frontiers
in Digital Humanities

ORIGINAL RESEARCH
published: 05 April 2018
doi: 10.3389/fdigh.2018.00004

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

Greta Franzini[1], Mike Kestemont[2], Gabriela Rotari[1], Melina Jander[1], Jeremi K. Ochab[3], Emily Franzini[1,4], Joanna Byszuk[5] and Jan Rybicki[3]*

[1] Georg-August-Universität Göttingen, Göttingen, Germany, [2] Universiteit Antwerpen, Antwerpen, Belgium, [3] Uniwersytet Jagielloński, Kraków, Poland, [4] Decoded Ltd., London, United Kingdom, [5] Instytut Języka Polskiego Polskiej Akademii Nauk, Kraków, Poland

OPEN ACCESS

Edited by:
Simonetta Montemagni,
Consiglio Nazionale Delle
Ricerche (CNR), Italy

Reviewed by:
Laurence Mellerin,
Centre national de la recherche
scientifique (CNRS), France
Hugh A. G. Houghton,
University of Birmingham,
United Kingdom
Federico Boschetti,
Istituto di Linguistica Computazionale
A. Zampolli (CNR), Italy

****Correspondence:***
Mike Kestemont
mike.kestemont@uantwerpen.be

Specialty section:
This article was submitted to
Digital Literary Studies,
a section of the journal
Frontiers in Digital Humanities

Received: *14 December 2017*
Accepted: *08 March 2018*
Published: *05 April 2018*

Citation:
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

This article presents the results of a multidisciplinary project aimed at better understanding the impact of different digitization strategies in computational text analysis. More specifically, it describes an effort to automatically discern the authorship of Jacob and Wilhelm Grimm in a body of uncorrected correspondence processed by HTR (Handwritten Text Recognition) and OCR (Optical Character Recognition), reporting on the effect this noise has on the analyses necessary to computationally identify the different writing style of the two brothers. In summary, our findings show that OCR digitization serves as a reliable proxy for the more painstaking process of manual digitization, at least when it comes to authorship attribution. Our results suggest that attribution is viable even when using training and test sets from different digitization pipelines. With regards to HTR, this research demonstrates that even though automated transcription significantly increases the risk of text misclassification when compared to OCR, a cleanliness above $\approx$ 20% is already sufficient to achieve a higher-than-chance probability of correct binary attribution.

Keywords: stylometry, authorship attribution, German literature, Grimm, digitization, OCR, HTR, digital humanities

1. INTRODUCTION

Studies show that researchers working on data analysis projects can spend up to 80% of project time pre-processing data and only 20% addressing their research question(s) (Wickham, 2014; Press, 2016). This imbalance suggests that scholars operate under the assumption that the time invested in pre-processing is directly proportional to the quality of the results—in other words, that any high quality analysis is intolerant to low quality data.

This research[1] builds upon studies that have explored acceptable degrees of digitization noise in computational, style-based authorship attribution,[2] in an effort to produce a model of Jacob and

[1] This article describes research conducted as part of a six-month pilot project at the University of Göttingen, Germany. The project, known as TrAIN (*Tracing Authorship In Noise*), sought to investigate the effect of noisy OCR'ed (Optical Character Recognition) and HTR'ed (Handwritten Text Recognition) data on computational text analysis. The project ran between July 2016 and January 2017. The project website is available at: http://www.etrap.eu/research/tracing-authorship-in-noise-train/ (Accessed: May 23, 2017).

[2] Since 2015, this topic has also been the subject of an EMNLP annual workshop. See: https://noisy-text.github.io/2017/ (Accessed: June 5, 2017).

Wilhelm Grimm's[3] handwriting, and identify authorship in a body of correspondence. Specifically, we ran the analyses on an uncorrected OCR'ed version of the letters derived from a print edition, and on an uncorrected HTR'ed version derived from digital images of the handwritten documents. The two versions introduce substantially different processing noise, giving us an effective means of comparison.

This article addresses the following research question: to which degree does the OCR and HTR noise affect the study of the individual stylome (Halteren et al., 2005) of the Grimm brothers? With this work, we explore the possibility offered by current technologies to better understand the writing styles of Jacob and Wilhelm Grimm. The challenges we face include the handling of textual noise generated through the digital transformation of the source texts, as well as the diversity and quantity of the data at our disposal.

This article is structured as follows: Sections 1 and 2 discuss the motivation behind the project and present related work; in Section 3 we introduce our materials, while in Section 4 we describe the digitization of these materials, the computational task of attributing authorship on the basis of the uncorrected digitized data, and discuss the results of the study; finally, Section 5 provides a summary of the results and suggestions for follow-up research.

2. RELATED WORK

Today, researchers working with texts available on the web are often confronted with unstructured and noisy data. Subramaniam et al. (2009) identified two types of noisy text: noise introduced during the generation process of a text (e.g., spelling errors, non-standard word forms, special characters, intentional abbreviations, grammatical mistakes, etc.) and noise introduced during the conversion process of a text to another form (i.e., through digitization or digital transformation) [(Subramaniam et al., 2009), p. 115]. Researchers working with *literary* texts available on the web typically deal with the latter type of noise—that caused by a text's transformation from its handwritten or printed form into machine-editable text *via* Optical Character or Handwritten Text Recognition (OCR and HTR, respectively). On *historical* texts, the recognition accuracy of OCR engines on the character level can reach 95% or more (Fink et al., 2017), but it may also perform poorly depending on the type of historical source (critical editions of classical texts as opposed to incunabula); for HTR, the rate ranges between 80 and 90% with respect to clean handwriting [(Agarwal et al., 2007), p. 5]. Lopresti (2009) discusses the effects of OCR errors on Information Retrieval (IR) and Natural Language Processing (NLP), and while there are methods in place to measure noise and recognition accuracy [(Subramaniam et al., 2009), p. 117–118], as well as semi-automatic applications to help correct the noise and tell historical spelling apart from machine errors (Vobl et al., 2014),[4] studies show that researchers working on data analysis projects can spend up to 80% of project time preparing data (Wickham, 2014).

In an effort to cut down on data pre-processing, scholars have tested algorithmic tolerance to noise. For instance, Agarwal et al. (2007) describe a series of experiments designed to better understand the effect of feature noise on automatic text classification algorithms, and have found that their classification accuracy tolerated up to 40% of introduced feature noise. In stylometry, Eder demonstrated the robustness of a number of authorship attribution methods across a variety of feature types in English, German, Polish, Ancient Greek, and Latin prose texts. His study showed that although noise tolerance varied across languages, even a 20% damage in the text did not significantly decrease the performance of authorship attribution [(Eder, 2013), p. 612]. Many present-day approaches to computational authorship attribution operate on very local features, such as the frequencies of word unigrams or character n-grams (Kjell, 1994; Stamatatos, 2009; Kestemont, 2014). These features are typically very common and well-distributed throughout texts (and much less sparse than, for instance, content words), which helps to explain why such feature categories are robust to the injection of significant levels of seemingly stochastic noise. Additionally, the use of regularization techniques (for example, with support vector machines) might prevent classifiers from overfitting on such noisy features. To the best of our knowledge, no systematic study thus far has "modelled out" such noise, although scholars routinely normalize their attribution through the use of NLP software (Juola, 2008; Koppel et al., 2009; Stamatatos, 2009).

In authorship attribution and stylometric analyses, a contributing factor to the reliability of the results is sample size or, in other words, the number of words analyzed to infer an author's style. In an article devoted to the topic, Eder, contrary to his previous claims (Eder, 2015), places the minimum acceptable sample size for authorship attribution as low as 2,000 words if the authorial fingerprint in the text is strong [(Eder, 2017), p. 223]. It follows that, while large samples are preferable, small samples can also, in certain cases, produce accurate results.

3. MATERIALS

3.1. The Grimm Correspondence

To investigate the effect of OCR and HTR noise on analyses in stylometric authorship attribution, we chose to work with a body of correspondence belonging to the Grimm family. The Grimm corpus, in fact, is well suited to this task for its diachrony and dual existence as handwritten documents and as a printed edition.

3.1.1. Handwritten Letters

In October 2015, we acquired a copy of a digitized corpus of roughly 36,000 *personal* letters belonging to the Grimm family

[3] Jacob (1785–1863) and Wilhelm Grimm (1786–1859) were German researchers, academics, and authors who collected and published folklore tales during the 19th century.

[4] For example, CIS-LMU's Post Correction Tool (PoCoTo), Available at: https://www.digitisation.eu/tools-resources/tools-for-text-digitisation/cis-lmu-post-correction-tool-pocoto/ (Accessed: May 24, 2017).

from the State Archives in Marburg.[5] Among these, there are many letters that Jacob and Wilhelm Grimm wrote to each other and to their acquaintances over a period of 70 years. Their letters vary in topic (from ailments to trips) and bear witness to the brothers' life and stylistic evolution. While large, the Marburg collection does not represent the complete Grimm corpus of letters. An additional 1,000 *professional* letters are held at the Humboldt University in Berlin,[6] but negotiations with the Center for Grimm's Correspondence to obtain a copy of those are still open.

As the purpose of the project was to study the handwriting of Jacob and Wilhelm Grimm, out of the complete Marburg collection we only selected those written by the two brothers.

3.1.2. Print Critical Edition of the Letters

The only critical edition of the entire corpus of Grimm letters in existence is Heinz Rölleke's 2001 *Jacob und Wilhelm Grimm, Briefwechsel*, which is made up of seven volumes (Rölleke, 2001).[7] The Editor's Note to the edition states that the editorial conventions follow those of a critical edition faithful to the original letters (*gedruckte Antiquitatexte*).[8] More specifically, Rölleke normalizes the text by reducing both single and double hyphens in word combinations to single hyphens; adds missing punctuation marks; italicizes unusual abbreviations in italic square brackets; places umlauts where these are missing, but does not record these omissions; adds incomplete characters without specifying where these occur in the handwritten document; and does not include stains and strike-through text.

