Methodenbeitrag: Digitale Manuskriptanalyse

Jan Horstmann [1]

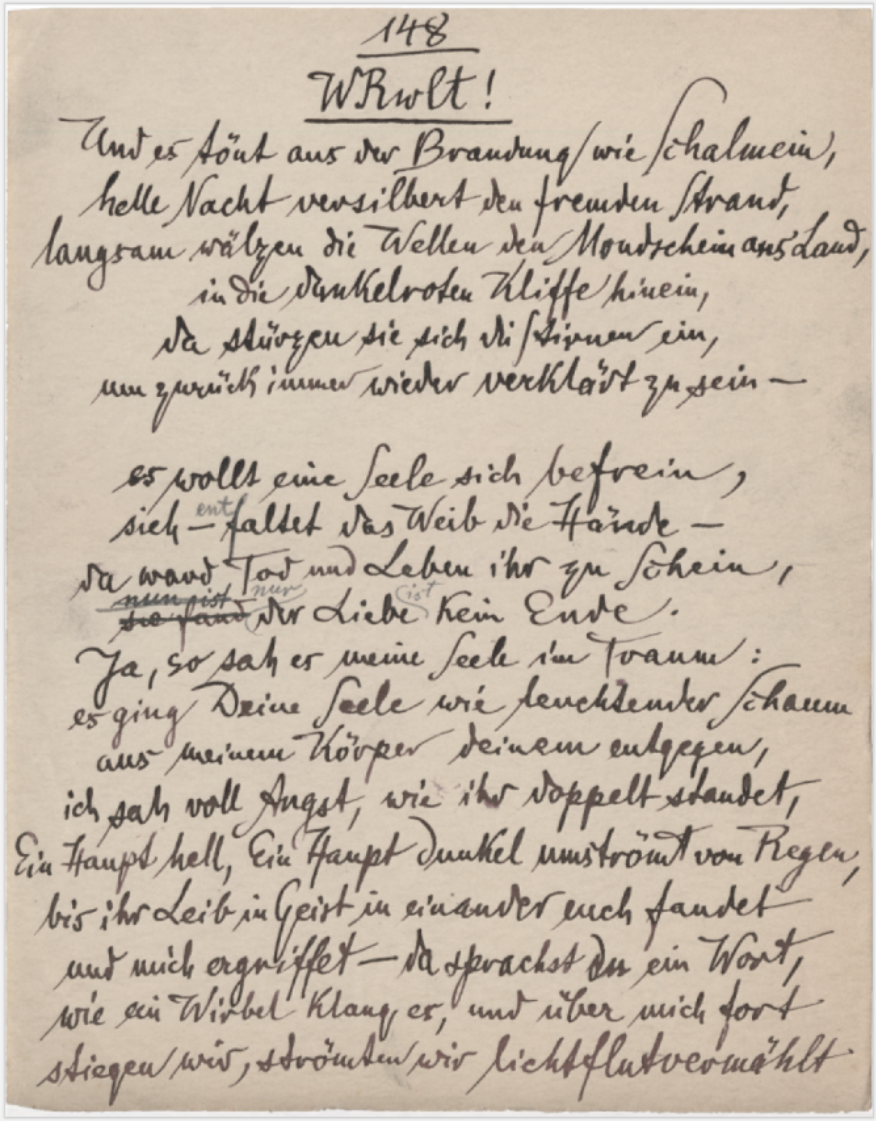
1. Universität Münster

forTEXT

Thema:	Textdigitalisierung und Edition	DOI:	10.48694/fortext.3744
Jahrgang:	1	Ausgabe:	3
Erscheinungsdatum:	12-06-2024	Erstveröffentlichung:	2018-07-12 auf fortext.net
Lizenz:	CC BY SA		open access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

1. Definition



148

WRwlt!

Und es tönt aus der Brandung wie Schalmein,
helle Nacht versilbert den fremden Strand,
langsam wälzen die Wellen den Mondschein ans Land,
in die dunkelroten Klippe hinein,
da stürzen sie sich in Strömen ein,
um zurück immer wieder verklärt zu sein —

es wollt eine Seele sich befrein,
sich — entfaltet das Weib die Hände —
da ward Tod und Leben ihr zu Schein,
~~nun ist~~ nur Liebe ist kein Ende.
Ja, so sah es meine Seele im Traum:
es ging Deine Seele wie leuchtender Schaum
aus meinem Körper deinem entgegen,
ich sah voll Angst, wie ihr doppelt standet,
Ein Haupt hell, Ein Haupt dunkel umströmt von Regen,
bis ihr Leib in Geist in einander euch fandet
und mich ergriffet — da sprachst du ein Wort,
wie ein Wirbel klang es, und über mich fort
stiegen wir, strömten wir lichtflutvermählt

