A Digital Edition between Stylometry and OCR

The *Klagenfurter Ausgabe* of Robert Musil

Simone Rebora

Abstract

*This article presents the digital edition of Robert Musil's work (*Klagenfurter Ausgabe*) and its role in a digital humanities project aimed at reconstructing Musil's activity in the WWI journal* Tiroler Soldaten-Zeitung. *First, the article reviews the ways in which the computational methods of stylometry are applied to attribute the anonymous texts published in the* Klagenfurter Ausgabe. *Second, it explores how optical character recognition (OCR) software is employed to expand the corpus. At the core of this methodology two machine learning algorithms are trained and revised using the transcriptions of the* Klagenfurter Ausgabe, *to reach an accuracy of about 99.9% in the digitization of the* Tiroler Soldaten-Zeitung *texts. The work of this project offers not only the possibility of expanding stylometric analysis to the whole journal, but also of improving the transcriptions of the* Klagenfurter Ausgabe.

Introduction. The edition that we needed.

When, in November 2009, the Robert-Musil-Institut of the University of Klagenfurt published the DVD containing the entire literary production of the Austrian author (Amann, Corino, and Fanta 2009), the event was saluted by Musil scholars as the long-awaited conclusion of an enterprise solicited by the inherently open-ended nature of Musil's work. In fact, his most important novel, *Der Mann ohne Eigenschaften*, is left unachieved with thousands of manuscript pages describing its potential conclusion. Given the extensiveness of such a manuscript legacy, scholars have traditionally concluded that it is problematic — if not impossible — to define a linear reading order for the work, "wenn der Hypertext die geeignete Form ist" (Salgaro 2014a, 8) ("when hypertext is the appropriate form").

Textual Cultures 12.2 (2019): 71–90. DOI: 10.14434/textual.v12i2.27689

The need for such a solution was felt so intensely by scholars that the history of digital editions of Musil began two decades before the Robert-Musil-Institut published the DVD in 2009. The first attempt dates to the year 1992, when a CD-ROM of Musil's *Literarische Nachlass* was issued by the publisher Rowohlt (Aspetsberger, Eibl, and Frisé 1993). However, both hardware and software limitations make it today "ein Dinosaurier der Informatik" ("a dinosaur of informatics" [Salgaro 2014a, 9]), with the typical interoperability issues that affect many of the earliest projects in digital editions. Yet while the 2009 DVD edition (also known as the *Klagenfurter Ausgabe*, from here on, *KA*) solved many of these issues, it also introduced further complications. With its tripartite hierarchical structure that juxtaposes the facsimiles of Musil's manuscripts with their transcriptions, while placing at the highest level an emended version of what the final texts might have been ("Lesetexte"), the *KA* seems to suggest "[eine] neuen Form der Nutzung, Navigation statt Lektüre ist angesagt" ("a new form of use, where navigation instead of reading is required" [Fanta 2010, 136]). Thus this new form of fruition that might be in line with the unsolvable fragmentariness of Musil's manuscript legacy also brings into light a fundamental theoretical issue. As noted by Aldo Venturelli, the risk here is that of "una *ideologizzazione* dell'edizione elettronica, che può comportare ricadute ermeneutiche da non sottovalutare. Di fatto è l'idea stessa di testo a essere messa in discussione" ("an *ideologization* of the electronic edition, which can entail hermeneutical repercussions that should not be underestimated. In fact, it is the very idea of text that is being questioned" [2010, 3]). By substituting the final version of the text with a provisional variant of it (the "Lesetext"), the very hermeneutic act of reading seems to be overpowered by the act of navigating, of interacting with an object that is no longer a text.

Together with these theoretical issues, a very practical issue arises from the fact that the software adopted to structure this extensive database, Folio Views, was originally developed for business companies and information publishers, but not for digital editions. In particular, the commercial nature of the software hindered both interoperability and content sharing, thus moving against some of the most fundamental aims of digital scholarly editing (Schmidt 2014). In addition, the DVD technology appears today as an already outdated support. To solve these problems, the most recent project of the Robert-Musil-Institut involves transforming the *KA* into an "hybrid edition", where 12 volumes (to be published by the year 2022) will host the sole "Lesetexte", while facsimiles, transcripts, and commentaries will migrate into the web portal *musilonline* following their adaptation to an

XML/TEI compliant format (Bosse et al. 2018; also http://musilonline.at/). Among the principal issues in this migration, as observed by the developers themselves, is the adaptation of "der chaotischen Struktur der FolioViews-Infobase" ("the chaotic structure of the FolioViews-Infobase" [Bosse et al. 2018, 99]), with its "zahlreichen Redundanzen, Inkonsistenzen, Fehlern und Ergänzungsbedarf" ("numerous redundancies, inconsistencies, errors, and additional requirements" [Bosse et al. 2018, 99]).

Notwithstanding its intrinsic limitations, the *KA* is still one of the most ambitious models of digital scholarly editing in Musil scholarship to date, and it has proved a powerful resource for multiple lines of research (see Salgaro 2014b, Bonacchi 2014), including one that brought together philological knowledge and computational methods to help solve one of the most complex attributive problems in Musil's production.

The Klagenfurter Ausgabe and stylometry

During the First World War, Musil fought in the Austrian army at the Italian front. Between 1916 and 1917, he was chief editor of the propagandistic journal *Tiroler Soldaten-Zeitung* (from here on, *TSZ*) in Bozen. While his role as editor is undisputed, it is an open question whether Musil also authored articles, and if so, how many.

In Musil studies, between 1960 and 2014, a growing number of articles has been attributed to the author. However, the surprising aspect of these attributions is the lack of evidence accompanying their assumptions. For example, Marie-Louise Roth lists 19 texts from *TSZ*, introducing them with the cryptic phrase, "Anonyme Schriften [. . .] die bis jetzt noch nicht mit Sicherheit identifiziert wurden" ("anonymous texts [. . .] which have not yet been identified with certainty" [Roth 1972, 528]). Subsequent studies, such as the one by Arntzen (1980), refer to Roth without highlighting the gaps in her argument. The Italian edition (Fontanari and Libardi 1987) simply includes all the texts previously indicated as Musil's production. And even the *KA* is no more accurate, since here the determination of attribution is defined as a "work in progress" (Amann, Corino, and Fanta 2009). Regina Schaunig, the author of the only monograph on Musil's activity in the *TSZ*, lists the 38 texts proposed by critics (Schaunig 2014, 356–7) and proposes 165 more for possible attribution.

