Zentrum für digitale Edition und quantitative Analyse an der Universität Würzburg

Zuwendungsempfänger:

Universität Würzburg

Förderkennzeichen:

01UG1415A

Vorhabenbezeichnung:

Verbundprojekt KALLIMACHOS – Aufbau eines Zentrums für digitale Edition und quantitative Analyse an der Universität Würzburg

Teilprojekt Würzburg: Aufbau, Koordination und Organisation eines Digital-Humanities-Zentrums an der Universität Würzburg

Teilprojekt Erlangen: Korpuslinguistische Methoden und statistische Auswertungen für den Workflow von KALLIMACHOS

Teilprojekt Kaiserslautern: Entwicklung eines OCR-Moduls für den Workflow von KALLIMACHOS

Laufzeit des Vorhabens:

01.10.2014 - 30.09.2017

Schlussbericht

Berichtszeitraum:
01.10.2014 – 30.09.2017

I. Kurze Darstellung

1. Aufgabenstellung

Die Förderung des Projekts KALLIMACHOS erfolgte im Rahmen der BMBF eHumanities Förderlinie 2, um Forschungsinfrastrukturen für die Geistes- und qualitativen Sozialwissenschaften unter maßgeblicher Einbeziehung der Informatik oder informatiknaher Fächer aufzubauen. Anhand mehrerer beispielhaft gewählter Untersuchungsthemen sollte die Fruchtbarkeit dieses interdisziplinären Ansatzes sichtbar gemacht und damit ein Digital Humanities Zentrum an der Universität Würzburg dauerhaft konstituiert werden.

Die Aufgabenstellung gliederte sich in die folgenden Arbeitspakete, denen wiederum geisteswissenschaftliche Use Cases zugeordnet sind. Narragonien digital (Use Case 1) erstellte ein Korpus ausgewählter Drucke des *Narrenschiffs* von Sebastian Brant, Anagnosis (Use Case 2) hatte die Verbindung von Text und Bildbereichen digitalisierter Papyri zum Gegenstand, im Rahmen von Schulwandbildern digital (Use Case 3) wurde Europas größte Sammlung an Schulwandbildern mit über 25.000 digitalisierten Wandbildern mit Metadaten versehen und durchsuchbar gemacht, die Provenienz-und Gattungsbestimmung (Use Case 4) hat ebenso wie die Narrativen Techniken (Use Case 5) und die Leserlenkung in Bezug auf Figuren (Use Case 6) automatisierte Verfahren zur semantischen Korpusanalyse entwickelt und die erstellten Werkzeuge in einem Werkzeugkasten gebündelt, der unter anderem auch die Identifzierung anonymer Übersetzer (Use Case 7) ermöglicht.

AP1: OCR-Modul

Use Cases: Anagnosis, Narragonien digital, Schulwandbilder digital
Aufgabe: OCR-Erfassung von Frühdrucken (15./16. Jh., Narrenschiffe), Textausgaben (Anagnosis) und Lehrerbegleitbänden zu den Schulwandbildern

AP2: Synoptischer Editor

Use Cases: Anagnosis, Narragonien digital
Aufgabe: Transkription von Bilddigitalisaten mittels Bild-Texts-Synopse; Korrektur von OCR -Ergebnissen

AP3: Wikimodul

Use Cases: Anagnosis, Schulwandbilder digital, Narragonien digital
Aufgabe: Präsentation von Texten und Bildern, Annotationen

AP4: Implementierung einer Schnittstelle Repositories-Datenanalyse

Use Cases: Provenienz- und Gattungsbestimmung; Leserlenkung in Bezug auf Figuren
Aufgabe: Entwicklung von Werkzeugen zur automatischen Datenanalyse

AP5: Prototypische Arbeitsabläufe in der Datenanalyse (best practice Modelle)

Use Cases: Identifizierung anonymer Übersetzer; Narrative Techniken und Untergattungen
Aufgabe: Entwicklung von Werkzeugen zur automatischen Datenanalyse

2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das “Zentrum für digitale Edition und quantitative Analyse der Universität Würzburg” wurde am bestehenden Digitalisierungszentrum der Universitätsbibliothek (UB) eingerichtet. Während der Projektlaufzeit konnten zwei mit Informatikern besetzte Projektstellen der UB in den Stellenplan übernommen und damit verstetigt werden. Wie in den Zwischenberichten näher ausgeführt, konnte an der UB nie die ursprünglich vorgesehene Zahl an Projektmitarbeitern erreicht werden, weshalb nicht alle vorgesehenen Ziele erreicht wurden. Die Rekrutierung wissenschaftlicher Mitarbeiter an den beteiligten Lehrstühlen verlief nach Plan.

3. Planung und Ablauf des Vorhabens

Unter der Leitung der UB wurde die Arbeit der beteiligten Lehrstühle und Professuren koordiniert und in regelmäßigen Arbeitssitzungen aufeinander abgestimmt. Der jährlich stattfindende PhilTag-Workshop diente der Vorstellung der jeweils erreichten Zwischenergebnisse vor einer interessierten Fachöffentlichkeit sowie dem Wissenstransfer durch angebotene Workshops und Tutorials. Jeweils im April 2016 und 2017 wurde ein Zwischenbericht erstellt.

4. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde, insbesondere

a. Angabe bekannter Konstruktionen, Verfahren und Schutzrechte, die für die Durchführung des Vorhabens benutzt wurden

Im Rahmen der Open Source Strategie des Projektes wurde darauf geachtet, keinerlei schutzrechtsbehaftete Konstruktionen und Verfahren einzusetzen, die die Freigabe der Projektergebnisse unter einer Open-Source-Lizenz behindern könnte.

b. Angabe der verwendeten Fachliteratur sowie der benutzten Informations- und Dokumentationsdienste

Die verwendete Fachliteratur ergibt sich aus den Zitaten der Veröffentlichungen. Neben den üblichen Informations- und Dokumentationsdiensten wurden Webdienste wie arxive.org (Preprint-Server), Google Scholar, Academia.edu sowie Researchgate.net genutzt.

5. Zusammenarbeit mit anderen Stellen.

Im Rahmen von Kallimachos wurde zusätzlich mit folgenden externen Stellen zusammengearbeitet:

Universitätsbibliothek:

- Centrum für Informations- und Sprachverarbeitung (CIS) der LMU München (Dr. Uwe Springmann) auf dem Gebiet der OCR früher Drucke

Narragonien:

- PD Dr. Michael Rupp
- Universitätsbibliothek Basel, Abt. Handschriften und Alte Drucke
- Bibliothek Otto Schäfer, Schweinfurt

- Projekt „Mittelniederdeutsch in Lübeck“ (MiL; WWU Münster); Projektleitung: Dr. Robert Peters, Norbert Lange
- Dr. Anne-Laure Metzger-Rambach, Université de Michel de Montaigne Bordeaux 3
- Dr. Olga Anna Duhl, Lafayette College

Anagnosis:

- Centro Internazionale per lo Studio dei Papiri Ercolanesi (CISPE), Neapel (TEI Epidoc-konforme Kodierung herkulanischer Papyri);
- Berliner Papyrussammlung (TEI Epidoc-konforme Kodierung Berliner Papyri und Verwendung der in der BerlPap-Datenbank verfügbaren hochauflösenden Abbildungen zur Implementierung der Alignment-Oberfläche);
- Institut für Papyrologie der Universität Heidelberg (Aufnahme der im Rahmen von Anagnosis kodierten Papyri in das Digital Corpus of Literary Papyri).

II. Eingehende Darstellung

1. Verwendung der Zuwendung und des erzielten Ergebnisses im Einzelnen, mit Gegenüberstellung der vorgegebenen Ziele

AP1: OCR-Modul

Aufgabe / Erreichungsgrad	voll	teilweise	nicht
TA 1.0: Pflichtenhefterstellung		X	
TA 1.1: Integration vorhandener OCR-Komponenten	X		
TA 1.2: Training neuer Modelle Drucke/Typeninventare; Handschrift	X		
TA 1.3: Sprachmodelle historische Orthografie			X
TA 1.4: Weiterentwicklung nichtparametrischer Methoden		X	
Use Case 1: Narragonien - Erstellung Kollationsvorlage layoutkonform	X		
Use Case 2: Anagnosis - Erstellung Kollationsvorlage schriftspiegelkonform		X	
Use Case 3: Schulwandbilder - Scan, OCR Begleitmaterialien, Lehrerhdb.; weitere Medientypen: Zeitschriften, Zeitungen 19. Jh. - OCR-Strategien		X	

Bericht Arbeitsgruppe DFKI Kaiserslautern

Im Arbeitspaket OCR-Modul sollte ein Vorverarbeitungstool auf Basis des etablierten, LSTM-basierten OCR-Systems OCRopus entwickelt werden. Dabei gab es zwei Zielsetzungen, die Verbesserung der Genauigkeit und damit zusammenhängend die Minimierung des manuellen Arbeitsaufwandes, speziell in Hinblick auf die im klassischen Workflow notwendige manuelle Transkription der Trainingsdaten und der Nachkorrektur der OCR-Ergebnisse.

OCRopus war schon vor dem Start des KALLIMACHOS Projektes in der Lage, mit ausgewählten historischen Dokumenten umzugehen. Als der limitierende Faktor des LSTM basierten Ansatzes wurden die große Menge an benötigten Trainingsdaten (Bilder der Textzeilen und der dazugehörige Text) identifiziert. Ein möglicher Lösungsansatz besteht darin, synthetische Trainingsdaten von repräsentativen Textzeilen zu erzeugen. Solch repräsentative Textzeilen stehen aber nicht immer zur Verfügung. In diesem Fall, wie z. B. bei den Narragonien, bräuchte man mindestens 100 Seiten manuell transkribierten Textes um ein gutes LSTM basiertes OCR System trainieren zu können.[1] Ein solcher Aufwand lohnt sich im Fall der Narragonien-Texte auf Grund der hohen Varianz in Schriftsatz und -art nicht. Der erarbeitete Lösungsansatz anyOCR umgeht die beschriebenen Limitierungen von auf LSTM basierten

1 Anmerkung der Projektleitung: Diese Aussage steht im Gegensatz zu den Erfahrungen von [Springmann et al., 2016], nach dem bereits etwa 150 Zeilen Trainingsmaterial ausreichen, um Zeichenerkennungsraten von bis zu 98% zu erreichen. Diese Erfahrungen wurden inzwischen in den Untersuchungen und Veröffentlichungen der Würzburger Arbeitsgruppe vielfach bestätigt ([Reul et al., 2017b], [Reul et al., 2017c]).

segmentierungsfreien OCR Modellen (welche manuell transkribierte Trainingsdaten benötigen) und segmentierungsbasierten OCR Modellen wie von Tesseract. Er kombiniert die Stärken von OCRopus und TesseRACT und wird daher auch als OCRoRACT bezeichnet [Ul-Hasan et al., 2016]. OCRoRACT nutzt die Vorteile des segmentierungsfreien Ansatzes, ohne dass es große Mengen manuell transkribierter oder synthetischer Trainingsdaten benötigt. Das Resultat sind Ergebnisse, die nahe denen des rein auf LSTM basierten und mit manuell transkribierten Daten trainierten Ansatzes sind. Für eine lateinische Version aus den Narragonien-Ausgaben aus Basel von 1497, deren Textzeilen halbautomatisch segmentiert wurden, erzielte der anyOCR Ansatz für einen Testbereich eine Genauigkeit von 95% im Vergleich zu 77% von Tesseract und 98% von dem traditionell trainierten LSTM Modell. Ein zusätzlicher Vorteil ist anyOCRs Robustheit gegenüber Pixelfehlern und ähnlichen Dokumentfehlern, welche typisch für historische Dokumente sind. In einem zweiten anyOCR Prototypen wurde TesseRACT, welches aufwendig manuell präparierte Trainingsdaten benötigt, durch eine Clusteranalyse ersetzt [Jenckel et al., 2016a]. Simultan wurden die Layoutanalyse Tools von OCRopus um einen neuen Prozess zur Text-Bild-Segmentierung sowie Verbesserungen der Textzeilen-Segmentierung erweitert. Diese erweiterte Version von OCRopus bezeichnen wir als anyBaseOCR.

Entsprechend des Zeitplans wurden im ersten Projektabschnitt die wissenschaftliche Fragestellung bezüglich der Anwendung LSTM basierter OCR-Verfahren für die Anwendung auf historische Dokumente konkretisiert und erste Prototypen entwickelt, sowie erste Tests anhand des Use Cases Narragonien durchgeführt.

Auch mit den neu entwickelten Prototypen ist eine “fast fehlerfreie” vollautomatische Texterkennung für die Anwendungsfälle Narragonien und Anagnosis mit dem heutigen Stand der Technik nicht möglich. Es gibt zwei Hauptlimitierungen: I) Vollautomatische Layout Analyse liefert keine 100%igen Ergebnisse und II) vollautomatische OCR Systeme produzieren keinen fehlerfreien Text. Zur weiteren Verbesserung der OCR Ergebnisse sind daher semi-manuelle Korrekturschritte notwendig.

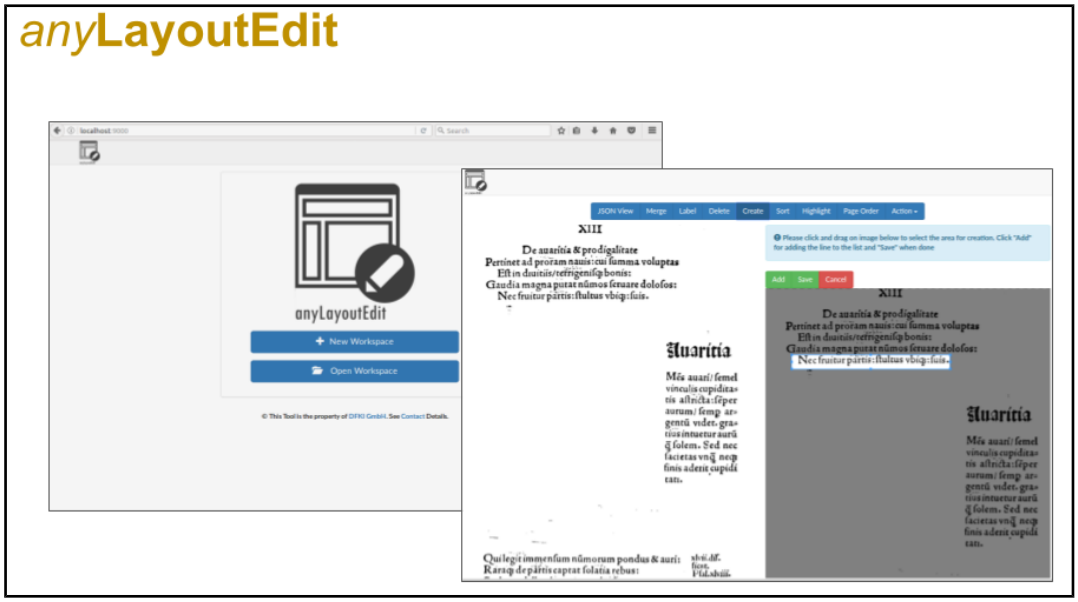

Abbildung 1: anyLayoutEdit Benutzeroberfläche

Mit anyLayoutEdit wurde am DFKI ein interaktiver Webservice entwickelt, bei dem der Benutzer die Ergebnisse automatischer Textzeilensegmentierung hochladen kann. Dem Nutzer stehen dann verschiedene Tools zur Korrektur von Segmentierungsfehlern zur Verfügung. Es können fehlende Zeilen eingefügt bzw. extra Zeilen entfernt werden, verschmolzene Zeilen getrennt werden, sowie die Lesereihenfolge angepasst werden. Anschließend können die Daten mit verbesserter Segmentierung heruntergeladen und für weitere OCR Methoden verwendet werden.

