table of contents

Use of the grant and enumeration of the most important scientific-technical and other results

A focus of the joint project was the OCR optimization and the support of the corresponding use cases Narragonia and Anagnosis. Various techniques were implemented and tested during the course of the project. A breakthrough was achieved with the provision of the semi-automated open source tool OCR4all, which for the first time allows the digitization of early prints with reasonable effort. OCR4all was not only used intensively in Callimachus, but has also become more widespread internationally (see WP1). The other sub-projects were also successfully completed. The following presentation of the results achieved adopts the structure of the application document, so that a comparison of the goals and the results for each work package is easy to understand.

AP1: OCR optimization

TA 1.1.1: Automatic segmentation

The OCR workflow can be divided into four main steps: pre-processing, segmentation, text recognition, post-correction. We describe segmentation in the context of the OCR workflow tool OCR4all (see TA 1.1.2).

TA 1.1.2: Expansion of the pharmacy approach and further development in OCR4all

The retail approach, which envisaged the tedious and time-consuming identification of individual shop-specific print types for OCR recognition, was dropped because OCR recognition with neural networks in LSTM architecture does not require segmentation of individual characters. Instead, prints are transcribed using mixed models trained on a variety of print types. A few pages are then corrected and a work-specific model is trained on this basis, with which the entire print is then transcribed and finally corrected. This procedure represents a very time- and cost-saving variant compared to a purely manual transcription and is currently the most efficient procedure for old prints. Accordingly, the comfortable workflow tool OCR4all [Reul et al.2019c] has already found a very good national and international response and distribution. The sub-steps of the workflow and the response are presented in more detail below.

Semi-automatic transcription tool OCR4all for old prints

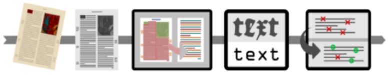
In order to make the presented approach available to as wide a user group as possible, the tool OCR4all (https://github.com/OCR4all)developed and made freely available on GitHub. The motivation behind OCR4all is that there are now some open source tools that provide excellent results (even on very old and sophisticated material), but the use of which can quickly become overwhelming for inexperienced, non-technical users. This is mainly due to the fact that many applications can only be operated via the command line and are sometimes difficult to install. The combination of different individual tools to form a coherent pipeline is often not trivial due to varying data formats. OCR4all tries to close this gap by encapsulating a complete OCR workflow in a single Docker application or alternatively Virtual Box that can be installed very easily.

The tool relieves the user of the management of the data and can be conveniently controlled via a clear graphical web interface. The aim is to also give non-technical users the opportunity to record even the oldest printed works independently, in a manageable amount of time and in the highest quality. In addition to the well-known OCRopus (https://github.com/tmbarchive/ocropy) and LAREX, which was created in the project, the open source OCR tool Calamari , developed by Christoph Wick at the Chair of Artificial Intelligence (https://github.com/Calamari-OCR)already fully integrated into OCR4all and the workflow contained there. Unlike OCRopus, Calamari uses a deep network structure with several hidden layers (deep learning) for character recognition, which results in significantly higher recognition rates ([Wick et al.2018]). In addition to this technical development, other methodological improvements such as voting [Reul et al.2018a], pretraining [Reul et al.2018b] and data augmentation were integrated, the use of which again significantly reduces the error rate [Reul et al.2018c, Wick et al. 2020].

workflow

A typical OCR workflow can basically be divided into four main steps (see Figure 1), whose goals, challenges and current implementation in OCR4all are presented below.

Preprocessing: In this first step, the input images are prepared for further processing. This includes both binarization (conversion to a black and white image) and straightening of the scanned page. It is also common to separate double pages that have been scanned together beforehand or to set up scan pages that have been scanned sideways. Binarization and straightening is done reliably by a script from the OCRopus Toolbox. In principle, OCR4all can also handle double pages or pages scanned sideways, but separation and uprighting is recommended, e.g. B. through the freely available and well-documented tool ScanTailor (https://scantailor.org/) (due to the lack of web compatibility, it cannot be meaningfully integrated into OCR4all).



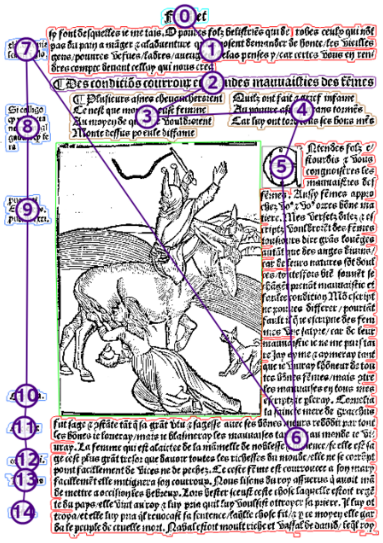
(/kallimachos/index.php/Datei:Abbildung_1_H

Figure 1: Main steps of a typical OCR workflow. From left to right: input image, pre-processing, segmentation, text recognition, post-correction

Segmentation: The task of this step is to subdivide the scanned page into smaller units. Depending on the material and the individual requirements of the user, very different characteristics are possible. So it can e.g. For example, it is sufficient to identify only the regions that contain text and to separate them from the rest (images, noise, etc.). At the other end of the spectrum there is a fine-grained semantic markup (see Figure 2), which not only differentiates between text and image regions, but also assigns further subtypes to the former (continuous text, heading, marginalia, etc.; again heavily dependent on the user and the material). Identified text regions must then be split into individual lines of text, as these represent the required input for modern OCR engines.

For this extremely demanding step, OCR4all currently provides two sub-modules in order to optimally adapt to the requirements of the respective user as well as the properties and challenges of the material at hand. On the one hand, the specially developed LAREX tool (see below) is used, whose semi-automatic approach is particularly suitable for users who are interested in a 100% correct result, including an individual and detailed semantic markup. On the other hand, a so-called dummy segmentation is offered for the fully automatic application, which does neither a semantic markup nor an explicit markup of images or other non-text regions, but concentrates directly on the detection of text lines.

Text recognition: The text shown in the text lines can now be extracted. To do this, OCR engines use so-called models. In general, a distinction is made between mixed and factory or type-specific models. The former are normally trained on a variety of similar types and can then be applied to unseen material out-of-the-box (without further work-specific training and thus without further creation of training data). This approach is all the more promising the more uniform the typography of the available material is. While with modern writing, but e.g. e.g. in Gothic script from the 19th century, one can hope for very low error rates (see evaluation in Table 1), but these can by no means be expected as the material ages and especially in the case of incunabula. This can be remedied by using work-specific models, which require work-specific ground truth to be created, which must be generated by manually correcting the texts transcribed using mixed models. Of course, this means additional effort, which, however, is necessary for sufficient quality in many applications due to the better recognition accuracy.



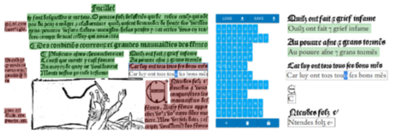
(/kallimachos/index.php/Datei:Abbildung_2_S

Figure 2: Segmentation of a complex page of the *Ship* of Fools including precise semantic markup and exact recording of the reading order

The OCR engine currently used in OCR4all is the specially developed Calamari, which is used both for recognition and for training your own models. With regard to operation by non-technical users, the training step in its implementation in particular posed a major challenge, since all of the methodological extensions mentioned above should be supported, but without overtaxing the users.

Correction: Since, despite great progress in recent years, an error-free OCR result on historical prints is not realistic, a final step is required in which the remaining errors are to be corrected or at least further reduced. This can be done automatically, e.g. B. by using language models, manually by a manual post-correction or by a combination of both methods. While an automatic post-correction is not yet available in OCR4all (but can be connected externally, e.g. PoCoTo [Vobl et al.2014] or PoCoWeb (https://github.com/cisocrgroup/pocoweb)), the integrated LAREX component now offers the option of conveniently correcting both the OCR text and the results of previous steps such as region and line polygons, reading order, semantic types, etc. due to extensive expansion (see Figure 3).

Due to the modular structure of the tool as well as the well-defined interfaces and the selected distribution channel via a container solution, the integration of further solutions is possible at any time.



(/kallimachos/index.php/Datei:Abbildung_3_T

Figure 3: Textual correction in LAREX: page-based view (left), configurable virtual keyboard (middle), line-based view (right).

Evaluation: In addition to the practical use of OCR4all at numerous institutions and in various projects (see below), the main publication [Reul et al. 2019c] carried out comprehensive evaluations in close cooperation with the designated humanities users.

The first evaluation relates to Gothic novels from the 19th century (with one exception from the late 18th century). In contrast to early modern incunabula and prints, e.g. B. the *Ship of Fools*, these have, in addition to the better state of preservation, a moderate layout and significantly more uniform typefaces, which enabled fully automatic indexing using OCR4all. The unified typography allowed the application of a mixed calamari model for 19th-century Gothic typefaces, which had previously been trained using the above accuracy-enhancing measures [Reul et al. 2019a]. The fully automatic OCR4all run was evaluated on ten pages from ten different works, with sometimes strongly fluctuating quality, as can be seen in Figure 4.

(/kallimachos/index.php/Datei:Abbildung_4_B

Figure 4: Example images of German Gothic novels. From left to right: F1870, F1781, F1818 (page in acceptable condition), F1818 (page in poor condition), F1803.

For comparison, the same evaluation was carried out with the commercial state-of-the-art tool ABBYY Finereader, which, in addition to a "Gothic" recognition for Fraktur writing, also offers a corresponding post-correction for "Old German". Table 1 summarizes the results.

Table 1: Comparison of the letter error rates in the fully automatic use of ABBYY Finereader and OCR4all, as well as the error reduction achieved by OCR4all (ErrRed.) and the improvement factor (Impr.).

plant	ABBYY	OCR4all	ErrRed.	imprint
F1781	2.9	0.60	79.3	4.8
F1803	27	4.89	81.9	5.5
F1810	3.8	0.61	84.0	6.2
F1818	10	1.35	86.6	7.5
F1826	1.1	0.06	94.4	18
F1848	0.93	0.20	78.5	4.7
F1851	1.0	0.16	84.0	6.3
F1855	4.0	0.33	91.8	12
F1865	1.6	0.18	88.8	8.9
F1870	0.48	0.13	72.9	3.7
Average	**5.3**	**0.85**	**84.2**	**7.8**

The values show that OCR4all delivers significantly better error rates than ABBYY Finereader for each individual work, resulting in an average error reduction of 84% and an improvement factor of almost 8. For both systems, the results vary greatly from book to book, which can be explained by the widely differing quality of the source material (see Figure 4). On average, OCR4all achieves a very low letter error rate (CER) of just 0.85% (ABBYY 5.3%), with this being less than 1% for eight of the ten works and even less than 0.5% for six. If only the top 50% of the works are considered, the CER even drops to an excellent 0.15%.

These experiments on Gothic novels from the 19th century show that a fully automatic application of OCR4all is not only possible but can also be extremely precise as long as the layout is moderate and a suitable OCR model is available. It should be noted that the extremely low error rates shown can only be achieved fully automatically if a high-performance mixed model is available. In this case, a model ensemble was available that was perfectly suited to the detection of the evaluation material at hand. Unfortunately, this is still an exception at the moment, since similarly specialized models are probably only available for recognizing modern English texts.

Since OCR4all was originally developed *digitally for the comparatively high requirements of the Narragonien* project (exact semantic marking already at layout level, error-free result text), works that are sometimes significantly more demanding in terms of layout and typography than the previously evaluated Gothic novels can also be edited without any problems will. This was evaluated in a second, very extensive, user-centered study:

25 works were edited, printed between 1474 and 1598, including numerous Ships of *Fools* and products by the universal scholar Joachim Camerarius the Elder (in cooperation with the DFG project 'Opera Camerarii', cf. [Hamm et al. 2019].) The editing took place for the most part by students, who were divided into two groups: Group 1 consisted of inexperienced users who had no prior experience worth mentioning with OCR4all or OCR in general. In contrast, the processors from Group 2 already had extensive experience. After an introduction by one of the experienced users, the inexperienced users had to independently edit the works assigned to them.

For all works, regions were semantically precisely recorded and marked at the layout level in order to later enable a complete reconstruction of the scanned page (see Figure 2 and Figure 5). Although all works should finally be available as citable full text, for the time being only a rough target character error rate of 1% was given for this evaluation. In addition to the time required for the segmentation, the error rate and the correction effort were recorded. Table 2 shows the results.

Table 2: Summary of the results achieved (mean values plus standard deviation, if applicable) when using OCR4all on early prints by users of different levels of experience.

	Inexperienced users	Experienced users
Achieved CE	0.47%±0.22%	0.49%±0.30%
Transcribed training material	988 lines	927 lines
Correction time per line	10s ± 5.2s	5.5s ± 2.4s
Segmentation time per page	1.2min ± 0.5min	0.6min ± 0.2min

As expected, user experience had no impact on the accuracy of the OCR, nor on the number of training lines needed to reach the predetermined CER of max 1%. Both user groups were able to achieve an excellent and almost identical average CER of less than 0.5%, especially with regard to the age of the material, and needed an average of almost 1,000 training lines. This underlines the effectiveness of the presented approach and the power of the integrated Calamari OCR software used for training and recognition. It is also not surprising that the experienced users work much more efficiently in terms of the time required, both in the segmentation and in the creation of the training data. A rough rule of thumb regarding the time required for precise recording (exact semantic labeling and a CER of less than 0.5% on average for 25 works with an average of more than 250 pages per work) can be derived from the collected values: Inexperienced users must Expect 150 minutes for GT creation and 1.1 minutes per page for segmentation. Experienced users can expect a much faster capture: 57 minutes effort for GT creation and 0.6 minutes for segmentation of each page. Inexperienced users should expect 150 minutes for GT creation and 1.1 minutes per page for segmentation. Experienced users can expect a much faster capture: 57 minutes effort for GT creation and 0.6 minutes for segmentation of each page. Inexperienced users should expect 150 minutes for GT creation and 1.1 minutes per page for segmentation. Experienced users can expect a much faster capture: 57 minutes effort for GT creation and 0.6 minutes for segmentation of each page.