In a recent series of studies, the methods and tools of stylometry have been adopted to help resolve this issue of attribution (see Herrmann et al. 2017, Salgaro et al. 2018, Rebora et al. 2019). The final goal of stylometry is as simple as it is far-reaching. Through statistical analyses of language, stylometry attempts to "measure" style, thus discerning authors' hidden "fingerprints" in a work. According to Patrick Juola (2006, 240–3), the origins of stylometry can be traced to the end of the nineteenth century, when Thomas C. Mendenhall (1887) first applied Augustus de Morgan's original theories — albeit inconclusively. While the history of stylometry has been marked by groundbreaking successes, such as Mosteller and Wallace's (1964) analysis of the Federalist Papers, epic failures, such as that of the Cusum technique by Andrew Morton (1978; cf. Holmes 1998, 114) have also occurred. The definitive affirmation of this field of research in literary studies, however, dates to the end of the twentieth century, when John F. Burrows proposed a surprisingly effective method for the attribution of authorship known from that moment on as "Delta distance" (Burrows 2002). During the last two decades, improvements have been proposed for Delta distance, but the statistical process has remained substantially the same (cf. Evert et al. 2017). Delta has proved a valid method for attributing authorship and has been applied to multiple disputes concerning contemporary blockbuster authors like J. K. Rowling (cf. Juola 2015), as well as authors like Dante and Shakespeare (see Canettieri 2016, Craig and Kinney 2009).

In the case of Musil and the *TSZ*, it has been demonstrated that a number of articles (at least ten out of the 38 proposed by Musil scholars) were more likely written by a lesser-known author, Albert Ritter, who was part of the *TSZ* editorial team (Rebora et al. 2019). The *KA* played a fundamental role in this discovery because it provided the digitized version of the texts for the stylometric analysis. It should be noted that the Österreichische Nationalbibliothek also provides an (almost) complete digitization of the *TSZ* articles (see http://anno.onb.ac.at/cgi-content/anno?aid=tsz). However, since the transcriptions were generated through Optical Character Recognition (OCR) software, they contain multiple errors and inconsistencies. Even if recent studies have demonstrated that stylometric analyses of noisy OCRed texts can be quite robust (Franzini et al. 2018), the brevity of the *TSZ* articles clearly called for the use of manually transcribed texts, namely the ones hosted in Section 11 ("Kleine Prosa") of the *KA*.

The Klagenfurter Ausgabe and OCR

As already noted, Regina Schaunig proposed a list of 165 texts (Schaunig 2014, 358–61), which may expand significantly the selection of possible candidates attributable to Musil. While this proposal has already been criticized by other scholars who have noted that many of these texts were actually plagiarized from previously-published articles (Gschwandtner 2015), further analysis of all 165 texts with specially developed stylometric methods is needed in order to verify the presence of Musil, Ritter, and other possible authors (cf. Urbaner 2001) among its pages.

Of the 43 issues of the *TSZ*, only 35 were digitized by the Österreichische Nationalbibliothek; the remaining eight were independently scanned at the Teßmann Library in Bozen. To evaluate the quality of the OCR, the *KA* transcriptions were compared with the OCRed versions provided by the Österreichische Nationalbibliothek. This operation reduced the selection to 30 texts because the remaining eight were not published in the digitized *TSZ* issues. Figure 1 provides an overview of the results.

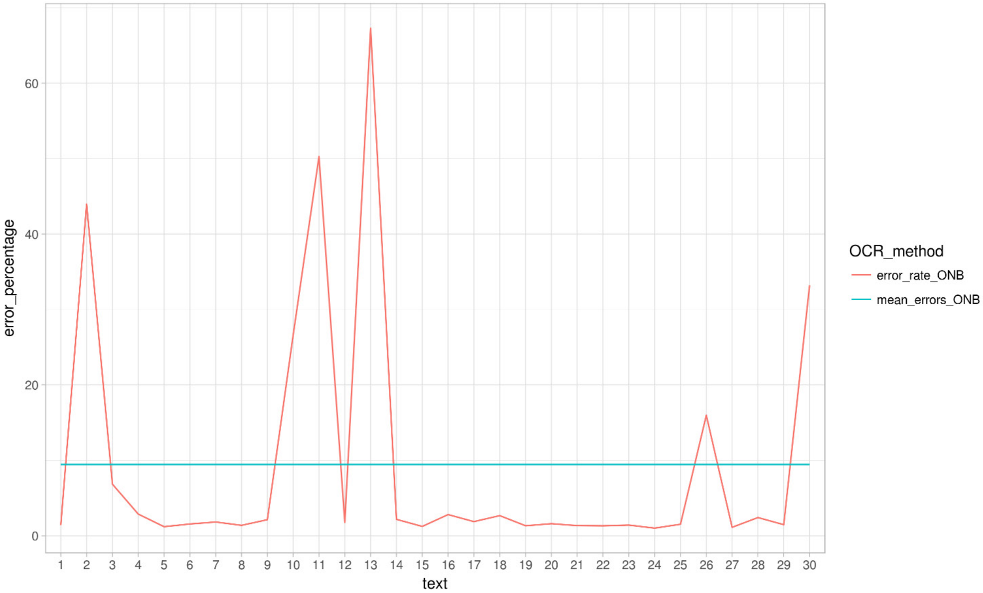


Figure 1. Character error rate for 30 OCRed *TSZ* texts.

The mean character error rate is calculated as about 10%. However, it is evident that some peaks, corresponding to specific texts, substantially increase its value. A closer analysis of the noisiest texts confirmed that these peaks issue primarily from errors in image segmentation: in many cases, the correct reading order was not respected, or text regions from different articles were incorrectly intermixed. Apart from these errors, however, the situation appeared quite promising, with a mean character error rate of 2–3%, which is generally considered as a high standard in OCR quality (Fink, Schulz, and Springmann 2017) and which may not influence significantly a stylometric analysis (Eder 2012). For these reasons, instead of proceeding with a manual transcription of the *TSZ* articles, I decided simply to re-apply the OCR process, while improving the quality of the process as much as possible.

After a consideration of the nature of the OCR errors, I selected and combined two main approaches: (1) defining a procedure for the improvement of automated page segmentation; and (2) training a machine learning algorithm to recognize the *TSZ* font, i.e. early twentieth-century Fraktur. The Österreichische Nationalbibliothek transcriptions were realized with the proprietary software *Abbyy Finereader* (see https://www.abbyy.com/finereader/). In order to avoid the restrictions associated with commercial software, I decided to select an alternative approach, which combines the functionalities of the server-based freeware *Transkribus* (Kahle et al. 2017; see also https://transkribus.eu/Transkribus/) — through which it is also possible to access the main features of *Abbyy Finereader* — with the open-source algorithms of *OCRopus/OCRopy*, a software that was developed specifically for the recognition of the Fraktur font (Breuel et al. 2013; see also https://github.com/tmbdev/ocropy).