Des Weiteren wurde am DFKI ein interaktiver und selbst lernender Webservice namens anyOCREdit entwickelt. Der Nutzer lädt dabei zunächst seine Bild- und dazugehörigen fehlerhaften OCR Text im .zip Format hoch. Der Webservice bietet dem Nutzer dabei zweierlei Hilfen: I) Alle notwendigen Tools, um schnell und effektiv OCR-Fehler zu korrigieren, wie zum Beispiel eine Vergrößerungslinse, die automatisch durch kurzes Hovern über dem Bild aktiviert wird, sodass auch kleinste Buchstaben oder beschädigte Textstellen gut sichtbar sind. II) Außerdem bietet anyOCREdit eine automatische OCR Fehlerkorrektur basierend auf Statistischer Maschinenübersetzung (SMT), welche aus den manuellen Korrekturen des Nutzers lernt und so den Vorgang insgesamt beschleunigt.

Zusammen mit anyBaseOCR und anyOCRModel komplettieren anyLayoutEdit und anyOCREdit das neue OCR System für nahezu fehlerfreies OCR. Das gesamte End-zu-End OCR System inklusive unserer vier neuen Module (anyBaseOCR, anyLayoutEdit, anyOCRModel und anyOCREdit) nennen wir anyOCR.

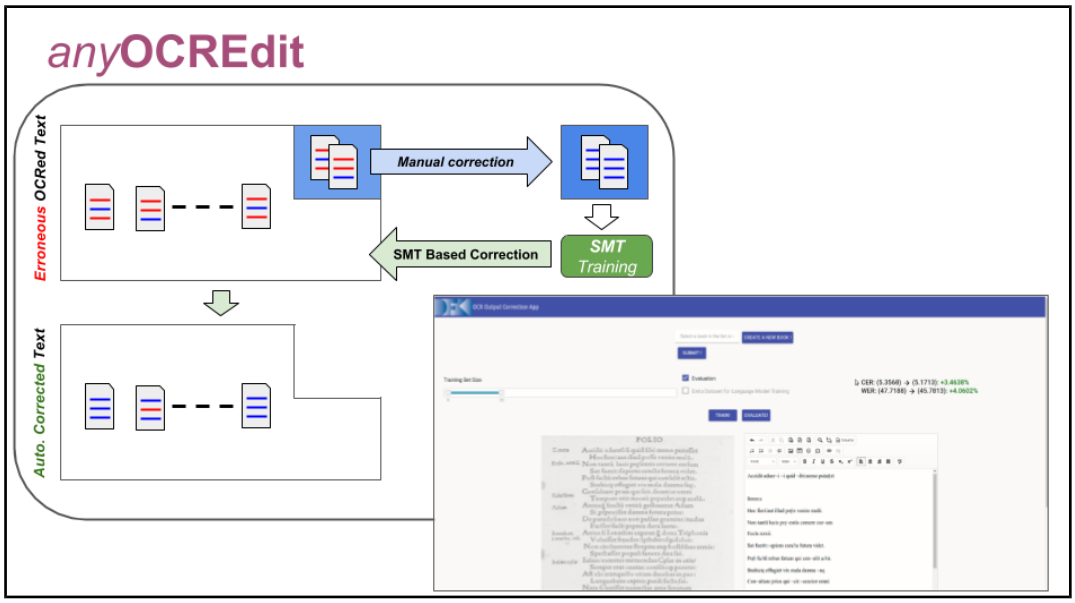

Abbildung 2: anyOCREdit Schema und Benutzeroberfläche

Die Entwicklungen im zweiten Projektabschnitt sind das Resultat regelmäßiger Treffen mit den beteiligten Projektpartnern und dem Schluss, dass trotz guter OCR-Ergebnisse eine Nachbearbeitung der OCR Ergebnisse im Use Case Narragonien unumgänglich ist.

In Kooperation mit den Projektpartnern wurde für den Use Case Anagnosis hinsichtlich des automatisierten Alignments eine OCR-freie Herangehensweise getestet, die sich Methoden der Computer Vision bedient. Nach einem Preprocessing, zu dem Binarisierung, Zeilensegmentierung und halbautomatische Festlegung der Lesereihenfolge gehören, wird die XML-Transkription als Basis für die Generierung eines synthetischen Bildes des Papyrusfragments benutzt. Die dafür nötigen Glyphen werden aus einem manuell erstellten Datensatz bezogen, welches aus Ausschnitten zahlreicher Bilder mit Varianten von Buchstabenformen besteht, bei denen die Entsprechung des zugehörigen Unicode-Zeichens bereits angegeben ist. Der Abgleich zwischen synthetischem Bild und originalem Bild erfolgt dann durch einen SIFT-Flow (Scale Invariant Feature Transformation)-Algorithmus. Dieser basiert auf der Erkennung von Ähnlichkeitszonen (key points) durch Pixelanalyse und -matching. Der Output erhält die Form von in Pixel ausgedrückten Letter-Spacing Informationen für jede Textzeile.

Ferner wurde ein neues, intuitives Interface erstellt, das den Benutzern erlaubt, sowohl in der Preprocessing-Phase als auch nach der Erzeugung der automatischen Bild-Text-Verknüpfung die Ergebnisse zu bewerten und ggf. zu korrigieren.

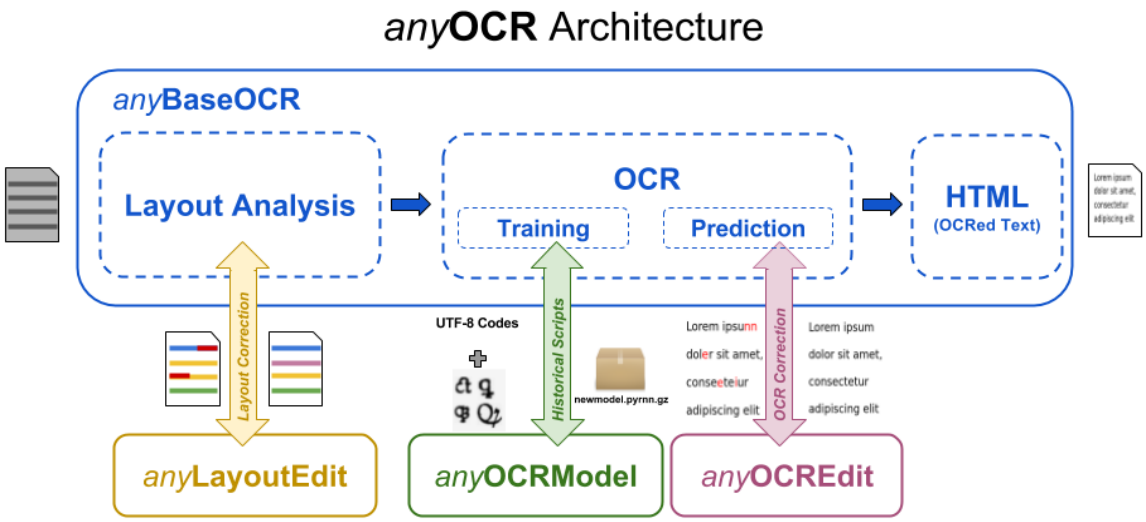

Abbildung 3: Gesamtarchitektur von anyOCR

Für die Generierung guter OCR Modelle mit Hilfe von neuronalen Netzen wird eine große Anzahl transkribierter Trainingsdaten benötigt. Im ersten Projektabschnitt haben wir für diese Problematik das anyOCRModel entwickelt, welche durch Kombination segmentierungsfreier und segmentierungsbasierter OCR Methoden ohne aufwendige manuelle Transkriptionen auskommt. Um dieses Modell weiter zu optimieren, wurden die Eigenschaften segmentierungsfreier OCR Systeme (LSTM) im Umgang mit fehlerhaften und ungenauen Transkriptionen untersucht. Dazu wurden im letzten Projektabschnitt zwei Studien durchgeführt. Bei der Kombination segmentierungsfreier und segmentierungsbasierter OCR Methoden wird die segmentierungsfreie OCR Methode (LSTM) auf das Ergebnis der segmentierungsbasierten OCR Methode (Clusteranalyse) angewendet. Dazu wurde untersucht, wie gut LSTM-basierte OCR Systeme mit fehlerhaften Transkriptionen in den Trainingsdaten umgehen können [Jenckel et al., 2017]. Die Ergebnisse zeigen, dass vor allem Fehler durch falsche Segmentierung oder falsche Annotation der Cluster im segmentierungsbasierten OCR Verfahren ausschlaggebend für die Qualität des Gesamtsystems sind. Fehler in der Clusteranalyse dagegen können gut vom LSTM-basierten OCR Verfahren kompensiert werden.

In der zweiten Studie wurde untersucht, wie sich LSTM-basierte OCR Systeme mit ungenauen Transkriptionen trainieren lassen [Jenckel et al., 2018]. Diese entstehen vor allem im Kontext beschädigter Buchstaben in historischen Dokumenten, welche sich nicht eindeutig identifizieren lassen. Dafür haben wir angenommen, dass anstatt einer einzelnen korrekten Transkription alle möglichen Buchstaben zur Verfügung stehen. Unsere Ergebnisse zeigen, dass LSTM-basierte OCR Systeme auch mit ungenauen Transkriptionen gute OCR Ergebnisse liefern können. Dies könnte neue und effizientere Möglichkeiten in der Transkription historischer Dokumente durch Laien ermöglichen.

Eine weitere Hauptproblematik vollautomatischer OCR Systeme liegt darin, dass sie keine fehlerfreien Ergebnisse liefern. Aufgrund dieser Problematik wurden im zweiten Projektabschnitt anyLayoutEdit und anyOCREdit als zwei interaktive Tools für eine schnelle semi-manuelle Korrektur automatischer Ergebnisse entwickelt. Um dem Ziel eines fehlerfreien OCR Systems näher zu kommen, wurde die Annotations-Oberfläche durch eine Möglichkeit zur Korrektur von Segmentierungsfehlern ergänzt. Die überarbeitete Annotations-Oberfläche erlaubt es dem Benutzer während der Annotation der Cluster einen Einblick in die einzelnen Buchstaben und ermöglicht eine einfache Korrektur von Segmentierungsfehlern, sowie Fehlern in der Clusteranalyse. Für eine leichtere Annotation sieht der Benutzer außerdem einzelne Buchstaben im Kontext ihrer Textzeile.

Ein weiteres Ziel des Projektes war die Bereitstellung der entwickelten Module als Open Access Software, sowie eine detaillierte Dokumentation zu dessen Nutzung. Dazu wurden alle im anyOCR Ansatz

entwickelten Lösungen in ein komplettes, web-basiertes, Open Access, End-zu-End OCR System integriert [Bukhari et al., 2017]. Der komplette Code des OCR-Moduls ist außerdem als Open Source Verfügbar. Nach der semi-automatischen Layoutanalyse durch anyBaseOCR und anschließender interaktiver Korrektur mit Hilfe von anyLayoutEdit wird mit dem anyOCRModel auf Basis der segmentierten Textzeilen ein OCR Modell generiert. Alternativ kann ein bereits vorhandenes OCR Modell verwendet werden. Abschließend können die vom Modell generierten OCR Ergebnisse in anyOCREdit interaktiv und semi-automatisch korrigiert werden. Um eine möglichst genaue Darstellung des Originaldokuments zu gewährleisten, werden alle OCR-Ergebnisse im standardisierten hOCR Format ausgegeben. Die Verwendung des gesamten End-zu-End OCR Systems, sowie für die Einzelmodule, wurde in Form von Tutorialvideos dokumentiert. Für den Use Cases Anagnosis wurde im letzten Projektabschnitt ebenfalls ein Open Access Webservice entwickelt, welcher Nutzern das automatische Alignment von Textzeile und Transkription ermöglicht [Bukhari et al., 2018].

Bericht Universitätsbibliothek

Wegen der in den Zwischenberichten 1 und 2 beschriebenen Zeitverzögerungen am DFKI wurde zeitweise parallel ein Ansatz auf Basis von Tesseract 3 an der UB implementiert, der auf einer Segmentierung von Einzelzeichen und ihrer Erkennung mittels eines Trainings auf Features für diese Zeichen beruht. Dieser als “Offizin-Ansatz” bezeichnete Workflow baute auf Vorarbeiten des EMOP-Projektes an der Texas A&M-University (Laura Mandell) unter Verwendung des dort erstellen FRANKEN+ Werkzeugs auf und konnte bis dahin mit Tesseract nicht für möglich gehaltene Zeichenerkennungsraten von bis zu 95% für eine Inkunabel erreichen ([Kirchner et al., 2016]). Jedoch bedeutete dieser Ansatz nicht nur einen erheblichen Arbeitsaufwand wegen der nötigen manuellen “Wiederbelebung” eines historischen Zeichensatzes, sondern das Verfahren wurde auch durch das Aufkommen segmentierungsfreier OCR-Verfahren mittels rekurrenter neuronaler Netze in LSTM-Architektur überholt, mit dem auch auf Inkunabeldrucken Erkennungsraten von über 98% erreicht werden konnten ([Springmann et al., 2015], [Springmann et al., 2016], [Springmann und Lüdeling, 2017]). Daher wurde der Offizinansatz wieder fallengelassen und die bereits im Antrag favorisierte Lösung auf Basis der Open-Source-Software OCRopus erneut aufgegriffen. Durch die Zusammenarbeit mit Dr. Springmann (CIS, LMU München) konnte mit tatkräftiger Unterstützung des Lehrstuhls Informatik VI (Prof. Dr. Frank Puppe, Christian Reul) ein freier Workflow auf Basis von OCRopus und dem von Christian Reul neu erstellten Segmentierungstool **LAREX** ([Reul et al., 2017a]) entwickelt werden, der es erlaubt, innerhalb weniger Tage ein Narrenschiff-Digitalisat manuell zu segmentieren (semantische Aufteilung einer Bildseite in diverse Bestandteile wie Bild, Haupttext, Rand- und Fußnoten etc.), eine Transkription mit dem von Felix Kirchner entwickelten **Transkriptionseditor** vorzunehmen, ein werkspezifisches OCR-Modell zu trainieren und damit eine OCR-Erkennung des Gesamttextes mit Erkennungsraten um die 98% als Ausgangspunkt für eine weitergehende Korrektur und Annotation zu erzeugen.

Das ursprünglich in der Vorhabenbeschreibung zusätzlich vorgesehene Training von OCR-Modellen für Handschriftenerkennung sowie die Entwicklung von Sprachmodellen für die Berücksichtigung historischer Schreibweisen bei der OCR-Korrektur hat sich als zu ambitioniert erwiesen bzw. wird mittlerweile durch andere Projekte gefördert,[2] weshalb innerhalb von Kallimachos davon abgesehen wurde, zu solchen Projekten in Konkurrenz zu treten.

Use Case 1: Narragonien digital

Im Arbeitspaket 1 (OCR-Modul) hat die notwendig gewordene Entkoppelung der Entwicklungsarbeit des DFKI Kaiserslautern von Use Case 1 (“Narragonien digital”) zu einer Verschiebung zu Lasten des Würzburger Partners geführt. Die geplanten Arbeiten mussten in Eigenleistung erbracht und die Ziele entsprechend modifiziert werden. Zum Ende der Projektlaufzeit hat der Lehrstuhl Puppe seine Unterstützung angeboten. In enger Zusammenarbeit mit dem Lehrstuhlmitarbeiter Herrn Reul wurden meh-

2 Z.B. durch das EU-geförderte READ-Projekt (https://read.transkribus.eu/).

rere Narrenschiffe segmentiert und mittels OCR eingelesen, dabei wurde insbesondere für das Segmentierungs- und das OCR-Korrekturtool in enger Abstimmung mit den Mitarbeitern des Use-Case Narragonien eine benutzerfreundliche GUI entwickelt.

Personelle Situation: Neben den Antragstellern, den wissenschaftlichen Mitarbeiterinnen Christine Grundig (50%, Hamm) und Viktoria Walter (50%, Burrichter) und zwei aus BMBF-Mitteln finanzierten Hilfskräften wurden aufgrund der gestiegenen Arbeitslast zusätzliche Hilfskraftstunden aus Lehrstuhlmitteln für das Projekt eingesetzt.

