

Attributing authorship in the **noisy** digitised correspondence of Jacob and Wilhelm Grimm

G Franzini, M Kestemont, G Rotari, M Jander
JK Ochab, E Franzini, J Byszuk and J Rybicki

Marian Smoluchowski Institute of Physics

Jagiellonian University, Cracow



Based on: DOI:10.3389/fdigh.2018.00004

5th July 2018

How digitisation methods impact computational text analysis

Digitisation:

- Optical Character Recognition (OCR)
- Handwritten Text Recognition (HTR)
- Gold standard: human expert transcription

Analysis:

- Authorship attribution
(within and between the above)

Manual transcription

The screenshot shows the Transkribus interface with a handwritten document on the right and its digital transcription on the left.

Handwritten Text:

14 handlangen scheint es gar nicht.
15 Lamprechts Tochter Syon verdient sicherlich eine ausgabe und
in erlangung bequemerer verleger steht die Quedlinburger
nationalbibliothek dafür offen. Basse gewährt auch andrän-
dige honorare.
19 Mich hochachtungsvoll empfehlend
20 Jac. Grimm
Cassel 29 mai 1840

Digital Transcription:

14 handlangen scheint es gar nicht.
15 Lamprechts Tochter Syon verdient sicherlich eine ausgabe und
in erlangung bequemerer verleger steht die Quedlinburger
nationalbibliothek dafür offen. Basse gewährt auch andrän-
dige honorare.
19 Mich hochachtungsvoll empfehlend
20 Jac. Grimm
Cassel 29 mai 1840

Left Panel (Toolbox and Properties):

- Tags: add, blackening, date, div, oao, organization, person, place, sic, signature, speech, swathed, textStyle, unclear, work
- Color palette: black, blue, cyan, magenta, green, purple, yellow, grey, red, orange, brown, light blue, pink, light green, light grey, white.
- Tags under cursor: person (offset:0; length:10; fir...)
- Buttons: Clear tags for selection, Add attribute..., Delete selected attribute.
- Properties of 'person' tag:
 - Add attribute...
 - Delete selected attribute.

Property	Value
offset	0
length	10
continued	false
notice	
occupation	
firstname	Lamprecht
dateOfDeath	nach 1250
dateOfBirth	1215
lastname	von Regensburg

[Image reproduced with permission of the Hessisches Staatsarchiv Marburg].

Jander, M. (2016). *Handwritten Text Recognition – Transkribus: A User Report*. Göttingen, Germany: eTRAP Research Group, University of Göttingen.

Data set

HTR and human trans.:

- collection of 36,000 *personal* letters belonging to the Grimm family [1]
- Selected letters:
Jacob 50 (**44**)
(1793-1863)

OCR:

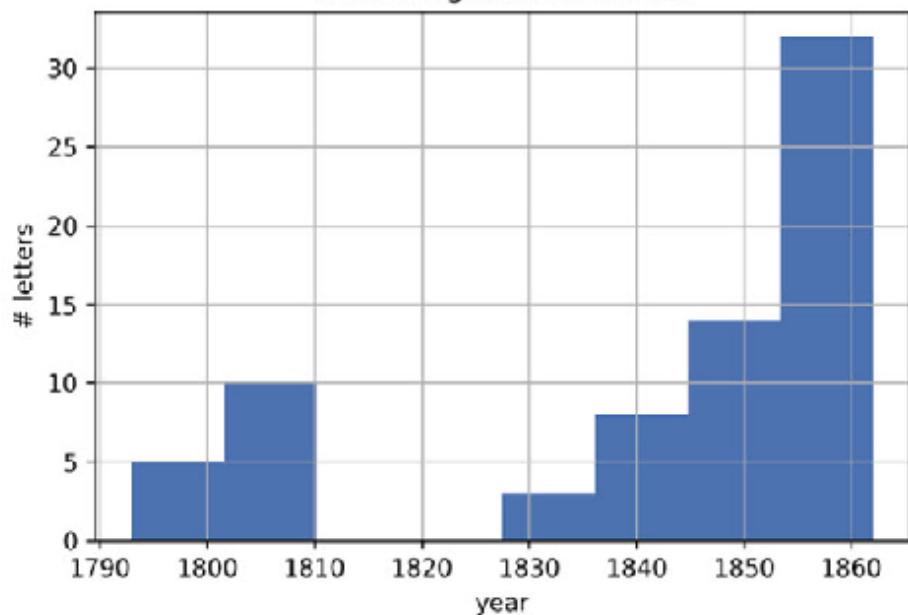
- seven-volume critical edition* [2]
Wilhelm 35 (**28**)
(1793-1859)