As for the automated page segmentation, a software pipeline was implemented that combined the most efficient features of different software, including (1) the image binarization of *OCRopus/OCRopy*, which uses an adaptive thresholding approach (Shafait, Keysers, and Breuel 2008), where local anomalies such as shadows and light variations are automatically compensated; (2) the page region segmentation — and semi-automated reordering — of *Transkribus*; and (3) the automated de-skewing functionalities of *ScanTailor* (see http://scantailor.org/). The backbone of the whole pipeline was a series of R scripts that worked both on the images and on the XML/PAGE files generated by *Transkribus*. All scripts and instructions are freely accessible on *Github* (see https://github.com/SimoneRebora/page_segmentation_pipeline).

By applying the pipeline to the 30 *TSZ* texts currently under investigation, all errors in the segmentation were solved, and only character recognition errors persisted (see Fig. 2).

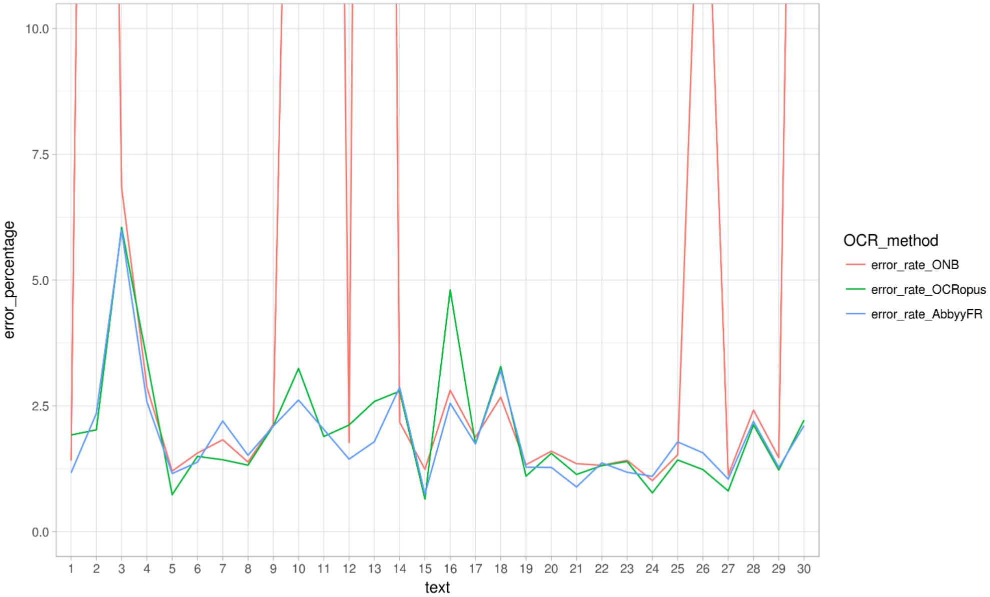


Figure 2. Comparison of character error rates for 30 OCRed TSZ texts: (1) Österreichische Nationalbibliothek versions; (2) *OCRopus/OCRopy* and (3) *Abbyy Finereader* after having applied the segmentation pipeline.

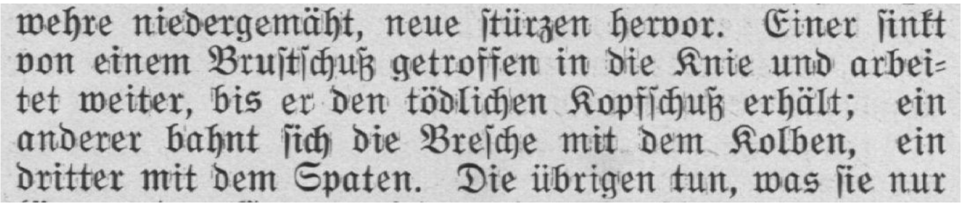
How training a machine learning algorithm can help improve a digital edition

In order to further reduce the OCR errors, the "training" functionalities of *OCRopus/OCRopy* and *Transkribus* were used to generate a model for the recognition of the *TSZ* font. Both softwares implement long short-term memory networks (Hochreiter and Schmidhuber 1997) in *OCRopus/OCRopy* and recurrent neural networks (El Hihi and Bengio 1996) in *Transkribus*. At the risk of oversimplification, it can be stated that (supervised) machine learning algorithms generally work as follows: (1) the algorithm "trains" itself on a training set, i.e. on a selection of documents that have been previously annotated by humans (in the case of OCR, these can be lines of text that have been manually transcribed); (2) the algorithm defines a "model" (i.e., a setup for its internal features) to optimize the task on which it was trained; and (3) the model is tested on a test set, i.e., on another selection of annotated documents, which have not yet been analyzed by the algorithm. If the testing produces good results, the algorithm has "learned" how to accomplish its task. As is evident from this brief explanation, "training" as an iterative process is at the core of the entire procedure: at each iteration, the algorithm analyzes one document from the training set and, if it generates an output that coincides with the human annotation (also known as "ground truth"), it simply moves on to the next document. If the output differs from the human annotation, the algorithm modifies its internal features in order to meet more closely the expected output. After a certain number of iterations, if the training has been successful, the features will converge towards a specific setup. At this stage of the process, the *KA* played once again a determinant role because it provided the transcriptions for 36 of the 38 *TSZ* articles attributed to Musil.[1] In other words, it provided both training and test set for the OCR machine learning algorithms.