Stand der Edition: Ediert werden zehn bedeutende Ausgaben und Bearbeitungen des 'Narrenschiffs', die in deutscher, lateinischer, französischer, englischer, niederländischer und niederdeutscher Sprache im 15. Jahrhundert (bzw. 1509) gedruckt worden sind ([Grundig et al., im Druck]). Jeder dieser Texte umfasst ca. 350 Seiten. Den Stand der Arbeiten dokumentiert die nachfolgende tabellarische Übersicht. Zur Erläuterung:

1. Die ausgewählten Druckausgaben liegen inzwischen als zeichengetreue Transkriptionen und Lesetextfassungen vor. Die Transkriptionen wurden händisch oder per OCR (ausschließlich durch die Würzburger Projektpartner) erstellt.
2. Die Korrektur der Transkriptionen ist abgeschlossen. Gleiches gilt für die händische Auszeichnung der Layoutzonen (Titel, Mottoverse, Holzschnitte, Spruchgedichte, Marginalien usw.), die die Grundlage für die übergreifende Suchfunktion der späteren Online-Präsentation bildet.
3. Zu den geplanten Editionen wurden behutsam normalisierte Lesetexte erstellt.
4. Die Teileditionen wurden bzw. werden zurzeit zur weiteren Bearbeitung in das Wiki-System eingespielt.
5. Das Personen- und Ortsregister wird die Namenssuche über alle edierten Ausgaben und Bearbeitungen des ‘Narrenschiffs’ hinweg ermöglichen. Die zeitaufwändige Auszeichnung der Lesetexte durch Registereinträge ist weitgehend abgeschlossen. Die zahlreichen Orts- und Personennamen wurden in den Editionen mit dem Register verlinkt.
6. Die Varianz der Basler Drucke (Edition 1) und der lateinischen Ausgaben (Edition 4) wurde zunächst analog erfasst. Für diejenigen Editionen, die nur durch eine Druckausgabe repräsentiert werden, entfällt die Erfassung der Varianz.
7. Die Marginalien mit Quellenhinweisen, die in der lateinischen Bearbeitung eingeführt und in deren Übersetzungen weitgehend übernommen wurden, werden zurzeit aufgelöst (bisher sind gut 70% der Marginalien bearbeitet). Die Druckausgaben der Editionen 1, 2, 3, 7, 9 und 10 haben keine Marginalien.

Bis jetzt sind damit die Texte der geplanten Ausgaben vollständig (inklusive Registereinträge) erfasst: Die drei Basler Ausgaben des ‘Narrenschiffs’ (GW 5041, 5046, 5047), der Nürnberger Raubdruck (GW 5042), das niederdeutsche Narrenschiff (GW 5053), die Straßburger interpolierte Fassung (GW 5048), die zweite Ausgabe der lateinischen Bearbeitung (GW 5061) sowie die drei französischen Übertragungen (GW 5058, 5060 und 5065), die niederländische Übertragung (GW 5066) und die englische Übertragung von Alexander Barclay (1509). OCR- und Lesetexte sind für die lateinische Erstausgabe (GW 5054) erstellt.

Einen Überblick über den Stand der Arbeiten bietet die folgende Tabelle:

UseCase „Narragonien digital“ (Stand: 30.9.2017) Legende: GRÜN = Erledigt – BLAU = z.Z. in Arbeit – GELB = Agenda – WEISS = nicht vorhanden oder nicht erforderlich

Nr.	GW-Nr.	Benutztes Exemplar	OCR / händ. Transkription	Textkorrektur / Markup Layout	Lese-fassung	Text im Wiki	Register-einträge	Erfassung von Varianz	Auflösung Marginalien	Vorläufiger TEI-Export	Endkorrektur begonnen
1	**Sebastian Brant, Narrenschiff (Basel 1494)**										
a	GW 5041	Berlin									
b	GW 5046	Regensburg									
c	GW 5047	Basel									
2	**Die Nürnberger Bearbeitung (1494)**										
	GW 5042	Berlin							.		
3	**Die Straßburger Bearbeitung (1494/5)**										
	GW 5048	Colmar							.		
4	**Die lateinische Bearbeitung (Jakob Locher, Stultifera navis, Basel 1497)**										
a	GW 5054	Berlin									
b	GW 5061	Basel									
c	GW 5062	Würzburg									
5	**Die französische Versbearbeitung (Pierre Rivière, La nef des folz du monde, Paris 1497)**										
	GW 5058	Wien									
6	**Die französische Prosabearbeitung (Jean Drouyn, La grant nef des fols, Lyon 1498 / 1499)**										
	GW 5060	Jena									
7	**Die anonyme französische Prosabearbeitung (La grant nef des folz du monde, Paris 1499)**										
	GW 5065	Wien									
8	**Die englische Bearbeitung (Alexander Barclay, The Shyp of Folys, London 1509)**										
		Cambridge									
9	**Die niederdeutsche Fassung (Lübeck 1497)**										
	GW 5053	Stockholm London									
10	**Die niederländische Fassung (Paris 1500)**										
	GW 5066	Paris									

Testen der neuen Tools und des Workflows: Die Projektgruppe hat die von Kallimachos neu entwickelten Verfahren und Instrumente (z.B. Transkriptionseditor, Dokumentenverwaltung, studentisches Crowdsourcing, Semantic MediaWiki, neue Segmentierungs- und OCR-GUI) ausführlich in der Praxis getestet. Sie war insbesondere an Anpassung und Ausbau des Semantic MediaWiki (mit Funktionen zur Auszeichnung von Varianz, Auszeichnung von Registereinträgen, Auflösung der Marginalien usw.) und an der Entwicklung des automatischen Segmentierungstools (LAREX) maßgeblich beteiligt. Die Weiterentwicklung des OCR-Tools und der Aufbau einer Ground-Truth-Datenbank ([Reul et al., 2017c]) wurde von der Projektgruppe durch Transkription und Korrekturen ausgiebig unterstützt. Ein detailliertes Konzept für die geplante Online-Präsentation der 'Narrenschiffe' wird, nachdem sich die UB aus diesem Bereich zurückgezogen hat, Use-Case-intern erstellt. In einer von der Projektgruppe geleiteten Forschungsübung im WS 2016/17 haben Studierende die Transkription und den Lesetext des Augsburger Narrenschiffs (GW 5045) erstellt und die hierfür bereitgestellten Tools in der universitären Praxis getestet.

Vorträge und Publikationen (siehe unten): Im Berichtszeitraum haben die Antragsteller und Mitarbeiterinnen das Projekt in Vorträgen auf Tagungen und Workshops der Digital Humanities (Stuttgart 2016, Erlangen 2017, Würzburg 2017), Altgermanistik (Hesselberg 2016, Castellabate 2016), Romanistik (Turin 2016, Brixen 2017, Warschau 2017) vor- und zur Diskussion gestellt. Mehrere wissenschaftliche Aufsätze sind erschienen bzw. im Erscheinen. Die Antragsteller und Mitarbeiter haben Lehrveranstaltungen zum Thema abgehalten und betreuen mehrere studentische Abschlussarbeiten zu den 'Narrenschiffen' des 15. Jahrhunderts.

Kooperationen: Die Würzburger Projektgruppe steht in engem Kontakt mit Kollegen der Buchwissenschaft und Frühneuzeitforschung und in Kooperation mit:

- Universitätsbibliothek Basel, Abt. Handschriften und Alte Drucke
- Bibliothek Otto Schäfer, Schweinfurt
- PD Dr. Michael Rupp, German. Mediävistik und Frühneuzeitforschung, Univ. Karlsruhe
- Dr. Anne-Laure Metzger-Rambach, Université de Bordeaux
- Projekt „Mittelniederdeutsch in Lübeck" (Dr. Robert Peters, Norbert Lange, Univ. Münster)

Publikationen, Vorträge, Präsentationen

Publikationen

- Brigitte Burrichter: Rahmen und intendiertes Publikum. Die Paratexte in Sebastian Brants 'Narrenschiff' und seinen Übersetzungen. In: Ajouri, Philip / Kundert, Ursula / Rohde, Carsten (Hg.): Rahmungen. Präsentationsformen und Kanoneffekte. Berlin 2017 (Beiheft zur Zeitschrift für deutsche Philologie), S. 207-222.
- Christine Grundig, Joachim Hamm, Viktoria Walter: Narragonien digital. Mit einer Analyse von Kapitel 4 des ‚Narrenschiffs' in Ausgaben und Bearbeitungen des 15. Jahrhunderts. Erscheint 2018 in: Wolfenbütteler Notizen zur Buchgeschichte (im Druck)
- Christine Grundig: Theologische Überformung des ‚Narrenschiffs' – Geiler von Kaysersberg und die sogenannte ‚Interpolierte Fassung'. In: Archiv für das Studium der neueren Sprachen und Literaturen 254 (2017), S.1-16.
- Joachim Hamm: Zu Paratextualität und Intermedialität in Sebastian Brants *Vergilius pictus* (Straßburg 1502). In: Diesseits des Laokoon. Intermedialität in der Frühen Neuzeit. Formen, Funktionen, Konzepte. Tagung an der Univ. Eichstätt, 28.-31.3.2012. Hg. v. Jörg Robert. Berlin, Boston 2017, S. 236-259.
- Joachim Hamm: Intermediale Varianz. Sebastian Brants 'Narrenschiff' in deutschen Ausgaben des 15. Jahrhunderts. In: Überlieferungsgeschichte transdisziplinär. Neue Perspektiven auf ein germanistisches Forschungsparadigma. In Verbindung mit Horst Brunner und Freimut Löser hg. v. Dorothea Klein. Wiesbaden 2016 (Wissensliteratur im Mittelalter 52), S. 223-240.

Studentische Abschlussarbeiten

- Rena Buß: „Lübecker Unbekanntheiten. Ein Verfasserprofil anhand paratextueller Konzepte in ‚Dat narren schyp' und anderen Mohnkopf-Drucken" (Zulassungsarbeit zum 1. Staatsexamen für das Lehramt am Gymnasium, Fach Deutsch)
- Maximilian Wehner: „Topographie der „Verkehrten Welt". Zur Ausgestaltung literarischer Räume in Sebastian Brants ‚Narrenschiff'" (Zulassungsarbeit zum 1. Staatsexamen für das Lehramt am Gymnasium, Fach Deutsch)
- Amelie Eileen Laut: „Das ‚Narrenschiff'. *Gedruckt zuo Augspurg; Gedruockt zu Nueremberg.* Die Offizin Schönsperger und deren Nachdruck der *editio princeps"* (Zulassungsarbeit zum 1. Staatsexamen für das Lehramt an Realschule, Fach Deutsch)
- Christine Grundig: Text und Paratext. Konzepte von Paratextualität in deutschsprachigen Werken Sebastian Brants. Masch. Magisterarbeit. Würzburg 2012.

Vorträge

- Brigitte Burrichter: La Nef des fous de Sebastian Brant dans le context européen, Paris, Ecole normale supérieure, 05.02.2018.
- Brigitte Burrichter: Patrice et les Dernydes. Les versions française de la Nef des Fous de Sebastian Brant, Vortrag bei der Tagung von fabula, Warschau 18.-20.10.2017.
- Joachim Hamm: Eine integrierte digitale Edition der 'Narrenschiffe' vor 1500. Vortrag in der Vortragsreihe des Akademieprojekts "Der Österreichische Bibelübersetzer", Univ. Augsburg, 30.11.2017.
- Brigitte Burrichter: Sebastian Brants Narrenschiff und seine europäische Rezeption. Vortrag bei der Tagung der Internationalen Oswald-von Wolkenstein- Gesellschaft, Brixen, vom 13.-15. September 2017.
- Joachim Hamm: Gelehrte Narreteien. Das 'Narrenschiff' von Sebastian Brant und das Würzburger Projekt "Narragonien digital". Vortrag im Alten Rathaus von Miltenberg in der Vortragsreihe des Unibundes, 16.1.2017.
- Joachim Hamm: Textuelle und intermediale Varianz im digitalen Kontext am Beispiel des Editionsprojekts "Narragonien digital". Vortrag im Workshop "Textvarianten in der digitalen Edition" des Instituts für Musikforschung (Univ. Würzburg). 19.-20.1.2017.
- Brigitte Burrichter, Joachim Hamm: Narragonien digital. Vortrag im Workshop "Digitale Paläographie" (Interdisziplinäres Zentrum Editionswissenschaften, IZED), Univ. Erlangen, 12.-13.01. 2017.
- Joachim Hamm: Die digitale Edition von 'Narrenschiffen' des 15. Jahrhunderts ("Narragonien digital"). Gastvortrag an der Univ. Stuttgart, Digital Humanities (Prof. Dr. Gabriel Viehhauser), 15.12.2016.
- Christine Grundig: Theologische Überformung des 'Narrenschiffs' - Geiler von Kaysersberg und die sog. "Interpolierte Fassung". Vortrag beim 13. Altgermanistischen Kolloquium am Hesselberg, 4.-6.10.2016.
- Brigitte Burrichter, Raphaëlle Jung: Les Nefs des fols en ligne. Présentation d'un projet d'édition en ligne des "Nefs des fols" du XVe siècle. Vortrag bei der Jahrestagung der Association Internationale pour l' Étude du Moyen Français in Turin, 28.9.-1.10. 2016.
- Brigitte Burrichter, Joachim Hamm: Narragonien digital. Vortrag beim XLIV. Internationalen Mediävistischen Colloquium in Castellabate (IT), 10-17.9.2016.
- Christine Grundig: *Narren en mouvance*. Adaptationen des *Narrenschiffs* im 15. Jahrhundert. Vortrag beim Workshop Wissen von Mensch und Natur. Tradierung, Aktualisierung und Vermittlung in frühneuzeitlichen Übersetzungen des DFG-Graduiertenkollegs *1876 Frühe Konzepte von Mensch und Natur* an der Universität Mainz,19.2.-20.2.2016.
- Brigitte Burrichter: Rahmen und intendiertes Publikum. Die Paratexte in Sebastian Brants 'Narrenschiff' und seinen Übersetzungen. Vortrag bei dem Theorie-Workshop *Rahmungen. Präsentationsformen kanonischer Werke* des Forschungsverbundes Marbach Weimar Wolfenbüttel, Projekt *Text und Rahmen*, vom 29.-31.7.2015 an der Herzog August Bibliothek Wolfenbüttel. (Publikation siehe oben).
- Brigitte Burrichter, Joachim Hamm: Narragonien digital. Vortrag bei der Tagung *Inkunabeln und Überlieferungsgeschichte* des Wolfenbütteler Arbeitskreises für Bibliotheks-, Buch- und Mediengeschichte an der Universität Mainz, 29.6.-1.7.2015.
- Christine Grundig: Sebastian Brants 'Narrenschiff': Zur Bild-Text-Relation in deutschsprachigen und europäischen Ausgaben des Werkes. Vortrag beim 10. Altgermanistischen Kolloquium am Hesselberg vom 1.-3.10.2013

Use Case 2: Anagnosis - Erstellung Kollationsvorlage schriftspiegelkonform

Der Beitrag von Anagnosis zum AP1 begrenzte sich, aufgrund der Entscheidung, ab Januar 2016 ein OCR-unabhängiges Text-Bild-Alignment zu entwickeln, auf die Erstellung einer Ground Truth zur weiteren Anwendung hinsichtlich des – in den Projektzielen jedoch nicht eingeschlossenen – Aufbaus eines spezifischen OCR-Verfahrens für griechische Papyri und vergleichbare historische Dokumente (s. dazu die Ausführungen unter AP2 und AP3). Die Erstellung der Ground Truth erfolgte durch eine schriftspiegelkonforme (d.h. dem für literarische Papyri typischen, in Kolumnen strukturierten Layout entsprechende) Kodierung der Texte in TEI/EpiDoc-Format anhand sowohl der jeweiligen Standardeditionen als auch – v.a. bei immer noch korrekturbedürftigen Fällen – photographischer Abbildungen. Die auf diese Weise digitalisierten Texte konnten dann im Rahmen des geplanten Arbeitsablaufs als Vorlage für die Automatisierung der buchstabengenauen Zuordnung mit den dazugehörigen Abbildungen verwendet werden.

Use Case 3: Schulwandbilder - Scan, OCR Begleitmaterialien, Lehrerhdb.; weitere Medientypen: Zeitschriften, Zeitungen 19. Jh. - OCR-Strategien

Die zu den Schulwandbildern gehörigen Lehrerbegleithefte wurden im Digitalisierungszentrum der UB gescannt. Es wurden 9215 Digitalisate aus 337 Begleitheften erzeugt. Verfahren zur automatischen Verschlagwortung wurden erprobt, jedoch noch nicht abgeschlossen. Weitere Ausführungen zu den Schulwandbildern siehe unter AP4.