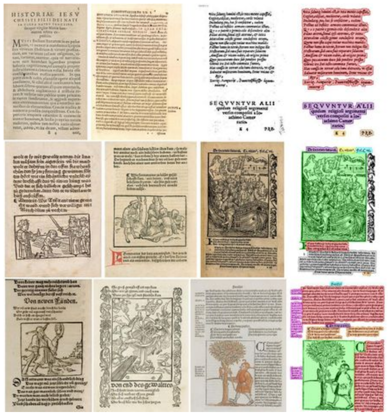
Distribution and use by the community:

OCR4all has been extremely well received both within the University of Würzburg and on a national and international level. The ZPD is making great efforts to further promote distribution and usability, as will be explained below.

Own publications of the project group: In the rather technical main publication [Reul et al. 2019c], the focus was on the description of the tool, the components used and the designed workflow. Furthermore, OCR4all was extensively evaluated (above evaluation shows an excerpt).

Another publication [Wehner et al. 2020] along with the associated workshop was published at this year's annual conference of Digital Humanities in German-speaking countries (DHd). The all-day workshop advertised for 25 participants met with great interest and was fully booked within a very short time.

At the request of the KulturBetrieb magazine, which is delivered twice a year to more than 2,200 culture-preserving institutions such as museums, archives and libraries, another article [Wehner, 2019] was written, which, in addition to a detailed description of the OCR workflow, has a special focus on the Development of the software in cooperation with the humanities and thus shows the benefits and opportunities of the software for the culture and art preservation institutions addressed.



(/kallimachos/index.php/Datei:Abbildung_5_B

Figure 5: Sample pages of the early prints used for evaluation and some of the expected segmentation (right).

Reporting (selection): In April 2019, the online magazine of the University of Würzburg einBLICK published an article (https://www.uni-wuerzburg.de/aktuelles/einblick/single/news/modernes-tool-fuer-alte-texte) on OCR4all in German and English and distributed it via various national and international communication channels. The article met with great interest, which, in addition to reporting in numerous online portals and daily newspapers (including Der Tagesspiegel, Augsburger Allgemeine and Der Standard), also resulted in a radio interview (https://www.swr.de/swr2/wissen/Digitalisierung-Mittelalterliche-Handschriften-werden-Textdokumente,aexavarticle-swr-62524.html) with SWR2 Impuls. Furthermore, there was the opportunity to introduce OCR4all to an even larger group of humanities users as part of a tool presentation (https://fortext.net/tools/tools/ocr4all) of the DFG-funded project forText . (https://fortext.net/ueber-fortext)

Lectures given (selection):

- Input workshop of the DFG-funded specialist information service philosophy of the University and City Library of Cologne
- Annual meeting of the Working Group on Provenance Research eV in Düsseldorf
- Guest lecture at the Trier Center for Digital Humanities
- Colloquium corpus linguistics and phonetics at the HU Berlin
- Already agreed: Conference on Digital Medieval Studies, January 2021 in Bremen

Instructions, workshops, internships and other teaching activities (selection): In order to make it easier for users to get started and continue working with OCR4all, extensive and clear instructions (https://github.com/OCR4all/getting_started) for installation and use were created and published in German and English.

Furthermore, the use of OCR4all is explained step-by-step using two supplied example works. The instructions are continuously maintained and updated regularly. A Semantic MediaWiki, which, in addition to the instructions, will also provide and link numerous definitions of terms, technical backgrounds and frequent problems as well as their solutions, is currently under construction. Other training activities are listed below. The following workshops were held at the University of Würzburg:

- Regular workshops for professors, employees and students of all faculties.
- Regular internships for students of the master's program "Middle Ages and Early Modern Times".
- Sub-module of the additional certificate Digital Competence (http://www.anglistik.uni-wuerzburg.de/studium/im-studium/zusatzzertifikat-digitale-kompetenz) , which offers students of modern philology the opportunity to acquire and demonstrate skills in handling digital data beyond their studies.

The following workshops were offered nationally and internationally:

- Workshop at the annual conference of Digital Humanities in German-speaking countries 2020 in Paderborn. The feedback from more than 20 participants was extremely positive.
- Train-the-Trainer workshop in summer 2019, at which interested parties who want to offer OCR4all workshops themselves or already offer them, were trained separately. They also had the opportunity to exchange feedback from the community with the developers and to discuss how to proceed. Due to the excellent feedback, further events of this kind are planned. The DHd workshop mentioned above was organized by the Würzburg working group in cooperation with some participants of the train-the-trainer workshop.
- Two workshops (Würzburg 2018 and Budapest 2019) as part of the COST Action Distant Reading for European Literary History.
- Seminar on historical corpus linguistics at the Humboldt University in Berlin: In cooperation with the ZPD, various works from the 17th and 18th centuries on the subject of herbs were transcribed by master's students. The necessary calculations ran on the Würzburg servers, while the students were able to make the necessary corrections remotely and conveniently via a web interface after a short briefing.
- Workshop as part of the "Digital Visual Studies" teaching project at the Art History Institute of the University of Zurich in the program "Strengthening digital skills in education"
- Workshop at the Swiss Idiotikon (https://www.idiotikon.ch) in Zurich.
- Workshop at the Trier Center for Digital Humanities.

The following courses have already been/are offered by participants of the above-mentioned train-the-trainer workshop (apart from the already mentioned DHd workshop and the numerous events at the University of Würzburg):

- Various courses at the Institute for Information and Language Processing, LMU Munich.
- Numerous workshops at the Leopoldina Center for Science Studies.
- Exercise at the history seminar of the LMU Munich.
- In summer 2020, two of the participants will offer two multi-day OCR4all workshops as part of the European Summer University in Digital Humanities at the University of Leipzig. (http://esu.culintec.de/?q=node/1216)

Well-known users and application scenarios: OCR4all has been downloaded more than 1,600 times via the main distribution channel DockerHub (as of March 2020). Since a user makes several downloads, an instance can be used by any number of users and the tool can also be installed completely by yourself using the code provided on GitHub, it is not possible to estimate the number of users more precisely. Instead, the following is an overview of well-known users and application scenarios. Only secured assignments (by publication or direct contact) are listed and the numerous other activities that are e.g. B. via GitHub or various social media platforms are ignored.

In addition to the applications in the Kallimachos sub-projects Narragonien digital and Anagnosis, OCR4all is also used in the following chairs and projects at the University of Würzburg:

- “Camerarius digital” (successor project to the DFG project Opera Camerarii (http://www.camerarius.de) , cf. [Hamm et al. 2019].): Collection of 303 Latin and Greek prints by the German humanist Joachim Camerarius (cf. section Evaluation; a DFG research grant application was submitted) . Applicants are the Chair of Classical Philology (Latin Studies), the Chair of Artificial Intelligence and Knowledge Systems, the Chair of History of Medicine and the Professorship of German Philology.
- Chair of Computer Philology and Modern German Literary History:
  - Mass collection of Gothic novels from the 19th century (over 800 novels already processed).
  - Recognition of poetry anthologies with poems of realism/naturalism (SPP Computational Literary Studies (https://dfg-spp-cls.github.io) , sub-project "Modern Poetry").
  - Structure of a corpus of notebook novels (preliminary work for a project application).
  - Building a corpus of novellas and stories (Julian Schröter’s habilitation project).
- Chair of German Linguistics:
  - Collection of various sources (focus on the 19th century) to enrich the Würzburg database of linguistic cases of doubt ([www.zweidat.germanistik.uni-wuerzburg.de ZweiDat]) (project application in preparation).
  - Transcription of German Gothic prints from the 16th century in the Greifswald Digital (http://www.stadtsprachgeschichte.germanistik.uni-wuerzburg.de) project .
  - Collection of travel guides for discourse linguistic studies (PhD project by Miriam Reischle).
  - Recognition of German-language printed texts (mostly 17th century) containing alchemical and astrological symbols (Jonathan Gaede's dissertation project).
- Chair of French and Italian Literature: Collection of early modern French manuscripts (feasibility study for project application).
- Chair of Modern German Literary History I: Transcription of selected libretti from the Hamburg Opera from the period 1670-1728 (preparation for project application).
- Jean Paul Portal (http://www.jean-paul-portal.uni-wuerzburg.de/startseite/) : OCR of original prints by Jean Paul as part of the sub-project "Flegeljahre" (starts on 04/01/2020).
- Chair of Comparative Linguistics: OCR of 19th-century Armenian texts.

- Chair of English Linguistics: Compilation of English newspaper texts and letters from the 19th and 20th centuries.

National and international projects and applications of OCR4all include:

- Project MiMoText (https://kompetenzzentrum.uni-trier.de/de/projekte/projekte/m) at the Competence Center of the University of Trier: Recording French novels of the 18th century
- Monumenta Germaniae Historica: Lexica from the incunable period (cooperation with ZPD to prepare a LIS project application).
- Max Planck Institute for European Legal History: Collection of legal historical sources (mainly early modern prints in various languages).
- Deutsches Historisches Museum Berlin: OCR of archival materials from the 19th and 20th centuries
- Department of English, University of Bristol: The Literary Heritage of Anglo-Dutch Relations, 1050-1600.
- Universidad Nacional de Educación a Distancia (Madrid): [www.incunabula.uned.es project] for the compilation of Latin texts from the 15th and 16th centuries.
- Older German Philology / Medieval Studies, Heidelberg University: OCR of various texts by Sebastian Brants around 1500.
- Commission for Bavarian State History at the Bavarian Academy of Sciences: Collection of various prints and typewriter products (including history books, yearbooks, place name books, ...).
- Project WiTTFind (http://wittfind.cis.uni-muenchen.de) at the CIS of the LMU Munich: Processing of different materials, including reinforced typewriter pages.
- Martin Luther University Halle-Wittenberg: Digitization of early modern encyclopedias.
- Project Heinrich Wölflin - Collected Works (https://www.woelfflin.uzh.ch) of the Art History Institute of the University of Zurich and the Max Planck Institute for Art History, Bibliotheca Hertziana: Collection of unpublished manuscripts (proof of concept in cooperation with the ZPD).
- Project "Epigrāphia Carnāṭica digital" of the University of Cologne and the LMU Munich (OCR of the Dravidian language Kannada; DFG application submitted, cooperation with ZPD intended).
- Humboldt University of Berlin, Institute for Archaeology: Compilation of Coptic texts from the 19th and 20th centuries
- City Museum Bingen: OCR of different historical documents.

Summary and Outlook: With OCR4all, a tool was created with which even non-technical users can capture extremely demanding material independently and with the highest accuracy via OCR. The tool is already being used productively in a large number of application scenarios far beyond the borders of Würzburg. The fact that it is firmly anchored in the ZPD ensures long-term further development, maintenance and dissemination, and actively promotes networking and training in the OCR4all community. Of course, a special focus must always be placed on the demands and needs of non-technical users. In addition to a constant optimization of the existing instructions, va a further simplification of the installation process through the use of VirtualBox and the integration of more user-friendly output formats in the foreground. The latter is currently being developed as part of a DHd-funded small project founded in the summer of 2019DHd working group OCR (https://dig-hum.de/ag-ocr)implemented by converting the developer format PageXML into the user formats TEI, ALTO and PDF. The University of Würzburg is prominently represented in the AG and, in the form of the ZPD, is directly involved in the implementation of the project. From a technical point of view, in addition to the intended integration of the solutions created within the framework of OCR-D (see section on the differentiation from and synergy effects with OCR-D), the expansion to a "real" server application represents the next big step the development of the application, due to the sometimes very computationally intensive processes, initially the use by a single user was clearly in the foreground, there is still potential for improvement in this regard. This is how the ZPD currently works in a multi-level user administration as well as a complex resource management, which promises a simple and secure collaborative work with ideal utilization of the existing server infrastructure. Apart from the instance of the ZPD, which is already available, several institutions have already expressed a concrete interest in setting up a similar system. Furthermore, the development and integration of a more robust, fully automatic segmentation method and a (semi) automatic post-correction are being pushed ahead. Apart from the instance of the ZPD, which is already available, several institutions have already expressed a concrete interest in setting up a similar system. Furthermore, the development and integration of a more robust, fully automatic segmentation method and a (semi) automatic post-correction are being pushed ahead. Apart from the instance of the ZPD, which is already available, several institutions have already expressed a concrete interest in setting up a similar system. Furthermore, the development and integration of a more robust, fully automatic segmentation method and a (semi) automatic post-correction are being pushed ahead.

Use Case 1: Naragonia 2.0

Supporting the development of OCR4all

The use case has continuously supported the development of OCR4all. In addition to providing Ship of *Fools* digital copies and manually created ground truth data, the Narragonien team has contributed its literary expertise to the further development of OCR4all: Questions of user-friendliness and user perspectives were discussed with the IT colleagues, workflow optimizations and usability discussed checked the user interface. In cooperation with the computer scientists, setup and user guides were written in English and German for non-computer users and guidelines (https://github.com/OCR4all/getting_started)created for OCR4all (layout segmentation, ground truth generation, etc.). In addition, the Narragonien team created several elaborate evaluation datasets (text and layout ground truth) that were used for the OCR4all publications, including for [Reul et al. 2019c], in which the Narragonien employees M. Wehner and Chr. Grundig were involved.