Staats- und Universitätsbibliothek Hamburg, Richard-Dehmel-Archiv, Mark Emanuel Amtstätter (CC BY-SA 4.0)

Die digitale Manuskriptanalyse beschäftigt sich mit der Auszeichnung bzw. Annotation kultureller Artefakte in Form eingescannter Handschriften. Sollten diese Bilddigitalisate aufgrund schlechter Lesbarkeit oder individueller und uneinheitlicher Handschrift nicht für eine automatisierte Texterkennung (vgl. Möglichkeiten der Textdigitalisierung (Horstmann 2024a)) in Frage kommen, werden die Manuskripte als Bilddateien gespeichert und als solche ausgezeichnet.

2. Anwendungsbeispiel

Haben Sie beispielsweise vor, ein Archiv von 50.000 Briefen Martin Bubers zu digitalisieren, könnten Sie nach dem Einscannen der Manuskripte (für das es unterstützende Workflows gibt, s. u.) entweder eine automatisierte Handschriftenerkennung anstreben, um den Inhalt der Texte digital analysierbar zu machen, oder die Briefscans in ihrem Zustand als Bilddateien belassen und als solche mit zusätzlichen Informationen (sog. Metadaten) bestücken. Die Briefe wären im letzteren Fall dann zwar nicht als Volltexte, wohl aber auf der Ebene der Metadaten digital analysierbar.

3. Literaturwissenschaftliche Tradition

Alle Textwissenschaften und Teilbereiche der Kulturwissenschaften beschäftigen sich mit dem Lesen, Annotieren, Analysieren oder Interpretieren von geschriebenen oder verschriftlichten Texten. Häufig werden dafür diese Texte übertragen (z. B. auf Papier) und nicht direkt auf dem Original bearbeitet – schon gar nicht, wenn es sich um historisch wertvolle Manuskripte handelt, die z. B. auf Pergament oder Papyrus verfasst sind, oder um Grabinschriften, Münzgravuren, etc. Während ein Manuskript als (autografes) Blatt vor allem Gegenstand materialwissenschaftlicher Betrachtungen wird, ist es als (allografe) Seite interessant für literaturwissenschaftliche Lektüren und Analysen (Benne 2015).

Bereits durch das Lesen eines Textes wird dieser funktional von seinem Trägermaterial unabhängig gemacht und dadurch immaterialisiert. Der Computer kann dies jedoch nicht implizit machen, sondern muss explizit lernen, wie dieser Schritt funktioniert (Rehbein 2017, 181). Der Text als immaterielle Größe ist deshalb in literaturwissenschaftlicher Hinsicht wie beim alltäglichen Lesen das Ziel der Digitalisierung. Dennoch vernachlässigen auch Literaturwissenschaftler*innen den autografen Aspekt nicht gänzlich, denn häufig können Informationen über die Schrift und das verwendete Material der literaturwissenschaftlichen Interpretation wesentliche Impulse geben. Nicht zuletzt für die Zitierbarkeit eines Textes ist die Angabe der jeweiligen Seite, auf der eine Textpassage zu finden ist, ein wichtiges Kriterium und sollte im Zuge der Umwandlung in einen elektronischen Text nicht verloren gehen.

Wie für die Möglichkeiten der Textdigitalisierung (Horstmann 2024a) allgemein lassen sich somit auch für die Manuskriptanalyse im Speziellen drei literaturwissenschaftliche Traditionslinien aufzeigen: (1) die Editionsphilologie und Textkritik, (2) die Paläografie und (3) das Setzen von Manuskripten seit der Erfindung des Buchdrucks.

4. Diskussion

Besonders in der Lehre bietet es sich an, nicht das Originalmanuskript, sondern ein Surrogat zu verwenden, um Zeit und Kosten zu sparen und um das Originalmanuskript zu schonen. Als Digitalisat lassen sich Manuskripte verlustfrei kopieren und vervielfältigen – ein Vorteil, den die klassische analoge Reproduktion nicht bietet (Rehbein 2017, 179). Die einzelnen Digitalisate können untereinander vernetzt und so in vielfältige Beziehungen zueinander gesetzt werden.