Ein prototypischer OCR-Workflow für ABBYY FineReader Recognition Server wurde anhand der außerhalb des Projektkontextes i.e.S. digitalisierten Zeitschrift „Daheim“ (ca. 66.000 Seiten, knapp 3 TB Daten, komplett gescannt, beschnitten und binarisiert) durchgeführt. Der angestrebte Wert von 98% Zeichengenauigkeit wurde erreicht, im Fließtext konnte eine Erkennungsgenauigkeit von über 99% erzielt werden, die Wortgenauigkeit liegt bei ca. 95%. Font-problematische Überschriften müssen allerdings oft manuell verbessert werden. Außerdem wurde ab dem Jahrgang 1880 eine stärkere manuelle Nachkorrektur insbesondere bei der Segmentierung und korrekten Lesereihenfolge der Textregionen notwendig, da die Komplexität des Layouts mit den Erscheinungsjahren ansteigt. Seit Juli 2016 konnten die Jahrgänge von 1865 bis einschließlich 1906 (ca. 47.000 Seiten) vollständig verarbeitet werden.

AP2: Modul Synoptischer Editor

Aufgabe / Erreichungsgrad	voll	teilweise	nicht
TA 2.0: Pflichtenhefterstellung		X	
TA 2.1: Anpassungen 3-fach-Lupe für Use-Cases			X
TA 2.2: Integration CK-Editior	X		
TA 2.3: Optimierung Usability, Exportformate	X		
TA 2.4: Integration Workflowsystem, TextGrid, Auslieferung		X	
Use Case 1: Narragonien - Manuelle Korrektur Volltext	X		
Use Case 2: Anagnosis - Manuelle Korrektur Volltext, synopt. Darstellung	X		

Bericht UB

Der von Mitarbeitern des Digitalisierungszentrums der UB (Herrn Felix Kirchner) erstellte Transkriptionseditor ermöglicht die verteilte, kollaborative Erstellung einer Transkription in einer Seitensynopse von Bild und Text. Besonders hervorzuheben ist die Möglichkeit, in einem Workflow einzelnen Transkriptoren (z.B. Hilfskräfte) bestimmte Seiten zuzuordnen und die Ergebnisse der Transkription in einer Qualitätsschleife zu inspizieren, ggfs. nachzubearbeiten und dann freizugeben. Gegenüber dem in Arbeitsbericht 2 darstellten Stand wurde aus Zeit- und Kapazitätsgründen keine weiteren Funktionen implementiert (keine 3-fach-Lupe).

Bericht Use Case 1: Narragonien

Die händische Nachkorrektur des OCR-Outputs wird durch den von KALLIMACHOS entwickelten synoptischen Transkriptionseditor erleichtert, der u.a. über eine eigene Benutzerverwaltung zur Planung und Aufgabenverteilung verfügt und die Korrektur und Auswahl der aus heutiger Sicht ungewohnten Drucktypen durch die Einbindung von Typentabellen unterstützt. Die für das frühneuzeitliche Druckbild typischen Sonderzeichen können in den Editor geladen werden und stehen bei der Korrektur schnell parat. Die aufwändige und fehleranfällige Suche nach den korrekten Unicode-Zeichen und die bei der Arbeit in externen Editoren oft auftretenden Probleme bei der Wahl der Textkodierung entfallen. Um die transkribierten Texte in einer vollwertigen digitalen Edition mit synoptischer Funktionalität zu vereinen, sind umfangreiche Auszeichnungen von Text und Bild nötig. Layoutelemente wie Textspalten, Überschriften und Marginalien, aber auch semantische Komponenten wie die argumentative Struktur der Spruchgedichte werden verzeichnet und sollen auch über mehrere Ausgaben des *Narrenschiffs* hinweg auffindbar und vergleichbar sein. Im Wintersemester 2016/17 wurde das Tool im Rahmen einer Lehrveranstaltung von B. Burrichter und J. Hamm unter Mitarbeit der Studierenden erfolgreich zur Erfassung eines kompletten Narrenschiff-Textes genutzt.

Bericht Use Case 2: Anagnosis

Das AP2 Synoptischer Editor wurde im Laufe der ersten Projektphase (November 2014 - Dezember 2015) zuerst durch Erstellung einer eigens entwickelten Testversion eines Online-Editors durchgeführt, der eine parallele Anzeige von Bildern und Texten aus griechischen Papyri ermöglichte. Durch den Einsatz einer Tesseract-basierten OCR und eines Word-Counting-Algorithmus konnte man Digitalisate binarisierter Kupferstiche (d.h. anhand von Abzeichnungen erstellter Faksimiles) von herkulanischen Papyri mit den entsprechenden Transkriptionen verknüpfen.[3]

Da der Use Case jedoch auf die Implementierung von Parallelanzeige und Alignment nicht nur bei Kopien, sondern in erster Linie bei Abbildungen originaler Dokumente grundsätzlich abzielt, wurde durch die Januar 2016 angefangene Kooperation mit dem DFKI das Spektrum der zu erfassenden Abbildungen auf photographische Aufnahmen von Papyri ausgedehnt. Bis Projektende wurde hinsichtlich des automatisierten Alignments hauptsächlich eine OCR-freie Herangehensweise getestet und entwickelt, die sich eines im Bereich der Computer Vision angewandten Verfahrens bedient – wodurch der Beitrag von *Anagnosis* im Rahmen des AP1 OCR-Modul auf eine automatisierte Bereitstellung der Ground Truth durch Bild-Text-Alignment für eine künftige Anwendung von OCR-Verfahren auf Papyri begrenzt werden musste. Nach einem Preprocessing, zu dem Binarisierung, Zeilensegmentierung und halbautomatische Festlegung der Zeilenreihenfolge gehören, wird die XML-Transkription als Basis für die Generierung eines synthetischen Bildes des jeweiligen Papyrusfragments benutzt. Die dafür nötigen Glyphen wurden zunächst aus einem manuell erstellten Datensatz bezogen, das aus Ausschnitten zahlreicher Bilder mit Varianten von Buchstabenformen besteht, bei denen die Entsprechung des zugehörigen Unicode-Zeichens bereits angegeben ist. Der Abgleich zwischen synthetischem Bild und originalem Bild erfolgt durch einen ORB-Algorithmus (*Oriented FAST and rotated BRIEF*). Dieser basiert auf der Erkennung von Ähnlichkeitszonen (*keypoints*) durch Pixelanalyse und

3 S. [Damiani, 2016].

-matching. Das Output erhält die Form von in Pixel ausgedrückten Letter-Spacing Informationen für jede Textzeile.

Die Software ist zuerst prototypisch als lokale Anwendung für Unix-Systeme und dann als Weboberfläche implementiert worden. Diese erlaubt nun den Benutzern, sowohl in der Preprocessing-Phase als auch nach der Erzeugung der automatischen Bild-Text-Verknüpfung die Ergebnisse zu bewerten und ggf. zu korrigieren. Das durch die manuelle Korrektur angereicherte Dataset trägt somit zur Erstellung eines Corpus im Goldstandard bei, welches wiederum für den Einsatz von Lernsoftware zur automatisierten Erkennung von Buchstaben (OCR) in griechischen literarischen Papyri verwendet werden kann.

AP3: Wiki-Modul

Aufgabe / Erreichungsgrad	voll	teilweise	nicht
TA 3.0: Pflichtenhefterstellung		X	
TA 3.1: Integration JAMWiki-Modul in Gesamtworkflow, Archiv-Anschluss		X	
TA 3.2: Anpassungen Use Cases	X		
TA 3.3: Formatintegration Volltext Wiki, Editor, OCR, Archivierung	X		
TA 3.4: Integration Workflowsystem, TextGrid, Auslieferung		X	
Use Case 1: Narragonien	X		
Use Case 2: Anagnosis - Kommentierung extern			X
Use Case 3: Schulwandbilder - Metadatenanreicherung, Komment. extern	X		

Für die automatische Extraktion von Schlüsselworten aus Begleitheften wurde ein regelbasiertes Verfahren entwickelt, das mit Anleitung zur Nutzung im Git des Instituts für Informatik bereitgestellt wird[4]. Die Grundidee der Architektur ist folgende: Das Bild beschreibende Schlüsselwörter sind in der Regel Substantive. Demzufolge ist jedes im Text erwähnte Substantiv ein Kandidat für ein Schlüsselwort. Mittels Regeln werden diesen Kandidaten in mehreren Pässen über den Text Scores zugeordnet. Diese Scores werden letztendlich dazu genutzt, die Kandidaten zu sortieren, so dass Kandidaten mit einem hohen Score am wahrscheinlichsten Schlüsselworte sind. Es fehlt allerdings eine qualitative Aussage der Güte dieser Komponente.

Bericht Use Case 1: Narragonien digital

Eine auf Semantic MediaWiki (SMW) beruhende Editionsumgebung für den Use Case 1 (“Narragonien digital”) wurde gemäß den Anforderungen der Projektgruppe erstellt und nach einer initialen Testphase in den Editionsworkflow integriert. Die separate Testversion des Wikis blieb im Folgenden für die Entwicklung zusätzlicher Auszeichnungsmöglichkeiten in Betrieb. Per CSV-Import können die im synoptischen Transkriptionseditor korrigierten Editionstexte in die Wiki-Datenbank übertragen werden. Das Semantic Media Wiki wurde so in der Abschlussphase der ersten Projektlaufzeit weiter ange-

4 https://gitlab2.informatik.uni-wuerzburg.de/mak28ma/KeyWordExtraction

passt. Es wurden alle geplanten Exemplare in das Wiki eingestellt und Korrektur gelesen (GW 5062 und 5064 wurden zurückgestellt, die Varianten wurden analog erfasst). Die Registereinträge (insgesamt mehr als 1100) sind komplett abgeschlossen. Die Marginalien wurden zur Hälfte aufgelöst. Die Varianz in GW 5061 wurde zum großen Teil bereits analog erfasst In enger Zusammenarbeit mit dem Informatiker der UB wurde ein Varianzmodul in das Wiki implementiert, es zeigte sich allerdings, dass zur Auszeichnung der Varianz das Wiki nicht geeignet ist. Sie wird in der zweiten Projektphase in TEI ausgezeichnet werden.

Bericht Use Case 3: Schulwandbilder

Ausführungen zu den Schulwandbildern siehe unter AP4.

Bericht Use Case 2: Anagnosis

Die ursprünglich vorgesehene Beteiligung von *Anagnosis* am AP3 Wikimodul hinsichtlich der externen Kommentierung der Ergebnisse wurde dadurch aufgehoben, dass selbst in dem als Weboberfläche verwendbaren Editor sowohl ein System zur Benutzerverwaltung auf verschiedenen Ebenen (Administrator, Mitarbeiter, Gast) als auch ein Issue-Reporting Service integriert wurden: Diese Maßnahmen sollen ebenso wie die Verwendung der anfangs dazu gedachten Open Source JAMWiki-Plattform eine auf Crowdsourcing basierte, nachhaltige Erweiterung des durch die Software bearbeiteten Datensatzes ermöglichen.

AP4: Schnittstelle Repositories-Datenanalyse

Aufgabe / Erreichungsgrad	voll	teilweise	nicht
TA 4.0: Pflichtenhefterstellung		X	
TA 4.1: Korpuszusammenstellung via Metadaten	X		
TA 4.2: Aufbereitung Texte für UIMA	X		
TA 4.3: Entwicklung Textanalysewerkzeuge in UIMA	X		
TA 4.3.1 Entwicklung regelbasiertes NE-Annotationswerkzeug	X		
TA 4.3.2 Erstellung statistischer NE-Modelle	X		
TA 4.3.3 Einbeziehung Coreference Resolution	X		
TA 4.3.4 Identifizierung, Strukturierung von NE	X		
TA 4.3.5 Stilometrische Analyse, Autoren- und Kontextspezifik	X		
TA 4.3.6 Evaluationskomponente	X		
TA 4.4: Einbindung Analyseergebnisse nach TextGrid		X	
TA 4.5: Anbindung UIMA-Pipeline an Python, R		X	

TA 4.6: Demonstrationsprojekte als Webservices, Auslieferung	X		
Use Case 3: Schulwandbilder - Metadatenanreicherung, SW automatisiert		X	
Use Case 4: Provenienz- und Gattungsbestimmung		X	
Use Case 6: Leserlenkung in Bezug auf Figuren		X	

Es wurden für ein Korpus aus Mittel- und Hochliteratur ausführliche Metadaten, wie beispielsweise Autornamen, Veröffentlichungsjahr, Genre und Happyend-Informationen erhoben.

In der Arbeitsgruppe Jannidis wurde ein Korpus aus 3800 Romanen, erschienen zwischen 1800 und 1930, zusammengestellt. Die Texte stammen aus dem Textgrid Repository[5], sowie dem deutschen Ableger der Initiative Projekt Gutenberg[6]. Die Romane sind vollständig in TEI/P5 annotiert und stark am DTA-Basisformat angelehnt. Dieses wurde lediglich um die Möglichkeit zusätzliche Schlagwörter zu ergänzen erweitert, z.B. Happyend/kein Happyend (siehe Usecase 4/5). Alle Texte sind mit einem eigens gefertigten Tool (siehe TA 4.2) nach Apache UIMA protierbar. 90 Fragmente dieser Texte der Sammlung bilden das Korpus DROC, welches zusätzlich manuelle Annotationen zu Figurenreferenzen, Koreferenzen und direkter Rede enthält. Während die Texte aus dem Textgrid Repository zuverlässige Metadaten enthalten, sind die aus Projekt Gutenberg stammenden Metadaten teils falsch oder unvollständig. Der Überprüfungsprozess dieser Daten dauert über Projektende hinweg an.

TA 4.2: Aufbereitung Texte für UIMA

Es wurde ein Konverter[7] entwickelt, der Texte, mit TEI-Xml als Eingabe automatisch nach UIMA konformen .xmi-Dateien konvertieren kann.

TA 4.3: Entwicklung Textanalysewerkzeuge in UIMA

Die in Use Case 4-6 (TA 4.3.1 - 4.3.6) entwickelten Werkzeuge lassen sich mit UIMA ansteuern.

TA 4.3.1 Entwicklung regelbasiertes NE-Annotationswerkzeug

Es wurde ein Skript in Apache UIMA Ruta entwickelt, dass zu Beginn der Laufzeit dazu genutzt wurde, Named Entities im Text zu markieren und später in ATHEN zu korrigieren. Dieses Skript liefert aktuell wesentlich schlechtere Ergebnisse, als die in 4.3.2 entwickelte statistische Methode.

TA 4.3.2 Erstellung statistischer NE-Modelle

Auf Basis des Entwickelten Korpus DROC, wurden mehrere Experimente mit maschinellem Lernen mit dem Ziel der Erkennung von Figurenreferenzen (NE`s). Als dabei bestes Verfahren stellte sich ein Maximum Entropy Klassifikator mit semantischer Generalisierung über Word2Vec heraus, der über die in Kallimachos entwickelte Pipeline-Jar auf beliebige Texte angewendet werden kann.[8] Zusätzlich kann basierend auf DROC ein Modell für das NE-Modul des Stanford-NE Taggers heruntergeladen und eingesetzt werden. Dazu ist sonst keine weitere Vorverarbeitung nötig.

5 https://textgrid.de/digitale-bibliothek

6 https://www.gutenberg.org/wiki/DE_Hauptseite

7 https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/TextConversion

8 https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/KallimachosEngines

TA 4.3.3 Einbeziehung Coreference Resolution

Die Koreferenzauflösung der NE's, die mit dem in TA 4.3.2 entwickelten Werkzeug automatisch annotiert werden können stellt sich als sehr große Herausforderung dar. Der Vergleich statistischer Methoden mit regelbasierten Methoden ergab, dass für diese Domäne regelbasierte Verfahren besonders gut geeignet sind. Daher wird auch die Regelbasierte Methode, stark angelehnt an das Modul aus Stanford in der Pipeline ausgeliefert. Die Qualität des Moduls ist allerdings lediglich bei etwa 85% MUC-F1 Score und 53% B^3-F1 Score.