OCR recognition of Ship of *Fool* outputs

Thanks to the advances made by OCR4all, the transcription of *Ship* of Fools prints, which had to be done mainly by hand due to the OCR problems in the first project phase, could be largely automated. By using manually corrected text material as a basis, new so-called “mixed models” were trained, which greatly facilitated the recognition of new *Ship* of Fools editions ([Springmann and Lüdeling, 2017]). With comparatively little effort, further editions / edits and several individual copies of the *Ship of Fools could be made with OCR4all*be tapped. Accuracies of up to 99.8% were achieved. For the early printing period, OCR is now recommended as a standard text-reading technology that can also be used successfully by non-computer scientists.

Collaborative work with Semantic MediaWiki

In the "Narragonien" project, the developed Ship of Fools texts were recorded in a Semantic MediaWiki (SMW), corrected in the digital copy and normalized to reading texts. These were linked to an overarching register of places and names as well as to a source register (see below for TA 1.2.1). In order to ensure the longevity of this data and its cross-platform reuse, a transformation routine was developed in cooperation with the ZPD (H. Baier), which automatically generates XML files in the valid TEI-P5 format from the Semantic MediaWiki data according to defined rules. This TEI-P5 subset is based on the common basic format of the Deutsches Textarchiv (http://www.deutschestextarchiv.de/doku/basisformat/)and has been expanded to include the TEI variance award. The SMW has also proven itself as an intuitive annotation tool for non-computer users. Due to its flexibility, ease of use and TEI export option, it is recommended as a tool for collaborative text entry and basic text markup. The SMW is already being reused in the humanities, including in the DFG project 'Opera Camerarii' (cf. http://wiki.camerarius.de (http://wiki.camerarius.de) , [Baier, Hamm, Schlegelmilch 2019]).

Text corpus: European *Ship* of Fools editions before 1500

The UseCase has set itself the literary goal of making important European *Ship* of Fools editions before 1500 digitally accessible (published project description and example analysis in [Grundig / Hamm / Walter 2017]. For later Ships of Fools cf. [Hamm 2019]). The following editions were published:

- Sebastian Brant, 'Ship of Fools', Basel 1494 (GW 5041): digital copy + transcription + reading text
- Sebastian Brant, 'Ship of Fools', Basel 1495 (GW 5046): digital copy + transcription + reading text
- Sebastian Brant, 'Ship of Fools', Basel 1499 (GW5047): digital copy + transcription + reading text
- Adaptation of the 'Ship of Fools', Nuremberg 1494 (GW 5042): digital copy + transcription + reading text
- Adaptation of the 'Ship of Fools', Strasbourg 1494/5 (GW 5048): digital copy + transcription + reading text
- Jakob Locher, 'Stultifera Navis', Basel 3/1/1497 (GW 5054): digital copy + transcription
- Jakob Locher, 'Stultifera Navis', Basel 8/1/1497 (GW 5061): digital copy + transcription + reading text
- Jakob Locher, 'Stultifera Navis', Basel 1498 (GW 5062): digitized + transcription
- Low German adaptation of the 'Ship of Fools', Lübeck 1497: transcription + reading text
  - The transcription of this edition was kindly provided by the project "Middle Low German in Lübeck" (MiL; WWU Münster; Dr. Robert Peters, Norbert Lange). It was slightly revised in 'Narragonien digital' and a reading version was added. The digital copy will be integrated into the presentation promptly.
- Pierre Riviere. La Nef des folz, Paris 1497 (GW 5058): digital copy + transcription + reading text

The transcriptions or reading texts each comprise approx. 350 printed pages, were coded in TEI-P5 and have been corrected manually. "Narragonia digital" thus has a text corpus of 18 TEI text versions with a total volume of 5950 p. (approx. 3500 p. in Early New High German, approx. 350 p. in Low German, approx. 1400 p. in Latin and approx. 350 in French language) developed. Due to the very time-consuming work on registers, source references and variant coding (see above) and due to the departure of Chr. Grundig (cf. [Grundig 2012], [Grundig 2016]. Christine Grundig's dissertation supervised by J. Hamm, which among other covers the German, Latin and English Ship of Fools, should be submitted in summer 2020) the final work and final corrections to the remaining Ship of Fools editions are not yet finished. Transcriptions and reading texts are available for the French editions GW 5060 and GW 5065, for the Dutch version GW 5066 and for the English edition (A. Barclay), but TEI export and final correction are still pending. They will be made up for by the end of 2020. The finished Ship of Fools texts are already or will be implemented in the digital text presentation (see below) and made available as an XML download. They will be made up for by the end of 2020. The finished Ship of Fools texts are already or will be implemented in the digital text presentation (see below) and made available as an XML download. They will be made up for by the end of 2020. The finished Ship of Fools texts are already or will be implemented in the digital text presentation (see below) and made available as an XML download.

Digital text presentation on the Internet

The project results are made available at http://kallimachos.uni-wuerzburg.de/exist/apps/narrenapp/ (http://kallimachos.uni-wuerzburg.de/exist/apps/narrenapp/) . The working version of this digital text presentation is currently being continuously improved and supplemented, then finally corrected and should be released at the end of 2020.

This digital presentation of the ships of fools was designed and programmed by the 'Narragon digital' team (scientific collaborator D. Heublein, student assistant Y. Herbst). The aim was to make some previously unedited Ship of Fool texts accessible for the first time and to use them to show historical text, image and layout transformations of the European Ship of Fool tradition ([Burrichter 2017]; [Hamm 2017]). This Narragonia homepage is hosted on the Kallimachos server for data availability. It is based on an eXist database in which the TEI files of the Ship of Fools editions exported from the wiki are stored. An application developed in the database bundles all components that are necessary for a web presentation. This app enables the data to be displayed in a synoptic online viewer,

1. **Paratexts and metadata:** Introduction to the project, metadata of the ships of *fools* , research bibliography, transcription and edition guidelines, index of places and names and sources cited in margine, download area (linking of the digital copies and provision of the XML/TEI files with transcription and reading text).
2. **Presentation of the reading texts:** reading texts with the respective digital copies; Linking of place names and personal names to the general register; interlinear display of edition variants, press corrections and emendations of the editors; chapter index; search function.
   1. from Brant's *Ship* of Fools (Basel 1495, GW 5046) with the text variants of the first (Basel 1494; GW 5041) and the third edition (Basel 1499; GW 5047).
   2. the Nuremberg processing (GW 5042)
   3. the Strasbourg version (GW 5048)
3. **Synoptic representation:** A freely configurable two-window synopsis is used to present the individual texts. It allows you to compare any two Ships of *Fools* or two media representations of a *Ship* of Fools chapter by chapter. *For example, the digitized version of the Ship of Fools* in Strasbourg can be displayed in the left window and its reading text in the right window, so that the comparative reading of image and text is possible. You can also put the German ship of fools in the left window and the Latin *ship in the right window, for example*load so that the Latin version can be compared with its German version. The synopsis contains, each with transcription and reading texts: GW 5041, GW 5042, GW 5046, GW 5047, GW 5048. The other text versions that have been prepared will be posted in the near future. Due to the modular structure of the eXist application, the synopsis can also be reused for other digital edition projects.
4. The **search function** includes the text search in a single text, but also enables a layout search across editions in the sense of filtering by layout areas. For example, the woodcuts for *Chapter* 3 or the motto verses for Chap. 12 show. Changes in the layout and the "intermedial variance" ([Hamm 2016]) of the *Ship* of Fools tradition can thus be represented and examined more easily.

The presentation of the European Ships of *Fools* , which is to be expanded to include the remaining texts by the end of 2020, goes far beyond the usual format of a "digital edition". It is an integrated tool that not only displays texts, but also presents them in various media forms and at the same time makes their historical movement representable and understandable. Results from this comparative Ship of Fools analysis were presented by the project group in lectures and published in essays, see most recently [Burrichter 2019a], [Burrichter 2019b], [Burrichter 2019c], [Hamm 2019], [Hamm i.Dr.] . With the viewer and the online presentation of the Ship of Fools, research not only has access to previously unedited *Ship of Fools*editions, but also an innovative tool for the philological comparison of texts and for the analysis of early modern fool literature in Europe.

Project proposals, conferences, lectures, etc.

(Publications see bibliography)

Follow-up project applications, further use of the project results

- Paratextuality and translation practice in instructional literature in the early modern period / Paratextualité et littérature didactique au début de l'ère modern. German-French application to the DFG in the DFG-ANR program, applicants Brigitte Burrichter and Anne-Laure Metzger-Rambach (Université Bordeaux-Montaigne), submitted on March 12, 2020.
- Camerarius digital. Application to the DFG in the research grant program. Applicant Thomas Baier, Joachim Hamm, Frank Puppe, Ulrich Schlegelmilch, submitted on February 7, 2020 [successor project to 'Opera Camerarii'; Use of Semantic MediaWiki and OCR4all for Greek and Latin early prints in Humanism]
- Internships for OCR4all in the master's program “Middle Ages and Early Modern Times” at the Univ. Wuerzburg
- Inclusion of the corrected OCR transcriptions from several Ships of Fools in the "GT4HistOCR dataset", which provides trained OCR models for text recognition of early modern prints [Springmann et al. 2018].

Conferences, workshops, excursions by the project group

- Bordeaux, 31.5. to 1.6. 2018: The international conference Les Nefs des folz en Europe (http://kallimachos.de/kallimachos/index.php/Datei:Narrenschifftagung_Bordeaux_2018.pdf) , organized by Brigitte Burrichter and Anne-Laure Metzger-Rambach, took place from 31.5. until June 1st, 2018 at the University of Bordeaux. The subjects were the adaptations of the Ship of Fools and the 'Stultifera navis' in early modern Europe.
- Würzburg, 24.5. until 25.5. 2019: Workshop on the current status of the planned online edition. Anne-Laure Metzger-Rambach (Bordeaux), Micheal Rupp (Leipzig) and Thomas Wilhelmi (Heidelberg) as well as all employees of the Würzburg working group took part as guests.
- Upper Rhine, 31.10. until 2.11. 2019: Excursion to Strasbourg and Basel, organized by Thomas Wilhelmi (Heidelberg). The employees of the Würzburg working group took part.

Lectures:

- Joachim Hamm: "Gen Naragonien". Sebastian Brant's 'Ship of Fools' (1494) and its German language adaptations in the 16th century. Guest lecture at the Univ. Brunswick, April 16, 2019.
- Brigitte Burrichter: Sebastian Brant in context. Workshop at the Ecole Normale Supérieure de Paris, February 11, 2019.
- Brigitte Burrichter: "Les Nefs des fous dans le contexte européen", lecture at the conference "À la recherche de Sébastian Brant (1457-1541), Strasbourg, February 8, 2019
- Joachim Hamm: Narragonia digital. Guest lecture at the Univ. Saarbrucken, June 14, 2018.
- Brigitte Burrichter: Les Nefs des folz en numerique. A line edition of the Nefs européennes. Lecture at the conference "Les Nefs des folz en Europe" (Bordeaux), May 31 - June 1, 2018.
- Raphaëlle Jung: Le chapter B99 de la Nef des fous - une analyses. Lecture at the conference "Les Nefs des folz en Europe" (Bordeaux), May 31 - June 1, 2018.
- Christine Grundig: « Here maketh myne Autour a specyall mencion - Concepts of adaptation and authorship in the English 'Ship of Fools' by Alexander Barclay and Henry Watson. Lecture at the conference "Les Nefs des folz en Europe" (Bordeaux), May 31 - June 1, 2018.
- Joachim Hamm: Variance and authorship. On the Basel editions of the 'Stultifera navis'. Lecture at the conference "Les Nefs des folz en Europe" (Bordeaux), May 31 - June 1, 2018.
- Dominika Heublein: Argument marking in TEI. Lecture at the conference "Les Nefs des folz en Europe" (Bordeaux), May 31 - June 1, 2018.
- Julius Goldmann: The image in the text - references between narration and woodcut. Lecture at the conference "Les Nefs des folz en Europe" (Bordeaux), May 31 - June 1, 2018.
- Joachim Hamm (together with Frank Puppe, Nico Balbach): Internal collation and analysis of variance in Narragonia digital 2.0. Lecture at Philtag 15, Univ. Wuerzburg. 10.4.2018.
- Brigitte Burrichter: Sebastian Brant in context. Workshop at the Ecole Normale Supérieure de Paris, February 5, 2018.
- Joachim Hamm: Scholarly Foolishness. The 'Ship of Fools' by Sebastian Brant and the Würzburg project "Narragon digital". Lecture in the old town hall of Miltenberg in the lecture series of the Unibun-es, January 16, 2018.
- Joachim Hamm: An integrated digital edition of the 'Ships of Fools' before 1500. Lecture in the lecture series of the academy project "The Austrian Bible Translator", Univ. Augsburg, November 30, 2017.
- Brigitte Burrichter: Patrice et les Dernydes. Les versions françaises de la Nef des fous de Sebastian Brant. Lecture at the Translatio et histoire des idées conference at the University of Warsaw, October 19-21, 2017.
- Joachim Hamm: Inconsistent texts? Reflections on the 'ships of fools' of the early modern period. Lecture at the International Symposium "The 15th Century", Melanchthon-Akademie Bretten, October 12th to 14th, 2017.