[1] Hessisches Staatsarchiv Marburg , *340 Grimm*, <http://www.unimarburg.de/uniarchiv/grimm>

[2] Rölleke, H. (2001). *Briefwechsel zwischen Jacob und Wilhelm Grimm* (Stuttgart: Hirzel Verlag)

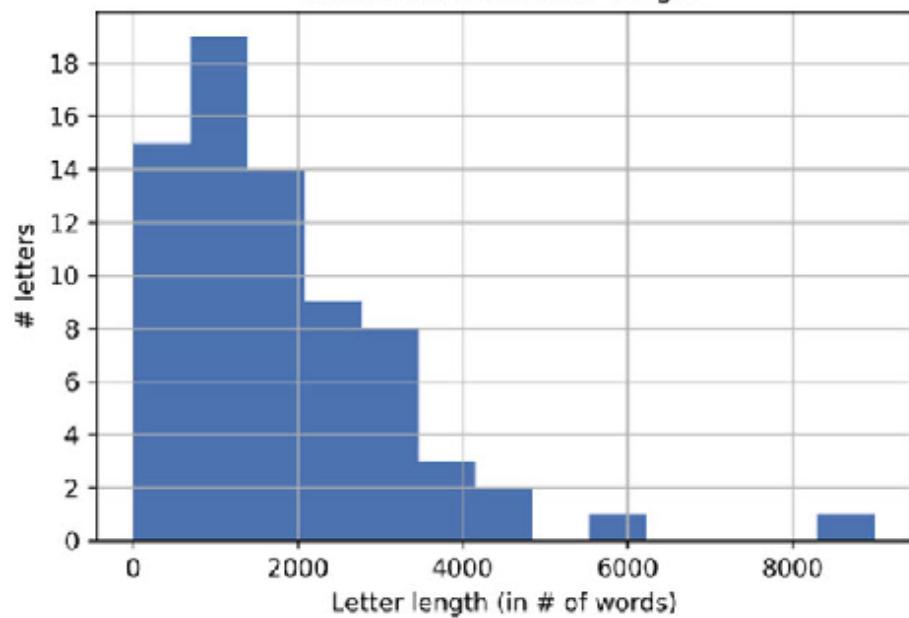
Data set

Chronological distribution



Jacob 50 (44)
(1793-1863)

Distribution of letter length



Wilhelm 35 (28)
(1793-1859)

- [1] Hessisches Staatsarchiv Marburg , 340 Grimm, <http://www.unimarburg.de/uniarchiv/grimm>
[2] Rölleke, H. (2001). *Briefwechsel zwischen Jacob und Wilhelm Grimm* (Stuttgart: Hirzel Verlag)

Data set

HTR training:

- 72 letters + 11 other documents
- 28,936 words

Error rate:	HTR	OCR
CHARACTERS	5.59%	2.21%
WORDS	19.15%	11.75%

Legibility and cleanliness

Wilhelm's letters:

- **very low** legibility
(Br 5993, 7 years old)
 - **low** legiability
(Br 2680, 45 years old)
 - **medium** legibility
(Br 2743, 73 years old)
 - **high** legibility
(Br 2736, 69 years old)

J.J. u. ich müssen abfliegen. Dr. Bx. Doctor kann ich auf der
Bx. können kommt in die Kino (wir haben uns zugesagt?) gleich
am anderen Tag auf der Bx. abgezogen, ist es so möglich und
wird ich mich zugesagt, obwohl mein neuer Sohn Margon auf
der Bibliothek ersten Stock. Ich füllte Ihnen darüber geschrieben,
mein Sohn

Der Klang ist leider sehr guter und gefüllt vom Liedesabfallen, doch stimmt die Stimmung dieses Liedes ab in die Übergangszeit nach dem Krieg nicht ausgenutzt wird. Heute haben wir uns wieder auf die Hoffnung gesetzt, sondern Hoffnungslosigkeit über alle hinaus.

Wir führen vor der Zukunft wie vor einem verschlossenen Thor, daß man an einen Krieg denkt und fügt dem rüstet ist natürlich, in der letzten Zeit schreit es mir vieler etwas zuviel getötet zu sein, immer aber hat man das gefühlt als sei es uns aufgelebt.