3.2. Letters Selected for the Study

The total number of letters selected from the Marburg corpus was 85, 50 written by Jacob and 35 by Wilhelm. For the most part, these letters are either addressed to each other, to their other relatives or to Karl Weigand, an author and philologist, who collaborated with the Grimms on the creation of their dictionary, the *Deutsches Wörterbuch*. **Tables 1** and **2** split the 85 letters between Jacob and Wilhelm Grimm and list them in chronological order.

3.2.1. Categorization of the Letters: Epoch and Readability

The 85 letters were manually categorized into epochs and according to their degree of readability.

3.2.1.1 Epochs

TABLE 1 | Overview of letters written by Jacob Grimm.

Letters written by Jacob Grimm: 50			
Epoch	**Letter ID**	**Year**	**Readability**
1. 1793	Br 5995	1793	Low [to very low]
2. 1800	Ms 237	1800	Low
3. 1805–1806	Br 2164	1805	Low
	Br 2165	1805	Low
	Br 2169	1805	Low
	Br 2163	1805	Low [to very low]
	Br 2166	1805	Low
	Br 2167	1805	Low
	Br 2168	1805	Low [to very low]
	Br 2170	1805	Low
	Br 2176	1805	Very low
	Br 2174	1806	Low
4. 1814–1863	Br 2175	1833	Low
	Br 2171	1833	Medium [to high]
	Br 2172	1838	Medium [to high]
	Br 5996	1838	Medium
	Br 2237	1840	Medium
	Br 2238	1840	Low
	Br 2239	1840	Low [to medium]
	Br 2240	1841	High
	Br 2241	1844	Very low [to medium]
	Br 2242	1846	Very low [to medium]
	Br 2243	1847	Medium
	Br 2173	1848	Low
	Br 2269	1848	Low
	Br 2244	1849	Low [to medium]
	Br 2245	1849	Low
	Br 2268	1850	Low
	Ms 131	1850	High
	Br 2246	1852	Low
	Br 2247	1853	Low
	Br 2248	1854	Low
	Br 2249	1855	Low
	Br 2250	1856	Low
	Br 2266	1857	Low
	Br 2251	1858	Low
	Br 2252	1858	Low
	Br 2253	1859	Low
	Br 2254	1859	Low
	Br 2255	1859	Low
	Br 2267	1859	Low
	Br 2256	1860	Low
	Br 2257	1860	Low
	Br 2258	1861	Medium [to low]
	Br 2259	1861	Medium [to low]
	Br 2260	1861	Low
	Br 2261	1862	Very low
	Br 2262	1862	Low
	Br 2263	1862	Low
	Br 2264	1863	Very low
	Br 2265	1863	Very low

Expert evaluation of the readability is provided in square brackets. The second epoch 1800 was added for the second HTR training model.

With age the individual writing style of the brothers changed. These changes are noticeable when studying the letters side-by-side. For example, Jacob's calligraphy between 1805 and 1806 (when he was 20–21 years old) differs from that of his final years. **Figure 1** is the earliest letter in the collection, dating back to 1793, written by Jacob at the age of 8.

Accordingly, we grouped letters by handwriting periods or epochs. This task produced four groups of letters per brother:

[5] The TIFF files and the publishing rights were purchased from the Hessen State Archive in Marburg, Germany. The specific collection name is *340 Grimm*. See: https://landesarchiv.hessen.de/ (Accessed: November 16, 2015).

[6] See: http://www.grimmbriefwechsel.de/arbeitsstelle/arbeitsstelle.html (Accessed: May 23, 2017).

[7] For more information about the critical edition, and to learn more about the contents of each volume, see: http://www.grimmbriefwechsel.de/ (Accessed: May 23, 2017).

[8] The Editor's Note can be read in full (in German) at: http://www.grimmnetz.de/bw/download/Band1-1_Edition.pdf (Accessed: May 24, 2017).

TABLE 2 | Overview of letters written by Wilhelm Grimm.

Letters written by Wilhelm Grimm: 35

Epoch	Letter ID	Year	Readability
1. 1793	Br 5993	1793	Low
	Br 5994	1793	Low
	Br 2678	1793	Low
	Br 2679	1793	Low
2. 1802–1805	Br 2677	1805	Low
3. 1831–1843	Br 2680	1831	Low
	Ms 426 Bl 7	1833	Very low [to low]
	Ms 428 Bl 7b	1833	Very low [to low]
	Ms 426 Bl 10	1833	Very low
	Ms 426 Bl 11	1833	Very low
	Ms 426 Bl 13	1833	Very low
	Ms 426 Bl 15	1833	Very low
	Br 1687	1843	Low
	Br 2681	1843	Very low [to medium]
	Br 1688	1843	Low
4. 1846–1859	Br 2734	1846	Medium [to high]
	Br 2682	1847	Low
	Br 2683	1848	Medium
	Ms 161	1850	Low
	Br 2735	1851	High [low to medium or high]
	Br 2736	1855	High [low to medium or high]
	Br 2684	1856	High [low to medium or high]
	Br 2685	1856	Medium
	Br 2687	1856	Medium
	Br 2686	1856	High [low to medium or high]
	Br 2688	1856	Medium
	Br 2689	1856	Medium
	Br 2737	1857	Medium
	Br 2690	1858	Low
	Br 2738	1858	Medium
	Br 2739	1858	Low
	Br 2740	1859	Low
	Br 2741	1859	Low
	Br 2742	1859	Low
	Br 2743	1859	Medium

Expert evaluation of the readability is provided in square brackets.

1793, 1800, 1805–06, 1814–63 for Jacob,[9] and 1793, 1802–05, 1831–43, and 1846–59 for Wilhelm.

3.2.1.2 Readability

The letters were also sorted into four levels of readability: *very low*, *low*, *medium,* and *high*, as shown in **Figure 2**. Readability is affected by the quality of the paper (low quality paper reveals ink-spots) and by the legibility of the script. In consultation with Grimm experts,[10] the readability categories of the Grimm handwriting were defined based on regularity, clearly distinguishable character length, and on the arbitrary omissions of final letters. One observation resulting from this analysis is that the early writing of both Jacob and Wilhelm is generally harder to read.

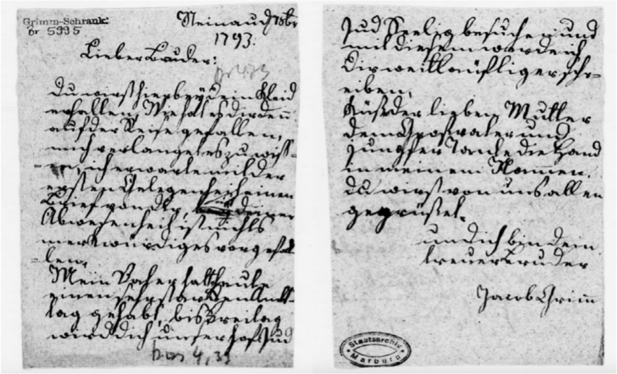


FIGURE 1 | Letter written by Jacob Grimm at the age of 8 (Br 5995). A complete transcription of the letter can be read in the Appendix [Image reproduced with permission of the Hessisches Staatsarchiv Marburg].