Ist ein Manuskript aufgrund seines Zustandes oder einer problematischen Handschrift nicht mit digitalen Methoden zu erfassen (d. h. vom Computer in einen elektronischen Text umzuwandeln), oder ist der Aufwand eines *Keyings* (vgl. Keying) (d. h. des manuellen Transkribierens/Abtippens) im Rahmen des jeweiligen Projektes nicht zu rechtfertigen, können einige digitale Methoden die Arbeit mit den Texten zumindest unterstützen. In diesem Fall bleiben die eingescannten Manuskripte als Bilddateien gespeichert und werden als solche annotiert (vgl. Annotation) bzw. mit Metadaten versehen. Eine solche Operation lässt sich z. B. im Laboratory von TextGrid (Horstmann 2024c) ausführen.

Wenn Sie beispielsweise in den einzelnen Dokumenten der gesammelten Briefe der Familie Mann und etwaigen Gegenkorrespondenzen jeweils die Korrespondenzpartner*innen oder auch die erwähnten Personen und Orte als Metadaten speichern, lässt sich die gesamte Textsammlung anhand dieser strukturellen Metadaten jeweils neu sortieren. Zudem kann eine Visualisierung als Netzwerk (Textvisualisierung (Horstmann und Stange 2024), Netzwerkanalyse (Schumacher 2024)) etwa Einblicke in die globale Vernetzung der Schriftstellerfamilie gewähren und somit wertvolle Beiträge zur Diskussion über Welt- oder europäische Literatur leisten.

Die Paläografie (die Lehre von alten Schriften) gilt derzeit als ein Arbeitsbereich für Expert*innen. Die Zusammenarbeit mit Fachleuten aus dem Bereich der Computer Vision verspricht jedoch, im Zuge einer Automatisierung des Lesens von Handschriften zunehmend auch Nicht-Expert*innen (vgl. Domäneadaption) die textanalytische Arbeit mit Handschriften – deren Wert als Quellen des Wissens über die menschliche Kultur, Gesellschaft, Geschichte und nicht zuletzt regionale, nationale und übernationale Identität nicht genug betont werden kann – zu ermöglichen (Hassner u. a. 2014).

Einen wichtigen Schritt in diese Richtung geht beispielsweise das EU-geförderte Transkribus-Projekt (Horstmann 2024b). Gerade durch seine Aufspaltung in einen *simple mode* und einen *expert mode* und seine grafische Nutzeroberfläche (vgl. GUI) ermöglicht es auch Einsteigern in der automatischen Handschriftenerfassung eine digitale Arbeit mit Manuskripten auf hohem Niveau. Entwicklungen im Bereich der automatischen Handschriftenerkennung werden beispielsweise diskutiert in (Sánchez u. a. 2015; Sánchez u. a. 2014; Sánchez u. a. 2016).

5. Technische Grundlagen

Haben Sie ein eingescanntes Manuskript als Datei vorliegen, testen Sie zunächst, ob es sich um eine Bilddatei oder ein computerlesbares Dokument handelt, indem Sie versuchen, einzelne Zeilen des Dokumentes zu markieren. Wird die gesamte Seite blau/ausgewählt, handelt es sich um eine Bilddatei (d. h. die Schrift ist für den Computer noch nicht lesbar); werden nur die von Ihnen markierten Zeilen blau bzw. ausgewählt, hat bereits eine OCR (*optical character recognition*) bzw. eine HTR (*handwritten text recognition*) stattgefunden (d. h. der Text des jeweiligen Dokumentes ist digitalisiert und als elektronischer Text vom Computer les- und analysierbar).

OCR und HTR sind sich sehr ähnlich, nur dass HTR nicht auf der Erkennung einzelner Buchstaben, sondern gesamter Wörter bzw. Zeilen basiert, da wir es bei Manuskripten in der Regel mit Schreibschriften zu tun haben, bei denen die einzelnen Buchstaben ineinander übergehen und je nach Vorgänger- oder Folgebuchstabe anders aussehen können.

Es gibt neben der OCR bzw. HTR etliche Möglichkeiten, Manuskripte in Digitalisierungsprojekten zu bearbeiten, wie z. B. eine forensische Dokumentanalyse, eine Quantifizierung schriftlicher „Fingerabdrücke“, metrische Analysen, der Einsatz von DNA-Analysemethoden oder Techniken multispektraler Bilddigitalisierung (Hassner u. a. 2014, 18).