According to Uwe Springmann (2015, 13), the ideal quantity of training material for *OCRopus/OCRopy* is between 1,000 and 5,000 text lines, while just one-tenth of these lines might suffice for testing. The 4,809 lines of the 38 *TSZ* articles thus seemed appropriate for training *OCRopus/OCRopy*. Before testing, however, three further adaptations were implemented to promote greater accuracy. First, transcriptions were segmented into lines based on the typographic layout of the *TSZ* articles. Second, numerous spelling normalizations were reversed: for example, the diphthongs "Ae", "Oe", and "Ue" were contracted by the *KA* transcriber into the capital letters "Ä", "Ö", and "Ü", while the abbreviation "z. B." was unfolded into "zum Beispiel" ("for example"). All these modifications had to be emended to preserve the closest possible correspondence between images and transcriptions. Third, and most important, some errors in the *KA* transcriptions were corrected, the most evident being three skipped lines and a series of misinterpreted words. For example, the passage shown in Figure 3 was transcribed in the *KA* as follows: "Einer sinkt von einem Brustschuß getroffen in die Knie und arbeitet weiter, bis er den tödlichen Kopfschuß erhält; ein dritter mit dem Spaten" ("One sinks to his knees hit by a pectoral shot and keeps working until he receives the deadly head shot; a third with the spade" [Amann, Corino, and Fanta 2009]). However, the transcription misses the connecting line: "[ein] anderer bahnt sich die

[1] The two missing texts were attributed to Musil after or separately from the publication of the *KA* (cf. Corino 2003, Corino 2010). Transcriptions were taken from Schaunig 2014.

Bresche mit dem Kolben" ("another breaks the breach with the butt [of the rifle]"). Among the misinterpreted words, see "Heeresstreifen" ("army strips") instead of "Heereskreisen" ("army circles"); "vollständig" ("completed") instead of "volkstümlich" ("popular"); "Durchführung" ("execution") instead of "Buchführung" ("accounting"). After having examined all the 38 texts, 55 errors were identified.[2]



wehre niedergemäht, neue ſtürzen hervor. Einer ſinkt von einem Bruſtſchuß getroffen in die Knie und arbeitet weiter, bis er den tödlichen Kopfſchuß erhält; ein anderer bahnt ſich die Breſche mit dem Kolben, ein dritter mit dem Spaten. Die übrigen tun, was ſie nur

Figure 3. Example of a skipped line. Source: http://anno.onb.ac.at/cgi-content/anno?aid=tsz&datum=19160726&seite=7 (accessed 15 September 2018).

With the adapted and emended transcriptions in place, the training procedure could finally start. However, this procedure generated some unexpected results. Figure 4 shows the "learning curve" of the first round of training on the entire corpus: on the x axis is the number of iterations, while on the y axis is the percentage of errors caused by the different models. For example, the model generated after 2,000 iterations caused an error rate of about 6.5% (meaning that 6.5% of the characters in the training set were incorrectly recognized), while after 4,000 iterations the error rate decreased to 3%, and so on. In an ideal setup, the error rate should decrease smoothly and reach its minimum after a certain number of iterations. However, this did not happen for the *TSZ* articles: at least two main peaks appeared at around 40,000 and 75,000 iterations, while the average quality of the models decreased substantially in the second part of the training process. This phenomenon may be caused by many factors, most of which are internal — such as "overfitting" the models (Dietterich 1995) with an excessive number of iterations —, but it may also be caused by external factors, such as inconsistent annotations. In the case of OCR, there is the

[2] For a detailed list, see https://github.com/SimoneRebora/OCRFraktur/blob/master/KA_transcriptions/KA_typos.csv (accessed 15 September 2018).

possibility that some minor errors persisted in the transcriptions, thus generating the "chain reactions" shown by Figure 4.[3]

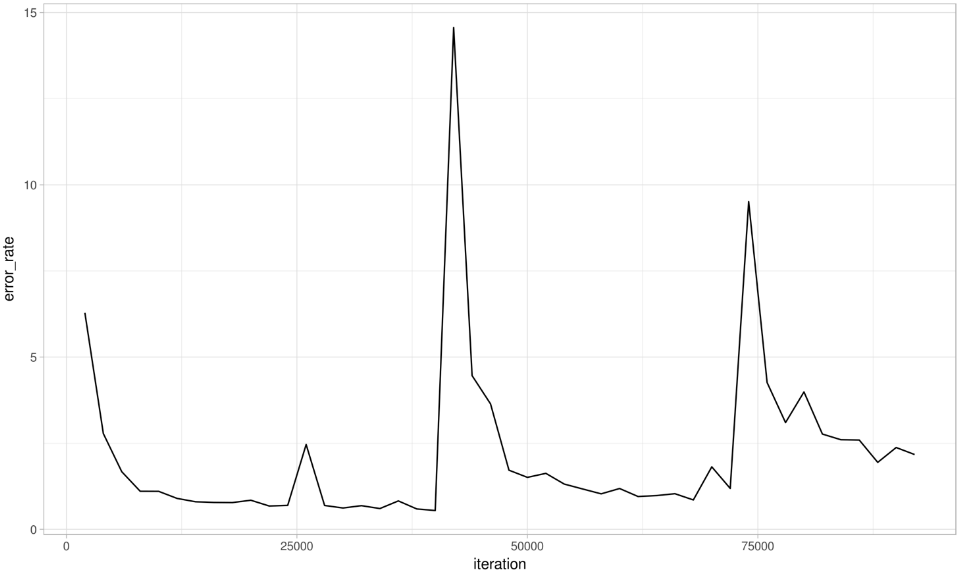


Figure 4. Learning curve of the first round of training with *OCRopus/OCRopy*.

Luckily, the training procedure can also help in identifying these errors. During the training, *OCRopus/OCRopy* produces a "log" file exemplified by Table 1, where three different outputs are generated at each iteration: the "TRU" (i.e. truth) line shows the correct (manual) transcription; the "OUT" (i.e. output) line shows the transcription produced by the software; and the "ALN" (i.e. alternative) line shows a different transcription that is generated after the software has modified its internal features. If there is an error in the transcription, there is a high probability that the "TRU" and "OUT" lines will always differ — in fact, the software would have learned

[3] It is probable that, in the attempt to adapt its output to an erroneous transcription, the algorithm produced a chain of errors in the subsequent iterations. However, machine learning algorithms are well known for the "opaqueness" of their internal processes, so it is not possible to find a unique explanation for the phenomenon.

how to correctly transcribe the text during the previous iterations on correct transcriptions and will (almost) always produce an output that differs from the incorrect transcription.

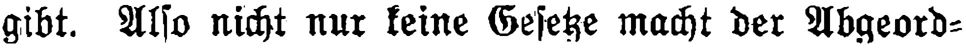
Table 1. Sample of the *OCRopus/OCRopy* log file for two iterations. The file has been edited to facilitate reading.