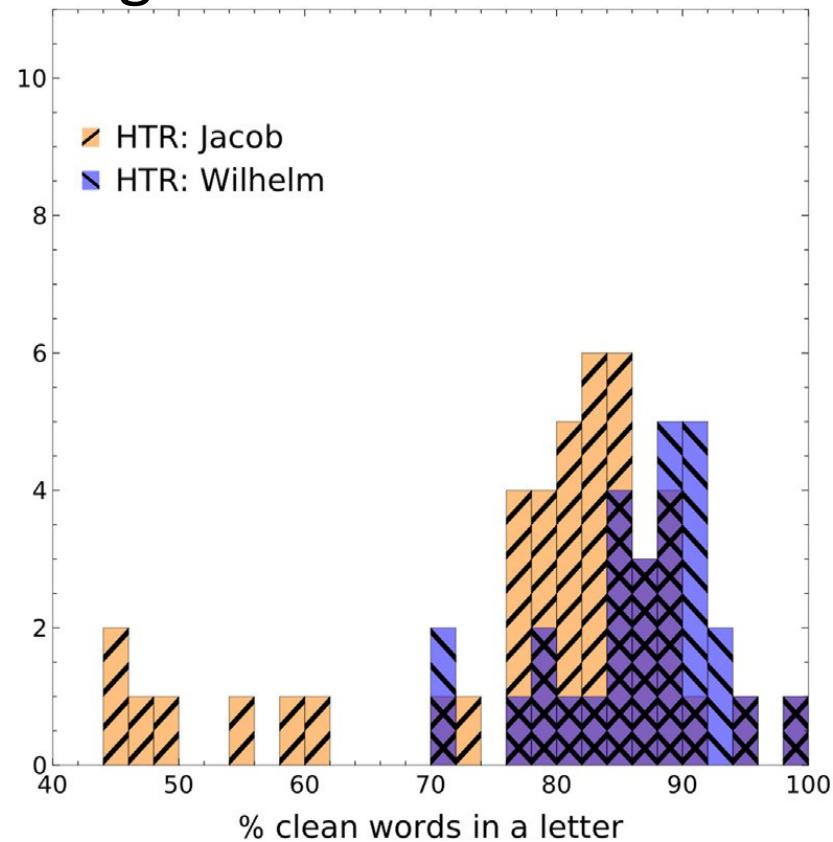
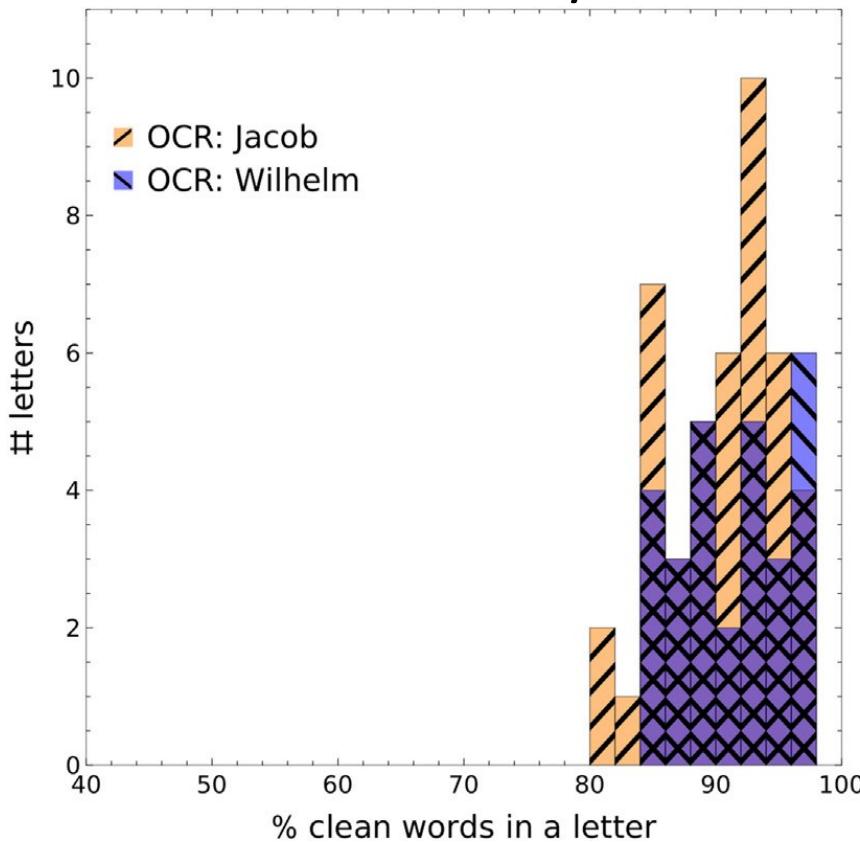
Zuerst meinen daas, hochgeeholte Herr professor, für die freundlichen
Wünsche zu meinem Geburtstag, mit dem ich auf jeden Fall Recht,
mit in diese Zeit die Grippe, die mich den winter über sehr geplagt
hat, sehr erfreut war.