Im Zuge einer Bilddigitalisierung findet in technischer Hinsicht folgender Prozess statt: Aus dem analogen optischen Signal eines Manuskriptes wird zunächst eine Rastergrafik erstellt, die durch Bildgröße und Farbtiefe charakterisiert ist. Dafür wird jedem Punkt eines Bildes ein bestimmter Wert zugeordnet. Bei Schwarz-Weiß-Bildern ist das eine 0 für jeden weißen Punkt und eine 1 für jeden schwarzen Punkt. Schwarzweißbilder haben nur diese beiden Helligkeitswerte, komplizierter wird es bei Graustufen- oder gar Farbdigitalisaten. Bei Graustufen wird der Farbkanal schwarz/weiß genauer ausdifferenziert, bei farbigen Scans kommen mehrere Farbkanäle hinzu wie z. B. Rot, Grün und Blau im weit verbreiteten RGB-Modell, deren Kombinationen eine riesige Palette an Mischfarben erzeugen können (Rehbein 2017, 182).

Ob die korrekte Wiedergabe von Farben im Digitalisat wichtig ist, ist projektspezifisch. Geht es ausschließlich darum, den Text eines Manuskriptes computerlesbar zu machen, unabhängig davon, ob einige Passagen beispielsweise in einer anderen Farbe geschrieben sind, reicht eine Schwarz-Weiß-Digitalisierung völlig aus und spart zudem Speicherplatz. Denn je nach Menge und (Farb-)Qualität der Digitalisate, die Sie für Ihr Projekt erstellen wollen, können schnell große Datenmengen entstehen. Die DFG (Deutsche Forschungsgemeinschaft 2013, 6) empfiehlt daher in ihren *Praxisregeln Digitalisierung* für eine Bildkompression das Speicherformat TIFF, das bei Bedarf eine verlustfreie Reproduktion des ursprünglichen unkomprimierten Scans ermöglicht, wodurch eine Nachhaltigkeit und Langzeitarchivierung sichergestellt ist.

Die Qualität des Digitalisats ist entscheidend für die weitere wissenschaftliche Bearbeitung. So kann eine bestimmte Stelle im Manuskript tatsächlich unlesbar sein, oder sie ist die Folge einer zu geringen Auflösung oder Farbtiefe des Digitalisats. Soll das Digitalisat mit dem bloßen Auge gut lesbar sein, empfiehlt sich eine Mindestauflösung von 300ppi (pixel per inch). Ist das Ziel jedoch eine Computerlesbarkeit der Handschrift, werden von der digitalen Paläografie deutlich höhere Auflösungen gefordert, damit z. B. detailliertere Analyseverfahren der Computer Vision angewandt werden können (Hassner u. a. 2014, 20f.).

Um ideale Bedingungen für einen Scanvorgang zu schaffen, hat das Transkribus-Projekt das sog. ScanTent entwickelt, das den Scanprozess beschleunigt und die jeweilige Manuskriptseite perfekt ausleuchtet. Dadurch werden Fehler, die einer mangelhaften Digitalisierung zuschulden kommen können, minimiert. Die mit Hilfe des ScanTents und der mobilen Scan-App DocScan (für Android) erzeugten PDF-Dokumente können über die App direkt auf einen Transkribus-Account hochgeladen werden (vgl. das umfangreiche Transkribus-WIKI für weitere Informationen und Anleitungen).

Es ist möglich, individuelle Handschriften zu „trainieren“ (vgl. Machine Learning), d. h. dem Computer durch manuelle Transkription ausgewählter Manuskripte zu ermöglichen, weitere Manuskripte des gleichen Autors

(bzw. des gleichen Schriftstils) automatisch erkennen bzw. „lesen“ zu können. Die Erkennung funktioniert besser, je umfangreicher dieses Modell zuvor trainiert wurde. In einem Transkriptionsprojekt für venezianische Handschriften des 18. Jahrhunderts wurde beispielsweise ein Modell trainiert, das weitere Manuskripte der Sammlung mit einer geringeren Fehlerrate transkribiert als Amateur-Transkribierende das zu leisten vermochten (Oliveira und Kaplan 2018). Das langfristige Ziel von HTR-Initiativen ist es, so viele unterschiedliche Schreibstile zu trainieren, dass es zukünftig möglich sein wird, die meisten handschriftlichen Dokumente ohne vorheriges individuelles Training zu erkennen – ähnlich wie es auch schon im OCR-Verfahren für die meisten Druckschriften funktioniert.