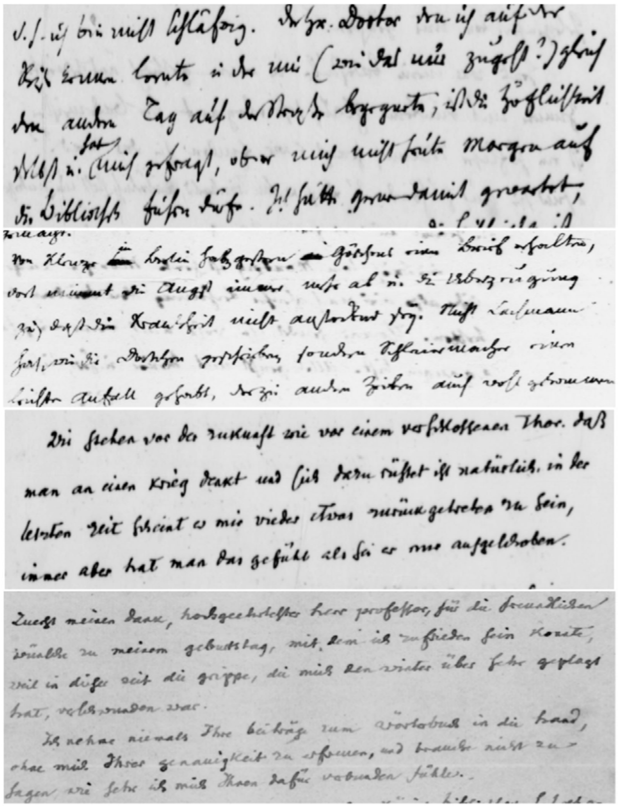
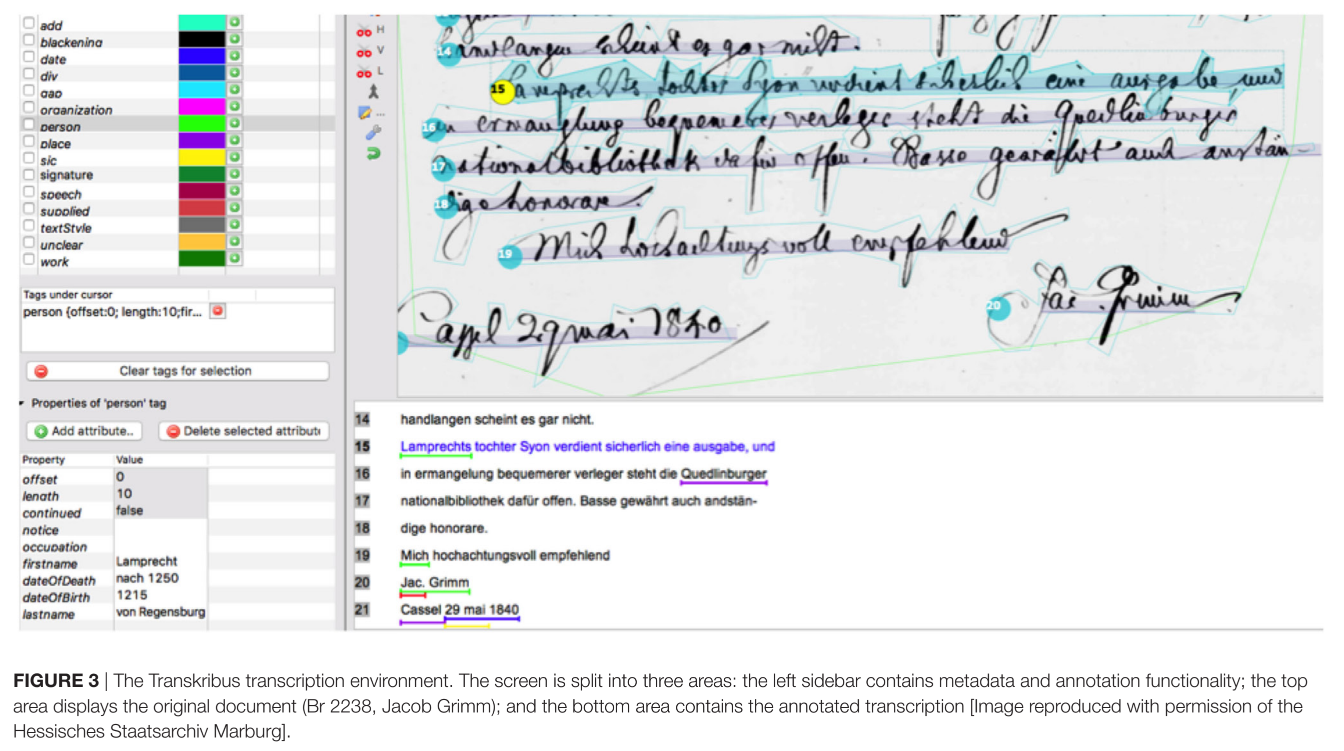


FIGURE 2 | Four letters written by Wilhelm Grimm. From top to bottom: very low readability (Br 5993, 7 years old); low readability (Br 2680, 45 years old); medium readability (Br 2743, 73 years old); high readability (Br 2736, 69 years old) [Images reproduced with permission of the Hessisches Staatsarchiv Marburg].

As will become clear later, the division of these groups of letters was essential to generate an HTR model that would produce the lowest possible error rate.

[9] The last group (i.e., 1814-63) contains small calligraphic changes within it, but they are not significant enough for this group to be further split.

[10] Bernhard Lauer and Rotraut Fischer of the Brüder Grimm-Gesellschaft e.V in Kassel, Germany.



□ add
□ blackening
□ date
□ div
□ gap
□ organization
□ person
□ place
□ sic
□ signature
□ speech
□ supplied
□ textStyle
□ unclear
□ work

Tags under cursor
person {offset:0; length:10;fir...

Clear tags for selection

Properties of 'person' tag

Add attribute.. Delete selected attribut

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

handlangen scheint es gar nicht.
Lamprechts tochter Syon verdient sicherlich eine ausgabe, und
in ermangelung bequemerer verleger steht die Quedlinburger
nationalbibliothek dafür offen. Basse gewährt auch andstän-
dige honorare.
Mich hochachtungsvoll empfehlend
Jac. Grimm
Cassel 29 mai 1840

FIGURE 3 | The Transkribus transcription environment. The screen is split into three areas: the left sidebar contains metadata and annotation functionality; the top area displays the original document (Br 2238, Jacob Grimm); and the bottom area contains the annotated transcription [Image reproduced with permission of the Hessisches Staatsarchiv Marburg].

4. METHODS AND RESULTS

4.1. Digitization: Transcription, HTR, and OCR

This section describes the digitization of the Grimm correspondence using three different methods and compares the differing degrees of cleanliness of the latter two.

4.1.1. Manual Transcription

Both the manual and automatic transcriptions of the letters were carried out using the state-of-the-art software *Transkribus*.[11] To build an HTR model that is able to automatically transcribe handwritten documents, Transkribus requires a minimum of 100 pages of manually transcribed text.[12] The 85 manually transcribed letters[13] varied in length—some are one page in length, some span multiple pages. The 50 letters by Jacob correspond to 90 pages of text, while the 35 letters by Wilhelm correspond to 64 pages, for a total of 154 manually transcribed pages.

As well as producing a diplomatic transcription of the letters,[14] we annotated meta-information about the brothers' calligraphy and the contents of the letters, as illustrated in **Figure 3** (Jander, 2016).

4.1.2. An HTR Model for the Handwritten Letters

The manual transcriptions were used to train a handwritten text recognition model that would be able to recognize and automatically transcribe other letters or documents written by the brothers (e.g., the 1,000 letters held in Berlin).

A reliable Transkribus HTR model needs a starting training set of, ideally, 100 manually transcribed pages of text.[15] As the 85 letters of Jacob and Wilhelm Grimm added up to 154 pages, we were faced with a choice between running a model on the brothers combined in order to meet the HTR 100-page requirement or training two separate models, one per brother, on fewer pages (90 pages for Jacob and 64 pages for Wilhelm). A decision was made to test both scenarios and compare the results.[16]

4.1.2.1. First HTR Model

The first HTR run was performed on the 85 manually transcribed letters (154 pages) of Jacob and Wilhelm *combined*, for a total of 26,983 words. As well as meeting the minimum page count, our hypothesis behind combining two handwritings into the same training model was that a mixed model would prove more resistant to diachronic changes in Grimm handwriting. However, the character error rate (CER) obtained from this mixed model

[11] For more information about Transkribus, see: https://transkribus.eu/Transkribus/ (Accessed: May 25, 2017).

[12] For more information, see the Transkribus Wiki: https://transkribus.eu/wiki/index.php/Questions_and_Answers#What_is_needed_for_the_HTR_to_work.3F (Accessed: May 25, 2017)

[13] The transcriptions were produced by three student assistants over a period of three months.

[14] For definitions of 'diplomatic transcription', see the *Lexicon of Scholarly Editing* at: http://uahost.uantwerpen.be/lse/index.php/lexicon/diplomatic-transcription/ (Accessed: May 25, 2017).

[15] For more information about training data-sets required for HTR, see: http://read.transkribus.eu/wp-content/uploads/2017/01/READ_D7.7_HTRbasedonNN.pdf (Accessed: June 5, 2017).

[16] The HTR model was trained by Dr Günter Mühlberger of Transkribus.

TABLE 3 | Letters discarded from the HTR training corpus due to their very poor readability, which negatively affected the first HTR run.

Letters discarded from HTR corpus			
Author	**Letter ID**	**Year**	**Readability**
Jacob	Br 2176	1805	Very low
	Br 2241	1844	Very low
	Br 2242	1846	Very low
	Br 2261	1862	Very low
	Br 2264	1863	Very low
	Br 2265	1863	Very low
Wilhelm	Ms 426 Bl 7	1833	Very low
	Ms 426 Bl 7b	1833	Very low
	Ms 426 Bl 10	1833	Very low
	Ms 426 Bl 11	1833	Very low
	Ms 426 Bl 13	1833	Very low
	Ms 426 Bl 15	1833	Very low
	Br 2681	1843	Very low

TABLE 4 | The 11 documents added for the second run of the HTR to compensate for the loss of 13 letters from the previous run.

Letters added to the HTR corpus					
Author	**Epoch**	**Document ID**	**Year**	**Readability**	**HTR word-count**
Jacob	2. 1800	Ms 237 (song)	1800	Low	343
	4. 1814–1863	Ms 239 (diary entry)	1815	High	1218
		Br 2231	1829	High	699
		Br 2230	1839	High	245
		Br 2232	1841	High	246
		Br 2235	1850	Medium	316
		Br 2233	1860	High	107
		Ms 242 (dictionary entry draft)	n.d.	High	485
Wilhelm	2. 1802–1805	Ms 245 (poem)	1802	Medium	177
	3. 1831–1843	Br 2579	1833	Medium	726
	4. 1846–1859	Br 2580	1854	Medium	1226

Documents included letters and also poems and songs.

amounted to 18.83%, which means that every fifth character in the text was not correctly recognized. To improve the CER, we transcribed an additional 2,000 words of Grimm handwriting (corresponding to 17 pages). Against expectations, the new CER increased to approximately 40%, that is, an error for every 2.5 characters in the text. Upon closer inspection, we noticed that the high CERs were heavily influenced by 13 letters whose readability was very low. To reduce the CER, we, therefore, prepared a second HTR run in which we replaced those problematic 13 letters with 11 other documents written by the brothers (35 pages for a total of 5,788 words).[17] **Tables 3** and **4** list the discarded letters and the added documents, respectively.