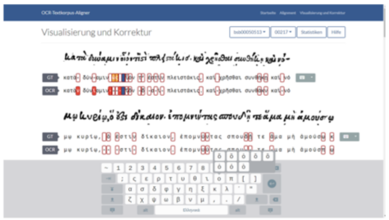
Use Case 5: Anagnosis 2.0

With the training of an OCR on Greek prints by Aldus Manutius, an attempt was made for the first time to record the early modern prints of Greek classics, which were heavily interspersed with handwritten ligatures. The first phase of the work on the Use Case Anagnosis was dedicated to an estimation of the implementation effort for the planned work steps based on a selection of representative outputs. The prerequisite for this was the procurement of the required material in the form of high-resolution images from the respective host institutions (Bavarian State Library, Jena University Library) at the beginning of October 2017. When selecting the outputs to be tested, care was taken that these are at least already available in full-text databases or are considered the oldest source for the text in question (since the handwritten original has been lost). In addition, the selection had to be representative of the different typefaces that were used for typesetting Greek script during the activity period of the workshop of Aldus Manutius (both during his lifetime and after his death). Accordingly, the following prints were selected as the basis for the text:*Epistolae diversorum philosophorum, oratorum, rhetorum* (Manutius 1499; ISTC ie00064000, GW 09367). Font 7:114Gr according to GW; *Galeni opera omnia* (Manutius, Andreas Asolanus 1525), see (perilli2012); Font 9:84Gr according to GW.

Two mutually dependent goals were pursued with the work on the text recognition of the early prints: (1) achievement of a minimum recognition rate, for which the sequence alignment suffices (see below); (2) Refinement of the OCR results so that they can in turn be used as ground truth for the further development of the recognition algorithm. Both goals were achieved in the reporting period. The refinement of the OCR recognition was done by manually entering the ground truth, with a total of approx. 550 lines GT being created, then by correcting the results created in each case. The input of the ground truth for the training of the various models was initially carried out using OCR4All (see Chapter 1.1). The recognition accuracy achieved varies greatly depending on the quality of the originals, but is usually sufficient for sequence alignment, ie finding any texts that may have already been transcribed from a data collection. An overview can be found in the poster "Aligning extant transcriptions of documentary and literary papyri with their glyphs" (links (https://d-scribes.philhist.unibas.ch/en/events-179/neo-paleography-conference/poster-session-copy-1-237/))

The sequence alignment tool workflow [Bald et al. 19] consists of the following steps. Before the sequence alignment, the lines created by text recognition are normalized (removing the diacritics) in order to increase the chance of finding suitable equivalents in the comparison text. By measuring the similarity of the beginning of the text, the comparison document in the full-text database with the highest correspondence is initially preselected. Each line to be transcribed (referred to below as "OCR line") is then segmented into n-grams of five characters and this is searched for in the comparison text (referred to below as ground truth or GT line). From local clusters of hits in the n-gram search, candidates are generated, which are evaluated with regard to the number of n-grams found and a similarity measurement. This assessment results in the best candidate in each case. The global (aligning over the full length of the lines) Needleman-Wunsch algorithm for comparing two character strings aligns the OCR line and the best GT candidate in such a way that as many characters as possible match. Missing characters (e.g. a comma) in the OCR line compared to the GT line are filled with hyphens (-) that mark gaps. The length of the OCR line is adjusted to the length of the GT line. The results of the alignment then serve as input for the correction tool (see Fig. 6). This shows the original line of text with the (incorrect) OCR transcription and the best line of comparison text found by means of alignments, with the differences being highlighted. Users can then decide which is the correct transcription by selecting letters (or accept the comparison text as a whole). Fig. 6 shows an example of the correction interface of the alignment tool.

In view of the positive results in the processing of early modern prints, the synergy of text recognition using OCR4all and the alignment tool was used especially in the last project phase to produce ground truth in a Greek manuscript from the 16th century (BML, Plut. 75.7), the has never been fully transcribed before. Such a historical document thus represented an interesting field of experimentation for the functions developed in the sub-project. The main criterion for the selection of this manuscript was the fact that the form of the letters in the manuscript does not differ significantly from the printed image of the Manutius editions used in the initial phase. The results of this first attempt give reason to hope that



(/kallimachos/index.php/Datei:Abbildung_6_K
Textkorpus-Aligner.png)

Figure 6: Correction view of the OCR text corpus aligner for two lines of Greek text: the original line at the top, the comparison line ("GT") selected by alignment below it, and the line transcribed by OCR at the bottom, with the ground truth being determined by colored markings (completely below a virtual editor with domain-specific special characters).

Work on integrating the alignment tool in OCR4all was started but could not be completed in the reporting period. For reasons of software compatibility, an extensive re-implementation of the alignment tool is required. Currently, the step of aligning an OCR transcription with selected comparison texts has to be carried out in a separate application. The result can be uploaded and post-processed in OCR4all's LAREX editor for correction.

WP2: Information Extraction

TA 1.2.1: Development of processes for semi-automatic indexing

The conceptual preparatory work in the “Narragonia” use case was continued in such a way that a (manual and analogue) marking of suitable terms was tested on a text. The extension to other texts and above all the planned technical support are complex, since the ship of *fools*-Texts are written in an older language level for which there are no ontologies and the translation of the terms into other languages cannot be based on dictionaries. Considered solutions to this problem - such as self-generated and edited dictionaries based on the texts as a starting point - were not feasible due to the human resources. In the end, the decision was made to focus all efforts on the development of the viewer so that it would be available at the end of the project. The work on the indexing should be continued after the end of the project. Irrespective of the fact that the subject index is still missing, all ships of *fools* were indexed by an index of people and places.

For this purpose, the reading texts (see 1.1 AP1 “UseCase 1: Narragonien”) were linked to a comprehensive register that was created in the Semantic MediaWiki and contains a total of 975 **place and personal names** . Each register entry was linked to the GND and to subject-specific person and place dictionaries and described in detail by the 'Narragonia' team. This digital register goes far beyond the usual referring paratext of the print medium:

- On the one hand, the digital explanations form a commentary on the Ships of *Fools* in nuce, as they make the complex allusions of the fool poets to biblical, historical and literary people and places transparent and comprehensible. This follows the example of the print edition of the German *Ship* of Fools [Knape 2005], whose number of lemmas has admittedly been significantly expanded.
- On the other hand, the online register includes not just a single Ship of Fools, but all digitally indexed Ships of *Fools* , i.e. integrating the German original edition as well as the German, Low German, English, French, Dutch and Latin versions. Thanks to wiki technology, the document locations in these Ships of *Fools* are automatically listed in every register entry and displayed in the viewer, so the user can jump to the relevant text passage and see immediately which Ships of *Fools* contain the respective personal or place name - and which do not take it from their template wanted to.

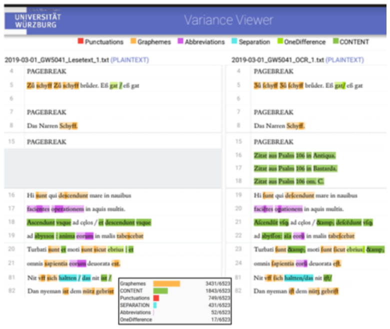
The index of places and names is suitable for subsequent use by other humanities Semantic MediaWiki projects that deal with the early modern period. This also applies to the **reference point register** : The Latinists Rena Buß and Helena Wächter from the 'Narragonien' team individually identified the approx. 1400 (!) Latin source references (*loca concordantia*) that Sebastian Brant had inserted into the 'Stultifera navis' and linked in the wiki with the full text of the quoted passage (mainly the Bible, ancient classics, legal texts). This extensive collection can be used in two directions. Follow the links in the *Ship of Fools*-Text, the respective source reference is resolved and one reads the Bible or classic citation in full text. If you start from the register, you can find out how many and which ships of *fools* refer to a certain reference point, for example where the biblical story of the poor Lazarus is reminiscent or quotes from Cicero's Verres speeches.

The two registers are the first building blocks of a comprehensive commentary on the early modern Ships of *Fools* , which has been a desideratum since the beginning of research in the early 19th century. They use the technique of semantic linking provided by a Semantic MediaWiki. These links were transferred to the TEI format using the transformation routine developed in the project and included in the digital presentation of the Ships of *Fools* .

TA 1.2.2: Workflow for analysis of variance

The “Variance Viewer” was developed in cooperation with the Use Case “Narragonia digital” (see the supplementary description below for “Use Case 1: Narragonia digital 2.0”). It is an open source tool that finds differences between two texts at the character level and differentiates and visualizes them into different categories that users can define themselves via a configuration file [Balbach et al. 2020]. This should enable them to quickly get an overview of the essential differences between texts and to recognize systematically caused differences (e.g. with regard to the normalization of texts) at a glance. Currently, the Variance Viewer can recognize the following categories of differences based on a generic concept of insertions, replacements and deletions (see Fig. 7):

- Punctuation: The change only applies to one punctuation mark (. , ; - ? ! etc.).
- Graphemes (Graphemics): The change only applies to certain spellings (ae ä; ue ü; oe ö; ss ß; upper/lower case; th t; etc.).
- Abbreviations: The change relates only to abbreviations (eg Dr. Doctor; Mr. Herr Herr; etc.).
- Spaces in the word that separate a word into two or more words. (separation). This option is technically more complex because not individual words but groups of words have to be compared with each other.
- Content changes with only one character difference (OneDifference) that are not included in the grapheme list and are evaluated differently than more complex changes.
- Content: All other changes that do not fall into any of the above categories, including additions or deletions, and changes where more than one change of the above types occurs at the same time.



2019-03-01_GW5041_Lesetext_1.txt (PLAINTEXT)

```
4   PAGEBREAK
5   Zů schyff Zů schyff brůder. Eß gat / eß gat
6
7   PAGEBREAK
8   Das Narren Schyff.
15  PAGEBREAK
16  Hi sunt qui descendunt mare in nauibus
17  facientes operationem in aquis multis.
18  Ascendunt vsque ad cęlos / et descendunt vsque
19  ad abyssos : anima eorum in malis tabescebat
20  Turbati sunt et moti sunt sicut ebrius : et
21  omnis sapientia eorum deuorata est.
81  Nit vff sich halten / das nit ist /
82  Dan nyeman ist dem nütz gebrist
```

2019-03-01_GW5041_OCR_1.txt (PLAINTEXT)

```
4   PAGEBREAK
5   Zů ſchyff Zů ſchyff brůder. Eß gat/ eß gat
6
7   PAGEBREAK
8   Das Narren Schyff.
15  PAGEBREAK
16  Zitat aus Psalm 106 in Antiqua.
17  Zitat aus Psalm 106 in Bastarda.
18  Zitat aus Psalm 106 om. C.
19  Hi ſunt qui deſcendunt mare in nauibus
20  faciētes opationem in aquis multis.
21  Aſcendūt vſq̄ ad cęlos / &amp; deſcēdunt vſq̄
22  ad abyſſos: aīa eorū in malis tabeſcebat
23  Turbati ſunt &amp; moti ſunt ſicut ebrius: &amp;
24  omnis ſapientia eorū deuorata eſt.
81  Nit vff ſich halten/das nit iſt/
82  Dan nyeman iſt dem nütz gebriſt
```

```
Graphemes      3431/6523
CONTENT        1843/6523
Punctuations    749/6523
SEPARATION      431/6523
Abbreviations    52/6523
OneDifference    17/6523
```

(/kallimachos/index.php/Datei:Abbildung_7_V

Figure 7: Comparison of the edited reading text of the Ship of Fools edition GW5041 with the result of the OCR on the original text of another print edition, with the changes including both OCR errors and font normalizations in the reading text. This is easier to understand by highlighting the change types in different colors (explanation of the types in the text). Below is a statistic that breaks down the 6523 changes found by change type. The entire text comprised 150 pages with 4200 lines, 26000 words and 121000 characters and the associated configuration file ("Settings") about 100 lines. For this analysis, the Variance Viewer needed about 25 seconds in the server-side demo mode on the web (for intensive use, the open source code should be installed locally:https://github.com/cs6-uniwue/Variance-Viewer (https://github.com/cs6-uniwue/Variance-Viewer)

In addition, prototypical typographical changes were also analyzed that do not affect the content but the layout, with initially only the TEI element “rend” with the corresponding attributes (italic; blocked; etc.) being analyzed and displayed.

The "Variance Viewer" has proven itself in the practice of variance analysis, especially in the UseCase "Narragonien digital" (see below for "UseCase 1"), not least because it can automatically assign the detected character differences to previously defined types of variance (e.g variance in punctuation, graphematics / abbreviations, etc.) and makes the variance display considerably clearer by marking these types.Because of its simple operation, the variance viewer was also used outside of the Kallimachus project, including in the Richard Wagner fonts project (http://www.musikwissenschaft.uni-wuerzburg.de/forschung/richard-wagner-schriften/) .

TA 1.2.3: Sequence Alignment

The work on the sequence alignment was carried out in the context of use case 5 anagnosis (sa)

TA 1.2.4: Development of a tool for automatic scene recognition

After initial difficulties in creating guidelines, it was nevertheless possible to create both a data set (which is currently being developed further and beyond the end of the project) and to find a neural network architecture that automatically detects the scenes with great accuracy.

TA 1.2.4.1: Conceptual modeling of action sequences

As part of this project, it became apparent that even manual annotation of scenes is very difficult from a literary perspective. Finally, several different definitions of scenes and their operational feasibility were examined in joint efforts with Prof. Gius (Hamburg) and Prof. Reiter (Darmstadt). As part of this collaboration, three different guidelines with associated annotated data were created, which were then reduced together to a set of annotation guidelines. This collaboration will continue after the end of the project, so that more manual data can be annotated and the automatic component can become more reliable with more training data.