Ihre noch niewals Thore beihörig zum Vorlesung in die hand,
ohne mir Ihre Genauigkeit zu sperren, und braucht nicht zu
sagen, wie sehr ich mir Ihnen dafür verbunden fühle.

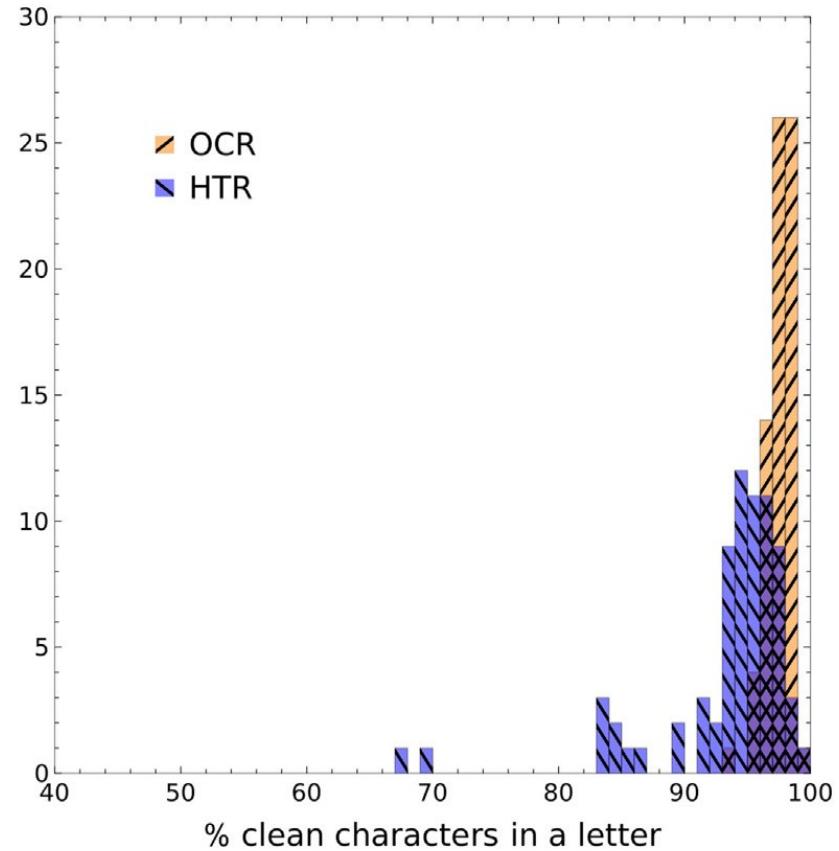
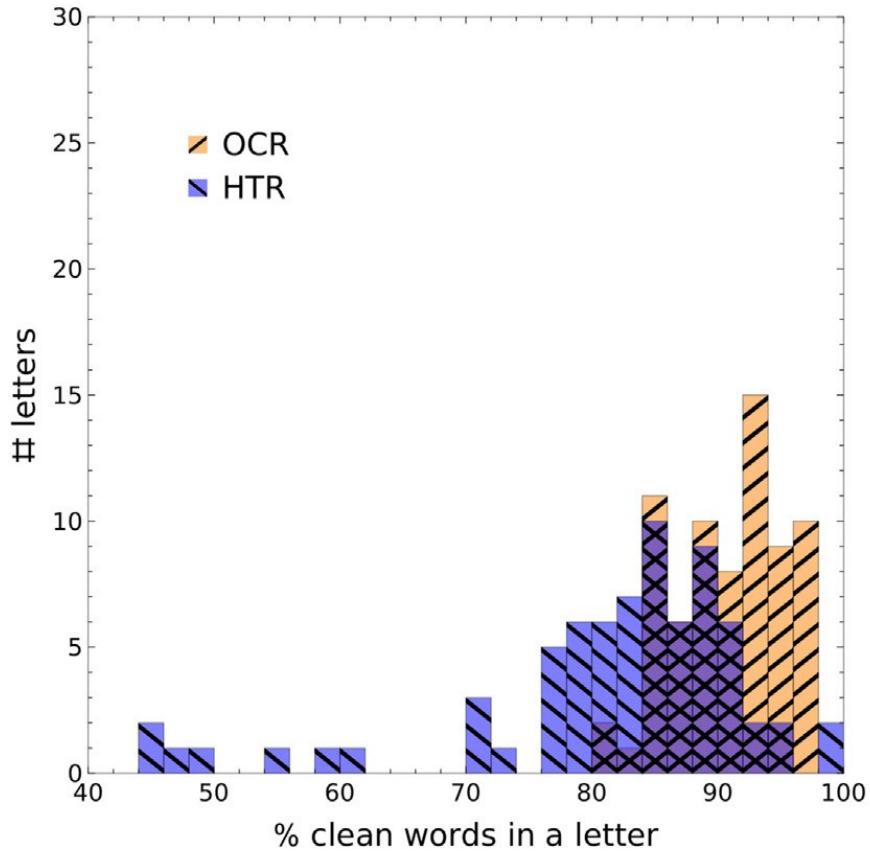
Legibility and cleanliness