4.1.2.2. Second HTR Model

The second HTR run was thus performed on 83 documents, totaling 28,936 words (10,250 for Wilhelm and 18,686 for Jacob) and 128 pages (44 for Wilhelm and 84 for Jacob). In this run, Jacob and Wilhelm were trained separately with different training and test sets in a *divide and conquer* strategy, producing eight different runs (one per epoch per author). The intention in this run was to verify if smaller data-sets would produce more stable models. The results were encouraging, with an overall CER across the runs of <10%.

Below is an excerpt of the HTR transcription of Jacob's letter "Br 2238" (dating to 1840) using the model trained on Jacob's handwritten letters dating to the 1814–1863 period:

> Text of the Original Letter
>
> …handlangen scheint es gar nicht. Lamprechts tochter Syon verdient sicherlich eine Ausgabe, und in ermangelung bequemerer verleger steht die quedlinburger nationalbibliothek dafür offen. Rasse gewährt auch andständigehonorare. Mich hochachtungsvoll empfehlend Jac. Grimm

> HTR Transcription [Errors Underlined]
>
> …handlangen scheint es gar nicht. Lamprechts tochter von verdient sicherlich eine ausgabe und in ermangelung bgineneber verleger steht die quedlinburger natüoalbiblittchke der für offen. Rasse gewährt auch wurdendigehonorare mich hochachtungsvoll empfehlend Ihr. Grimm

4.1.3. OCR of the Critical Edition

The digitization and OCR of the seven volumes of the *Grimm Briefwechsel* edition produced seven output files.[18] The text below is an example of noisy OCR output (letter Br 2238 by Jacob Grimm):

> Handlangen scheint es gar nicht.
>
> Lamprechts tochter Syon verdient sicherlich eine ausgabe, und in er-manglung bequemerer Verleger steht die Quedlinburger nationalbibliothek dafür offen. Basse gewaährt auch anständige honorare. Mich hochachtungsvoll empfehlend
>
> Jac. Grimm.

As the reader will notice, the OCR preserved the edition's carriage return hyphen in the word *er-manglung*, and the word *Verleger* is capitalized while in the print edition it is in lower case.

As **Tables 5** and **6** show the accuracy of the OCR to be high with a median letter reaching above 91% clean words (above 98% correct characters).

4.1.4. Evaluating the Cleanliness of the HTR'ed and OCR'ed Data-Sets

Next, the cleanliness of the letters for which we had manual, OCR and HTR transcriptions was compared. The number of letters

[17] These were taken from the online portal of the Hessisches Staatsarchiv Marburg at: https://arcinsys.hessen.de/arcinsys/detailAction.action?detailid=g195109&icomefrom=search (Accessed: June 5, 2017).

[18] The digitization and OCR was done by the Göttingen Digitisation Centre with Abbyy Fine Reader. See: https://www.sub.uni-goettingen.de/en/copying-digitising/goettingen-digitisation-centre/ (Accessed: April 04, 2017).

available in these three formats was 72. We considered the manual transcriptions a gold standard against which the cleanliness of the other versions was measured (see **Table 5**). While for HTR the differences only come from recognition errors, for OCR it may be both recognition errors *and* possible editorial interventions by Rölleke.

TABLE 5 | Cleanliness of the collection of 72 letters.

Mean collection cleanliness		
	Clean words in %	**Clean characters in %**
OCR	88.25	97.79
HTR	80.85	94.41
Letter cleanliness (three quartiles)		
OCR	86.80; **91.69;** 94.06	97.95; **98.70;** 99.18
HTR	79.28; **84.29;** 88.39	94.09; **95.89;** 97.44

Numbers in bold are medians of the distributions of letters. Standard errors for the means over the collection are negligible.

TABLE 6 | Cleanliness of the collection of 72 letters by author.

Mean collection cleanliness				
	Clean words in %		**Clean characters in %**	
	Jacob	Wilhelm	Jacob	Wilhelm
OCR	87.10	91.12	97.60	98.26
HTR	79.44	84.21	94.24	94.81
Letter cleanliness (three quartiles)				
OCR	86.65; **91.69;** 93.87	87.51; **91.98;** 94.24	98.29; **98.86;** 99.17	97.49; **98.43;** 99.19
HTR	76.93; **81.93;** 85.68	83.61; **87.30;** 90.41	94.00; **95.50;** 96.96	95.22; **96.77;** 98.39

Wilhelm's letters have consistently higher scores, even though there are fewer of them. Numbers in bold are medians of the distributions of letters. Standard errors are <0.0026% (words) and <0.00016% (characters).

Depending on the stylometric type of analysis, cleanliness is assessed either on the percentage of misrecognized words (if we use words, word n-grams or lemmas as features in the authorship attribution/classification task) or of characters (if we use character n-grams), since any error in that respect changes the word/character/n-gram frequencies and can consequently alter the distances measured between texts (Burrows, 2002) (see **Figure 4**). A differing error rate between authors might also be problematic (see **Table 6** and **Figure 5**).

In the first step, it might not be viable to scrutinize this influence on all the features. Instead, we report the influence of recognition errors on the lexical richness of the texts, and only then we continue with authorship attribution (Section 4.2). Lexical richness is not necessarily a good authorship indicator (Hoover, 2003), but it has a straightforward stylistic interpretation and it can illustrate possible issues with noisy data. Out of numerous richness scores (Tweedie and Baayen, 1998; Wimmer and Altmann, 1999) we used two: *Shannon entropy*, $H = -\sum_{t=1}^{T} p_t \log p_t$, and *Simpson's index* $D = \sum_{t=1}^{T} p_t^2$ (also called *Inverse Participation Ratio*, IPR) are both special cases of diversity indices (Hill, 1973), where T is the number of types and p_t is the probability of occurrence of a type t (i.e., the number of occurrences of t divided by the number of all words in a text). Despite some criticism (Holmes, 1985; Thoiron, 1986) and the fact that they are very strongly (though non-linearly) correlated to one another, these indices are the simplest, the least arbitrary, and the theoretically best understood. For N being the number of tokens in the text, IPR ranges from $1/N$ to 1 (maximal richness and zero richness, respectively), it focuses more on the core of the word frequency distribution (the most frequent words), and so it stabilizes quickly with text length N; entropy, visualized in **Figure 6**, ranges from 0 to log N (zero and maximal richness), focuses on the tails of the word distribution

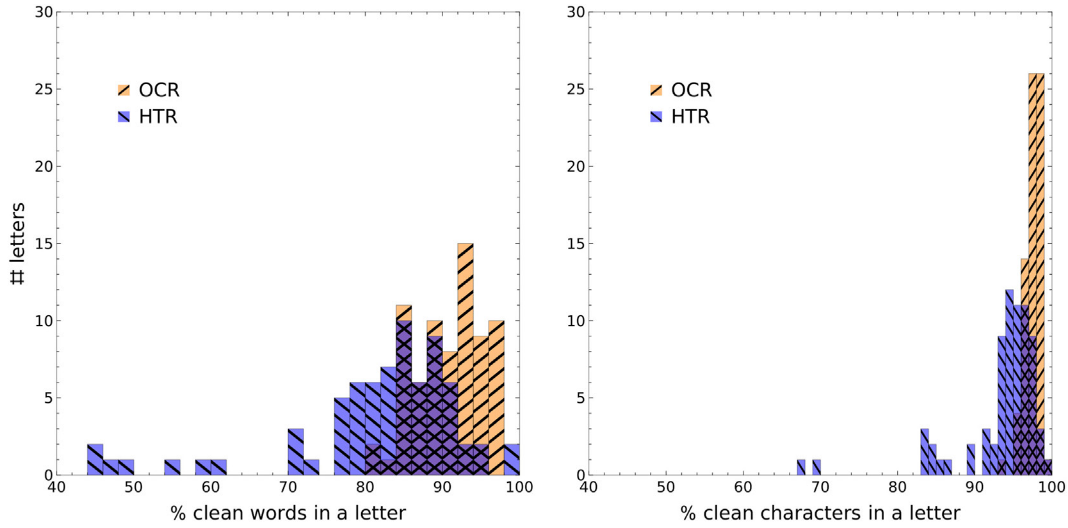


FIGURE 4 | Histograms show the number of letters that have a given percentage of correctly recognized words/characters. Noticeably, HTR produces rather unstable results (the outliers to the left). For visibility, the overlap between two histograms is indicated by mixing their colors and adding a cross-pattern.

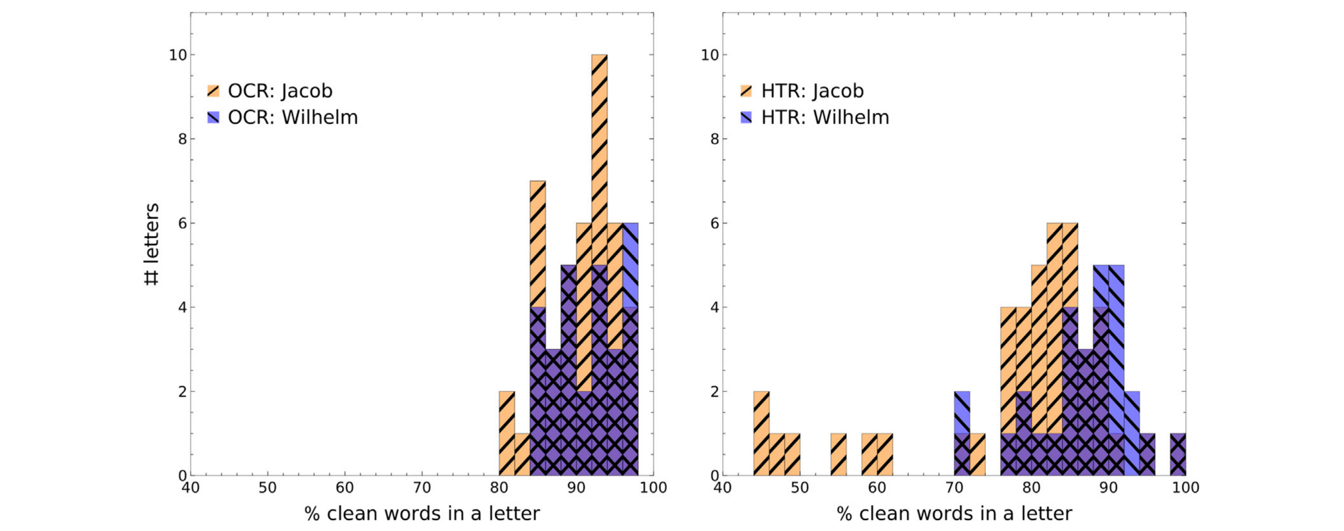


FIGURE 5 | Histograms of letters divided by Jacob and Wilhelm. The right panel shows that it was Jacob's writing that was more problematic for HTR.