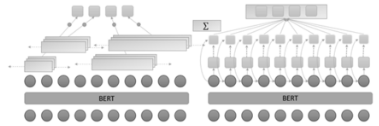
The resulting guidelines and the annotations were presented to a specialist audience as part of a panel at the DH 2019 and discussed there (Gius et al. 2019).

TA 1.2.4.2: Annotation and automatic recognition of action sequences

Regardless of the difficulty of data acquisition, an algorithm based on deep learning and voting has been developed that is able to automatically predict scene boundaries with approximately 90% accuracy. It is pleasing that this algorithm does not require a particularly large amount of data for this recognition accuracy (approx. 10 notebooks were sufficient for this), nor is it specially adapted to a guideline. Since we had data sets for several guidelines available, we were able to empirically verify that the same algorithm, given suitable training data, is also able to internalize these guidelines and then recognize scenes with a high level of accuracy.

The algorithm works in a sliding window method with a window size of at least 10 sentences. Each sentence is BERT coded and then features of a sentence are extracted with the help of different convolution and recurrent layers (see Fig. 8).

These sentence representations are then processed by another LSTM, from which a prediction is then made for each sentence within the window as to whether it represents a scene boundary or not. Since each sentence is contained in several of these windows, all predictions are aggregated and the final statement is made using a voting process.



(/kallimachos/index.php/Datei:Abbildung_8_A

The high level of accuracy without the further need for external features, such as locations or a fine-grained statement about figures that are currently present and absent, helped to find a good method despite the delay in annotation. The resulting paper (Herud et al 2020) is currently under review for the

ACL, a high-level natural language processing conference.

Figure 8: Illustration of the sentence representation in the neural architecture for automatic scene recognition.

TA 1.2.4.3: Integration of the developed methods into a common prototypical end-to-end UIMA workflow

Due to the use of deep learning for the automatic recognition of scenes, this step was carried out in Python. This code works on a JSON format for which there is a converter from and to UIMA-xmi, so that at least the results can be used in the UIMA workflow. Furthermore, a component for Python was already developed in Callimachus I, which allows Python users to work directly with xmi documents.

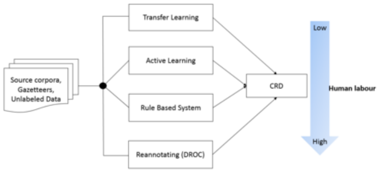
TA 1.2.5: Refinement of character analysis (named entity recognition, coreference resolution, relationship recognition between characters, sentiment recognition)

Within the framework of this work package, various experiments for the development of a character recognition based on the data set DROC annotated in Callimachus were carried out, as well as the co-reference was extended by relation and the possibility of integrating user-specific knowledge. Furthermore, an architecture of a neural network was developed that can follow the emotional course of characters in novels.

TA 1.2.5.1: Fine-grained attribution of characters through knowledge modeling

With the help of existing data from Callimachus I, various investigations and extensions could be carried out for this work package. The recognition of characters in novels is initially divided into two parts: a) the recognition and resolution of character references and b) the extraction of relations between the characters.

Since usually, as at the beginning of Callimachus I, there is no annotated data of the domain, various investigations were carried out for an automatic domain adjustment when recognizing figure references (see Fig. 9).



(/kallimachos/index.php/Datei:Abbildung_9_A

Figure 9: Different approaches for developing high quality figure reference detection.

It turned out that a recognition quality of over 80% F1 score can only be achieved with data from another domain and lists compiled from the Internet. If you want more than 90% F1 score, you either need a very good rules engineer, or you have to annotate data of the target domain. It could also be shown that a combination of the rule-based component and the learned component achieves results of 92% and thus performs best overall.

The co-reference has been extended with global constraints and a component for integrating user-specific knowledge from summaries (see TA 1.5.3).

In the field of relation recognition, the automatic recognition of speaker and addressee was expanded using deep learning methods, so that good recognition accuracy of over 90% is also possible on English-language texts. The use of deep learning to detect family relations could not surpass the results from Callimachus I, so a combination of a rule-based and a machine learning component is still used for this component. By means of summaries annotated with ATHEN, various investigations for the evaluation of automatically extracted figure networks could be made. The core results show that the core figures are recognized with a high accuracy, but the relations are not recognized well,

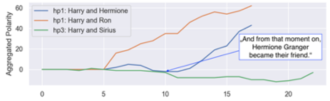
There are currently two journal papers under review (Krug et al 2020a; Krug et al 2020b)

TA 1.2.5.2: Expansion of relations to include dimensions of emotionality and polarity)

In order to expand the extracted relations to include emotional relationships, a bachelor thesis was supervised, in which, based on preliminary work by Kim and Klinger [Kim et al. 2019] improved the analysis of emotions in interactions between characters. To this end, a model for the classification of emotions described in short text excerpts containing two figures was first developed, which almost always improves the results of Kim and Klinger. A method was then developed to aggregate the extracted emotions via the text and thus to recognize the overall relationship of the characters. Furthermore, the course of the relationships between figures can be visualized with this method, as shown in Figure 10 as an example. To evaluate this procedure, relationships from the Harry Potter novels by JK Rowling were manually annotated. The resulting paper (Zehe et al 2020) is currently under review for KONVENS. The annotated corpus will also be released for further research at the latest when the paper is published.

TA 1.2.5.3: The integration of global constraints to improve coreference resolution

A component was integrated into the existing rule-based algorithm, which generates consistent results with the help of automatically recognized family relations between the characters. Even if this is an advantage, the recognition rate can only be increased by about 1-2% F1 score. In order to further exploit this, a possibility was designed to give user-specific knowledge about a text directly to the algorithm. This is done via a JSON file. This knowledge includes general relationships between characters, nicknames, and metadata such as the gender of individual characters. In an empirical study, this meta-knowledge was manually extracted from summaries and given to the algorithm. Since this meta-knowledge is manually collected from summaries, it doesn't take much time. Overall, improvements of up to 10% per document could be achieved, an overall average improvement of 4% F1.



(/kallimachos/index.php/Datei:Abbildung_10_

Figure 10: Course of sentiments between different main characters about Harry Potter

TA 1.2.5.4: Integration of the developed methods into a common prototypical end-to-end UIMA workflow

The above components are available for use and integrated into a unified workflow.

TA 1.2.6: Collecting metadata for the literary corpora

The metadata of the corpus of German-language novels of the 19th century were completely supplemented with information on the publication date or the narrow publication period. In cases where the publication can no longer be reconstructed, an estimate based on the age of the author was made. The information comes from relevant databases and from bibliographies digitized for this purpose. This completes the title, author, and publication date metadata for 3800 novels. For a sub-corpus, narrative perspective and genre were also determined.

OCR advances made in WP1 have made it possible to further expand the body of novels. 400 novels from the holdings of the Bavarian State Library were transformed into TEI and added to the collection. The process will continue after the end of the project with the developed pipeline.

The construction of a collection of contemporary literature has progressed through the purchase of e-books (paid for with funds from the Chair of Computer Philology) and the creation of a pipeline for transformation into valid TEI. The collection is divided into the categories of high and schematic literature. While there are currently only 200 high literature novels in the corpus, there are already 1000 so-called booklet novels of schematic literature (romance, horror and science fiction novels). This imbalance can be attributed both to the increased heterogeneity in the layout of the high literature, which makes the transfer to TEI more difficult, and to the fundamental problems of classifying a novel as high literature. To avoid a subjective imbalance within the collection,*German Book Prize* and *Büchner Prize*). The metadata could be taken from the online presence of the publishers across the board. Through cooperation with the German National Library, a much larger collection of novels could be made accessible. There are currently 25,000 paperback novels of various genres, 2000 high literature novels that were determined using the criterion described above, and 8000 novels from the paperback segment. The texts are available in TEI and in various formats required for analysis. In order to extract the novels from their source format (ebooks), a component has been developed which will be released in the near future. While the collection of acquired novels is used to develop questions and methods, the large corpus of the DNB can be used to generate statistically representative results (see Jannidis et al. 2019a; 2019b, 2020). From the cooperation, a requirement profile for future cooperation between research groups and the DNB could be developed. On the basis of these empirical values, it is currently arriving thereCall for Projects (https://www.dnb.de/DE/Professionell/Services/WissenschaftundForschung/wissenschaftundforschung_node.html) .

Use Case 1: Narragonia digital 2.0

On the contributions of the UseCase "Narragonia digital" to the development of processes for semi-automatic indexing and register creation (cf. TA 1.2.1. above)

In addition, aspects of variance were examined in the 'Narragon' UseCase . *In this context, any kind of discrepancy between related text editions of the Ship* of Fools is understood as “variance”.(cf. [Hamm i.Dr.], [Hamm 2016]): i.e. differences at the letter, word or sentence level, in the woodcuts, in the chapter stock or in the chapter arrangement. These variants can be encountered in the early printing period in successive editions and can go back to a revision, a revision. However, they can also arise as a result of direct interventions in the ongoing printing process and thus lead to differences between the copies of an edition ("press corrections"). These variance phenomena (known from early printing research) are important in terms of media history, since they document the "mobility" of texts not only in the handwritten age, but also in the printing age and refute the popular view that a printed text is unchangeable.

In editing studies, the collation of related text versions is a standard technique of textual criticism. It is supported by numerous text comparison tools already on the market. **The 'Variance Viewer'** (see TA 1.2.2) developed at the Chair of Artificial Intelligence offers the advantage of being able to be run online in any browser and to be able to compare two texts in a classifying manner. The idea developed in the 'Narragonia' project was to have the differences (characters, words, sentences, etc.) identified in the text comparison classified by the automatic application of configurable rules. In this way, the 'Variance Viewer' can, for example, detect deviations that are in the previously specified character group **,;!?;**. fall, classify as "variance in punctuation" and highlight it in the result display with its own color marking (can be shown / hidden). This extension of the 'Variance Viewer' and its various tests were carried out in close cooperation with Nico Balbach.

The classifying text collation was tested on the Latin *Ship* of Fools , as there are several editions and digitized printed copies of this. It yielded two results:

1. The collation with the 'Variance Viewer' determined far more variants between the first and second edition of the 'Stultifera navis' than previously known. The reason for this lies in the fact that research has so far been based on the printed partial edition of the 'Stultifera navis' ([Hartl 2001]), which was based on the Münster copy of the first edition. This copy, however, turned out to be a mixed version of the first and second edition after its collation (which had previously gone unnoticed), so that correspondingly fewer variants occurred between the first and second *Ship* of Fools editions. The main result of the 'Narragonia' collation is that Sebastian Brant's revision between the first and second edition was far more extensive and far-reaching than previously known. This finding was made on the*Ship* of Fools conference in Bordeaux 2018 and will now be published in a study accepted for publication [Hamm i. dr].
2. The hypothesis that the copies of the Latin ship of fools (as well as the German one) show textual differences that indicate press corrections could not be confirmed. With the help of OCR4all, several specimens of the first and second edition of the 'Stultifera navis' were recognized (with an accuracy of approx. 99.8%) and collated in the 'Variance Viewer'. On the one hand, this "internal collation" confirmed the already known finding that individual copies of the first edition were expanded by pages of text or an entire quire during printing or binding, which are only present in all copies of the second edition. On the other hand, the collation showed that in the examined copies of the first and second edition no variance at character, word or sentence level could be determined.*Ships of Fools* : Because in the Basel first edition of the German *Ship* of Fools several press corrections could be determined, and also in the French first edition, which has been examined more closely so far, there are copy variants, albeit to a small extent (so far three copies have been compared), which are listed in the text presentation.

The 'Variance Viewer' accommodates the standard procedure of every edition project that has to deal with variance phenomena. Since the viewer runs online in the browser and has an automatic classification system that the philological user can configure independently, it is an essential tool for digital editions that can be reused many times over.

Use Case 2: Quantitative Analysis of Narrative Texts

As a result of these use cases, an annotation environment is now available with ATHEN, with the focus on creating annotations of entities, their co-references and relations between the entities. The annotation of scenes was done with the web version "WebATHEN". The pipeline developed from Callimachus I has been revised with automatic components and the features mentioned in 2.5 have been added. The use of BERT and deep learning also enabled a component capable of automatically predicting scene boundaries with approximately 90% accuracy. The annotated algorithms are publicly available, but the data cannot be published due to restricted licenses.

Use Case 5: Anagnosis 2.0

The aim of AP2 is the automatic assignment of text fragments recognized by OCR against a full-text database (sequence alignment) with the output of the canonical citation style. The alignment has already been achieved in TA 1.2. The canonical output is still pending, but should be considered as a corollary to this result.

AP3: Stylometry

All work described in this section was carried out in close cooperation between sub-project 01UG1715B (Prof. Evert, FAU Erlangen-Nuremberg) and the Chair of Computer Philology (Prof. Jannidis, in sub-project 01UG1715A at the University of Würzburg). In most cases, a clear demarcation between the contributions of the two working groups is not possible. In the following sections, it is highlighted if individual sub-tasks were primarily or primarily located at one of the locations (FAU or JMU).

TA 1.3.1: Robust methods for recording lexical complexity (in connection with sub-project FAU 01UG1715B, Prof. Evert)

In a first phase, statistical models for type token distributions (so-called LNRE models) were further developed and made usable for simulation experiments. The results of this work were integrated into an open source software package (https://zipfr.r-forge.r-project.org/ (https://zipfr.r-forge.r-project.org/)) and communicated to a broader specialist public in tutorials, lectures and summer school courses:

- “Type-token distributions, Zipf's law, and quantitative productivity.”