Iteration	Label	Output
4016		
	TRU	gleitung des Beschauers, der für diese lebenswahre er-
	OUT	gleitung des Geschauers, der für diese lebenswahre er-
	ALN	gleitung des Beschauers, der für diese lebenswahre er-
4017		
	TRU	Oh Nöraler, wer ist heute so benörgelt wie Du. oh
	OUT	Df raler, wer ist heute so benörgelt wie Du oh
	ALN	Of Nöraler, wer ist heute so benörgelt wie Du. oh

To test this assumption, a simple statistical analysis was performed on the *OCRopus/OCRopy* log file to identify the text lines that produced more inconsistencies during the training. Indeed, among the first positions in the list, it was possible to identify some more transcription errors. In this case, the mistakes were less evident, but not less significant: for example, the verb “sieht” (“looks”) was transcribed as “steht” (“stands”), the word “letzten” (“last”) appeared as “letzen” (a typo), “Lesens” (“reading”) as “Lebens” (“living”), and so on. In particular, two letters were incorrectly transcribed more than once: the letter “k”, mistaken for a long “s”, and the letter “x”, mistaken for an “r”. These errors were caused by the similarity between the characters in the Fraktur font but caused some significant changes in the meaning of sentences, when for example the pronoun “kein” (a negation) was transcribed as “sein” (a possessive). In the example shown in Figure 5 (below), the *KA* transcription reads “Also nicht nur seine Gesetze” (“Not only his laws” [Amann, Corino, and Fanta 2009]), while the correct transcription should be “Also nicht nur keine Gesetze” (“Not only no laws”).[4] The *KA* transcription of Figure 6 (below) reads “Herr Hanotaur”

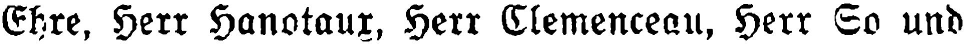
[4] In this passage, the author complains about the fact that deputies not only produce no laws but also destroy the existing laws.

(Amann, Corino, and Fanta 2009), while the correct transcription should be “Herr Hanotaux” (“Mr. Hanotaux”).[5]



gibt. Alſo nicht nur keine Geſetze macht der Abgeord=

Figure 5. Example of a mistakenly transcribed “k”. Source: http://anno.onb.ac.at/cgi-content/anno?aid=tsz&datum=19160827&seite=3 (accessed 15 September 2018). The image is the final output of the segmentation pipeline (cleaned, de-skewed, and binarized)



Ehre, Herr Hanotaux, Herr Clemenceau, Herr So und

Figure 6. Example of a mistakenly transcribed “x”. Source: http://anno.onb.ac.at/cgi-content/anno?aid=tsz&datum=19170211&seite=3 (accessed 15 September 2018)

After having repeated the procedure three times on the entire corpus (with a decreasing number of typos identified after each repetition), 29 further mistakes were corrected.[6] In addition, most of the text lines that generated the highest numbers of inconsistencies without containing typos appeared as “dirty” or poorly printed, so they were finally excluded from the training process. With a total of 4,287 lines (reinforced by 3,000 artificially-generated lines)[7] in the training set and 410 lines in the test set, *OCRopus/OCRopy* generated the learning curves shown in Figure 7, with a minimum error rate for the test set of 0.48%.

[5] This reference is to the French historian and politician Gabriel Albert Auguste Hanotaux (1853–1944).

[6] For a detailed list, see https://github.com/SimoneRebora/OCRFraktur/blob/master/KA_transcriptions/KA_typos.csv (accessed 15 September 2018).

[7] This is a procedure suggested by the *OCRopus/OCRopy* developers who generated their Fraktur model by using only artificial text lines. The function that generates these lines is included in *OCRopus/OCRopy* and was already adopted to generate hybrid training sets (composed by real plus artificial lines): see https://github.com/tmbdev/ocropy/blob/master/ocropus-linegen (accessed 15 September 2018); https://github.com/jze/ocropus-model_fraktur/ (accessed 15 September 2018).

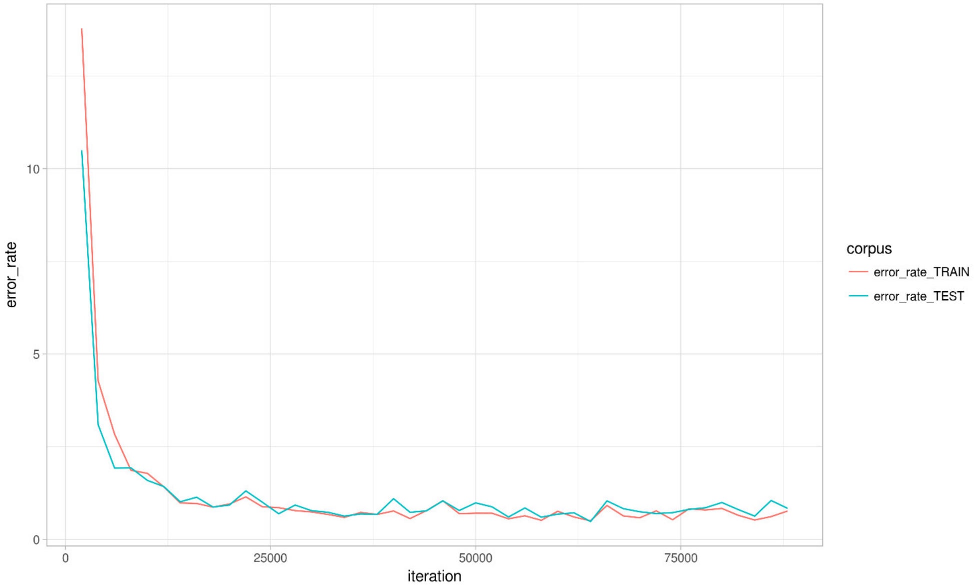


Figure 7. Learning curve of the final round of training with *OCRopus*/*OCRopy*.

The same transcriptions (excluding the artificial lines) were used to train *Transkribus*'s recurrent neural network.[8] The learning curves appeared as equally stable (see Fig. 8) and the quality of the results increased further, with a final error rate for the test set of 0.11%. This percentage represents a potentially crucial improvement when compared to the results of the untrained algorithms (see Fig. 2). However, the most important outcome of this work was the significant improvement of the transcriptions, which are now published online and made available for the future editions of the *KA* (see https://github.com/SimoneRebora/OCRFraktur/tree/master/KA_transcriptions).

[8] For this experiment, the latest (and still experimental) version of *Transkribus*'s machine learning algorithm was used. All details about its architecture are available at https://read.transkribus.eu/wp-content/uploads/2017/12/Del_D7_8.pdf (accessed 15 September 2018).