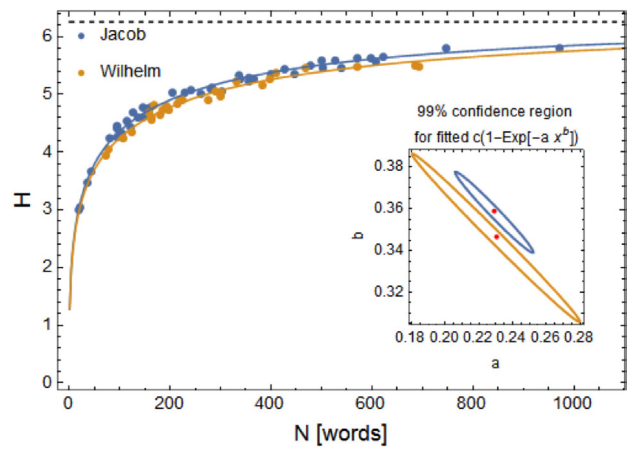


FIGURE 6 | Entropy of Jacob's and Wilhelm's manually transcribed letters (points) shows that Jacob has a greater lexical diversity. The result is statistically significant at a 0.99 confidence level (see inset). The black dashed line represents the fitted asymptote *c* common to both curves.

(least frequent words like *hapax legomena*), which stabilize slowly and make it more sensitive to small changes as text progresses.

One observation is that HTR produces enough errors to statistically and significantly yield lower richness per letter (p-values of paired T-tests are $p = 1.04 \times 10^{-7}$ and $p = 8.03 \times 10^{-6}$ for entropy and IPR, respectively); in some short letters this seems to be caused by HTR omitting or merging some words, which results in lower *N*, and—in the case of entropy—in lower log *N*. Otherwise, there are no statistical correlations between text richness and cleanliness of HTR or OCR. However, taking all the above into consideration, OCR seems a more viable option for stylometric measurements.

In order to test for the difference between the lexical richness of Jacob's and Wilhelm's letters we must circumvent the problem of the dependence of IPR and of the entropy on text length. We model it with an exponential function approaching a constant from below, which upon visual inspection, as shown in **Figure 6**, works well. Then, one may test for the difference between parameters of the fitted curves (inset in the Figure shows that they do differ).

4.2. Authorship Attribution

4.2.1. Rationale and Setup

In this section, we report on the authorship attribution task of modeling the individual writing styles of the Grimm brothers in the noisy digitized correspondence discussed above. We cast the modeling of their authorship as a categorization or classification experiment using Machine Learning (Sebastiani, 2002; Juola, 2008; Koppel et al., 2009; Stamatatos, 2009). We resorted to a standard binary classification routine, i.e., a Support Vector Machine (SVM) with a linear kernel, with its default settings as implemented in the well-known *scikit-learn* library (Pedregosa et al., 2011). Studies have shown that SVMs typically offer a strong baseline for authorship attribution (Stamatos, 2013), even in the face of extremely sparse input vectors. Because our data-set was small, we adopted a leave-one-out cross-validation procedure, where we iterated over all individual letters: in each iteration, one letter was set aside as a test instance, while the classifier was trained on the remaining training instances. Next, the trained model's prediction as to the authorship of the held-out sample was recorded. We describe the performance of individual models using the established evaluation metrics accuracy and F1-score. With respect to this experimental procedure, it should be noted that this setup can be considered as a relatively unchallenging attribution problem because the number of authors is very limited (Luyckx and Daelemans, 2011; Eder, 2015) and the genre of the texts involved is relatively stable (Stamatos, 2013). The size of the data-set might be considered small from a Machine Learning point of view, but it is highly representative of the specific field

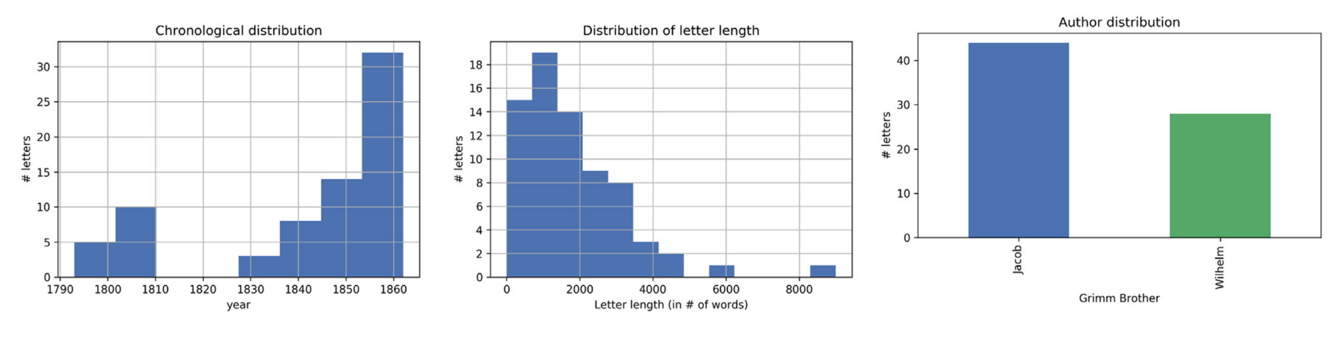


FIGURE 7 | General information about the subcorpus in this section: distribution of letters in terms of chronology, length, and authorship labels.

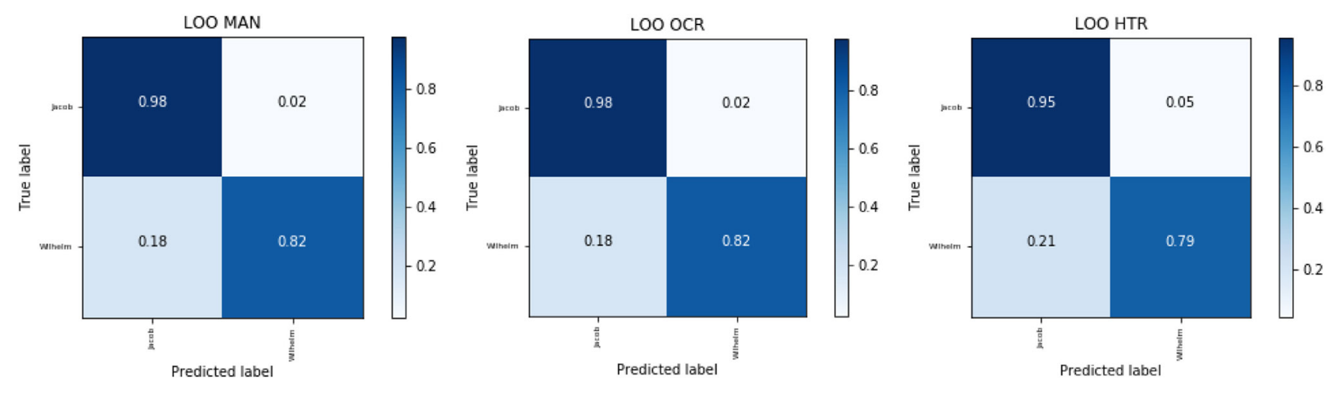


FIGURE 8 | Confusion matrices for the leave-one-out validation approach for each of the three data-sets.

TABLE 7 | Leave-one-out validation results for intra-modality attribution in terms of accuracy and F1.

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

of authorship studies, where small data-sets are very common, explaining the relative popularity of leave-one-out validation as an evaluation procedure. The length of the individual letters can be considered challenging, as most letters contain significantly fewer words that the minimal length thresholds previously discussed (Eder, 2015).

Here, we are especially concerned with the effect of different digitization modalities on attribution performance, i.e., the manually transcribed letters (MAN), automatic Handwritten Text Recognition on the original letters (HTR), and Optical Character Recognition on the original letters (OCR). Therefore, in our experiments we did not test our attribution models only within a single digitization modality, but also across the different modalities. As such, we aim to advance our understanding of *directionality artifacts*, i.e., the models built on the basis of one modality might scale relatively better to other modalities, which would make them more attractive for future projects. The base texts used in authorship studies typically have rather diverse origins and will often conflate materials coming from very different sources or editions or which result from different digitization modalities (e.g., OCR vs. HTR). An insight into these directionality effects would allow us to formulate very useful recommendations for future data collection efforts in the context of text classification.