Course at Corpus Linguistics Summer School 2018. Birmingham, UK.

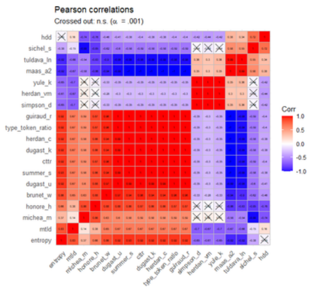
- “Measures of Productivity and Lexical Diversity,” Plenary Lecture at the 44th Annual Conference of the Japan Association for English Corpus Studies. Tokyo, Japan.
- "What Every Computational Linguist Should Know About Type-Token Distributions and Zipf's Law." Tutorial at the LREC 2018 Conference. Miyazaki, Japan. [1] (http://zipfr.r-forge.r-project.org/lrec2018.html)
- “What Every Corpus Linguist Should Know About Type-Token Distributions and Zipf's Law.” Tutorial workshop at the Corpus Linguistics 2019 Conference. Cardif, UK.
- "Corpus Statistics with R." Tutorial at KONVENS 2019. Erlangen, Germany.

Materials for the courses are provided online (https://zipfr.r-forge.r-project.org/lrec2018.html (https://zipfr.r-forge.r-project.org/lrec2018.html)) for self-study and can be reused under a Creative Commons license. Further insights from this work were published in [Diwersy et al. 2019].

The simulation experiments showed that the statistical models for relatively small samples (e.g. short stories or individual novel chapters) have a high level of uncertainty and are not suitable as a basis for robust complexity measures. Instead, an empirical approach was used that represents a combination of bootstrapping and cross-validation techniques (Evert et al. 2017). Complexity measures are calculated for non-overlapping windows of a fixed size and then averaged over the whole text, thereby (i) removing the problematic dependence of many measures on the text size and (ii) confidence intervals of all complexity measures for individual texts can be determined.

For the experiments carried out in work package 3, the corpus of contemporary novels from high and schematic literature described in TA 1.2.6 was used (hereinafter referred to as "novel corpus"). Since it cannot be taken for granted that high literature actually has a greater lexical complexity than schema literature, a corpus of articles from the journals GEO and GEOlino (Magazine for Children), which originates from readability research, was also used to validate the complexity measures (Weiß & Meurers 2018) and within the framework of a cooperation with Prof. Meurers (Tübingen).

Existing measures of lexical complexity based on type token distributions were collected, following the approach of Evert et al. (2017) modified, implemented in Python and empirically tested for robustness. Essential aspects were the dependency on the selected window size ("Do the complexity values remain similar across different window sizes?"; analogous to the text length dependency of classic measures), the ranking stability ("Does the ranking of the individual texts remain the same with regard to their complexity across different window sizes stable across the board?”) and the influence of the computational linguistic preprocessing (e.g. “How does a lemmatization affect the complexity values?”). It turned out, among other things, that it hardly makes a difference whether the measures are calculated using the original word forms or using lemmas, but that the accuracy of the tokenization plays a role (improved accuracy was achieved with the specially developed tokenizer SoMeWeTa, cf. Proisl 2018). It has also proven useful to exclude punctuation marks, since they can systematically distort comparisons between text types or genres in some dimensions.

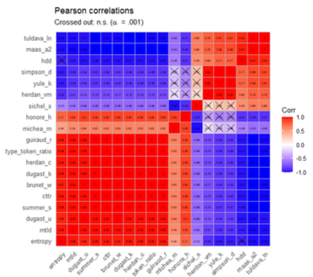


(/kallimachos/index.php/Datei:Abbildung_11_

Figure 11: Correlation matrix of the lexical complexity measures for the texts in the novel corpus, calculated for a window size of 7500 tokens

Finally, correlation analyzes were performed to determine similarities between the complexity measures (“Are there groups of measures that correlate very strongly and thus capture the same aspect of lexical complexity?”). As a result, the more than 20 measures proposed in the literature could be subdivided into a few groups, with measures within a group being very similar or even almost identical (see Fig. 11). However, the grouping depends on the window size: with small windows, the dimensions behave more and more similarly and groups begin to merge (Fig. 12). A summary publication of these findings is currently being prepared (cf. TA 1.3.2).

LNRE models were primarily researched and further developed at FAU. The compilation of the complexity measures, their empirical evaluation and correlation analyzes were carried out jointly by both working groups.



(/kallimachos/index.php/Datei:Abbildung_12_

Figure 12: Correlation matrix of the lexical complexity measures for the texts in the novel corpus, calculated for a window size of 250 tokens.

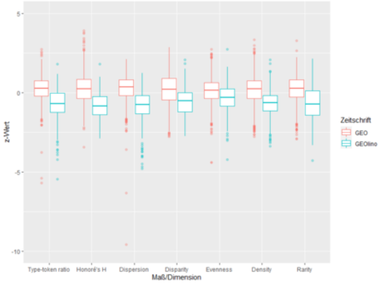
TA 1.3.2: Development of new complexity measures (in connection with sub-project FAU 01UG1715B)

In subtask 1.3.2, quantitative complexity measures were considered that describe the lexical level more comprehensively than pure type token measures or that go beyond this level. The focus of our investigations was complexity as a linguistic and literary construct or as a text-immanent phenomenon, in contrast to complexity concepts from readability/comprehensibility research and psychology. In order to empirically test the validity of the complexity measures examined, texts with the expected high complexity (high literature, journal articles for adults) were compared with texts with the expected lower complexity (schematic literature, journal articles for younger people) using the novel corpus and the GEO/GEOlino corpus reader) compared.

The aim of this study was to achieve a higher construct validity for the concept of "lexical complexity" as such by operationalizing various aspects of the intuitive concept of complexity as precisely as possible on an intermediate level. Mathematical complexity measures can then be related to this intermediate level and thus receive a clear interpretation.

A first comparison of works of high literature with those of the schema literature revealed less pronounced differences in complexity than expected for many measures - individual measures do not seem to fully reflect the complexity expected for high literature. At the same time, for all the measures examined, the variance between different texts in high literature is significantly higher than in schema literature, which is much more homogeneous in all genres. As the window size increases, the expected differences in complexity between high and schema literature become more apparent, while differences between genres of schema literature remain unchanged. An obvious - although yet to be tested - explanation would be that that repetitions within smaller sections of text in the schema literature are deliberately avoided (as recommended by common style guides and probably required by the editors). This is not necessarily the case in high literature: Repetitions may even be used as a stylistic device, and the assumed greater vocabulary of the authors only becomes apparent when examining larger sections. If so, the lexical measures at very small and very large window sizes measure different things, sort of micro- and macro-complexity. Repetitions may even be used as a stylistic device, and the assumed greater vocabulary of the authors only becomes apparent when examining larger sections. If so, the lexical measures at very small and very large window sizes measure different things, sort of micro- and macro-complexity. Repetitions may even be used as a stylistic device, and the assumed greater vocabulary of the authors only becomes apparent when examining larger sections. If so, the lexical measures measure different things, somewhat micro- and macro-complexity, at very small and very large window sizes.

Overall, it turns out that lexical-stylistic complexity is a multidimensional phenomenon that cannot be measured individually. Consequently, it is not sufficient to develop statistically sound and validated measures that are insensitive to artefacts (such as non-randomness). The combination of several measures in the form of a principal component analysis of the correlation pattern showed promising results, but at the same time it became obvious that the widespread complexity measures do not cover all potentially relevant dimensions of complexity. Based on Jarvis (2013a, 2013b), suitable measures for further dimensions (*dispersion, semantic disparity, evenness, lexical density* and *rarity*) searched and validated using the GEO/GEOlino corpus.

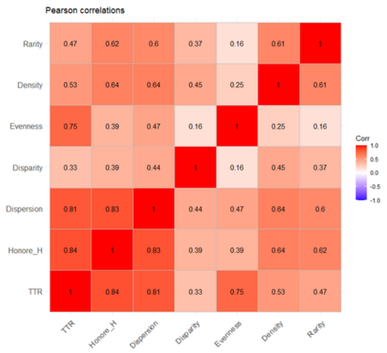


(/kallimachos/index.php/Datei:Abbildung_13_

Figure 13: Complexity measures for the GEO/GEOlino corpus (standardized).

Fig. 13 shows the lexical measures Type-Token-Ratio and *Honorés* H as well as selected reference measures for the additional dimensions in comparison; all values were z-standardized for this. The correlation matrix in Fig. 14 indicates that these are actually additional dimensions that measure different aspects of text complexity.

In order to record the additional dimensions of complexity, measures for the recognition of repetitions and formulaicity in texts were examined. The focus was on classical and neural language models, as well as compression algorithms. The calculation of perplexity values with the help of classic n-gram models has not turned out to be expedient. A particular problem is the pronounced sensitivity of the language models for typographical differences between texts and for words not contained in the language model. In addition, n-gram models require complex software libraries and require large amounts of memory. They are therefore not very suitable for integration into the toolbox intended for end users (TA 1.3.3). Neural language models (eg. BERT) offer a more accurate prediction of the text in comparison, are easier to use afterwards - although training the models is much more complex - and can better deal with words that are not in the vocabulary. Common compression algorithms were partially able to reproduce the expected differences in complexity between high-level and schematic literature and between journal articles for adult and younger readers (e.g. with gzip). Due to the complexity of such algorithms, however, it cannot be clearly clarified which complexity dimension is covered by them. Common compression algorithms were partially able to reproduce the expected differences in complexity between high-level and schematic literature and between journal articles for adult and younger readers (e.g. with gzip). Due to the complexity of such algorithms, however, it cannot be clearly clarified which complexity dimension is covered by them. Common compression algorithms were partially able to reproduce the expected differences in complexity between high-level and schematic literature and between journal articles for adult and younger readers (e.g. with gzip). Due to the complexity of such algorithms, however, it cannot be clearly clarified which complexity dimension is covered by them.



	TTR	Honore_H	Dispersion	Disparity	Evenness	Density	Rarity
Rarity	0.47	0.62	0.6	0.37	0.16	0.61	1
Density	0.53	0.64	0.64	0.45	0.25	1	0.61
Evenness	0.75	0.39	0.47	0.16	1	0.25	0.16
Disparity	0.33	0.39	0.44	1	0.16	0.45	0.37
Dispersion	0.81	0.83	1	0.44	0.47	0.64	0.6
Honore_H	0.84	1	0.83	0.39	0.39	0.64	0.62
TTR	1	0.84	0.81	0.33	0.75	0.53	0.47

(/kallimachos/index.php/Datei:Abbildung_14_

Figure 14: Correlation matrix of the complexity measures for the GEO/GEOlino corpus.

In addition to lexical dimensions, complexity measures for the syntactic structure of texts were examined, which are made explicit either via a dependency or a constituent structure analysis. Numerous syntactic complexity measures have already been extensively studied and applied (e.g. Pakhomov et al. 2011). However, in the collection and implementation of such complexity measures, research gaps have emerged, especially with regard to connections between different syntactic measures and between syntactic and lexical measures, which were not yet apparent at the time the application was submitted. Our experiments on this showed that syntactic complexity measures can also be divided into groups of measures that are strongly correlated with one another. At least in part, as hoped, they provide information

The integration of the dimension semantic disparity follows the intuition that redundancy in texts cannot be fully captured on the lexical level. Semantic disparity therefore also includes the semantics of the text. In order to determine the similarity of two segments, based on Cha et al. (2017) calculated and averaged the distance between all words of the first segment and all words of the second segment in a word-embedding space. The measure takes into account that although both segments contain different words, they can express similar content. The measure seems particularly interesting for further studies, since semantic distance, analogous to complexity, has to be seen as a multidimensional phenomenon.

The results of subtasks TA 1.3.1 and 1.3.2 will be published in the form of a best practice reference publication that is currently being prepared and will be submitted to Digital *Scholarship in the Humanities* .

The work on this sub-task was mainly carried out jointly by both working groups. New measures based on language models and semantic disparity were primarily developed at the JMU, principal component analyzes were primarily carried out at the FAU.

TA 1.3.3: Development of a common toolbox (in connection with sub-project FAU 01UG1715B)

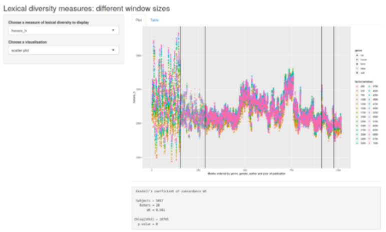
The measures and methods examined or developed in subtasks 1.3.1 and 1.3.2 were implemented in a freely available, open-source Python toolbox and made available via a GitHub repository (https://github.com/tsproisl/Linguistic_and_Stylistic_Complexity (https://github.com/tsproisl/Linguistic_and_Stylistic_Complexity)). As far as it makes sense and is possible, all dimensions are calculated using the window approach described in TA 1.3.1 (with a freely selectable window size), which means that standard deviations and confidence intervals can also be output. In addition to the implementation of the metrics, selected studies of the metrics are freely available within the toolbox. The toolbox has a modular structure and can therefore easily be expanded to include additional dimensions. It can be easily called from the command line, but can also be integrated into other program packages via a Python API. The measures implemented in the toolbox can be used, for example, for clustering tasks that build on preliminary work from the first project phase (authorship attribution, high literature vs. schema literature, literary vs. non-literary texts, etc.).