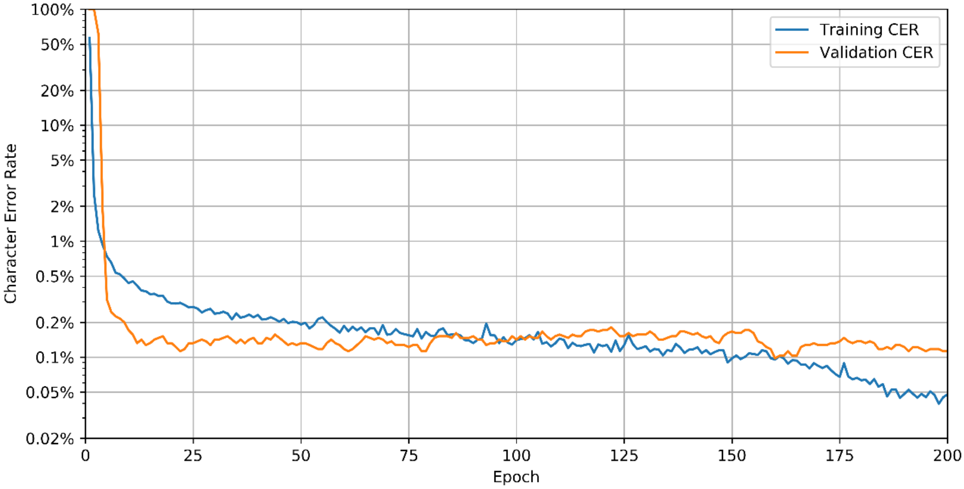


Figure 8. Learning curve with *Transkribus*. To highlight variations, the y axis is based on a logarithmic scale.

Conclusions and future perspectives

The main positive outcome of this work generated a potentially pivotal issue at the core of the whole project. In fact, the stylometric experiments were performed on the transcriptions provided by the *KA*; if these transcriptions contain a significant number of errors, then the results of the stylometric analyses might be unreliable. Mike Kestemont (2014) showed how function words (like articles, conjunctions, and prepositions) play a determinant role in stylometry-based authorship attribution. The fact that transcription errors in the *KA* concerned also pronouns such as “seine” and “keine” increased the probability of such a complication. In order to verify the validity of the results, the final experiment (“simplified design”) in Rebora et al. (2019) was repeated with the emended transcriptions. The results did not change substantially (cf. Fig. 9), while the level of confidence for some attributions was even increased (see the decreased p-values for texts no. 11 and 28 in Fig. 9b). This result confirmed the robustness of stylometric methods with (slightly) noisy texts, as already suggested by Eder (2012) and Franzini et al. (2018).

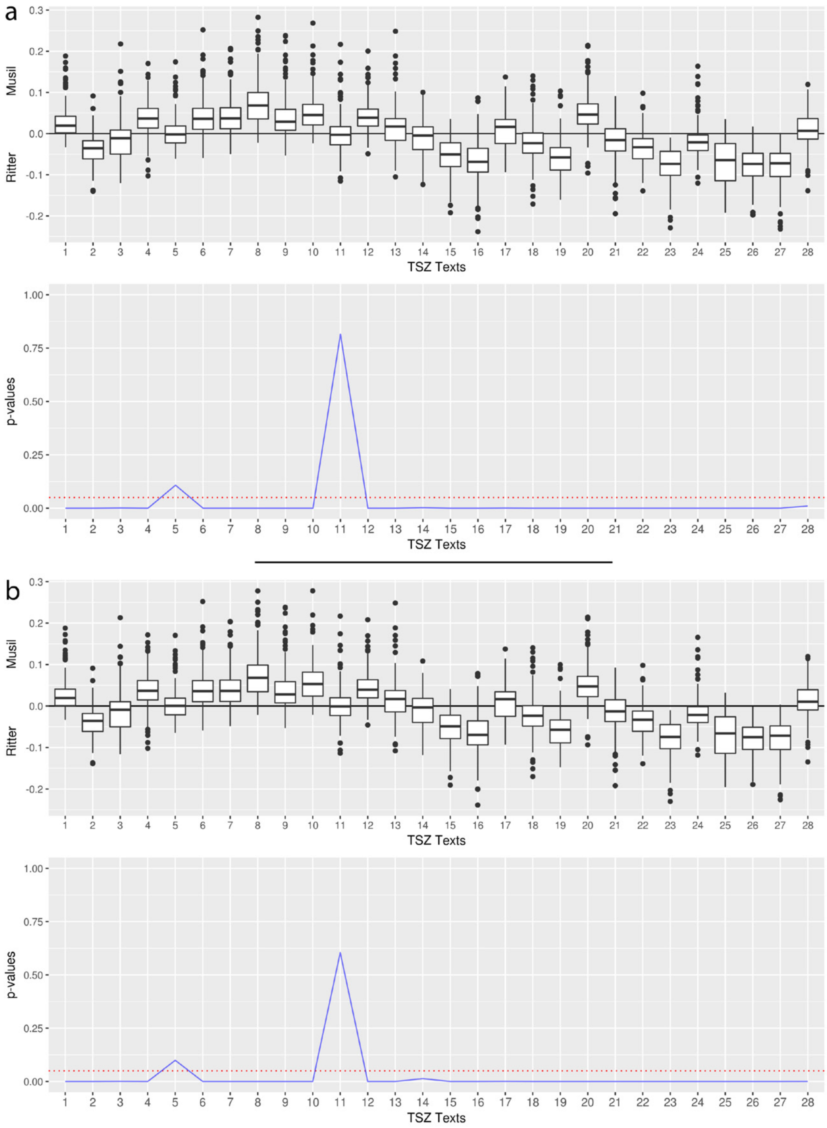


Figure 9. Comparison of the stylometric analyses of 28 TSZ texts* based on the *KA* transcriptions (a) and on the emended transcriptions (b).

* Ten texts were excluded from the stylometric analysis: nine because they were too short (under 500 words) and one because it had already been attributed to Musil on the basis of philological proof (cf. Corino 1973). For each text, a total of 160 measurements was performed: the box plot graph indicates if the text has been attributed to Musil or Ritter by the majority of the classifiers; the p-values indicate the levels of confidence for the attributions (where a low value means a high confidence).