TA 4.3.4 Identifizierung, Strukturierung von NE

Für die Beziehungen zwischen NE`s wurde eine hierarchische Ontologie mit insgesamt 57 Labels entwickelt. Ein Ausschnitt davon ist in Abbildung 4 (TODO fix Nr) gezeigt.Für Entitäten (Cluster von NE`s) kann mit Hilfe der Coreference Resolution ein Aussagekräftig Name gewählt werden. Bisher liegen keine Daten für eine feingranularere Attribuierung der Entitäten vor.

TA 4.3.5 Stilometrische Analyse, Autoren- und Kontextspezifik

Die Pythonbibliothek PyDelta[9] bietet eine Implementierung der wichtigsten stilomertischen Maße zur Identifizierung von Autorschaft und die Möglichkeit zur Evaluation der Ergebnisse. Die Ergebnisse der Untersuchungen wurde in den folgenden Publikationen dargestellt:

[Evert et al., 2015a], [Evert et al., 2015c], [Evert et al., 2016a], [Evert et al., 2016c].

TA 4.3.6 Evaluationskomponente

Eine Evaluationskomponente für die interaktive Gegenüberstellung bis zu n verschiedener Annotatoren (manuell oder automatisch) für die Mächtigkeit des Typsystems von UIMA wurde in ATHEN integriert.

TA 4.4: Einbindung Analyseergebnisse nach TextGrid

Die Einbindung nach TextGrid verzögert sich über das Projektende hinaus. Die zugehörige Veröffentlichung ist als DARIAH Working Paper geplant.

TA 4.5: Anbindung UIMA-Pipeline an Python, R

Für die in Python entwickelten Teile wurde ein Python-Modul[10] zum Lesen und Schreiben der von UIMA generierten xmi-Dateien entwickelt. Damit lassen sich auch diese in eine UIMA-Pipeline einbinden.[11]

TA 4.6: Demonstrationsprojekte als Webservices, Auslieferung

Zu Demonstrationszwecken wurden zu den veröffentlichten Techniken ausführliche Online-Tutorials entwickelt, mit denen sich die Ergebnisse reproduzieren und auf andere Datensätze übertragen lassen. Die Tutorials sind auf der Webseite des Kallimachos-Projektes abrufbar.[12]

Use Case 3: Schulwandbilder - Metadatenanreicherung, SW automatisiert

Im Rahmen des Teilprojekts ‚Schulwandbilder digital' wurden insgesamt über 25.000 Schulwandbilder gescannt und mit umfassenden Metadaten versehen (der Scanvorgang selbst wurde unabhängig vom Kallimachos-Projekt finanziert). Für die bibliographische Erschließung wurde der Metadatenedi-

9 https://github.com/cophi-wue/pydelta

10 https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/PyCAS

11 https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/KallimachosPythonUIMAEngines

12 http://kallimachos.de/kallimachos/index.php/Tutorial_Figurennetzwerke, http://kallimachos.de/kallimachos/index.php/Tutorial_Emotionsverläufe_und_Happy_Ends, http://kallimachos.de/kallimachos/index.php/KallimachosEngines

tor verwendet, der während einer früheren Kooperation aus UB Würzburg und Lehrstuhl II für Informatik entwickelt wurde (Java-Webstart Anwendung in Verbindung mit einer Postgres Datenbank). Zur Darstellung von Scans und Metadaten und zur optimalen Unterstützung der Forschenden durch eine komplexe Suche wurde mit Hilfe von Apache Solr und Apache Velocity eine Suchplattform vom Digitalisierungszentrum der UB entwickelt. Diese ermöglicht komplexe Suchanfragen inkl. Phrasensuche, Trunkierung, negierte und unscharfe Suche und individuelle Einstellung der Gewichtung. Darüber hinaus erlaubt die Suchplattform die Verwendung von Facetten zur schnellen Eingrenzung der gesuchten Objekte. Die Suchplattform steht jedem Nutzer innerhalb der Forschungsstelle für historische Bildmedien zur Verfügung und kann auch an den von der Forschungsstelle im Lagerraum bereitgestellten Touch-Monitoren verwendet werden. Dies ermöglicht ein schnelles Auffinden der Schulwandbilder innerhalb des weitläufigen Lagerraums. Die Entwicklung der Suchplattform mittels Apache Velocity erlaubte ein Template-artiges Design wodurch die Plattform binnen kürzester Zeit auch für andere Sammlungsbestände angepasst und verwendet werden kann.

Use Case 4/5: Provenienz- und Gattungsbestimmung/Narrative Techniken

Das Ziel dieses Use Case ist eine Analyse von Romantexten hinsichtlich ihrer Erzählstruktur. Dazu zählt beispielsweise die automatische Erkennung von literarischen Subgenres oder aber wichtiger Ereignisse im Verlauf des Romans. Wesentliche Ergebnisse wurden in den Bereichen der Subgenreklassifikation, Verwendung von Signifikanztests, Modellierung der Handlungsstruktur und der Sentiment Analysis erzielt.

In [Hettinger et al., 2015] und [Hettinger et al., 2016a] wurde die Verwendung von Features aus LDA ([Blei et al., 2003]), häufigen Wörter, sowie Figurennetzwerken zur automatischen Erkennung von Subgenres evaluiert. Die besten Ergebnisse wurden mit den häufigsten Wörtern erzielt, es zeigte sich allerdings auch, dass die Klassifikation für manche Subgenres deutlich höhere Genauigkeit liefert als für andere. So konnten etwa Abenteuerromane besser erkannt werden als andere Genres, wohl aufgrund der größeren thematischen Entfernung. Darauf aufbauend wurde in [Hettinger et al., 2016b] untersucht, inwieweit Segmentierung Abhilfe für zu kleine Datensätze schaffen kann. Hierzu wurden Romane in kleinere Abschnitte getrennt, um mehr Trainingsdaten verfügbar zu haben. Dieser Ansatz zeigte sich allerdings als wenig vielversprechend. Darüber hinaus wurde gezeigt, dass gerade auf kleinen Datensätzen die Verwendung von Signifikanztests unerlässlich ist, da sonst leicht eigentlich unwesentliche Änderungen in der Klassifikationsgüte überschätzt werden.

Da vollständige Romane eine sehr große Einheit für Klassifikationsaufgaben darstellen, wurde anschließend beschlossen, aus diesen zunächst wesentliche Handlungselemente zu extrahieren, um anhand deren Abfolge den Verlauf eines Romans besser beschreiben zu können. Als eines dieser wesentlichen Handlungselemente wurde das Vorhandensein oder Fehlen eines Happy Ends identifiziert. In [Zehe et al., 2016] wurde ein auf relativ einfacher Sentiment Analysis (die automatische Erkennung von Gefühlen in Text) basierendes System vorgestellt, das in der Lage ist, automatisch Happy Ends in Romanen zu erkennen.[8] In [Jannidis et al., 2016] wurde ausgehend von dieser Klassifikation einerseits untersucht, welche Features besonders stark zum Ergebnis beitragen, andererseits für welche Romane die Klassifikation mehr oder weniger gut funktioniert. Es zeigte sich, dass vorrealistische Romane leichter zu klassifizieren sind, was vermutlich mit der größeren Schemahaftigkeit der Romane dieser Zeit erklärt werden kann.

Um die Ergebnisse der Happy End-Erkennung weiter zu verbessern, wurden in [Zehe et al., 2017] einige state-of-the-art Systeme zur Sentiment Analysis für die Verwendung auf deutscher Literatur adaptiert. Dazu wurde ein annotierter Datensatz von Sätzen aus deutschsprachigen Romanen zusammengestellt.[13] Auf diesem Datensatz wurden die Systeme trainiert und evaluiert. Es konnte ge-

13 http://www.dmir.uni-wuerzburg.de/datasets/german-novel-dataset/

zeigt werden, dass die Methoden grundsätzlich im Bereich deutscher Literatur eingesetzt werden können, allerdings ist der vorhandene Datensatz noch zu klein, um gute Ergebnisse zu erzielen.

Use Case 6: Leserlenkung in Bezug auf Figuren

Die Hauptergebnisse die zur Laufzeit des Projektes Kallimachos für diesen Use Case sind:

- Die Entwicklung und Bereitstellung einer eigenstehenden Pipeline für die vollautomatische Vorverarbeitung literarischer Texte. Dies beinhaltet folgende Komponenten:
  - Tokenisierung, Satzendenerkennung, POS-Tagging und Parsing
  - Figurreferenzerkennung (NE)
  - Erkennung von Sprecher und Angesprochenen direkter Reden
  - Koreferenzerkennung
  - Relationserkennung
  - Interaktionserkennung

  Die Ergebnisse können dabei entweder als UIMA konforme .xmi-Datei oder als spaltenbasiertes Format ausgegeben werden. Die Anwendung ist in Java entwickelt und sollte auf jeder Plattform lauffähig sein.

- Die Entwicklung und Bereitstellung von ATHEN[14] (Annotation and Text Highlighting ENvironment), einer general-purpose Annotationsumgebung für die Annotation UIMA-konformer Dokumente. Die für diesen Use-Case entwickelte Pipeline kann direkt aus ATHEN ausgeführt werden und die Ergebnisse manuell nachgebessert werden. Darüber hinaus wurde ATHEN mit der Vision entwickelt, auch über Kallimachos hinaus erweitert zu werden, indem die OSGI-Architektur des Eclipse RCP Frameworks verwendet wurde. Es wurden zahlreiche Markdown-Tutorials erstellt.
- Die Visualisierung der Ergebnisse der Figurenerkennung und Koreferenzauflösung mittels Sozialer Netzwerke, mit verschiedenen Methoden [Krug et al., 2017a].
- Anschließend an die Arbeiten zur Figurenerkennung, Koreferenzauflösung und Erkennung des Sprecher und des Angesprochenen wurden Relationen zwischen Romanfiguren in den Fokus der Arbeit gestellt [Krug et al., 2017c]. Dies ist relevant, da solche Relationen bei der automatischen Erstellung von Figurennetzwerken oder bei der Charakterisierung von Figuren einbezogen werden können.
- Zunächst wurden Relationen zwischen zwei Figurenreferenzen in ausgewählten Sätzen nach einem hierarchischen Schema manuell annotiert, das in 4 Haupttypen von Relationen gegliedert ist (s. Abbildung). Über einen gesamten Roman hinweg sind Textstellen, an denen solche Relationen explizit genannt werden, jedoch rar. Daher wurden Methoden zur effizienteren Annotation untersucht [Krug et al., 2017b]: Zum einen wurde Active Learning verwendet, um Sätze zur Annotation vorzuschlagen. Außerdem wurde versucht, Zusammenfassungen zu den Romanen zu nutzen, da diese kurz sind und alle wichtigen Informationen gebündelt enthalten.

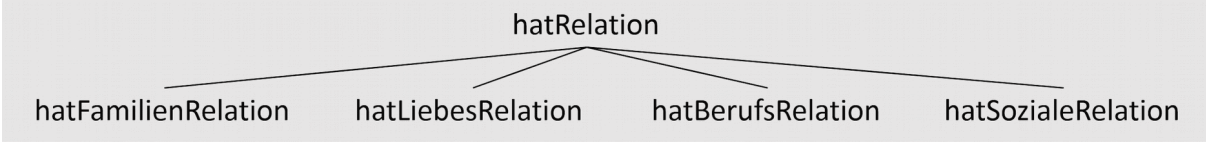

Abbildung 5: Schema der 4 Haupt-Relationstypen

14 https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/Athen.

- Für die automatische Erkennung der Relationen wurden regelbasierte und featurebasierte Ansätze verglichen, wobei eine Kombination aus beiden mit 73% für das Vorliegen einer Relation am besten abschneidet. Die vier Haupttypen können mit einer Genauigkeit von 61% erkannt werden. Ein erfreuliches Ergebnis ist, dass Familienrelationen eine hohe Präzision über 90% aufweisen, so dass, wenn eine Familienrelation gefunden wurde, diese mit einer hohen Wahrscheinlichkeit korrekt ist.
- Während der vergangenen Monate wurde das selbstentwickelte, generische Annotationstool ATHEN unter anderem um die Möglichkeit erweitert, effizient Relationen zwischen zwei Figurenreferenzen zu annotieren. Positiv ist festzuhalten, dass ATHEN bereits von weiteren Projekten[15] nachgenutzt wird. Die entwickelten NLP-Komponenten (Extraktion von Figurenreferenzen, Koreferenzauflösung inklusive Sprechererkennung [Krug et al., 2016], Relationserkennung und Klassifikation) stehen sowohl als UIMA Komponenten zur Verfügung,[16] als auch gepackt in einer .jar-Datei zum Ausprobieren für jedermann.[17] Die Gitlab-Repositories enthalten zu den jeweiligen Projekten Erklärungen in Form von Markdown-Tutorials, um deren Nutzung und Funktionalität zu beschreiben.

AP5: Aufbau prototypischer Arbeitsabläufe zur Datenanalyse

Aufgabe / Erreichungsgrad	voll	teilweise	nicht
TA 5.1: Integration von Best-Practice-Implementierungen zu Workflows	X		
TA 5.2: Methodologische Klärungen, use-case-spezifisch	X		
TA 5.3: Schulungskonzepte	X		
TA 5.4: Vorbereitung zur Nachnutzung, Vernetzung Zielportale	X		
Use Case 5: Narrative Techniken		X	
Use Case 7: Identifizierung anonymer Übersetzer		X	

Input Teilprojekt Erlangen:

Das Teilprojekt Erlangen hat seine eigenen Arbeitspakete definiert, die sich im Gesamtprojekt unter das Arbeitspaket 5 einordnen:

AP 1 Erlangen: Statistische Auswertung

TA 1.1: Aufarbeitung des aktuellen Forschungsstands

Die Aufarbeitung des Forschungsstands erfolgte planmäßig und konnte im Rahmen des Seminars „Fortgeschrittene statistische Methoden in der Korpuslinguistik“ auch mit der Lehre verzahnt werden.

15 DFG-Projekt “Redewiedergabe”, http://www1.ids-mannheim.de/lexik/redewiedergabe.html.

16 https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/KallimachosEngines.git.

17 http://ki.informatik.uni-wuerzburg.de/nappi/NLP-preprocessing/.

Durch regelmäßige Literaturrecherchen sowie die Teilnahme von Projektmitarbeitern auf Fachkonferenzen wurde die Entwicklung des Forschungsstands kontinuierlich weiterverfolgt.

TA 1.2: Bedarfsanalyse Use Cases

Die initiale Bedarfsanalyse der Use Cases ergab Anknüpfungspunkte zur Methodenentwicklung für die Use Cases 5 („Narrative Techniken“), 6 („Leserlenkung in Bezug auf Figuren“) und 7 („Identifizierung anonymer Übersetzer“). Im Projektverlauf wurden die in Kooperation mit den Erlanger Projektpartnern untersuchten Deltamaße hauptsächlich in den Use Cases 5 und 7 verwendet, während in Use Case 6 Verfahren zur Named-Entity-Recognition, Koreferenzauflösung und Relationserkennung zum Einsatz kamen um Figurennetzwerke aufzubauen und zu erkennen.

TA 1.3: Weiterentwicklung von Signifikanztests

In enger Anbindung an die Use Cases 5 („Narrative Techniken“) und 7 („Identifizierung anonymer Übersetzer“) wurde die Funktionsweise von Delta-Maßen zur Autorschaftszuschreibung untersucht. Dabei standen insbesondere die quantitativ-methodologischen Aspekte im Fokus. Ein Hauptergebnis dieser Untersuchungen ist die „Schlüsselprofilhypothese“, die eine mathematische Erklärung für die Funktionsweise von Deltamaßen zur Autorschaftsattribution liefert. Demnach manifestiert sich das stilistische Profil eines Autors in einer qualitativen Präferenz für bestimmte Wörter, also im Muster des Über- und Untergebrauchs von Vokabular. Ein Textabstandsmaß ist dann besonders erfolgreich in der Autorschaftszuschreibung, wenn es auf strukturelle Muster in den stilistischen Profilen reagieren kann, ohne allzu sehr von der Amplitude, d.h. den numerischen Absolutwerten, beeinflusst zu werden. Diese Ergebnisse sind ebenso exemplarisch wie richtungsweisend für die Weiterentwicklung korpuslinguistischer Signifikanztests und anderer quantitativer Verfahren.