In a web app (https://kallimachos.shinyapps.io/lexical_diversity_measures (https://kallimachos.shinyapps.io/lexical_diversity_measures)), which was developed using the R framework Shiny, the investigated measures of lexical diversity can be compared interactively (see Fig. 15). For example, correlation plots or box plots can be viewed depending on genre and window size. The Shiny app is constantly being further developed and supplemented with new displays and data.

The software packages were developed jointly by both working groups under the leadership of FAU.

Use Case 3: Complexity of literary works from a stylometric point of view (in connection with sub-project FAU 01UG1715B)

Literary studies usually assume that literary texts have different degrees of complexity and that the complexity of the language used - ie the richness or variety of the vocabulary - is one of many determining factors. The investigations from TA 1.3.2 have shown that this factor is to be considered as a multidimensional phenomenon. The selected and newly developed quantitative measures for the various complexity dimensions could be used to describe texts from the high and schematic literature (TA 1.2.6). A first finding (Fig. 16) shows several abnormalities. First, high literature is, contrary to what might be expected, not more complex in every dimension than schema literature. Secondly, it is noticeable that high literature scatters more widely within the aspects than schema literature, i.e. has a larger internal variance. When adding the values for individual aspects to a combined measure for the overall complexity (Fig. 17), science fiction novels turn out to be more complex in absolute terms. This effect is due to the different characteristics of the dimensions for both text groups, which opens up room for interpretation. The aspect here is comparatively clear*Rarity* , because science fiction novels usually put a lot of effort into constructing fictitious worlds including exclusive vocabulary, while high literature novels are more caught up in everyday scenarios. Furthermore, *density* is not only more pronounced in science fiction, but also in horror and crime novels than in high literature, which could be explained by a higher, action-driven narrative pace. The effect is less pronounced for *evenness* , but suggests more structured storytelling, such as the separation between plot and description of the fictional world in science fiction novels.


Lexical diversity measures: different window sizes

Choose a measure of lexical diversity to display: honore_h

Choose a visualization: scatter plot

Plot
Table
Books ordered by genre, gender, author and year of publication

```
Kendall's coefficient of concordance Wt

Subjects = 1017
  Raters = 28
      Wt = 0.941

Chisq(1016) = 26765
  p-value = 0
```

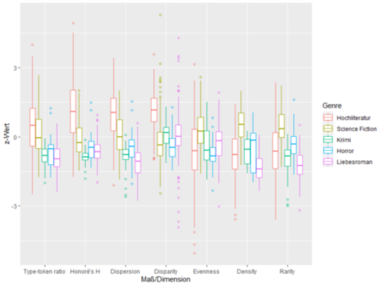
(/kallimachos/index.php/Datei:Abbildung_15_App_Komplexit%C3%A4tsma%C3%9Fe.png)

Figure 15: Shiny app for interactive visualization of complexity measures.

In order to check whether the high measured complexity of science fiction was only due to the text selection, a follow-up study was carried out with English-language e-books and large amounts of English-language fan fiction (Blombach, A. & Proisl, T. (2020): Unexpected Complexity and Romance in Disguise: The Case of Science Fiction Novels and Fanfiction Lecture at the 9th Hildesheim-Göttingen Workshop on DH and CL Blombach, A., Proisl, T., Evert, S., Heinrich, P. , & Dykes, N. (accepted): Into the Perryverse: A CL Journey to the Realm of Lexical Complexity. Presented at ICAME workshop To boldly go: Corpus approaches to the language of Science Fiction.)

Here, too, there was a significantly higher level of complexity compared to other genres.

The findings on complexity differences in high literature and science fiction point to characteristic, interpretable complexity profiles of literary text groups. In another study, using the DNB inventory (see TA 1.2.7), a similar profile was actually measured for subgenres of romance novels (doctor, homeland, nobility and family novels, see Fig. 18 in comparison to science fiction). will. This finding opens up follow-up questions about historical continuity, stability in genres and genres, and the general development of complexity in literary history.

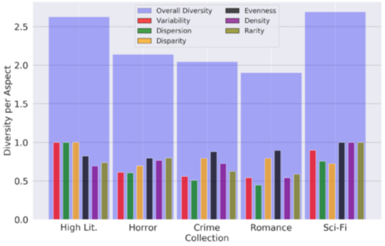


(/kallimachos/index.php/Datei:Abbildung_16_Genres.png)

Figure 16: Complexity measures for high literature and nonfiction genres (standardized).

The publication of the results is intended in the immediate future within a journal in the DH area. In addition, a project-related exchange of people (DAAD funding line (https://www.daad.de/de/infos-services-fuer-hochschulen/weiterfuehrende-infos-zu-daad-foerderprogrammen/ppp/)) was carried out with the Chair of Language Informatics at the University of Osaka during the project period. The aim of this cooperation is the evaluation and application of complexity measures in different languages.

All work on Use Case 3 was carried out in close cooperation between both working groups. The cooperation project with the University of Osaka is located at the JMU.



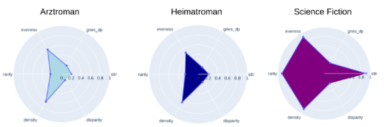
(/kallimachos/index.php/Datei:Abbildung_17_Genres.png)

Figure 17: Combined complexity of high fiction and nonfiction genres from 6 dimensions; the dimension variability is represented here by the type-token ratio.

WP4: User interface for quantitative analysis of Arabic-Latin translations

TA 1.4.1: Extension of correction tools for orthographic normalization

For the previous structure of the text corpus, a combination of rule-based replacements and manual corrections was used to standardize the spelling. With the constant growth of the text corpus, the manual part must be reduced on the one hand, and on the other hand the application of the rules to new vocabulary must not lead to incorrect substitutions. For this purpose, methods of deep learning (cf. e.g. [Kestemont et al. 2017]) are currently being discussed and, as part of our project, their suitability for normalizing the Arabic-Latin text corpus is being examined. Because orthographic normalization, along with OCR error correction and abbreviation resolution, is an important but later-in-workflow part of the “Use Case 4:



(/kallimachos/index.php/Datei:Abbildung_18_

Figure 18: Complexity profiles of different genres of novels.

TA 1.4.2: Use of Semantic MediaWiki as a digital research environment

A comprehensive data model for a Semantic MediaWiki could be developed on the basis of the experiences gained during the Narragonia project. In addition to entering and managing text metadata, the aim is to record the state of research on Arabic-Latin translations. This includes, on the one hand, a list of available text sources in manuscripts and editions, and, on the other hand, research into previous attributions of authorship and translation of the texts. A platform was created in the wiki for this purpose, which in particular facilitates the cooperation of several editors. The data previously available as XML documents or tables could be imported with the help of self-developed conversion scripts. The integration of the existing work environment, which bundles various functions for text analysis and text comparison is also being pursued. Since full integration would require the tools to be ported from Python to PHP and thus involve a considerable amount of development work, the best strategy has proven to be to connect the application underlying the existing web interface to the existing MediaWiki API interfaces in order to create a To enable data exchange between the two platforms and a flexible distribution of the individual user interfaces.

In day-to-day work, the wiki has proven itself as a platform for building a database of text sources. It now documents more than 400 texts by over 100 authors and translators as well as numerous sources in manuscripts, prints and editions. The structured metadata collection compiled in this way serves as the basis for planning the digitization of the Arabic-Latin translations as part of the [http://arabic-latin-corpus.philosophie.uni-wuerzburg.de/ (http://arabic-latin-corpus.philosophie.uni-wuerzburg.de/) Arabic-Latin Corpus Project] and for the necessary classification of the texts in their corpus-linguistic development.

Use case 4: Identification of translators

Previous work with Delta has shown that valid statements using stylistic methods are only possible with a sufficiently large body of text. The aim of the Arabic-Latin Corpus Project initiated with the Callimachus infrastructure is therefore to collect all Arabic-Latin translations of the 10th-14th centuries. century to digitize. The expansion of our text collection begins with the collection of the existing sources in the wiki, but in the second step above all requires a digitization workflow. The first pilot project was an early printing of the Latin version of Alhazens *Optik* (http://www.mdz-nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:bvb:12-bsb10197486-8) recorded with LAREX and Ocropus. In order to access not only Latin but also Arabic text sources, the OCR solutions already developed and tested in the other working groups were adapted accordingly. The advantages of the Kraken system based on Ocropus ([Romanov et al. 2017]) with regard to the recognition of bidirectional texts as well as the newly developed tool Calamari used in OCR4All and the performance of semi-automatic layout analysis are currently being used connected in LAREX. The feasibility of this approach could first be tested using a modern edition of the Arabic version of the optics.

The development of a web-based transcription tool "nashi" with OCR support was started in order to make the possibilities of OCR available to several users in parallel during operation. The application already on [2] (https://andbue.github.io/nashiGithub)is freely available, offers an interface for the transcription, correction and commenting of scanned texts based on the PAGE format, whereby a separate interface integrating the text into the image display was created for an ergonomically sensible and efficient workflow. The OCR text can be generated on a separate system and continuously updated in the background as the transcription progresses. In the last two years, around 300,000 lines of text from Arabic and Latin editions as well as early Latin prints have been segmented and transcribed. The texts are successively converted to TEI-XML, maintained in a Git repository, provided with metadata from the wiki (TA 4.2) and published on a dedicated website.

The digitization of the Arabic originals now opens up a new perspective on the different translator styles. A quantitative analysis requires an aligned version of the original and the translation. As a first step in this direction, a bilingual parallel view of selected texts implemented in Javascript was created, which relates both texts on the basis of existing glossary data. This tool is already being used productively for work on the Arabic-Latin Glossary , and the publication of bilingual digital texts is being planned. (http://www.arabic-latin-glossary.philosophie.uni-wuerzburg.de/)In order to also achieve alignment for texts for which no glossaries exist, initial experiments were carried out using tools from the field of statistical machine translation.

The ongoing work on the digitization and digital analysis of the Arabic-Latin translations was presented at lectures in Cordoba, Boston, London, Hamburg and Vienna:

- Hasse, DN: Boston, Tufts University, Classics Department: 'Identifying anonymous translators from Arabic into Latin: solving problems of philology and computational stylometry'.
- Hasse, DN: London, Institute of Historical Research: 'Using Digital Technologies for the Study of Medieval Arabic-Latin Translators'.
- Hasse, DN and Büttner, A.: Hamburg, Third PESHAT International Conference: 'Arabic and Latin Glossary and Arabic and Latin Corpus'.
- Hasse, DN and Büttner, A.: Vienna, Austrian Academy of Sciences: 'Creating ALGloss and ALCorpus: a Digital Lexicon and a Digital Corpus of Arabic-Latin Translations'.

The study of the translations of philosophical texts [Hasse & Büttner 2018] has now been published, as well as an article identifying the translator of some astronomical-astrological treatises [Hasse 2016]. Another article on texts that have been translated several times is already in print [Hasse, in press].

publications

Software directory' with download links

- OCR4all: OCR workflow software including comprehensive instructions, including installation via DockerHub or VirtualBox
  - https://github.com/OCR4all (https://github.com/OCR4all)
- Calamari: Automatic text recognition in the OCR workflow
  - https://github.com/Calamari-OCR (https://github.com/Calamari-OCR)
  - LAREX (Layout Analysis and Region Extraction): Segmentation in the OCR workflow
- https://github.com/OCR4all/LAREX (https://github.com/OCR4all/LAREX)

- VarianceViewer: Typed analysis of variance of similar texts:
  - https://github.com/cs6-uniwue/Variance-Viewer (https://github.com/cs6-uniwue/Variance-Viewer)
- ATHENS: UIMA-based information extraction pipeline
  - https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/Athen (https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/Athen)
- WebATHEN: UIMA-based web annotation tool for texts
  - http://webathen.informatik.uni-wuerzburg.de/ (http://webathen.informatik.uni-wuerzburg.de/)
- Nashi: transcription interface
  - https://github.com/andbue/nashi (https://github.com/andbue/nashi)
- Stylometric Toolbox
  - https://github.com/tsproisl/Linguistic_and_Stylistic_Complexity (https://github.com/tsproisl/Linguistic_and_Stylistic_Complexity)
- Statistical models for type token distributions
  - https://zipfr.r-forge.r-project.org/ (https://zipfr.r-forge.r-project.org/)