Da die Handschriftenerkennung nicht auf Grundlage einzelner Buchstaben, sondern ganzer Zeilen funktioniert, müssen (im Gegensatz zur OCR) bei einem HTR-Vorgang die Zeilen (und ihre Reihenfolge, was z. B. bei mehrspaltigen Texten oder Ergänzungen zwischen den einzelnen Zeilen sehr relevant wird) zunächst festgelegt werden. Transkribus bietet hier beispielsweise eine automatisierte Zeilenerkennung, die händisch manipuliert werden kann.

Sollte es nicht das Ziel Ihres Projektes sein, die Manuskripte in einen elektronisch lesbaren Text umzuwandeln, können die Scans auch lediglich als Bilddateien verarbeitet werden. Neben den strukturellen Metadaten, die Informationen über die inhaltliche Struktur eines digitalisierten Objektes verzeichnen, gibt es deskriptive, technische oder administrative Metadaten. Als deskriptive Metadaten lassen sich beispielsweise Informationen über Epochen- oder Textsortenzugehörigkeit oder Informationen über den Autor eines Manuskriptes speichern. Technische Metadaten verzeichnen den Zustand des digitalisierten Objektes (wie etwa seine Auflösung) und administrative Metadaten beispielsweise Zugriffsrechte unterschiedlicher Personengruppen (Rehbein 2017, 192). Metadaten werden meistens im XML-Dateiformat gespeichert. Informationen und Guidelines zur Erstellung von Metadaten bietet beispielsweise die Text Encoding Initiative (TEI).

Externe und Weiterführende Links

- ScanTent: https://web.archive.org/save/https://www.transkribus.org/scantent (Letzter Zugriff: 04.06.2024)
- TEI: Text Encoding Initiative: https://web.archive.org/save/http://www.tei-c.org/index.xml (Letzter Zugriff: 04.06.2024)
- Transkribus. Digitisation and Digital Preservation Group, Universität Innsbruck: https://web.archive.org/save/https://www.transkribus.org/de (Letzter Zugriff: 04.06.2024)

Bibliographie

Benne, Christian. 2015. *Die Erfindung des Manuskripts. Zur Theorie und Geschichte literarischer Gegenständlichkeit*. Berlin: Suhrkamp.

Deutsche Forschungsgemeinschaft. 2013. *Handreichung: Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora*. https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/information_fachwissenschaften/geisteswissenschaften/standards_recht.pdf (zugegriffen: 9. Januar 2020).

Deutsche Forschungsgemeinschaft und Digitalisierung. 2016. *DFG-Praxisregeln. „Digitalisierung“*. http://www.dfg.de/formulare/12_151/12_151_de.pdf (zugegriffen: 12. Juli 2018).

Hassner, Tal, Malte Rehbein, Peter A. Stokes und Lior Wolf. 2014. Computation and Paleography: Potentials and Limits. *Dagstuhl Manifesto* 2, Nr. 1: 14–35. https://drops.dagstuhl.de/opus/volltexte/2013/4167/pdf/dagman-v002-i001-p014-12382.pdf.

Horstmann, Jan. 2024a. Methodenbeitrag: Möglichkeiten der Textdigitalisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3741, https://fortext.net/routinen/methoden/moeglichkeiten-der-textdigitalisierung.

———. 2024b. Toolbeitrag: Transkribus. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3746, https://fortext.net/tools/tools/transkribus.

———. 2024c. Ressourcenbeitrag: TextGrid Repository. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 11. Bibliografie (29. November). doi: 10.48694/fortext.3794, https://fortext.net/ressourcen/textsammlungen/textgrid-repository.

Horstmann, Jan und Jan-Erik Stange. 2024. Methodenbeitrag: Textvisualisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 5. Textvisualisierung (7. August). doi: 10.48694/fortext.3772, https://fortext.net/routinen/methoden/textvisualisierung.

Oliveira, Sofia Ares und Frederic Kaplan. 2018. Comparing human and machine performancesiIn transcribing 18th century handwritten Venetian script. In: *DH 2018. Conference Abstracts*. https://dh2018.adho.org/en/comparing-human-and-machine-performances-in-transcribing-18th-century-handwritten-venetian-script/ (zugegriffen: 11. Juli 2018).

Rehbein, Malte. 2017. Digitalisierung. In: *Digital Humanities. Eine Einführung*, hg. von Fotis Jannidis, Hubertus Kohle, und Malte Rehbein, 179–198. Stuttgart: Metzler.

Sánchez, Joan Andreu, Verónica Romero, Alejandro Héctor Toselli und Enrique Vidal. 2016. ICFHR2016 competition on handwritten text recognition on the READ dataset. *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*: 630–635. https://api.semanticscholar.org/CorpusID:19239978.