Human-assessed legibility (very low excluded):

- Jacob: 36 low/9 medium–high
- Wilhelm: 15 low/13 medium–high



Legibility and cleanliness



Stylometry and authorship attribution based on:
words/word n-gram/character n-gram
frequencies.

How is the
word frequency distribution
affected by errors in HTR/OCR?

Lexical richness

Out of many diversity indices:

- Shannon entropy (**tails**)
- Simpson's index (**core**)
(inverse participation ratio)

$$H = - \sum_{t=1}^T p_t \log p_t$$

$$D = \sum_{t=1}^T p_t^2$$

Simple, least arbitrary, theoretically understood,
known limiting values

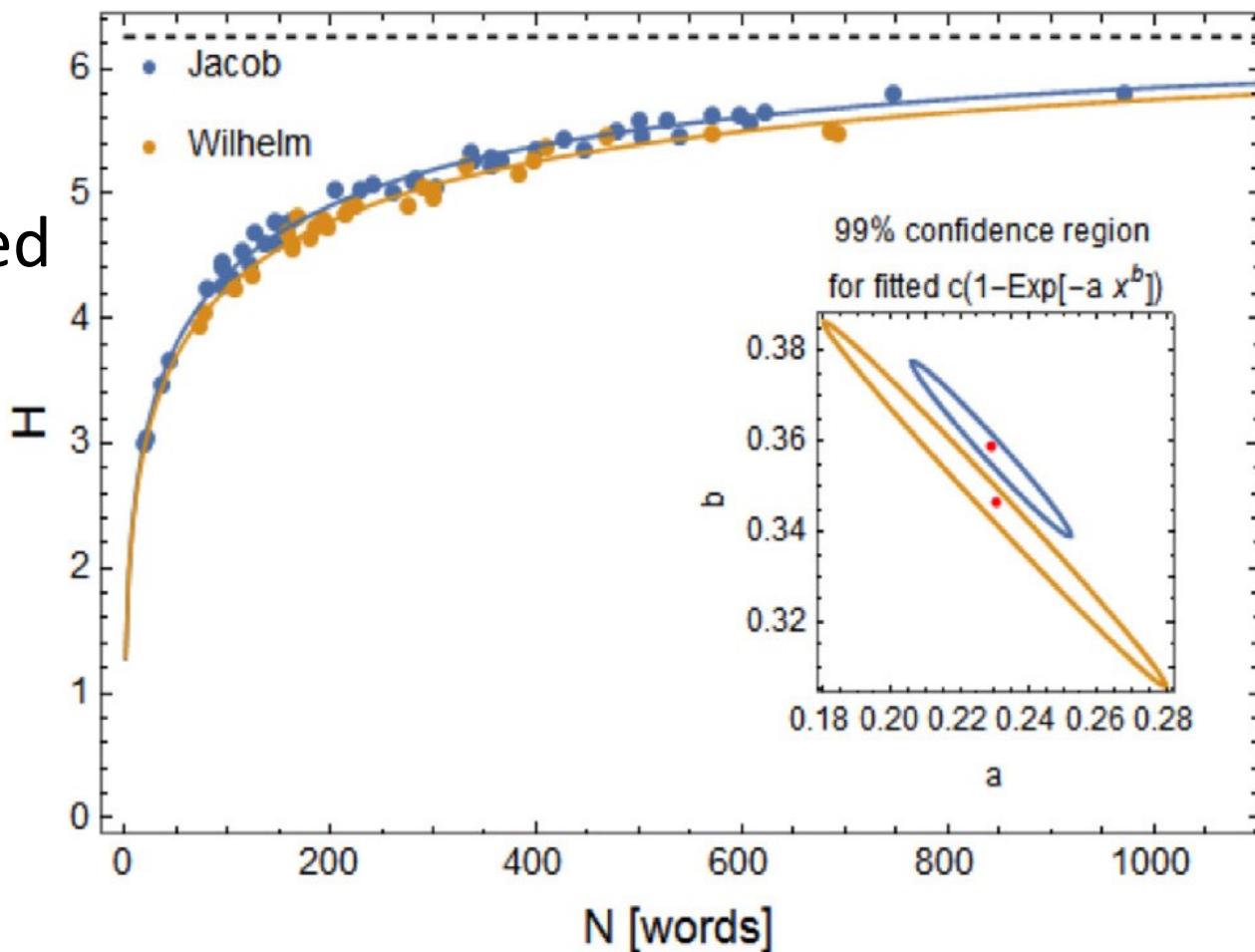
Lexical richness

- HTR produces **enough errors** to significantly yield **lower richness** per letter
- in short letters probably caused by HTR omitting or merging words
- no other correlations between text richness and cleanliness of HTR or OCR

OCR is more viable for stylometric measurements.

Lexical richness

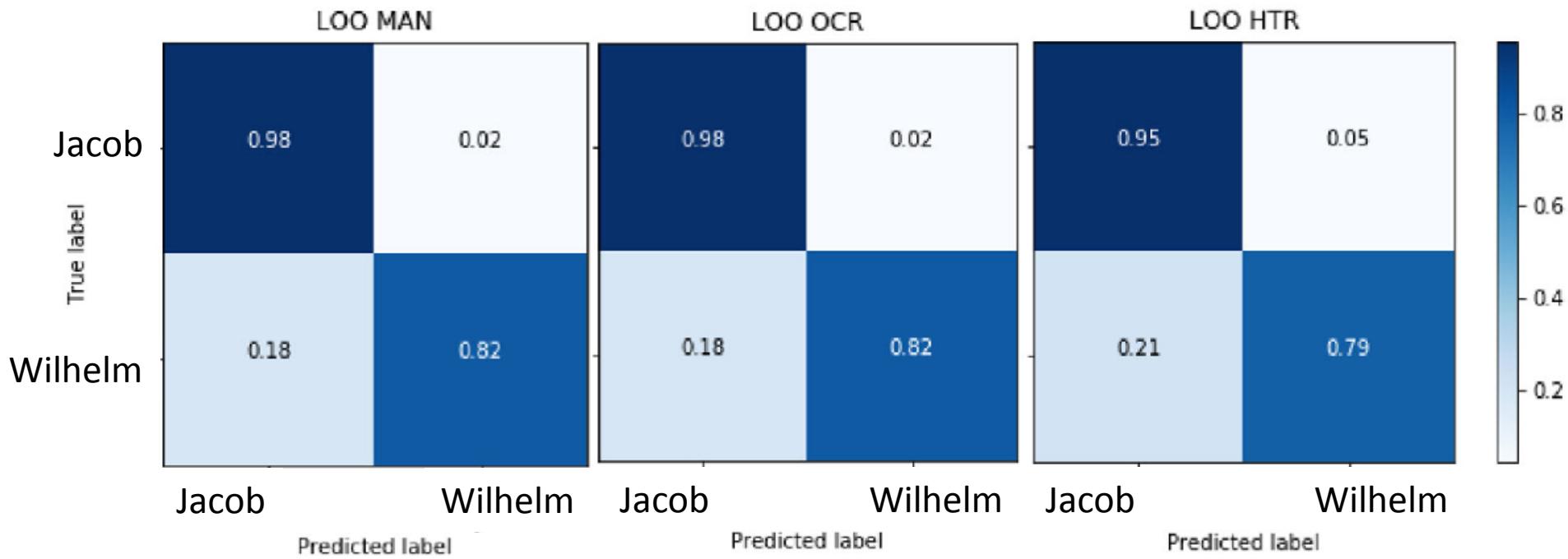
- can be authorial marker, but...
- depends on
text length
- can be modelled