In pre-processing, we lowercased all documents to reduce feature sparsity and we removed all salutations in the form of the brothers' names that might interfere with the authorship task (e.g., *W.* in a phrase like *Lieber W.*). Each modality contained the exact same set of 72 letters for which some aggregate statistics are visualized in the plots in **Figure 7**. Most letters date from the second half of the period covered, although a number of youth letters are included too. The average letter length (expressed in word length) is ≈1,832, but the lengths show a considerable standard deviation (SD) (≈1,464). Note that Wilhelm ($n = 28$) is quantitatively outnumbered by his more prolific brother ($n = 44$). Character n-grams are a state-of-the-art text representation strategy in authorship studies and allowed our models to capture fine-grained, sub-word level information (Kestemont, 2014; Sapkota et al., 2015). We vectorized the data-sets with a standard TF-IDF weighted vector space model using a character n-gram feature extraction method that took into account the 5,000 most frequent n-grams (bigrams, trigrams, and tetragrams) with a minimal document frequency of 2. Finally, we applied row-wise L1-normalization to the resulting vectorization matrix for the

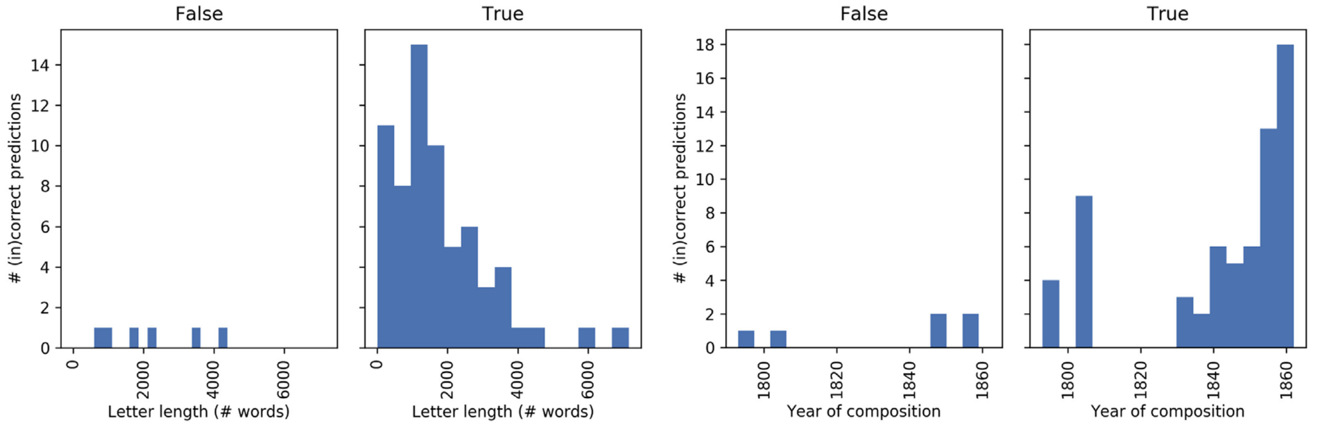


FIGURE 9 | Distribution of correct (true) and incorrect (false) predictions in the MAN data-set in terms of length, expressed in number of characters (left) and date (right).

entire data-set, followed by column-wise feature scaling as is customary in stylometry (Burrows, 2002).

4.2.2. Intra-Modality Attribution

To assess the general performance of our model, we start by reporting leave-one-out scores for the setup discussed above on each data-set independently, where each letter subset (MAN, OCR, HTR) is individually vectorized. The confusion matrices are presented in **Figure 8** and **Table 7**, which details the accuracy and F1-scores for each data-set. In general, the results are relatively high in terms of accuracy and F1, but certainly not perfect; the greedy SVM-tendency clearly over-attributes to Jacob, probably because of his relative prominence in the training data. The results for MAN and OCR are identical, whereas HTR seems to perform slightly worse already at this stage for both evaluation metrics.

Upon a closer inspection of the predictions for MAN—which should in principle be the most reliable data source—in relation to the length and date of composition of the letters, we see that misattributions are, perhaps counter-intuitively, not restricted to shorter letters (see **Figure 9**). Given the overall chronological distribution of letters, we also see that misattributions seem to occur throughout the letter subsets.

Interestingly, whereas date or length do not appear to be decisive factors at this stage, Table S1 in Supplementary Material shows how it is in fact the same letters that tend to get misattributed across the different modalities. Whereas the attributions are fully consistent for OCR and MAN, we see that HTR behaves erratically in a small number of instances. Note how the data-sets also contain a number of extremely short letters (as short as eight words), which are nevertheless correctly attributed by some models—possibly because of the bias toward Jacob.

4.2.3. Cross-Modality Attribution

In digital literary studies, researchers frequently have to conflate textual materials of different origins and mix documents which have been digitized in multiple and not necessarily compatible modalities. While this practice is clearly sub-optimal, it is also often unavoidable. To assess the impact of the digitization modality on the results of authorship attribution we turned to a cross-modality experiment. To align the data-sets in terms of feature extraction, we vectorized all three data-sets simultaneously using the previous vectorization and extracting the 5,000 most frequent character n-grams. We ran the leave-one-out validation approach in the following manner: (1) We looped over each letter in the collection (which is present in each of the three data-sets) and set its three versions (MAN, OCR, HTR) apart as held-out items; (2) next, we trained three distinct classifiers on the remaining 71 letters in each subset; (3) finally, we instructed each of the three classifiers to make an attribution for each of the three held-out samples (9 predictions in total).

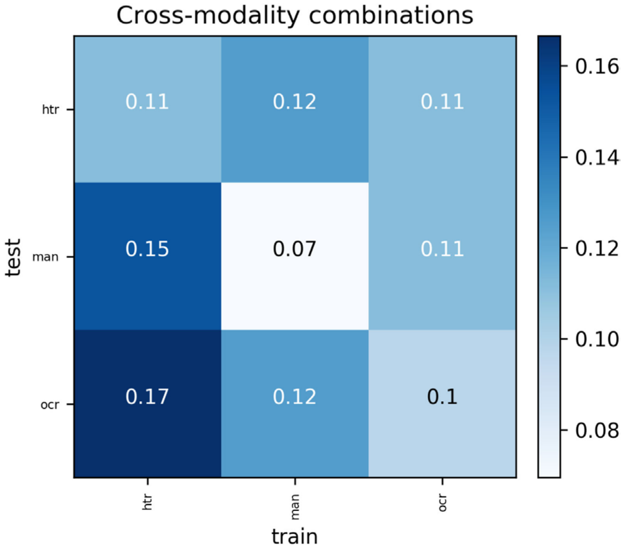
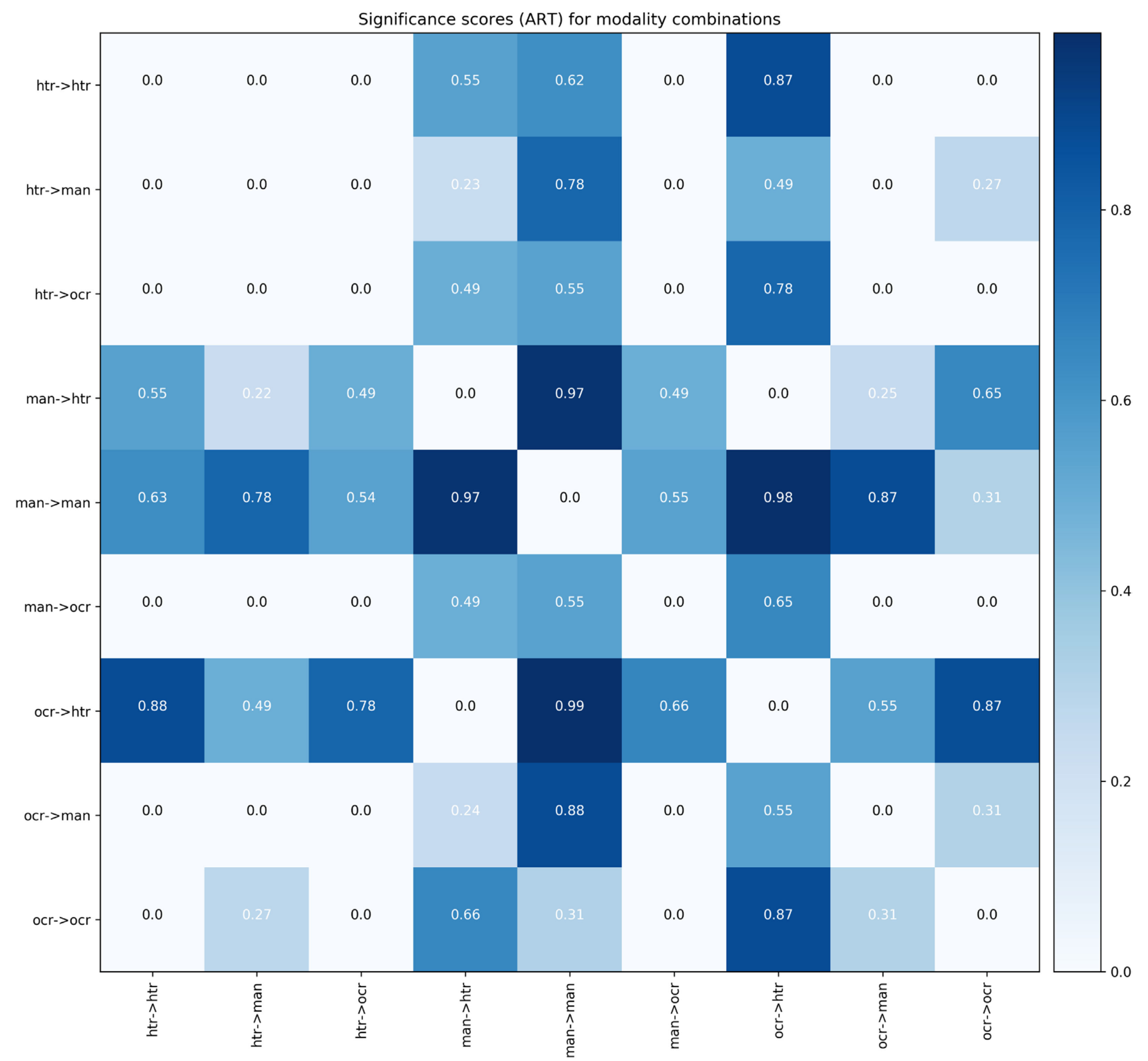