Die Weiterentwicklung von statistischen Signifikanztests war auch Gegenstand einer Podiumsdiskussion auf der Corpus Linguistics 2015, die von Stefan Evert mit organisiert wurde ([Evert et al., 2015e]). Neue Methoden zum Vergleich von Korpushäufigkeiten auf Basis statistischer Regressionsmodelle (GLM, *generalized linear models*) wurden im Rahmen eines praktischen Workshops beim Symposium Methods and Linguistics Theories (Bamberg, November 2015) vorgestellt und sind in den Online-Kurs SIGIL (http://sigil.r-forge.r-project.org/) eingeflossen.

In Zusammenarbeit mit Andreas Büttner konnte im Rahmen von Use Case 7 („Identifizierung anonymer Übersetzer“) auf Basis von arabisch-lateinischen Übersetzungen gezeigt werden, dass dieselben Merkmale, auf denen Deltamaße basieren, mit Hilfe von Merkmalsselektion dazu verwendet werden können, das Übersetzersignal vom Genresignal zu isolieren.

Erfreulich ist auch die Kooperation mit Friedrich Michael Dimpel von der FAU Erlangen-Nürnberg. Seine Arbeiten auf Basis mittelhochdeutscher Texte, die zum einen den Einfluss von nicht-normierten Schreibweisen auf die Autorschaftsattribution untersuchen und zum anderen zeigen, dass Deltamaße auch auf Basis einer metrischen Analyse erfolgreich angewendet werden können, sind ein gutes Beispiel für eine Nachnutzung unserer Projektergebnisse in den e-Humanities, von der beide Seiten profitieren ([Büttner et al., 2017]).

TA 1.4: Best-Practice-Lehrbuch

Da das ursprünglich geplante Best-Practice-Lehrbuch auch die für das Anschlussprojekt vorgesehenen Arbeiten zu Komplexitätsmaßen mit einbeziehen soll, wurde es in die zweite Projektphase verschoben. Die Hauptergebnisse der ersten Projektphase wurden in einem umfassenden Zeitschriftenartikel zur Autorschaftsattribution dargestellt ([Evert et al., 2017]).

AP 2 Erlangen: Korpus- und computerlinguistische Methoden

TA 2.1: Implementierung korpuslinguistischer Verfahren

Es wurde ein Shared Task zur Verbesserung von Tokenisierung und Part-of-Speech-Tagging deutscher Texte organisiert ([Beißwenger et al., 2016]). In diesem Rahmen wurde ein frei verfügbarer Tokenizer auf State-of-the-Art-Niveau entwickelt ([Proisl und Uhrig, 2016]; https://pypi.python.org/pypi/SoMaJo). Auf Basis der im Shared Task annotierten Daten wurde zusätzlich ein Part-of-Speech-Tagger entwickelt, der ebenfalls State-of-the-Art-Ergebnisse erzielt und frei verfügbar ist ([Proisl, 2018]; https://github.com/tsproisl/SoMeWeTa).

Außerdem wurden die untersuchten Deltamaße in Python implementiert und unter dem Namen PyDelta öffentlich zur Verfügung gestellt (https://github.com/cophi-wue/pydelta; Dokumentation: http://dev.digital-humanities.de/ci/job/pydelta-next/Documentation/index.html)

TA 2.2: Verbesserung computerlinguistischer Werkzeuge

TA 2.2.1: Sentiment Analysis

Im Bereich Sentiment Analyse waren keine Arbeiten mehr notwendig, da die Würzburger Kollegen im Rahmen von Use Case 6 („Leserlenkung in Bezug auf Romanfiguren“) ein entsprechendes System für Romantexte entwickelten ([Zehe et al., 2016]; [Jannidis et al., 2017]).

SemantiKLUE, das bestehende System zur Bestimmung der semantischen Ähnlichkeit von Texten, wurde zwar weiterentwickelt ([Plotnikova et al., 2015]), fand aber in den Use Cases aufgrund veränderter Zielsetzungen, die sich erst im Lauf der Projektarbeit ergeben haben, keine Anwendung.

Stattdessen wurde der Schwerpunkt der Forschungsarbeiten auf die Bestimmung von Textähnlichkeiten mit Delta-Maßen gelegt. In ausführlichen Untersuchungen konnte gezeigt werden, dass diese Maße eine Mischung aus inhaltlichen und stilistischen Ähnlichkeiten berechnen.

Kurz vor Projektende wurden erste Evaluationsexperimente zum Vergleich von N-Gram-Tracing, einem neuen Verfahren zur Autorschaftsattribution, mit Deltamaßen durchgeführt ([Proisl et al., 2018]). In derzeit laufenden Anschlussarbeiten soll N-Gram-Tracing zu einem Textähnlichkeitsmaß weiterentwickelt werden.

TA 2.3: Integration mit UIMA-Pipeline

Die im Rahmen des Teilprojekts entwickelte Software wurde als eigenständige, leicht zu installierende Software in Python implementiert, die auf dem UIMA-Output arbeitet.

Kommunikation/Vermittlung

Methodenwissen über Signifikanztests und andere statistische Verfahren der Korpuslinguistik wurde in einem einwöchigen Kurs bei der Corpus Linguistics Summer School in Birmingham (Evert, 06/2016) sowie in einem eintägigen Intensivworkshop mit der studentischen Projektgruppe Computer-based Analysis of Personal Style in Göttingen (Evert, 06/2016) vermittelt.

Methodenwissen im Bereich maschinelles Lernen wurde im Rahmen des DARIAH-Methodenworkshops unter anderem an die eHumanities-Nachwuchsgruppe „Computergestützte literarische Gattungsstilistik“ weitervermittelt ([Proisl, 2015]).

Methodenwissen über distributionelle Semantik wurde in einem einwöchigen Kurse bei der ESSLLI-Sommerschule 2016 (Evert, 08/2016) und in einem zweiteiligen Tutorial an der DFKI Saarbrücken und der Universität des Saarlandes (Evert, 04/2016 + 03/2017) an Nachwuchsforscher/innen aus e-Humanities, Sprachwissenschaft und Computerlinguistik vermittelt.

Publikationen

- [Evert et al., 2015d]
- [Plotnikova et al., 2015]
- [Beißwenger et al., 2016]
- [Proisl und Uhrig, 2016]
- [Büttner et al., 2017]
- [Evert et al., 2017]
- [Proisl, 2018]
- [Proisl et al., 2018]

Vorträge

- [Evert et al., 2015e]
- [Evert et al., 2015b]
- [Proisl, 2015]
- [Büttner und Proisl, 2016]
- [Evert und Proisl, 2016]
- [Evert et al., 2016b]
- [Evert, 2016]
- [Evert, 2017]

Use Case 7: Identifizierung anonymer Übersetzer

Die Identifizierung anonymer Übersetzer aus dem Arabischen ins Lateinische im 12. Jahrhundert mit einer Textsammlung von etwa 50 überwiegend philosophischen Texten begonnen. Da alle durch stilometrische Analysen getroffenen Aussagen zur Übersetzerattribution stets nur relativ zum verfügbaren Textkorpus möglich sind, wurde die Menge der Texte im Rahmen des Projekts mehr als verdoppelt, wobei die inhaltliche Ausrichtung auch auf weitere wissenschaftliche Disziplinen erweitert wurde.

Um die Analyse der Texte zu erleichtern, wurde ein einfaches Webinterface entwickelt (https://github.com/andbue/altusi), in dem die folgenden Funktionen gebündelt wurden:

1. Manuelle Korrektur der Texte mit Wörterbuchunterstützung (ohne historische Schreibvarianten).
2. Erstellung von Teilkorpora.
3. Anwendung stilometrischer Tools (zunächst stylo ([Eder et al., 2016]), später pydelta (https://github.com/cophi-wue/pydelta) zur Visualisierung von stilistischer Ähnlichkeit.
4. Erweiterte Suchfunktion mit automatischer Konkordanzerstellung und Suche der Hapax legomena.
5. Synopsenansicht zur Suche nach Übereinstimmungen zwischen Texten.

Zusätzlich zu diesem intern verwendeten Analysewerkzeug wurde eine Website erstellt[18], die durch eine einfache Suche die Texte auch für externe Nutzer zugänglich macht.

Dass mit dem Kosinus als Abstandsmaß ("Cosine Delta") auch die Übersetzer stilometrisch unterschieden werden können, wurde anhand des philosophischen Korpus deutlich. Um den mit der Ausweitung des Korpus auf andere Disziplinen einhergehenden Einflüssen des inhaltlichen Vokabulars auf die Analysen entgegenzuwirken, wurden zwei Strategien verfolgt. Erstens wurden mit einem Ma-

18 http://arabic-latin-corpus.philosophie.uni-wuerzburg.de/

schinenlernverfahren (recursive feature elimination) das disziplinenspezifische Vokabular identifiziert und aus den Wortlisten entfernt [Evert et al., DHd 2016.], [Büttner et al., 2017]. Zweitens konnte durch den Ausschluss bestimmter Wortarten eine einfache und effiziente Reduzierung der inhaltlichen Einflüsse erreicht werden [Hasse und Büttner, Erscheinen].

Neben der Analyse der häufigsten Wörter durch Delta wurden auch seltenere stiltypische Wörter und Wortgruppen isoliert, um dabei insbesondere bei kürzeren Texten ein philologisch anschauliches Merkmal zur Identifizierung der Übersetzer zu gewinnen ([Hasse und Büttner,], [Hasse, 2016]) Diese Methode erwies sich gerade bei Doppelübersetzungen bzw. Revisionen von Übersetzungen als sehr hilfreich ([Hasse,]). Dass die Erforschung der Übersetzungen auch weiterhin von der Fruchtbarkeit der stilometrischen Analysen profitieren kann, wird durch den Aufbau eines Textkorpus arabisch-lateinischer Übersetzungen des 10.-14. Jahrhunderts im Rahmen des Leibniz-Programmes gewährleistet.

Use Case 1: Narragonien

Der Workflow des Use-Cases umfasst nahezu alle Einzelprozesse von Kallimachos: Von der Segmentierung und per OCR erfassten zeichengetreuen Texterstellung über die Herstellung eines Lesetextes bis zur Auszeichnung verschiedener Einträge (Lemmata, Marginalien) im Semantic Media Wiki. Die ursprünglich vorgesehene Erstellung einer Präsentationsoberfläche (WebViewer) durch die UB wurde nicht umgesetzt. Der Workflow wird daher in der zweiten Projektphase um die projektinterne Weiterbearbeitung der Daten in TEI erweitert, was dann den Import in einen im Use-Case konzipierten Viewer ermöglicht. Dafür wurde im Use Case “Narragonien” für die zweite Förderphase eine Mitarbeiterin für Digital Humanities eingestellt.

Zur Datenanalyse wurde ein umfangreiches und übergreifendes Register für alle behandelten Narrenschiff-Ausgaben erstellt und mit den Texten verknüpft sowie um Links zu einschlägigen Online-Lexika ergänzt. Eine entsprechende Verknüpfung der Marginalien, deren Quellenhinweise vollständig aufgelöst und mit den entsprechenden Prätext verlinkt werden, ist in Arbeit.

2. Die wichtigsten Positionen des zahlenmäßigen Nachweises

Der entsprechende zahlenmäßige Nachweis wurde mit Datum vom 20.02.2018 von der Universitätsbibliothek (Dr. Schmidt) an die DLR übermittelt.

3. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die Arbeiten waren sowohl notwendig als auch, verglichen mit den in den Zwischenberichten dargestellten Problemen der Gewinnung geeigneten Personals in der UB, sehr erfolgreich. Ein Teil der durch Personalmangel verursachten Ausfälle konnten durch unerwartete Fortschritte, z.B. dem Gebiet der Layoutanalyse und OCR historischer Drucke, wettgemacht werden.

4. Voraussichtlicher Nutzen, insbesondere die Verwertbarkeit des Ergebnisses im Sinne des fortgeschriebenen Verwertungsplans

Wirtschaftliche Erfolgsaussichten nach Projektende

Es ist keine wirtschaftliche Ergebnisverwertung geplant. Unsere Open-Source-Strategie, die unsere Daten offen für die Wiederverwertung zur Verfügung stellt, schließt aber eine wirtschaftliche Nutzung durch Dritte nicht aus. Während die Datengrundlage frei verfügbar ist, steht es Firmen frei, auf Grundlage dieser Daten auch kostenpflichtige und marktfähige Produkte oder Dienste zu entwickeln, etwa um die Daten sehr komfortabel zu durchsuchen und auszuwerten.

Wissenschaftliche und/oder technische Erfolgsaussichten nach Projektende

Mit den im Zentrum entwickelten Softwarekomponenten, die Musterlösungen für Arbeitsschritte anbieten, die bei allen Digitalisierungsvorhaben in ähnlicher Weise anfallen (z.B. einer Anwendung zur örtlich verteilten Erstellung von Ground Truth und die Korrektur von OCR-Ergebnissen, die Abbildung einer Arbeitskette vom Digitalisierungsauftrag über den Scan bis zur Veröffentlichung in einem Portal, das Trainieren von gemischten und werkspezifischen OCR-Modellen auch für Inkunabeln und Frühdrucke, Methoden zur Named-Entity-Recognition etc.), kann unser Zentrum in den nächsten beiden Jahren nicht nur international zu den im Bereich der Digital Humanities führenden USA aufschließen, sondern insbesondere in der OCR von Frühdrucken sogar noch eine Führungsrolle einnehmen. Außerdem konnten prototypische Arbeitsabläufe im Bereich der computergestützten Analyseverfahren, also methodische und technische Lösungen wie z.B. Algorithmen und Routinen zur Transformation von Daten erarbeitet, als Open-Source-Tools beispielhaft implementiert und der Fachgemeinschaft zur Verfügung gestellt werden.

Das am 01.10.2017 begonnene Anschlussprojekt KALLIMACHOS II (Laufzeit: 01.10.2017 - 30.09.2019) wird die Koordination und Entwicklung einschlägiger Open-Source-Komponenten und prototypischer Arbeitsabläufe fortsetzen und der wissenschaftlichen Öffentlichkeit sowie Gedächtnisinstitutionen wie Archiven und Bibliotheken zur Verfügung stellen. Eine enge Zusammenarbeit mit CLARIN, der BBAW (Deutsches Textarchiv) sowie der OCR-D-Initiative der DFG ist initiiert.

5. Während der Durchführung des Vorhabens dem ZE bekannt gewordene Fortschritte auf dem Gebiet des Vorhabens bei anderen Stellen

DFKI:

Im Bereich der Digitalisierung historischer Dokumente gab es im Projektzeitraum für die projektspezifischen Anforderungen und Use Cases keine direkten Fortschritte.

Im Rahmen von HisDoc 2.0, einem Projekt an der Universität Freiburg, wurden verschiedene Arbeiten zur Layoutanalyse ([Chen et al., 2015], [Seuret et al., 2015], [Garz et al., 2017], [Garz et al., 2016]) historischer Schriften durchgeführt. Auch für dieses Projekt wurden Plattformen für die manuelle Bearbeitung konzipiert. Diese lassen sich jedoch nicht direkt auf die im KALLIMACHOS Projekt spezifizierten Use Cases übertragen. Weitere Arbeiten im Bereich Layoutanalyse für arabische historische Dokumente wurden von der Arbeitsgruppe um Prof. Jihad El-Sana veröffentlicht ([Kassis und El-Sana, 2016]).

LSTM basierte OCR Methoden haben sich im Projektzeitraum auch in anderen bekannten OCR Systemen, wie ABBYY Fine Reader und Tesseract durchgesetzt und sind auf dem Weg, OCR Standard zu werden.