The present work has developed the tools and resources for a significant extension of the research on Robert Musil's activity in the *TSZ*. By combining the segmentation pipeline and the machine learning algorithms of *OCRopus/OCRopy* and *Transkribus*, a digitization of the entire *TSZ* with an accuracy close to 99.9% can be generated. This is not, of course, the kind of material that can be directly integrated in a digital scholarly edition, but it is indeed a dataset that can have a high relevance for Musil studies in general. Much more extensive and detailed work should be dedicated to the analysis of this dataset to verify not only Musil's authorship of further articles, but also to verify the actual involvement of Albert Ritter, a still understudied author who might have inspired at least two characters in *Der Mann ohne Eigenschaften*, in the *TSZ*.[9]

Apart from these very practical outcomes, a fundamental methodological acquisition issues from this work. Although the current skepticism many scholars hold regarding the indiscriminate use of computational methods in the study of literature (Tomasin 2017) is still warranted, it is also true that the potential of such resources cannot be easily dismissed, especially when it offers the opportunity of observing well-known phenomena from a new, still unexplored perspective (Hammond 2017). In the case of digital editions, the expert eye of the editor cannot be substituted for the cold intelligence of algorithms, but — as this study might have demonstrated — it can and should be supported by such devices because no intelligence is infallible, be it human or artificial, while an open and critical confrontation between the two might actually lead towards an unexpected and unprecedented growth in knowledge.

[9] For a first introduction to Ritter, cf. Salgaro 2018. Ritter is explicitly cited in *Der Spion* (one of the preparatory works for *Der Mann ohne Eigenschaften* written by Musil between 1918 and 1922). According to Walter Fanta, the word "Spion" refers to the verb "Spähen" ("to scout"), to research the reasons that led to the WWI conflict (Fanta 2000, 138). Musil wanted to depict here the most representative human types of his era, and he acknowledged that it was necessary to "auch einen Alldeutschen zeichnen, der nicht überrascht wird. Zum Beispiel Ritter" ("draw also a Pangermanist, one that cannot be surprised. For example, Ritter" [Amann, Corino, and Fanta 2009]). This reference to the "Alldeutschen" (who wanted to reunite the German-speaking countries under the leadership of Prussia) connects directly to the character of Gerda Fischel in *Der Mann ohne Eigenschaften* and may also be an anticipation of the character of Hans Sepp (cf. Fanta 2000, 236). Apart from these preliminary notes, the subject requires more extensive research.

Acknowledgments

The research was made possible by an FSE fellowship at the Universities of Verona and Göttingen and is part of the project “Robert Musil, chief editor of the ‘Journal of the soldiers’ from the perspective of the Digital Humanities” (University of Verona, 02–2017/02–2019). I thank Massimo Salgaro (University of Verona) for his experienced support as a Germanist and Musil scholar, Johannes Andresen (director of the Teßmann library in Bosen) for providing the TSZ original copies, and Günter Mühlberger (University of Innsbruck) for granting me access to the most advanced functionalities of the *Transkribus* software.

Works Cited

Amann, Klaus, Karl Corino, and Walter Fanta. 2009. *Robert Musil: Klagenfurter Ausgabe: Kommentierte Edition Sämtlicher Werke, Briefe Und Nachgelassener Schriften, Mit Transkriptionen Und Faksimiles Aller Handschriften*. Klagenfurt: Robert Musil-Institut, Alpen-Adria Universität Klagenfurt.

Arntzen, Helmut. 1980. *Musil-Kommentar sämtlicher zu Lebzeiten erschienener Schriften ausser dem Roman “Der Mann ohne Eigenschaften”*. München: Winkler.

Aspetsberger, Friedbert, Karl Eibl, and Adolf Frisé, eds. 1993. *Robert Musil. Der literarische Nachlass*. Reinbek bei Hamburg: Rowohlt.

Bonacchi, Silvia. 2014. “Robert Musil’s Dissertation „Beitrag zur Beurteilung der Lehren Machs“ im Lichte der Klagenfurter Ausgabe”. In *Robert Musil in der Klagenfurter Ausgabe Bedingungen und Möglichkeiten einer digitalen Edition*, edited by Massimo Salgaro, 135–54. München: W. Fink.

Bosse, Anke, Walter Fanta, Katharina Godler, Gerrit Brüning, and Artur Boelderl. 2018. “Musilonline - Integral Lösen. Dialogfeld Digitale Edition”. In *DHd 2018 Konferenzabstracts*, 98–100. Köln: DHd. http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf.

Breuel, Thomas M., Adnan Ul-Hasan, Mayce Ali Al-Azawi, and Faisal Shafait. 2013. “High-Performance OCR for Printed English and Fraktur Using LSTM Networks”. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference On*, 683–87. IEEE.

Burrows, John. 2002. “‘Delta’: A Measure of Stylistic Difference and a Guide to Likely Authorship”. *Literary and Linguistic Computing* 17 (3): 267–87.

Canettieri, Paolo. 2016. “Chi Non Ha Scritto Il Fiore”. In *Sulle Tracce Del Fiore*, edited by Natascia Tonelli, 121–34. Firenze: Le Lettere.

Corino, Karl. 1973. “Robert Musil, Aus Der Geschichte Eines Regiments”. *Studi Germanici* 11: 109–15.

———. 2003. *Robert Musil: eine Biographie*. Reinbek bei Hamburg: Rowohlt.

———. 2010. “Klaviersonnen Über Schluchten Des Gemüts. Robert Musil Und Die Musik”. *Das Plateau* 120: 4–21.

Craig, Hugh, and Arthur F. Kinney. 2009. *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Dietterich, Tom. 1995. “Overfitting and Undercomputing in Machine Learning”. ACM *Comput. Surv.* 27 (3): 326–27: https://doi.org/10.1145/212094.212114.

Eder, Maciej. 2012. “Mind Your Corpus: Systematic Errors in Authorship Attribution”. In *Digital Humanities 2012: Conference Abstracts, (Hamburg, Grmany)*, 181–85. Hamburg: Hamburg University Press: https://sites.google.com/site/computationalstylistics/preprints/m-eder_mind_your_corpus.pdf?attredirects=0.

El Hihi, Salah, and Yoshua Bengio. 1996. “Hierarchical Recurrent Neural Networks for Long-Term Dependencies”. In *Advances in Neural Information Processing Systems 8 (NIPS 95)*, edited by David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, 493–99. Cambridge: MIT Press.

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. “Understanding and Explaining Delta Measures for Authorship Attribution”. *Digital Scholarship in the Humanities* 32 (suppl_2): ii4–ii16: https://doi.org/10.1093/llc/fqx023.

Fanta, Walter. 2000. *Die Entstehungsgeschichte des „Mann ohne Eigenschaften“ von Robert Musil*. Böhlau: Wien.

———. 2010. “Robert Musil–Klagenfurter Ausgabe”. *Editio* 24: 117–48.