bibliography

- [Balbach et al. 2020] Balbach, N., Reul, C., Puppe, F. (2020): Typed variance analysis of texts. In: DHd 2020 Scope: Digital Humanities between modeling and interpretation, 235-238.
- [Bald et al. 2019] Bald, M., Damiani, V., Essler, H., Eyeselein, B., Reul, C., Puppe, F. (2019]: Correction of erroneous OCR results by automatic alignment with texts of a corpus, DHd (Digital Humanities, 6th Annual Meeting), 309-311.
- [Burrichter 2017]: Burrichter, B. (2017): Framework and intended audience. The paratexts in Sebastian Brant's 'Ship of Fools' and its translations. In: Framing. Forms of presentation and canon effects. Ed. Ph. Ajouri, U. Kundert and C. Rohde. Berlin, 107-122.
- [Burrichter 2019a]: Burrichter, B. (2019): Sebastian Brant's 'Ship of Fools' and its French translations. In: Etudes Germaniques 3, 505-521.
- [Burrichter 2019b]: Burrichter, B. (2019): Sebastian Brant's 'Ship of Fools' and its European reception in the 15th century. Presentation of the digital edition of important editions (German, Latin, French and English) and first results of a comparison. In: B. Bastert, S, Hartmann: Romania and Germania. Cultural and literary exchange processes in the late Middle Ages and early modern period, Wiesbaden, 311-323.
- [Burrichter 2019c] Burrichter, B. (2019): Patrice et les Dernydes. Les versions françaises de la Nef des fous de Sebastian Brant. In: A. Kukulka-Wojtasik (ed.): "Translatio" et Histoire des Idées: "Translatio" and the History of Ideas, Frankfurt aM.
- [Diwersy et al. 2019] Diwersy, S., Evert, S., Heinrich, P., & Proisl, T. (2019): Means of Productivity - on the Statistical Modeling of the Restrictedness of Lexico-Grammatical Patterns. In: EUROPHRAS 2019. Productive Patterns in Phraseology, pages 20–21. Santiago de Compostela, Spain.
- [Evert et al. 2017] Evert, S., Wankerl, S., & Nöth, E. (2017): Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch. In: Proceedings of the Corpus Linguistics 2017 Conference. Birmingham, UK.
- [Fischer 2017] Fischer, E. (2017). Automatic extraction of interactions between two people in literary texts. Master's thesis, University of Würzburg.
- [Grundig / Hamm / Walter 2017] Grundig, C., Hamm, J., Walter, V. (2017): Narragonia digital. With an analysis of Chapter 4 of the 'Ship of Fools' in 15th Century Editions and Arrangements. In: Wolfenbüttel Notes on Book History 42, 97-120.
- [Grundig 2017] Grundig, C. (2017): Theological transformation of the 'Ship of Fools'. Geiler von Kaysersberg and the so-called 'interpolated version'. In: Archive for the study of modern languages and literatures 254, 1-16.
- [Hamm 2016] Hamm, J. (2016): Intermedial variance. Sebastian Brant's 'Ship of Fools' in German editions of the 15th century. In: History of transmission transdisciplinary. New Perspectives on a German Studies Research Paradigm. In connection with H. Brunner and F. Löser ed. v. D. Small. Wiesbaden, 223-240.
- [Hamm 2017] Hamm, J. (2017): On paratextuality and intermediality in Sebastian Brant's Vergilius pictus (Strasbourg 1502). In: Intermediality in the early modern period. Forms, functions, concepts. Ed. J Robert. Berlin, Boston, 236-259.
- [Hamm 2019] Hamm, J. (2019): Fools with Interpretation. On the ›World Mirror or Fools Ship‹ (Basel 1574) by Nikolaus Höniger von Königshofen. In: Traditional and Innovative in the Spiritual Literature of the Middle Ages. Edited by J. Haustein et al., Stuttgart, 407-426.
- [Hamm et al. 2019]: T. Baier, J. Hamm, U. Schlegelmilch (eds.; 2019): Opera Camerarii. A semantic database for the printed works of Joachim Camerarius d.Ä. (1500 - 1574). edit of M. Gindhart, M. Huth and J. Schultheiss, Würzburg, http://wiki.camerarius.de (http://wiki.camerarius.de) .
- [Hamm i.Dr.]: Hamm, J. (i.Dr.): Auctor and interpres in dialogue. Sebastian Brant's contributions to the 'Stultifera navis' (1497). In: The 15th Century. International Symposium at the Melanchthon Academy in Bretten, 12.-14. October 2017. (Manuscript accepted for printing at http://www.camerarius.de/wp-content/uploads/2020/03/Hamm_Narrenschiff_im_Druck.pdf (http://www.camerarius.de/wp-content/uploads/2020/03/Hamm_Narrenschiff_im_Druck.pdf)
- [Hasse & Büttner 2018] Hasse, D. and Büttner, A.: Notes on Anonymous Twelfth-Century Translations of Philosophical Texts from Arabic into Latin on the Iberian Peninsula, in: DN Hasse and A. Bertolacci, eds., The Arabic, Hebrew and Latin Reception of Avicenna's Physics and Cosmology (Berlin / Boston: de Gruyter, 2018), pp. 313-369.
- [Hasse 2016] Hasse, D.: 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 C. Burnett, P. Mantas-Espana, eds. , Ex Oriente Lux. Translating Words, Scripts and Styles in Medieval Mediterranean Sociecty (Córdoba / London: UCOPress, CNERU / The Warburg Institute: 2016), 19-43.
- [Hasse in print] Hasse, D.: Three Double Translations from Arabic into Latin by Gerard of Cremona and Dominicus Gundisalvi, in: Dragos Calma (ed.), Reading Proclus and the Book of Causes, Volume 2 (Leiden: Brill, im Print).
- [Jannidis et al. 2018] Jannidis F, Konle L, Zehe A, Hotho A, Krug M (2018). Analyzing Direct Speech in German Novels. In DHd 2018.
- [Jannidis et al. 2019a] Jannidis, F., Konle, L., Leinen, P. (2019). A distant view of 10,000 magazine novels. In DHd 2019.
- [Jannidis et al 2019b] Jannidis, F., Konle, L., Leinen, P. (2019). Thematic Complexity. In DH 2019.
- [Jannidis et al 2020] Jannidis F, Konle L, Leinen P (2019). Confounding variables in sub-genre classification: intrusive problems. In DHd 2020.
- [Krug et al. 2017] Krug, M., Reger, I., Jannidis, F., Weimer, L., Madarász, N., and Puppe, F. (2017). Overcoming Data Sparsity for Relation Detection in German Novels.
- [Krug et al. 2020a] Krug, M., Schmidt, D., Wehner, N., Jannidis, F., Puppe, F. (2020). Evaluation of state of the art methods for coreference resolution and quotation attribution on German literary novels, Under Review for Journal of Natural Language Engineering.
- [Krug et al. 2020b] Krug, M., Schmidt, D., Jannidis, F., Puppe, F. (2020). Techniques for High Quality Character Reference Detection on German Historical Novels. Under Review for De Gruyter Open Linguistics.
- [Gius et al. 2019] Gius E, Jannidis F, Krug M, Zehe A, Hotho A, Puppe F, Krebs J, Reiter N, Wiedmer N & Konle L (2019). Detection of Scenes in Fiction. Proceedings of Digital Humanities 2019.

- [Hartl 2001] Hartl, N. (2001): The »Stultifera navis«. Jakob Locher's translation of Sebastian Brant's "Ship of Fools", 2 vols., Münster 2001 (Studies and texts on the Middle Ages and early modern times 1).
- [Herud et al. 2020] Herud, K., Zehe, A., Krug, M., Puppe, F. & Hotho, A. (*2020). SceneIt - End-to-End Neural Scene Detection in Fictional Texts, Under Review for ACL 2020.
- [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), pages 665–70. Miyazaki, Japan.
- [Proisl et al. 2019] Proisl, T., Konle, L., Evert, S., Jannidis, F. (2019): Dependency-based syntactical complexity measures. In: DHd 2019 Conference Abstracts. Frankfurt am Main, Germany.
- [Reul et al. 2017] Reul, C., Springmann, U., Puppe, F. (2017). 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, 137-142, New York, NY, USA. ACM.
- [Reul et al. 2018a] Reul, C., Springmann, U., Wick, C., Puppe, F. (2018). Improving OCR Accuracy on Early Printed Books by Utilizing Cross Fold Training and Voting. In: 13th IAPR International Workshop on Document Analysis Systems (DAS), 423-428.
- [Reul et al. 2018b] Reul, C., Wick, C., Springmann, U., Puppe, F. (2018). Transfer Learning for OCRopus Model Training on Early Printed Books. In: 027.7 Journal for Library Culture 5,1, 38-51.
- [Reul et al. 2018c] Reul C, Springmann U, Wick C, Puppe F: Improving OCR Accuracy on Early Printed Books by combining Pretraining, Voting, and Active Learning. In: JLCL (Special Issue on Automatic Text and Layout Recognition) 33,1, 3-24.
- [Reul et al. 2019a]: Reul, C., Springmann, U., Wick, C., Puppe, F. (2019). State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines. In: DHd 2019 Digital Humanities: multimedial & multimodal, 212-215.
- [Reul et al. 2019b] Reul, C., Göttel, S., Springmann, U., Wick, C., Würzner, KM., Puppe, F. (2019). Automatic Semantic Text Tagging on Historical Lexica by Combining OCR and Typography Classification. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, 33-38.
- [Reul et al. 2019c] Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., Puppe, F (2019). OCR4all - An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. In: Applied Sciences. 9 (22) 4853. https://doi.org/10.3390/app9224853 (https://doi.org/10.3390/app9224853)
- [Wehner 2019] Wehner, M. (2019). Text recognition software for historical prints. In: KulturBetrieb 25 (2019), 42-43.
- [Wehner et al, 2020] Wehner M, Dahnke M, Landes F, Nasarek R, Reul C (2020). OCR4all - A semi-automated open source software for the OCR of historical prints. In: DHd 2020 Scope: Digital Humanities between modeling and interpretation. Conference Abstracts, 43-45. http://doi.org/10.5281/zenodo.3666690 (http://doi.org/10.5281/zenodo.3666690)
- [Weiß & Meurers 2018] Weiß, Z. & Meurers, D. (2018): Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In: Proceedings of the 27th International Conference on Computational Linguistics, pages 303-317, Santa Fe, New Mexico, USA.
- [Wick et al. 2018] Wick C, Reul C, Puppe F (2019). Comparison of OCR Accuracy on Early Printed Books using the Open Source Engines Calamari and OCRopus. JLCL (Special Issue on Automatic Text and Layout Recognition) 33,1, 79-96.
- [Wick et al. 2020] Wick C, Reul C, Puppe F, (2020). Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. In: Digital Humanities Quarterly (finally accepted for publication).
- [Zeh et al. 2020] Zehe, A., Arns, J., Hettinger, L. & Hotho, A. (*2020). HarryMotions - Classifying Relationships in Harry Potter based on Emotion Analysis. Under Review for KONVENS 2020

Literature cited by third-party authors (not in the project)

- [Boenig et al. 2020] Boenig M, Engl E, Baierer K, Hartmann V, Neudecker C (2020). Full text transformation of early modern prints - results and perspectives of the OCR-D project. In: DHd 2020 Scope: Digital Humanities between modeling and interpretation. Conference Abstracts, 244-247.
- [Cha et al. 2017] Cha M, Gwon Y & Kung HT (2017): Language modeling by clustering with word embeddings for text readability assessment. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.
- [Grundig 2012]: Grundig, C. (2012): Text and Paratext. Concepts of paratextuality in the German-language works of Sebastian Brant. Masch. Master's thesis. Wuerzburg 2012.
- [Hart 2001]: Hartl, N. (ed., 2001): The 'Stultifera navis'. Jakob Locher's translation of Sebastian Brant's 'Ship of Fools'. Vol. 1.1: Investigation and Commentary; Vol. 1.2: Partial edition and translation. Munster, New York, Munich.
- [Jarvis 2013a] Jarvis, S. (2013a): Capturing the Diversity in Lexical Diversity. In: Language Learning, 63 (1): 87-106.
- [Jarvis 2013b] Jarvis S (2013b): Defining and Measuring Lexical Diversity. In: Jarvis, S. & Daller, M. (eds.): Vocabulary Knowledge. Human Ratings and Automated Measures. Amsterdam: John Benjamins. (= Studies in Bilingualism 47)
- [Kestemont et al. 2017] Kestemont M, de Pauw G, van Nie R and Daelemans W (2017). Lemmatization for variation-rich languages using deep learning. Digital Scholarship in the Humanities, 32(4):797–815.
- [Kim et al. 2019] Kim E & Klinger R (2019). Frowning Frodo, Wincing Leia, and a Seriously Great Friendship: Learning to Classify Emotional Relationships of Fictional Characters, NAACL 2019
- [Knape 2005]: Knape, J. (ed., 2005): Sebastian Brant. The 'Ship of Fools'. study edition. With all 114 woodcuts from the print Basel 1494. Stuttgart 2005.
- [Neudecker et al. 2019] Neudecker, C., Baierer, K., Federbusch, M., Würzner, KM., Boenig, M., Herrmann, E., Hartmann, V. (2019). OCR-D: An end-to-end open-source OCR framework for historical documents. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, Brussels, pp. 53–58.
- [Pakhomov et al. 2011] Pakhomov S, Chacon D, Wicklund M, & Gundel J (2011): Computerized assessment of syntactic complexity in Alzheimer's disease: A case study of Iris Murdoch's writing. Behavior research methods, 43(1): 136-144.
- [Pletschacher and Antonacopoulos 2010] Pletschacher, S., Antonacopoulos, A. (2010). The PAGE (page analysis and ground-truth elements) format framework. In Pattern Recognition (ICPR), 2010 20th International Conference on, pp. 257-260. IEEE.
- [Romanov et al. 2017] Romanov M, Miller MT, Bowen S, Kiessling B (2017). Important New Developments in Arabographic Optical Character Recognition (OCR). CoRR, abs/1703.09550.
- [Springmann and Lüdeling, 2017] Springmann, U., 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).
- [Springman et al. 2018] Springmann, U.; Reul, Chr.; Dipper, St.; Baiter, J. (2018): Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin. Archive e-prints. https://arxiv.org/abs/1809.05501 (https://arxiv.org/abs/1809.05501)

- [Strötgen and Gertz, 2013] Strötgen, J., Gertz, M. (2013). Multilingual and cross-domain temporal tagging. Language Resources and Evaluation, 47(2):269-298.
- [Vobl et al. 2014] Vobl, T., Gotscharek, A., Reffle, U., Ringlstetter, C., and Schulz, K. (2014). PoCoTo - an Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts. In Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 57-61. DATECH '14. New York, NY, USA: ACM.

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung (https://www.bmbf.de/)

FAU (https://www.fau.de)

register (/kallimachos/index.php?title=Spezial:Anmelden&returnto=Kallimachos+II+%28Eingehende+Darstellung%29)