FIGURE 10 | Outcome of the cross-modality experiments, presented as a square matrix showing misclassifications (error rate).



	htr->htr	htr->man	htr->ocr	man->htr	man->man	man->ocr	ocr->htr	ocr->man	ocr->ocr
htr->htr	0.0	0.0	0.0	0.55	0.62	0.0	0.87	0.0	0.0
htr->man	0.0	0.0	0.0	0.23	0.78	0.0	0.49	0.0	0.27
htr->ocr	0.0	0.0	0.0	0.49	0.55	0.0	0.78	0.0	0.0
man->htr	0.55	0.22	0.49	0.0	0.97	0.49	0.0	0.25	0.65
man->man	0.63	0.78	0.54	0.97	0.0	0.55	0.98	0.87	0.31
man->ocr	0.0	0.0	0.0	0.49	0.55	0.0	0.65	0.0	0.0
ocr->htr	0.88	0.49	0.78	0.0	0.99	0.66	0.0	0.55	0.87
ocr->man	0.0	0.0	0.0	0.24	0.88	0.0	0.55	0.0	0.31
ocr->ocr	0.0	0.27	0.0	0.66	0.31	0.0	0.87	0.31	0.0

FIGURE 11 | Tabular representation of the scores yielded for the outputs of different modality combinations in attribution. Higher scores indicate that two systems produce significantly different outputs.

This procedure enables cross-modality performance, i.e., how well a classifier trained on one modality performs when tested on a different modality. In total, this yielded nine combinations of train-test modalities to compare (HTR→HTR, HTR→MAN, HTR→OCR, MAN→HTR, MAN→MAN, MAN→OCR, OCR→HTR, OCR→MAN, OCR→OCR), including the case where a single modality is used for training and testing (note, however, that the latter results might differ from those presented in the previous section, because of the different vectorization approach). Additionally, it is important to keep track of the directionality in these simulations, since the performance of moving from modality A in training to modality B in testing is not necessarily comparable to the inverse direction. Crucially, this setup might allow us to provide initial recommendations as to which modalities are to be preferred for the digitization and construction of training sets and test sets (which are not necessarily the same).

In the matrix in **Figure 10**, we present the outcome of this result in terms of misclassifications as a square matrix. Three observations can be made: the optimal combination is obtained when a system is trained and tested on manually transcribed data (0.07 error rate). Second, it is clear that systems trained on HTR'ed data do not perform well on either of the three test modalities

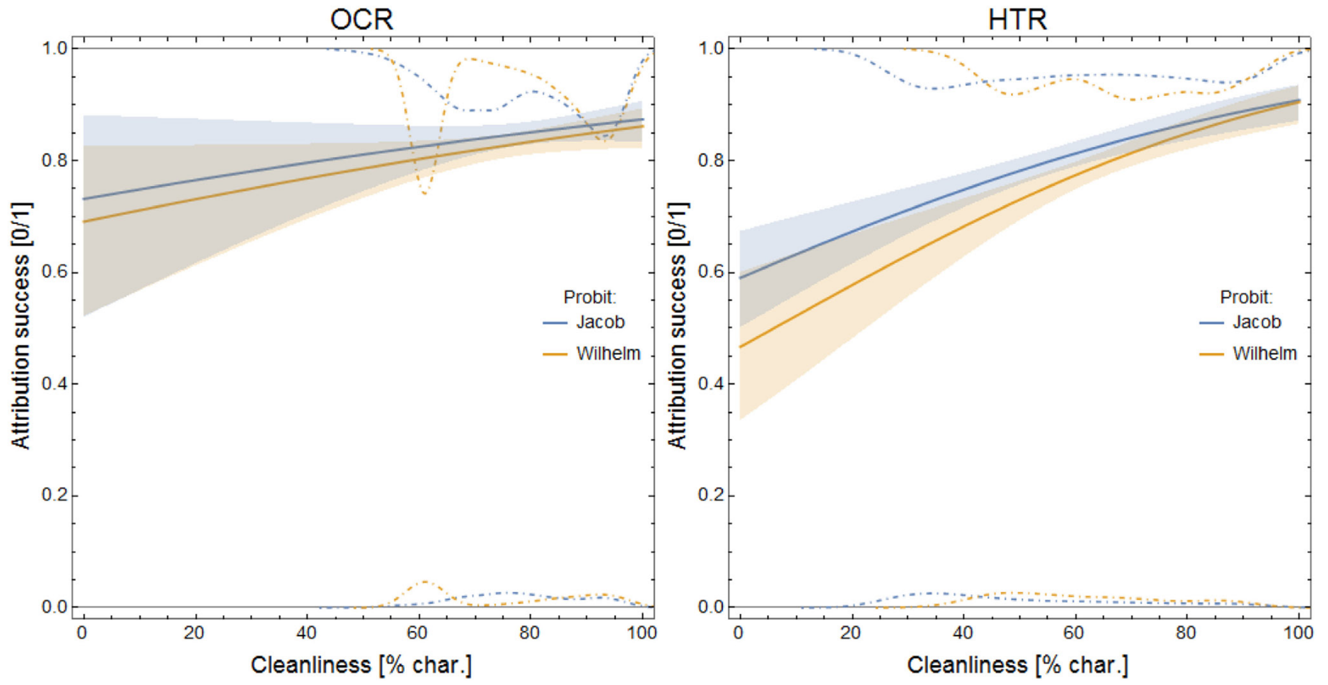


FIGURE 12 | Probit models (continuous lines) fitted to randomly sampled texts; the shaded regions are 0.95 confidence boundaries. Dot-dashed lines indicate relative number of samples attributed: (top) correctly and (bottom) erroneously. Note that the upper histograms are upside down. For instance, in the OCR case, most of Wilhelm's samples are 60 or 95% clean (peaks in the upper histogram), and at 60% cleanliness there are five times as many correct as incorrect attributions (upper relative to the lower peak height).

and especially OCR (0.17 error), hinting at an aggravated effect of combined recognition errors. Third, OCR seems to do relatively well as a training modality; counter-intuitively, a system trained on OCR and tested on HTR performs even slightly better than a system trained on MAN. These hypotheses are visually supported in the agreement in Table S2 in Supplementary Material for the individual letters. Again, we see that it is typically the same letters that are misattributed in the different cross-modality setups. Nevertheless, this table also shows that it is typically the systems that involve the HTR modality that introduce inconsistencies on this level.

The differences between the different modality combinations showed a number of valuable emerging patterns, but these differences at times also became very small. We, therefore, turned to a significance testing to assess whether the differences observed for the different modality combinations are, statistically speaking, robust. Assessing the statistical significance of the outputs of different classifiers is a debated topic in Machine Learning, especially for small data-sets like the ones under scrutiny here, where the distribution of class labels is not known or cannot be properly estimated. Because of this, we turned to an approach known as "approximate randomization testing" (Noreen, 1989). This non-parametric test is common in computational authorship studies, such as the PAN competition (Stamatatos et al., 2014), to compare the output of two authorship attribution systems, where we cannot make assumptions about the (potentially highly complex) underlying distributions. The test outputs a score that helps us to assess whether the output of two binary classification systems is statistically significant with respect to the F1-score metric. These individual scores represent the probability of failing to reject the null hypothesis ($H0$) that the classifiers do *not* output significantly different scores.

These scores are represented in a tabular format in **Figure 11**. This figure largely confirms our previous observations: systems trained on HTR perform comparatively worse and their outputs do not significantly diverge from each other. As shown by the darker rows and columns, combinations involving HTR typically invite decidedly different results from those involving (only) MAN or OCR. The lighter cells additionally justify the hypothesis that in many cases combinations involving OCR do not diverge strongly from the MAN subset. This suggests that OCR digitization at this stage can serve as a fairly reliable proxy for the more painstaking MAN digitization, at least when it comes to authorship attribution.

4.2.4. Binary Attribution vs. Text Cleanliness

The experiment reported here was designed to verify whether there is any relationship between the success of binary authorial attribution of the letters of the Grimm brothers (i.e., attributing authorship either to Jacob or to Wilhelm) and the quality of the automated text recognition (OCR or HTR). In this section, the attribution takes place across modalities, with the manual transcription as the training set and either OCR or HTR as the test set.

Instead of correlating the cleanliness (see Section 4.1.4) of the existing letters with the classification results, we decided to exploit more fine-grained information and to improve the statistical properties of the corpus by random sampling. The training set was constructed from the manually transcribed letters by random sampling individual lines, so that the resulting corpus included 72 texts randomly generated *via* an approach similar to "bag of words" (44 based on lines randomly sampled from Jacob's and 28 from Wilhelm's letters), each of at least 1,500 characters (roughly 250 words) in length. This allowed us to normalize the samples while concurrently retaining some realistic properties of the original data-set. Next, another random text (1,500 characters long) was sampled from lines of an automatically transcribed set (either OCR or HTR) in such a way that we could track its total cleanliness (ratio of correctly transcribed characters) and classify it (here, with Burrows delta and 100 most frequent words). The whole procedure (generation of training and test set and classification) was repeated for each author and automatic transcription method to reach a total of 3,800 random samples characterized by their cleanliness (0–100%) and classification result (0 or 1, for incorrect or correct attribution). These data points could then be analyzed by means of probit regression, as illustrated in **Figure 12**.