Fink, Florian, Klaus U. Schulz, and Uwe Springmann. 2017. “Profiling of OCR’Ed Historical Texts Revisited”. In *Proceedings of the 2Nd International Conference on Digital Access to Textual Cultural Heritage*, 61–66. New York, NY, USA: ACM: https://doi.org/10.1145/3078081.3078096.

Fontanari, Alessandro, and Massimo Libardi. 1987. *La guerra parallela*. Trento: Reverdito.

Franzini, Greta, Mike Kestemont, Gabriela Rotari, Melina Jander, Jeremi K. Ochab, Emily Franzini, Joanna Byszuk, and Jan Rybicki. 2018. “Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm”. *Frontiers in Digital Humanities* 5: https://doi.org/10.3389/fdigh.2018.00004.

Gschwandtner, Harald. 2015. “In Der Sperrgewalt Der Fackel? Karl Kraus, Robert Musil Und Die ‘Tiroler Soldaten-Zeitung’”. *Musil-Forum* 34: 157–76.

Hammond, Adam. 2017. “The Double Bind of Validation: Distant Reading and the Digital Humanities”Trough of Disillusionment’”. *Literature Compass* 14 (8): e12402.

Herrmann, J. Berenike, Gerhard Lauer, Simone Rebora, and Massimo Salgaro. 2017. “Short Texts in Authorship Attribution. The Case of Robert Musil’s War Articles”. In *AIUCD 2017 - Book of Abstracts*, edited by Fabio Ciotti and Gianfranco Crupi, 50–56: https://doi.org/10.6092/unibo/amsacta/5885.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. “Long Short-Term Memory”. *Neural Computation* 9 (8): 1735–80: https://doi.org/10.1162/neco.1997.9.8.1735.

Holmes, David I. 1998. “The Evolution of Stylometry in Humanities Scholarship”. *Literary and Linguistic Computing* 13 (3): 111–17: https://doi.org/10.1093/llc/13.3.111.

JUOLA, Patrick. 2006. "Authorship Attribution". *Foundations and Trends in Information Retrieval* 1 (3): 233–334.

———. 2015. "The Rowling Case: A Proposed Standard Protocol for Authorship Attribution". *Digital Scholarship in the Humanities* 30 (suppl. 1): 100–13: https://doi.org/10.1093/llc/fqv040.

KAHLE, Philip, Sebastian COLUTTO, Günter HACKL, and Günter MÜHLBERGER. 2017. "Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents". In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 4:19–24: https://doi.org/10.1109/ICDAR.2017.307.

KESTEMONT, Mike. 2014. "Function Words in Authorship Attribution. From Black Magic to Theory?". In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)* 59–66. Gothenburg, Sweden: Association for Computational Linguistics: http://aclweb.org/anthology/W/W14/W14-0908.pdf.

MENDENHALL, Thomas C. 1887. "The Characteristic Curves of Composition". *Science* 9 (214): 237–46: https://doi.org/10.1126/science.ns-9.214S.237.

MORTON, A. Q. 1978. *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents*. New York: Scribner's.

MOSTELLER, Frederick, and David L. WALLACE. 1964. *Inference and Disputed Authorship: The Federalist*. Reading, Mass.: Addison-Wesley.

REBORA, Simone, J. Berenike HERRMANN, Massimo SALGARO, and Gerhard LAUER. 2019. "Robert Musil, a War Journal, and Stylometry: Tackling the Issue of Short Texts in Authorship Attribution". *Digital Scholarship in the Humanities*, 34(3): 582–605: https://doi.org/10.1093/llc/fqy055.

ROTH, Marie Louise. 1972. *Robert Musil: Ethik und Ästhetik : zum theoretischen Werk des Dichters*. München: Paul List.

SALGARO, Massimo. 2014a. "Einleitung". In *Robert Musil in der Klagenfurter Ausgabe Bedingungen und Möglichkeiten einer digitalen Edition*, edited by Massimo SALGARO, 7–26. München: W. Fink.

———. 2014b. "Musils Rezeptionsästhetik im Spiegel der Klagenfurter Ausgabe". In *Robert Musil in der Klagenfurter Ausgabe Bedingungen und Möglichkeiten einer digitalen Edition*, edited by Massimo SALGARO, 111–34. München: W. Fink.

———. 2018. "Albert Ritter, der ghostwriter in der Redaktion der Tiroler Soldatenzeitung — ein biographisches Profil". In *Landsturm-Oberleutnant Dr. Robert Musil als Redakteur der Tiroler Soldatenzeitung*, edited by Mariaelisa DIMINO, Elmar LOCHER, and Massimo SALGARO, [in press]. Paderborn: Fink Verlag.

SALGARO, Massimo, Simone REBORA, Gerhard LAUER, and J. Berenike HERRMANN. 2018. "The Tiroler Soldaten-Zeitung and Its Authors. A Computer-Aided Search for Robert Musil". In *DHd 2018 Konferenzabstracts*, 315–20. Köln: DHd: http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf.

SCHAUNIG, Regina. 2014. *Der Dichter im Dienst des Generals: Robert Musils Propagandaschriften im Ersten Weltkrieg*. Klagenfurt; Wien: Kitab.

SCHMIDT, Desmond. 2014. "Towards an Interoperable Digital Scholarly Edition". *Journal of the Text Encoding Initiative* 7: https://doi.org/10.4000/jtei.979.

SHAFAIT, Faisal, Daniel KEYSERS, and Thomas M. BREUEL. 2008. "Efficient Implementation of Local Adaptive Thresholding Techniques Using Integral Images". In *Proceedings of the 15th Document Recognition and Retrieval Conference*, edited by Berrin A. YANIKOGLU and Kathrin BERKNER, 6815:681510: https://doi.org/10.1117/12.767755.

SPRINGMANN, Uwe. 2015. "Ocrocis. A High Accuracy OCR Method to Convert Early Printings into Digital Text. A Tutorial": http://cistern.cis.lmu.de/ocrocis/tutorial.pdf.

TOMASIN, Lorenzo. 2017. *L'impronta digitale: cultura umanistica e tecnologia*. Roma: Carocci.

URBANER, Roman. 2001. ". . . Daran Zugrunde Gegangen, Dass Sie Tagespolitik Treiben Wollte? Die '(Tiroler) Soldaten-Zeitung' 1915–1917". *EForum ZeitGeschichte* 3.4: http://www.eforum-zeitgeschichte.at/3_01a8.html.

VENTURELLI, Aldo. 2010. "Robert Musil, Klagenfurter Ausgabe". *Osservatorio Critico Della Germanistica* 12 (31): 1–4.