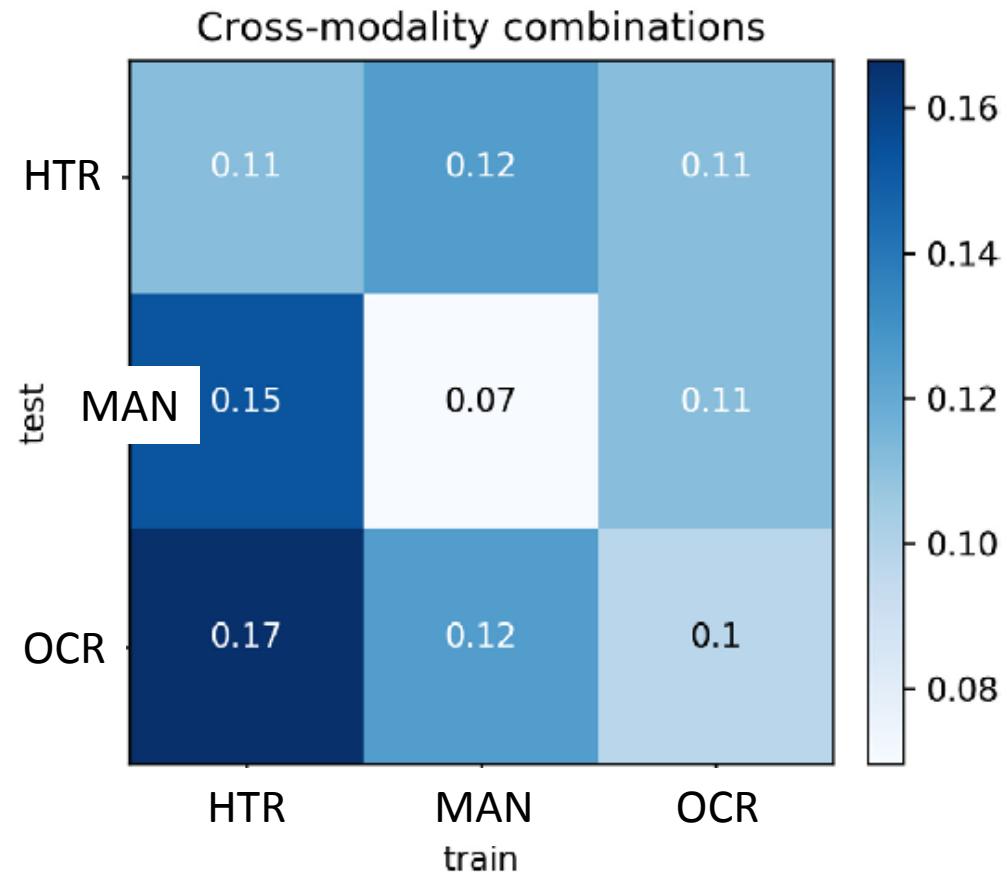
Authorship attribution

- 2- to 4-grams
- SVM (linear kernel)
- Leave-one-out cross-validation

	MAN	OCR	HTR
Accuracy	91.66	91.66	88.88
F1 score	88.46	88.46	84.61



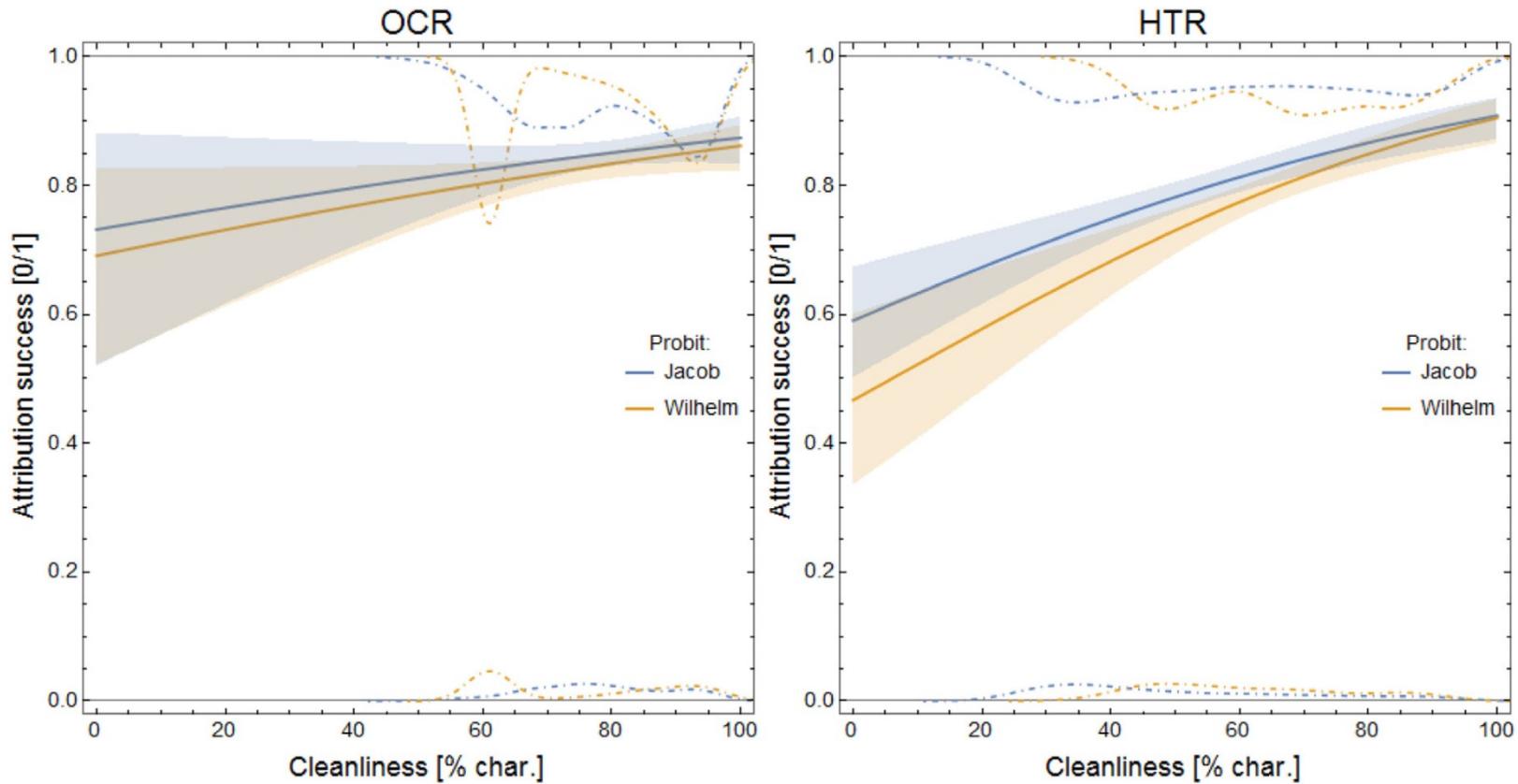
Authorship attribution



Authorship attribution

How OCR/HTR errors affect attribution?

- Texts (>1500 chars) made up from random original lines
- Burrows delta, 100 most frequent words



Conclusions

- human-assessed legibility agrees with HTR errors
- errors in **HTR** affect lexical richness
- **text-length corrected lexical richness** distinguishes Jacob and Wilhelm
- significant relation between auth. attr. **performance** and **cleanliness** for **HTR**
- auth. attr. **performs** as **well** on **OCR** as on human transcription



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN



Based on:

**Franzini G, Kestemont M, Rotari G, Jander M, Ochab JK,
Franzini E, Byszuk J and Rybicki J (2018)**

*Attributing Authorship in the Noisy Digitized Correspondence
of Jacob and Wilhelm Grimm.*

Front. Digit. Humanit. 5:4. doi:10.3389/fdigh.2018.00004



**Federal Ministry
of Education
and Research**

SPONSORED BY THE

Funded by the “Interessenbekundung zur Themenbesetzung im Campuslabor Digitalisierung” initiative of the University of Göttingen (Projekt: 392860 “Campuslabor”). GF, GR and MJ are funded by the German Federal Ministry of Education and Research (BMBF) grant no. 01UG1509.



M. Eder

M. Büchler

