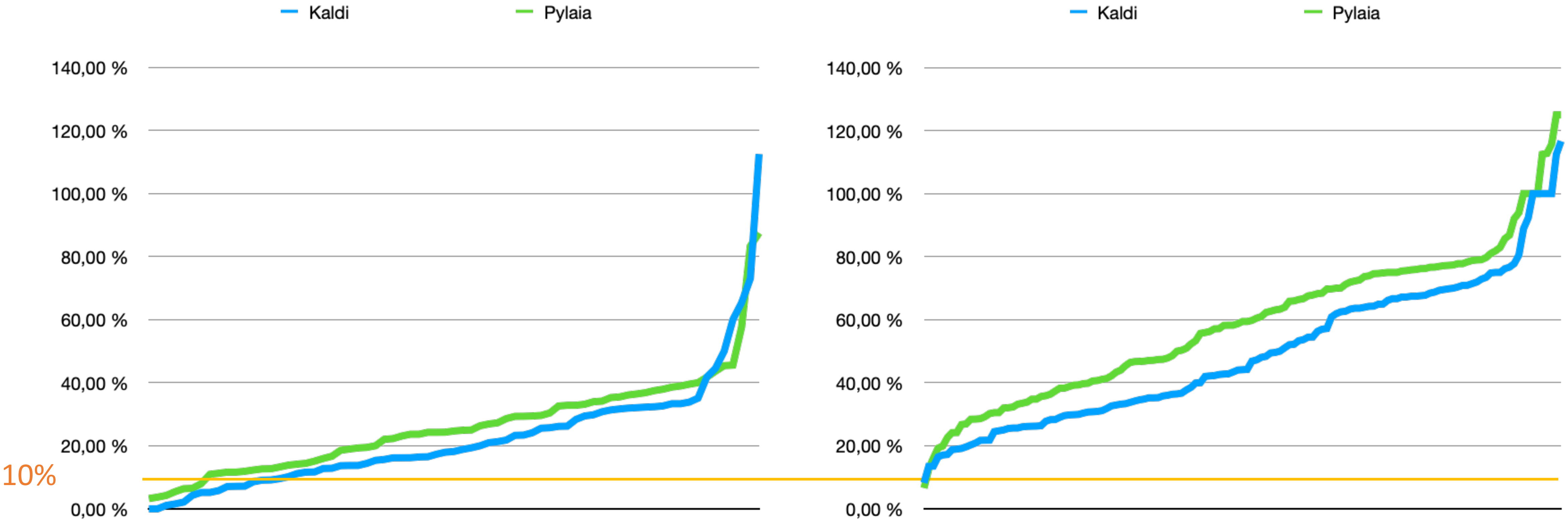
Recognition results with unseen writers split



Model	Height	Augm.	Train		Val		Test	
			CER	WER	CER	WER	CER	WER
Kaldi basic	40	no	4.90	11.34	12.57	28.10	24.24	44.49
Kaldi expert	40	no	4.37	10.48	11.03	25.79	**21.79**	**42.13**
PyLaia basic	128	yes	2.70	8.25	10.64	27.58	24.36	49.42
PyLaia expert	128	yes	1.64	5.40	9.53	25.90	22.74	47.95

- Training the best models with the writer split
- Lack of generalization, not enough different writers

Distribution of WER at document level



Random split

Writer split

Conclusions

- New challenging dataset for HTR
- Comparison of open source HTR libraries with software criteria and CER/WER
  - need to promote best practices in software development for HTR libraries
- Need to go beyond CER/WER analysis
- No Transformer: did not meet the criterion, but to be updated

Tusen takk !

TEKLIA | www.teklia.com

Christopher Kermorvant
kermorvant@teklia.com