The result depends on the training set (and possibly on the relative sizes of the brothers' subsets therein). We checked that the fitted models shown in **Figure 12** hold for equally sized subsets as well. For a small number of samples (a few hundred) the results were still unreliable, but the several thousands were enough to produce a convergent, stable result. The results show that there is only a marginally significant relation between cleanliness and attribution success for OCR (at 0.95, but not at 0.99 confidence level), and there is a significant (0.99 confidence level) relation for HTR in both Wilhelm's and Jacob's writing: the cleaner their texts are, the more probable correct attribution is. It is noteworthy that cleanliness above ≈20% is already enough for HTR to have a higher-than-chance probability of correct attribution (for OCR the probability is always higher-than-chance).

CONCLUSION

This article describes the impact of digitization noise on the automatic attribution of a body of letters to Jacob and Wilhelm Grimm.

Accordingly, to test all possible digitization scenarios, we prepared three different digitization outputs to compare: (1) manual transcriptions of the original letters, (2) the OCR of a 2001 printed critical edition of the Grimm letters, and (3) an HTR model for the automatic transcription of the original letters. The manual transcription was used as a gold standard for the evaluation of the cleanliness of the OCR and the HTR output. As expected, the HTR error rate was higher than OCR due to the instability of handwriting as opposed to the uniformity of print. Nevertheless, our experiments showed that despite our imperfect data-set for HTR processing, the generated models for the Grimm brothers averaged <6% character error rate (i.e., an error every 17 characters).

Such an error rate is already high enough to significantly lower the vocabulary richness of the HTR'ed letters. Since this measure is a distinctive factor the both brothers, we also tested the adequacy of the three different digitization outputs—manual transcription, noisy OCR, and noisy HTR—for authorship attribution. What we found was that OCR digitization served as a reliable proxy for the more painstaking manual digitization, at least when it comes to authorship attribution.

Interestingly, it appears that the attribution is viable even when the training and test sets were built from differently digitized texts. With regards to HTR, our research demonstrates that even though automated transcription significantly increases the risk of text misclassification in comparison to OCR, a cleanliness of above ≈20% is already enough for it to have a higher-than-chance probability with respect to a binary attribution task (for OCR the probability is always higher-than-chance).

While still preliminary, our results add further support to the argument that absolute text cleanliness is not a major prerequisite for authorship attribution (Eder, 2015), or at least not in the case of the letters written by the Grimm brothers discussed here. Our HTR model is the first model of the Grimms' handwriting to have been produced and one that can be refined if trained on more handwritten documents (e.g., the set of professional letters currently housed in Berlin). Looking ahead, a next research avenue might be the cross-genre authorship verification of the Grimm's *Kinder und Hausmärchen* to identify, if at all possible, which (parts of the) fairy tales are more markedly Jacob or Wilhelm.

AUTHOR CONTRIBUTIONS

EF and GF are co-responsible for data acquisition and project setup. GR coordinated the manual transcription and, together with MJ, transcribed the letters. GR also jointly oversaw the OCR and HTR of the letters. MK, JO, JB, and JR designed and performed the computational analyses. GF, GR, EF, MJ, MK, and JO wrote sections of the manuscript. GF put together the first draft of the manuscript and brought it to completion. All authors contributed to manuscript revision, read and approved the submitted version.

ACKNOWLEDGMENTS

The authors wish to thank colleagues Maria Moritz, Kirill Bulert, Marco Büchler, and former colleague Linda Brandt for their valuable contribution to the project. We also wish to thank Bernhard Lauer and Rotraut Fischer of the Brüder Grimm-Gesellschaft e.V in Kassel, Germany, for their expert advice and input on the calligraphy of the Grimm brothers. We thank Dr Stephan Tulkens for his kind support in the research for Section 4.2.3, and Dr Günther Mühlberger for his crucial contribution in HTR modeling of the Grimm handwriting. Finally, special thanks go out to Professor Gerhard Lauer for his continuous guidance and support.

FUNDING

This research project was funded by the "Interessenbekundung zur Themenbesetzung im Campuslabor Digitalisierung" initiative of the University of Göttingen (Projekt: 392860 "Campuslabor"). Authors affiliated with the University of Göttingen are employed under grant no. 01UG1509 awarded by the German Federal Ministry of Education and Research (BMBF).

REFERENCES

Agarwal, S., Godbole, S., Punjani, D., and Roy, S. (2007). How much noise is too much: a study in automatic text classification. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 3–12. Omaha, NE, USA: IEEE.

Burrows, J. (2002). "Delta": a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17: 267–87. doi:10.1093/llc/17.3.267

Eder, M. (2013). Mind your corpus: systematic errors in authorship attribution. *Literary and Linguistic Computing* 28: 603–14. doi:10.1093/llc/fqt039

Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Literary and Linguistic Computing* 30: 167–82. doi:10.1093/llc/fqt066

Eder, M. (2017). Short samples in authorship attribution: a new approach. In *Digital Humanities 2017: Conference Abstracts*, 221–224. Montréal, Canada.

Fink, F., Schulz, K.U., and Springmann, U. (2017). Profiling of OCR'ed historical texts revisited. In *In Conference on Digital Access to Textual Cultural Heritage (DATeCH '17)*, 59–66. Göttingen, Germany: EACL.

Halteren, H.V., Baayen, H., Tweedie, F., Haverkort, M., and Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics* 12: 65–77. doi:10.1080/09296170500055350

Hill, M.O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology* 54: 427–32. doi:10.2307/1934352

Holmes, D.I. (1985). The analysis of literary style – a review. *The Journal of the Royal Statistical Society* 148: 328–41. doi:10.2307/2981893

Hoover, D.L. (2003). Another perspective on vocabulary richness. *Computers and the Humanities* 37: 151–78. doi:10.1023/A:1022673822140

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

Juola, P. (2008). Authorship attribution. *Foundations and Trends in Information Retrieval* 1: 233–334. doi:10.1561/1500000005

Kestemont, M. (2014). Function words in authorship attribution. From black magic to theory? In *Third Computational Linguistics for Literature Workshop*, 59–66. Gothenburg, Sweden: European Chapter of the Association for Computational Linguistics.

Kjell, B. (1994). Discrimination of authorship using visualization. *Information Processing and Management* 30: 141–50. doi:10.1016/0306-4573(94)90029-9

Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology* 60: 9–26. doi:10.1002/asi.20961

Lopresti, D. (2009). Optical character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition (IJDAR)* 12: 141–51. doi:10.1007/s10032-009-0094-8

Luyckx, K., and Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing* 26: 35–55. doi:10.1093/llc/fqq013

Noreen, E.W. (1989). *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. New York: John Wiley and Sons.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12: 2825–30.

Press, G. (2016). *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says*. Available at: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#5f6ea6516f63 (Accessed: March 17, 2018).

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

Sapkota, U., Bethard, S., Montes, M., and Solorio, T. (2015). Not all character n-grams are created equal: a study in authorship attribution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 93–102. Denver, Colorado: Association for Computational Linguistics.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys* 34: 1–47. doi:10.1145/505282.505283

Stamatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy* 21: 421–39.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology* 60(3): 538–56. doi:10.1002/asi.21001

Stamatatos, E., Daelemans, W., Verhoeven, B., Stein, B., Potthast, M., Juola, P., et al. (2014). Overview of the author identification task at PAN 2014. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014*, 877–897. Sheffield.

Subramaniam, L.V., Roy, S., Faruquie, T.A., and Negi, S. (2009). A survey of types of text noise and techniques to handle noisy text. In *The Third Workshop on Analytics for Noisy Unstructured Text Data*, 115–122. New York, USA: ACM.

Thoiron, P. (1986). Diversity index and entropy as measures of lexical richness. *Computers and the Humanities* 20: 197–202. doi:10.1007/BF02404461

Tweedie, F.J., and Baayen, R.H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32: 323–52. doi:10.1023/A:1001749303137

Vobl, T., Gotscharek, A., Reffle, U., Ringlstetter, C., and Schulz, K.U. (2014). PoCoTo – an open source system for efficient interactive postcorrection of OCRed historical texts. In *First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14)*, 57–61. Madrid, Spain: EACL.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software* 59: 1–23. doi:10.18637/jss.v059.i10

Wimmer, G., and Altmann, G. (1999). Review article: on vocabulary richness. *Journal of Quantitative Linguistics* 6: 1–9. doi:10.1076/jqul.6.1.1.4148

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at https://www.frontiersin.org/articles/10.3389/fdigh.2018.00004/full#supplementary-material.

TABLE S1 | Overview of the prediction for the three intra-modality models.

TABLE S2 | Outcome of the cross-modality experiments, showing the agreement between different combinations for individual letters.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer FB and handling Editor declared their shared affiliation.

APPENDIX

Manual transcription of Jacob Grimm's letter dating 1793.

Montag

Steinau den 7 8br 1793.

Lieber Bruder!

Du wirst hierbey dein Kleid erhalten, Wie hatt es dir denn auf der Reise gefallen, mich verlanget es zu wissen, ich erwarte mit der ersten Gelegenheit einen Brief von dir, seit deiner Abwesenheit ist nichts merkwürdiges vorgefallen.

Mein Vater hatte heute einen sehr starken Amtstag gehabt, bis Freitag wird dich unser hofjud Jud Seelig besuchen und mit diesem werde ich dir weitläufliger schreiben, Küβe der lieben Mutter dem Groβvater und jungfer Tante die Hand in meinem Nammen. Du wirst von uns allen gegrüβet, und ich bin dein treuer Bruder

Jacob Grimm