6. Erfolgte oder geplante Veröffentlichungen des Ergebnisses nach Nr. 6 BNBest-BMBF 98.

DFKI:

Im Rahmen der im Projekt geleisteten Arbeit konnten am DFKI mehrere Publikationen bei namhaften Konferenz erreicht werden. Dazu zählen:

- "12th IAPR International Workshop on Document Analysis Systems (DAS'16)" ([Ul-Hasan et], [Jenckel et al., 2016b])

- "International Conference on Image Processing (ICIP2016)" ([Nunamaker et al., 2016])
- "2nd International Conference on Natural Sciences and Technology in Manuscript Analysis" ([UI-Hasan et al., 2016])
- “4th International Workshop Historical Document Imaging and Processing (HIP'17)" ([Jenckel et al., 2017])
- "14th IAPR International Conference on Document Analysis and Recognition (ICDAR2017)" ([Bukhari et al., 2017])
- "7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)" ([Jenckel et al., 2018])

Veröffentlichungen im Rahmen von Kallimachos:

[Bukhari et al., 2017] Bukhari, S. S., Kadi, A., Jouneh, M. A., Mir, F. M., und Dengel, A. (2017). anyOCR: An Open-Source OCR System for Historical Archives. In *The 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017)*. IEEE.

[Bukhari et al., 2018] Bukhari, S. S., Saha, M., Badimala Giridhara, P. K., Lohano, M. K., und Dengel, A. (2018). anyAlign: An Intelligent and Interactive Text-Alignment Web-Application for Historical Document. In *13th IAPR International Workshop on Document Analysis Systems DAS 2018*.

Brigitte Burrichter, Joachim Hamm: Narragonien digital. Vortrag bei der Tagung Inkunabeln und Überlieferungsgeschichte des Wolfenbütteler Arbeitskreises für Bibliotheks-, Buch- und Mediengeschichte an der Universität Mainz, 29.6.-1.7.2015.

Brigitte Burrichter: Rahmen und intendiertes Publikum. Die Paratexte in Sebastian Brants 'Narrenschiff' und seinen Übersetzungen. Vortrag bei dem Theorie-Workshop Rahmungen. Präsentationsformen kanonischer Werke des Forschungsverbundes Marbach Weimar Wolfenbüttel, Projekt Text und Rahmen, vom 29.-31.7.2015 an der Herzog August Bibliothek Wolfenbüttel.

Brigitte Burrichter, Raphaëlle Jung: Les Nefs des fols en ligne. Présentation d’un projet d’édition en ligne des "Nefs des fols" du XVe siècle. Vortrag bei der Jahrestagung der Association Internationale pour l’ Étude du Moyen Français in Turin, 28.9.-1.10.2016.

Brigitte Burrichter, Joachim Hamm: Narragonien digital. Vortrag beim XLIV. Internationalen Mediävistischen Colloquium in Castellabate (IT), 10-17.9.2016.

Brigitte Burrichter: Rahmen und intendiertes Publikum. Die Paratexte in Sebastian Brants 'Narrenschiff' und seinen Übersetzungen. In: Ajouri, Philip / Kundert, Ursula / Rohde, Carsten (Hg.): Rahmungen. Präsentationsformen und Kanoneffekte. Berlin 2017 (Beiheft zur Zeitschrift für deutsche Philologie), S. 207-222.

Brigitte Burrichter, Joachim Hamm: Narragonien digital. Vortrag im Workshop "Digitale Paläographie" (Interdisziplinäres Zentrum Editionswissenschaften, IZED), Univ. Erlangen, 12.-13.01. 2017.

Brigitte Burrichter: Sebastian Brants Narrenschiff und seine europäische Rezeption. Vortrag bei der Tagung der Internationalen Oswald-von Wolkenstein- Gesellschaft, Brixen, vom 13.-15. September 2017.

Brigitte Burrichter: Patrice et les Dernydes. Les versions française de la Nef des Fous de Sebastian Brant, Vortrag bei der Tagung von fabula, Warschau 18.-20.10.2017.

Brigitte Burrichter: La Nef des fous de Sebastian Brant dans le context européen, Paris, Ecole normale supérieure, 05.02.2018.

[Büttner et al., 2017] Büttner, A., Dimpel, F. M., Evert, S., Jannidis, F., Pielström, S., Proisl, T., Reger, I., Schöch, C., und Vitt, T. (2017). ‚Delta‘ in der stilometrischen Autorschaftsattribution. *Zeitschrift für digitale Geisteswissenschaften*.

[Büttner und Proisl, 2016] Büttner, A. und Proisl, T. (2016). Delta und Merkmalsselektion: Welche Wörter unterscheiden arabisch-lateinische Übersetzer? Presentation at ‹philtag n="13"/›. Würzburg. 2016-02-26.

[Damiani, 2016] Damiani, V. (2016). Anagnosis: automatisierte Buchstabenverknüpfung von Transkript und Papyrusabbildung. In *Altertumswissenschaften in a Digital Age. Egyptology, Papyrology and beyond, proceedings of a conference and workshop in Leipzig, November 4-6, 2015*.

[Evert, 2016] Evert, S. (2016). On the significance of multivariate models of linguistic variation. Presentation at Hildesheim-Göttingen Workshop. Hildesheim. 2016-10-14.

[Evert, 2017] Evert, S. (2017). Authorship Attribution with Delta Measures. Presentation at the SFB 732 colloquium. Stuttgart. 2017-07-13.

[Evert et al., 2015a] Evert, S., Proisl, T., Jannidis, F., Pielström, S., Schöch, C., und Vitt, T. (2015a). Explaining Delta, or: How do distance measures for authorship attribution work? In *Corpus Linguistics*.

[Evert et al., 2015b] Evert, S., Proisl, T., Schöch, C., Jannidis, F., Pielström, S., und Vitt, T. (2015b). Explaining Delta, or: How do distance measures for authorship attribution work? Presentation at Corpus Linguistics 2015. Lancaster. 2015-07-24.

[Evert et al., 2015c] Evert, S., Proisl, T., Schöch, C., Jannidis, F., Pielström, S., und Vitt, T. (2015c). Towards a better understanding of Burrows´s Delta in literary authorship attribution. In *4th Workshop on Computational Linguistics for Literature*, Denver.

[Evert et al., 2015d] Evert, S., Proisl, T., Vitt, T., Schöch, C., Jannidis, F., und Pielström, S. (2015d). Towards a better understanding of Burrows's Delta in literary authorship attribution. In Feldman, A., Kazantseva, A., Szpakowicz, S., und Koolen, C., Herausgeber, *Proceedings of the Fourth Workshop on Computational Linguistics for Literature (CLfL 2015)*, Seiten 79–88, Denver, CO. Association for Computational Linguistics.

[Evert et al., 2015e] Evert, S., Schneider, G., Brezina, V., Gries, S., Lijffijt, J., Rayson, P., Wallis, S., und Hardie, A. (2015e). Corpus statistics: key issues and controversies. Panel discussion at Corpus Linguistics 2015. Lancaster. 2015-07-24.

[Evert et al., DHd 2016.] Evert, S., Jannidis, F., Dimpel, F. M., Schöch, C., Pielström, S., Vitt, T., Reger, I., Büttner, A., und Proisl, T. "Delta" in der stilometrischen Autorschaftsattribution. *DHd 2016. Konferenzabstracts*, Seiten 61–74.

[Evert et al., 2016a] Evert, S., Jannidis, F., Proisl, T., Reger, I., Vitt, T., Schöch, C., und Pielström, S. (2016a). Outliers or Key Profiles? Understanding Distance Measures for Authorship Attribution. In *DH conference*, Krakow.

[Evert et al., 2016b] Evert, S., Jannidis, F., Proisl, T., Vitt, T., Schöch, C., Pielström, S., und Reger, I. (2016b). Outliers or Key Profiles? Understanding Distance Measures for Authorship Attribution. Presentation at Digital Humanities 2016. Krakow. 2016-06-13.

[Evert et al., 2016c] Evert, S., Proisl, T., Schöch, C., Jannidis, F., Pielström, S., Reger, I., und Vitt, T. (2016c). Burrows' Delta verstehen. In *DHd-Tagung*, Leipzig.

[Evert et al., 2017] Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., und Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32(suppl_2):ii4–ii16.

[Evert und Proisl, 2016] Evert, S. und Proisl, T. (2016). Burrows's Delta verstehen. Presentation at ‹philtag n="13"/›. Würzburg. 2016-02-26.

Christine Grundig: Narren en mouvance. Adaptationen des Narrenschiffs im 15. Jahrhundert. Vortrag beim Workshop Wissen von Mensch und Natur. Tradierung, Aktualisierung und Vermittlung in frühneuzeitlichen Übersetzungen des DFG-Graduiertenkollegs 1876 Frühe Konzepte von Mensch und Natur an der Universität Mainz,19.2.-20.2.2016.

Christine Grundig: Theologische Überformung des 'Narrenschiffs' - Geiler von Kaysersberg und die sog. "Interpolierte Fassung". Vortrag beim 13. Altgermanistischen Kolloquium am Hesselberg, 4.-6.10.2016.

Christine Grundig: Theologische Überformung des ‚Narrenschiffs' - Geiler von Kaysersberg und die sogenannte ‚Interpolierte Fassung'. In: Archiv für das Studium der neueren Sprachen und Literaturen 254 (2017), S.1-16.

[Grundig et al., im Druck] Grundig, C., Hamm, J., und Walter, V. (2018 im Druck). Narragonien digital. Mit einer Analyse von Kapitel 4 des ‚Narrenschiffs' in Ausgaben und Bearbeitungen des 15. Jahrhunderts. *Wolfenbütteler Notizen zur Buchgeschichte*.

Joachim Hamm: Intermediale Varianz. Sebastian Brants 'Narrenschiff' in deutschen Ausgaben des 15. Jahrhunderts. In: Überlieferungsgeschichte transdisziplinär. Neue Perspektiven auf ein germanistisches Forschungsparadigma. In Verbindung mit Horst Brunner und Freimut Löser hg. v. Dorothea Klein. Wiesbaden 2016 (Wissensliteratur im Mittelalter 52), S. 223-240.

Joachim Hamm: Die digitale Edition von ‘Narrenschiffen’ des 15. Jahrhunderts (“Narragonien digital”). Gastvortrag an der Univ. Stuttgart, Digital Humanities (Prof. Dr. Gabriel Viehhauser), 15.12.2016.

Joachim Hamm: Zu Paratextualität und Intermedialität in Sebastian Brants Vergilius pictus (Straßburg 1502). In: Diesseits des Laokoon. Intermedialität in der Frühen Neuzeit. Formen, Funktionen, Konzepte. Tagung an der Univ. Eichstätt, 28.-31.3.2012. Hg. v. Jörg Robert. Berlin, Boston 2017, S. 236-259.

Joachim Hamm: Gelehrte Narreteien. Das 'Narrenschiff' von Sebastian Brant und das Würzburger Projekt "Narragonien digital". Vortrag im Alten Rathaus von Miltenberg in der Vortragsreihe des Unibundes, 16.1.2017.

Joachim Hamm: Textuelle und intermediale Varianz im digitalen Kontext am Beispiel des Editionsprojekts “Narragonien digital”. Vortrag im Workshop "Textvarianten in der digitalen Edition" des Instituts für Musikforschung (Univ. Würzburg). 19.-20.1.2017.

Joachim Hamm: Eine integrierte digitale Edition der 'Narrenschiffe' vor 1500. Vortrag in der Vortragsreihe des Akademieprojekts "Der Österreichische Bibelübersetzer", Univ. Augsburg, 30.11.2017.

[Hasse, 2016] Hasse, D. N. (2016). Stylistic Evidence for Identifying John of Seville with the Translator of Some Twelfth-Century Astrological and Astronomical Texts from Arabic into Latin on the Iberian Peninsula. In Burnett, C. und Mantas-Exã, P., Herausgeber, *Ex Oriente Lux. Translating Words, Scripts and Styles in Medieval Mediterranean Society*, Seiten 19–43. Cordoba and London.

[Hasse, forthcoming] Hasse, D. N. (forthcoming). The Latin translation of the Liber de causis in comparison with double translations by Gerard of Cremona and Dominicus Gundisalvi. In Calma, D., Herausgeber, *Proclus and the Book of Causes: Insights into their Influence*, Band 2 Acculturation. Leiden.

[Hasse und Büttner, im Erscheinen] Hasse, D. N. und Büttner, A. (im Erscheinen). Notes on Anonymous Twelfth-Century Translations of Philosophical Texts from Arabic into Latin on the Iberian Peninsula. In Hasse, D. N. und Bertolacci, A., Herausgeber, *The Arabic, Hebrew and Latin Reception of Avicenna´s Physics and Cosmology*. Berlin and Boston.

[Hettinger et al., 2015] Hettinger, L., Becker, M., Reger, I., Jannidis, F., und Hotho, A. (2015). Genre classification on German novels. In *Proceedings of the 12th International Workshop on Text-based Information Retrieval*.

[Hettinger et al., 2016a] Hettinger, L., Jannidis, F., Reger, I., und Hotho, A. (2016a). Classification of Literary Subgenres. In *DHd 2016*.

[Hettinger et al., 2016b] Hettinger, L., Jannidis, F., Reger, I., und Hotho, A. (2016b). Significance Testing for the Classification of Literary Subgenres. In *DH 2016*.

[Jannidis et al., 2016] Jannidis, F., Reger, I., Zehe, A., Becker, M., Hettinger, L., und Hotho, A. (2016). Analyzing Features for the Detection of Happy Endings in German Novels. cite arxiv:1611.09028.

[Jenckel et al., 2016a] Jenckel, M., Bukhari, S. S., und Dengel, A. (2016a). anyOCR: A sequence learning based OCR system for unlabeled historical documents. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, Seiten 4035–4040. IEEE.

[Jenckel et al., 2016b] Jenckel, M., Bukhari, S. S., und Dengel, A. (2016b). Clustering benchmark for characters in historical documents. In *12th IAPR International Workshop on Document Analysis Systems DAS 2016*.

[Jenckel et al., 2017] Jenckel, M., Bukhari, S. S., und Dengel, A. (2017). Training LSTM-RNN with Imperfect Transcription - Limitations and Outcomes. In *The 4th International Workshop on Historical Document Imaging and Processing (HIP 2017)*.

[Jenckel et al., 2018] Jenckel, M., Parkala, S., Bukhari, S. S., und Dengel, A. (2018). Impact of Training LSTM-RNN with Fuzzy Ground Truth. In *The 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*.

[Kirchner et al., 2016] Kirchner, F., Dittrich, M., Beckenbauer, P., und Nöth, M. (2016). OCR bei Inkunabeln–Offizinspezifischer Ansatz der Universitätsbibliothek Würzburg. *ABI Technik*, 36(3):178–188.

[Krug et al., 2016] Krug, M., Jannidis, F., Reger, I., Weimer, L., Macharowsky, L., und Puppe, F. (2016). Attribuierung direkter Reden in deutschen Romanen des 18.-20. Jahrhunderts: Methoden zur Bestimmung des Sprechers und des Angesprochenen.

[Krug et al., 2017a] Krug, M., Puppe, F., Jannidis, F., Reger, I., Weimer, L., und Macharowsky, L. (2017a). Comparison of Methods for the Identification of Main Characters in German Novels. DH.

[Krug et al., 2017b] Krug, M., Reger, I., Jannidis, F., Weimer, L., Madarász, N., und Puppe, F. (2017b). Overcoming Data Sparsity for Relation Detection in German Novels.

[Krug et al., 2017c] Krug, M., Wick, C., Reger, I., Jannidis, F., Weimer, L., Madarász, N., und Puppe, F. (2017c). Comparison of Methods for Automatic Relation Extraction in German Novels.

[Nunamaker et al., 2016] Nunamaker, B., Bukhari, S. S., Borth, D., und Dengel, A. (2016). A Tesseract-based OCR framework for historical documents lacking ground-truth text. In *Image Processing (ICIP), 2016 IEEE International Conference on*, S. 3269–3273. IEEE.

[Plotnikova et al., 2015] Plotnikova, N., Lapesa, G., Proisl, T., und Evert, S. (2015). SemantiKLUE: Semantic Textual Similarity with Maximum Weight Matching. In Cer, D. M., Jurgens, D., Nakov, P., und Zesch, T., Herausgeber, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Seiten 111–116, Denver, CO. Association for Computational Linguistics.