Sánchez, Joan Andreu, Verónica Romero, Alejandro H. Toselli und Enrique Vidal. 2014. ICFHR2014 Competition on Handwritten Text Recognition on tranScriptorium Datasets (HTRtS). In:, 181–186.

Sánchez, Joan Andreu, Alejandro H. Toselli, Verónica Romero und Enrique Vidal. 2015. ICDAR 2015 competition HTRtS: Handwritten Text Recognition on the tranScriptorium Dataset. In:, 1166–1170. Tunis. doi: 10.1109/ICDAR.2015.7333944,.

Schumacher, Mareike. 2024. Methodenbeitrag: Netzwerkanalyse. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 6. Netzwerkanalyse (30. August). doi: 10.48694/fortext.3759, https://fortext.net/routinen/methoden/netzwerkanalyse.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

Commandline Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (GUI) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd“ + „space“, geben „Terminal“ ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R“, geben „cmd.exe“ ein und klicken Enter.

Domäneadaption Domäneadaption beschreibt die Anpassung einer in einem Fachgebiet entwickelten digitalen Methode an ein anderes Fachgebiet.

Double-keying Double-Keying ist eine Variante des Keying, bei der zwei Personen den Inhalt eines Dokumentes abtippen. Anschließend sucht ein Computerprogramm nach Differenzen zwischen den beiden Versionen. Gefundene Tippfehler werden dann von einer dritten Person korrigiert. So entstehen nahezu fehlerfreie Textdigitalisate.

GUI GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der Commandline zu umgehen.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

HTR HTR steht für *Handwritten Text Recognition* und ist eine Form der Mustererkennung, wie auch die OCR. HTR bezeichnet die automatische Erkennung von Handschriften und die Umformung dieser in einen elektronischen Text. Die Automatisierung beruht auf einem Machine-Learning-Verfahren.

Keying In den Bibliotheks- und Textwissenschaften beschreibt Keying das manuelle Erfassen, also das Abtippen, eines Textes im Zuge seiner Digitalisierung (siehe auch Double-Keying).

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup (Textauszeichung) Die Textauszeichnung (eng. *Markup*) fällt in den Bereich der Daten- bzw. Textverarbeitung, genauer in das Gebiet der Textformatierung, welche durch Auszeichnungssprachen wie XML implementiert wird. Dabei geht es um die Beschreibung, wie einzelne Elemente eines Textes beispielsweise auf Webseiten grafisch dargestellt werden sollen.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z.B. HTML, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

OCR OCR steht für *Optical Character Recognition* und bezeichnet die automatische Texterkennung von gedruckten Texten, d. h. ein Computer „liest“ ein eingescanntes Dokument, erkennt und erfasst den Text darin und generiert daraufhin eine elektronische Version.

PDF PDF steht für *Portable Document Format* . Es handelt sich um ein plattformunabhängiges Dateiformat, dessen Inhalt auf jedem Gerät und in jedem Programm originalgetreu wiedergegeben wird. PDF-Dateien können Bilddateien (z. B. Scans von Texten) oder computerlesbarer Text sein. Ein lesbares PDF ist entweder ein OCRter Scan oder ein am Computer erstellter Text.

POS PoS steht für *Part of Speech* , oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

TEI Die *Text Encoding Initiative* (TEI) ist ein Konsortium, das gemeinsam einen Standard für die Darstellung von Texten in digitaler Form entwickelt. Die TEI bietet beispielsweise Standards zur Kodierung von gedruckten Werken und zur Auszeichnung von sprachlichen Informationen in maschinenlesbaren Texten (siehe auch XML und Markup).

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.

WIKI Ein Wiki ist eine Webseite mit einer Sammlung von Informationen zu ausgewählten Themen, die i. d. R. von mehreren Nutzer*innen zusammengestellt werden. Zu jedem Eintrag in einem Wiki gibt es eine Diskussionsseite, die auch frühere Versionen des Eintrags anzeigt.

XML XML steht für *Extensible Markup Language* und ist eine Form von Markup Language, die sowohl computer- als auch menschenlesbar und hochgradig anpassbar ist. Dabei werden Textdateien hierarchisch strukturiert dargestellt und Zusatzinformationen i. d. R. in einer anderen Farbe als der eigentliche (schwarz gedruckte) Text dargestellt. Eine standardisierte Form von XML ist das TEI-XML.