[Proisl, 2015] Proisl, T. (2015). Maschinelles Lernen mit Python. Presentation at DARIAH-Methodenworkshop Natural Language Processing für Literaturwissenschaftler. Würzburg. 2015-09-16.

[Proisl, 2018] Proisl, T. (2018). SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki. European Language Resources Association.

[Proisl et al., 2018] Proisl, T., Evert, S., Jannidis, F., Schöch, C., Konle, L., und Pielström, S. (2018). Delta vs. N-Gram Tracing: Evaluating the Robustness of Authorship Attribution Methods. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki. European Language Resources Association.

[Proisl und Uhrig, 2016] Proisl, T. und Uhrig, P. (2016). SoMaJo: State-of-the-art tokenization for German web and social media texts. In Cook, P., Evert, S., Schäfer, R., und Stemle, E., Herausgeber, *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, Seiten 57–62, Berlin. Association for Computational Linguistics.

[Reul et al., 2017a] Reul, C., Springmann, U., und Puppe, F. (2017a). LAREX: A Semi-automatic Open-source Tool for Layout Analysis and Region Extraction on Early Printed Books. In *Proceedings of the 2Nd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2017, Seiten 137–142, New York, NY, USA. ACM.

[Reul et al., 2017b] Reul, C., Springmann, U., Wick, C., und Puppe, F. (2017b). Improving OCR Accuracy on Early Printed Books by utilizing Cross Fold Training and Voting. *ArXiv e-prints*.

[Reul et al., 2017c] Reul, C., Wick, C., Springmann, U., und Puppe, F. (2017c). Transfer Learning for OCRopus Model Training on Early Printed Books. *027.7 Zeitschrift für Bibliothekskultur / Journal for Library Culture*, 5(1):38–51.

[Ul-Hasan et al., 2016] Ul-Hasan, A., Bukhari, S. S., und Dengel, A. (2016). Meaningless text ocr model for medieval scripts. In *2nd International Conference on Natural Sciences and Technology in Manuscript Analysis*.

[Ul-Hasan et al., 2016] Ul-Hasan, A., Bukhari, S. S., und Dengel, A. (2016). OCRoRACT: A Sequence Learning OCR System Trained on Isolated Characters. In *DAS*, Seiten 174–179.

[Zehe et al., 2016] Zehe, A., Becker, M., Hettinger, L., Hotho, A., Reger, I., und Jannidis, F. (2016). Prediction of Happy Endings in German Novels. In Cellier, P., Charnois, T., Hotho, A., Matwin, S., Moens, M.-F., und Toussaint, Y., Herausgeber, *Proceedings of the Workshop on Interactions between Data Mining and Natural Language Processing 2016*, Seiten 9–16.

[Zehe et al., 2017] Zehe, A., Becker, M., Jannidis, F., und Hotho, A. (2017). Towards Sentiment Analysis on German Literature.

Literaturverzeichnis

[Beißwenger et al., 2016] Beißwenger, M., Bartsch, S., Evert, S., und Würzner, K.-M. (2016). EmpiriST 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora. In Cook, P., Evert, S., Schäfer, R., und Stemle, E., Herausgeber, *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, Seiten 44–56, Berlin. Association for Computational Linguistics.

[Blei et al., 2003] Blei, D., Ng, A., und Jordan, M. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.[Chen et al., 2015] Chen, K., Seuret, M., Liwicki, M., Hennebert, J., und Ingold, R. (2015). Page segmentation of historical document images with convolutional autoencoders. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, Seiten 1011–1015. IEEE.

Rena Buß: „Lübecker Unbekanntheiten. Ein Verfasserprofil anhand paratextueller Konzepte in ‚Dat narren schyp' und anderen Mohnkopf-Drucken" (Zulassungsarbeit zum 1. Staatsexamen für das Lehramt am Gymnasium, Fach Deutsch)

[Eder et al., 2016] Eder, M., Rybicki, J., und Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *R journal*, 8(1):107–121.

[Garz et al., 2017] Garz, A., Seuret, M., Fischer, A., und Ingold, R. (2017). A User-Centered Segmentation Method for Complex Historical Manuscripts Based on Document Graphs. *IEEE Transactions on Human-Machine Systems*, 47(2):181–193.

[Garz et al., 2016] Garz, A., Würsch, M., Fischer, A., und Ingold, R. (2016). Simple and fast geometrical descriptors for writer identification. *Electronic Imaging*, 2016(17):1–12.

Christine Grundig: Text und Paratext. Konzepte von Paratextualität in deutschsprachigen Werken Sebastian Brants. Masch. Magisterarbeit. Würzburg 2012.

Christine Grundig: Sebastian Brants 'Narrenschiff': Zur Bild-Text-Relation in deutschsprachigen und europäischen Ausgaben des Werkes. Vortrag beim 10. Altgermanistischen Kolloquium am Hesselberg vom 1.-3.10.2013

[Kassis und El-Sana, 2016] Kassis, M. und El-Sana, J. (2016). Scribble Based Interactive Page Layout Segmentation Using Gabor Filter. In *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*, Seiten 13–18. IEEE.

Amelie Eileen Laut: „Das ‚Narrenschiff'. *Gedruckt zuo Augspurg*; *Gedruockt zu Nueremberg*. Die Offizin Schönsperger und deren Nachdruck der *editio princeps*" (Zulassungsarbeit zum 1. Staatsexamen für das Lehramt an Realschule, Fach Deutsch)

[Seuret et al., 2015] Seuret, M., Chen, K., Eichenbergery, N., Liwicki, M., und Ingold, R. (2015). Gradient-domain degradations for improving historical documents images layout analysis. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, Seiten 1006–1010. IEEE.

[Springmann et al., 2015] Springmann, U., Fink, F., und Schulz, K. U. (2015). Workshop: OCR & postcorrection of early printings for digital humanities.

[Springmann et al., 2016] Springmann, U., Fink, F., und Schulz, K. U. (2016). Automatic quality evaluation and (semi-) automatic improvement of OCR models for historical printings. *ArXiv e-prints*.

[Springmann und Lüdeling, 2017] Springmann, U. und Lüdeling, A. (2017). OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. *Digital Humanities Quarterly*, 11(2).

Maximilian Wehner: „Topographie der „Verkehrten Welt". Zur Ausgestaltung literarischer Räume in Sebastian Brants ‚Narrenschiff'" (Zulassungsarbeit zum 1. Staatsexamen für das Lehramt am Gymnasium, Fach Deutsch)

III. Anlage - Ergebniskontrollbericht

1. Beitrag des Ergebnisses zu den förderpolitischen Zielen

Das förderpolitische Ziel ergibt sich aus der BMBF eHumanities Förderlinie 2, Zentren:

"Zweck der zu fördernden Zentren ist es, Forschungsinfrastrukturen für die Geistes- und qualitativen Sozialwissenschaften unter maßgeblicher Einbeziehung der Informatik oder informatiknaher Fächer aufzubauen. Die Zielgruppe dieser Förderung sind Hochschulen und außeruniversitäre Forschungseinrichtungen, die Zentren für einen Standort oder für eine Region für die Gesamtheit der Geistes- und qualitativen Sozialwissenschaften oder überregional für eine Disziplin etablieren wollen. Es ist möglich, sowohl neue Zentren aufzubauen, als auch bestehende zu verstärken oder weiterzuentwickeln. Damit soll der dezentrale Aufbau von Kompetenzen und Kapazitäten unterstützt werden."

Durch die Einrichtung eines "Zentrums für digitale Edition und quantitative Analyse an der Universität Würzburg" konnte in einem vorher nicht möglichen Ausmaß die Zusammenarbeit mehrere geisteswissenschaftlicher Lehrstühle, der Informatik und der Universitätsbibliothek erprobt und erfolgreich umgesetzt werden. Die langfristige Etablierung dieses Ansatzes erfolgt in Zukunft außerhalb der Bibliothek im Rahmen des ebenfalls vom BMBF geförderten "Zentrums für Philologie und Digitalität".

2. Wissenschaftlich-technische Ergebnis des Vorhabens, die erreichten Nebenergebnisse und die gesammelten wesentlichen Erfahrungen

Das Ergebnis des Projekts besteht einerseits in Softwarewerkzeugen, die für Projekte ähnlichen Zuschnitts frei wiederverwendet werden können. Dazu gehören die OCR-Komponenten für die Erkennung historischer Drucke, der Transkriptionseditor, das Wiki-Modul einschließlich der Routine zum Export der im Wiki annotierten Werke nach TEI sowie die Werkzeuge für die Untersuchungen an Textkorpora. Als Nebenergebnis wurden annotierte Korpora erstellt (Deutsches Romankorpus DROC[19], Narrenschiff-Korpus, Trainingsdaten-Korpus für historische OCR), die unter einer Open Source Lizenz für die weitere Verwendung zur Verfügung gestellt werden.

Als Haupterfahrung bleibt festzuhalten, dass sich die Zusammenarbeit zwischen Geisteswissenschaftlern und Informatikern als sehr fruchtbar erwiesen hat, wenn es darum geht, geisteswissenschaftliche Daten in quantifizierbarer Art und Weise zu erheben, zu annotieren und mit Hilfe von Softwarewerkzeugen zu bearbeiten.

3. Fortschreibung des Verwertungsplans

a. Erfindungen/Schutzrechtsanmeldungen und erteilte Schutzrechte, die vom Zuwendungsempfänger oder von am Vorhaben Beteiligten gemacht oder in Anspruch genommen wurden, sowie deren standortbezogene Verwertung (Lizenzen u.a.) und erkennbare weitere Verwertungsmöglichkeiten

19 http://kallimachos.de/kallimachos/index.php/DROC

Es sind keine Schutzrechtsanmeldungen oder wirtschaftlichen Verwertungen geplant, da alle Ergebnisse unter Open Source Lizenzen gestellt wurde. Die Verwertung bezieht sich daher auf die freie Nachnutzung durch Forschungseinrichtungen und Bibliotheken, die uneingeschränkt möglich ist.

b. Wirtschaftliche Erfolgsaussichten nach Projektende (mit Zeithorizont) - z.B. auch funktionale/wirtschaftliche Vorteile gegenüber Konkurrenzlösungen, Nutzen für verschiedene Anwendergruppen/-industrien am Standort Deutschland, Umsetzungs- und Transferstrategien (Angaben, soweit die Art des Vorhabens dies zulässt)

Es ist keine wirtschaftliche Ergebnisverwertung geplant. Unsere Open-Source-Strategie, die unsere Daten offen für die Wiederverwertung zur Verfügung stellt, schließt aber eine wirtschaftliche Nutzung durch Dritte nicht aus. Während die Datengrundlage frei verfügbar ist, steht es Firmen frei, auf Grundlage dieser Daten auch kostenpflichtige und marktfähige Produkte oder Dienste zu entwickeln, etwa um die Daten sehr komfortabel zu durchsuchen und auszuwerten.

c. Wissenschaftliche und/oder technische Erfolgsaussichten nach Projektende (mit Zeithorizont) - u.a. wie die geplanten Ergebnisse in anderer Weise (z.B. für öffentliche Aufgaben, Datenbanken, Netzwerke, Transferstellen etc.) genutzt werden können. Dabei ist auch eine etwaige Zusammenarbeit mit anderen Einrichtungen, Firmen, Netzwerken, Forschungsstellen u.a. einzubeziehen.

Die Nachnutzung unserer Methoden und Tools ist zunächst einmal ein großer Vorteil für alle Institutionen und Forschungsinstitute, die sich in ähnlicher Weise darum bemühen, die Datengrundlage für weitergehende Forschungen der Digital Humanities zu legen. Diese Datengrundlage besteht sowohl in den Scans von Manuskript- und Druckseiten als auch insbesondere und zunehmend in der OCR der gesamten modernen Druckgeschichte. Daraus ergibt sich eine Ersparnis gegenüber der bisherigen Praxis, interessante Werke entweder transkribieren (abschreiben) zu müssen oder selbst ähnliche Softwarekomponenten für jedes Digital-Humanities-Projekt noch einmal entwickeln zu müssen. Die Zurverfügungstellung der Datengrundlage ermöglicht schließlich auch allen staatlichen Institutionen sowie jedem interessierten Bürger, sich kostenlos Einblick in wesentliche Teile unseres kulturgeschichtlichen Erbes zu verschaffen und diese Daten quantitativ auszuwerten. Hier ist mehr als nur Open Access gemeint, nämlich auch Open Data (Möglichkeit, die Daten herunterzuladen und lokal weiterzuverarbeiten anstatt sie nur über einen Webbrowser anschauen zu können) sowie Open Science, bei der die Primärdaten unserer Veröffentlichungen und typische Anwendungsdaten unserer Tools zur Verfügung gestellt werden, sodass unsere Ergebnisse und die Anwendungsfälle der Tools nachvollzogen werden können.

Neben den Daten kann man auch auf die von uns entwickelten prototypischen Arbeitsabläufe (im Sinne von "best practice"-Methoden) zurückgreifen und sie mit den von uns bereitgestellten Open-Source-Tools umsetzen, was ebenfalls ein kostensparendes Synergiepotential ergibt.

d. Wissenschaftliche und wirtschaftliche Anschlussfähigkeit für eine mögliche notwendige nächste Phase bzw. die nächsten innovatorischen Schritte zur erfolgreichen Umsetzung der Ergebnisse

In der zweiten Phase von Kallimachos (01.10.2017-30.09.2019) werden die Fortschritte auf dem Gebiet historischer OCR weiter ausgebaut und die erstellten Werkzeuge in einem Werkzeugkasten gebündelt, dokumentiert und verfügbar gemacht.

DFKI:

Neben den im Abschnitt 2.4 dargestellten Verwertungsmöglichkeiten, prüft das DFKI eine mögliche Weiterentwicklung des im Projekt erarbeiteten OCR-Verfahrens, sowie der erstellten Services und Tools, zu einem kommerziellen Produkt.

4. Arbeiten, die zu keiner Lösung geführt haben

DFKI:

Zu Beginn des Projektes und wie in der Planung angedeutet, wurden zunächst Ansätze mit synthetischen Trainingsdaten verfolgt. Dafür wurden Texte mit historischen Fonts gerendert und so Trainingsdaten generiert. Für viele Dokumente lagen allerdings keine passenden Fonts vor und das manuelle Erstellen solcher Fonts ist mit einem hohen Arbeitsaufwand verbunden. Die Ergebnisse unter Verwendung ähnlicher Fonts waren ebenfalls nicht zufriedenstellend. Für die weitere Bearbeitung des Projekts wurde daher der im Abschlussbericht erwähnte anyOCR Ansatz entwickelt.

UB Würzburg:

Auch die ursprünglich vorgesehene Entwicklungen von historischen Sprachmodellen für die OCR-Nachkorrektur und die Entwicklung eines Systems zu Erkennung von Handschriften hat sich als allzu ambitioniert herausgestellt, als dass man diese Bereiche quasi nebenbei noch hätte erledigen können.

5. Präsentationsmöglichkeiten für mögliche Nutzer - z.B. Anwenderkonferenzen (Angaben, soweit die Art des Vorhabens dies zulässt)

Nutzer können sich sowohl der Tutorials auf der Projektwebseite kallimachos.de bedienen, als auch z.B. auf den jährlichen Philtag-Konferenzen an der Universität Würzburg über die erstellten Werkzeuge und Arbeitsketten informieren und mit ihnen vertraut machen. Darüber hinaus werden auf den jährlichen DHd-Konferenzen sowie in COST-Aktionen Workshops angeboten.

6. Einhaltung der Ausgaben- und Zeitplanung. Im Erfolgskontrollbericht kann auf Abschnitte des Schlussberichts (Nrn. I. und II.) verwiesen werden.

Für die Ausgabenplanung verweisen wir auf die Darstellung der Universitätsbibliothek (zahlenmäßiger Nachweise von Dr. Schmidt mit Datum vom 20.02.2018, sh. Abschnitt II.2). Demnach wurde der vorgegebene Finanzrahmen exakt eingehalten. Die Zeitplanung gemäß Balkenplan wurde weitgehend eingehalten, wobei Abweichungen in den Zwischenberichten bzw. in den entsprechenden Abschnitten von Teil II dargestellt wurden.

Würzburg, 23.03.2018

Dr. Uwe Springmann
Verbundkoordinator