Michael Fröstl, Stefan Zathammer, Sarah Lang

Zur Transkription von Alchemica mithilfe der Transkribus-Software

Zu Handschriften, Drucken und dem NOSCEMUS GM 6 Modell

Alchemische Labore. Alchemical Laboratories, Sarah Lang (Hg.), unter Mitarbeit von Michael Fröstl & Patrick Fiska, Graz 2023, S. 363–378, DOI: https://doi.org/10.25364/978390337404119

Michael Fröstl, frostlmichael@gmail.com
Stefan Zathammer, stefan.zathammer@uibk.ac.at, ORCID: 0000-0001-9460-3299
Sarah Lang, sarah.lang@uni-graz.at, ORCID: 0000-0002-4618-9481

Zusammenfassung

Digitalisierungsinitiativen wie VD17 (Verzeichnis der im deutschen Sprachraum erschienenen Drucke des 17. Jahrhunderts) ist es zu verdanken, dass eine Vielzahl an Alchemica mittlerweile in Form digitaler Faksimiles verfügbar ist. Die Edition und (editions-)philologische Behandlung der Texte stellt jedoch weiterhin ein dringendes Desiderat der Alchemieforschung dar. Die Transkribus-Software bietet vielversprechende Möglichkeiten zur automatisierten Transkription auf Basis von Bilddaten. Forschende können auf bereits existente Modelle zurückgreifen, wie etwa das NOSCEMUS GM 6, das für lateinische Druckwerke des 15.–19. Jahrhunderts sehr gute Ergebnisse erzielt. Mit dessen Hilfe wurde im NOSCEMUS-Projekt ein Korpus von 82 lateinischen Texten in der Kategorie Alchemie/Chemie transkribiert. Sowohl das transkribierte Korpus als auch das Modell sind frei verfügbar und nachnutzbar.

Schlagwörter: Transkribus, NOSCEMUS, maschinelle Transkription, Editionsphilologie, gedruckte Alchemica

Abstract

Thanks to digitization initiatives like VD17 (Union Catalogue of Books Printed in German-Speaking Countries in the 17th Century) many alchemical texts have become available as digital facsimiles. Editing those texts is an important desideratum in the historiography of alchemy and chemistry. The Transkribus software offers promising features for the automated transcription of historical text based on image data. Researchers can freely access a number of pre-trained models, for instance the NOSCEMUS GM 6, a high-performing model for Latin print of the 15th to 19th centuries. In the Innsbruck NOSCEMUS project, about 80 alchemical texts have been machine-transcribed using the aforementioned model. They are available as open access publications and can be reused freely by the research community.

Keywords: Transkribus, NOSCEMUS, machine transcription, edition, alchemical printed works

Einleitung

Wer sich der Alchemie und Chymie mit textbasierten Methoden annähert, gleich, ob mit historischem oder philologischem Schwerpunkt, wird schnell feststellen, dass die Anzahl relevanter Quellen so groß ist, dass sie von einer Einzelperson kaum mehr überblickt, geschweige denn bearbeitet werden können. So wie bei Werken der frühen Neuzeit im Allgemeinen, so hat auch bei der Alchemie und der Chymie der Buchdruck zur massenhaften Produktion und zu relativ weiter Verbreitung einschlägiger Werke beigetragen. Dass die Digital Humanities hier Abhilfe zu schaffen versuchen, indem sie große Textmengen alchemischer Literatur nicht bloß als Bildfaksimiles digitalisieren, mitunter in TEI-XML transkribieren und bereitstellen, sondern auch bestrebt sind, sie mit digitalen Werkzeugen und Methoden zu analysieren, darf als bekannt vorausgesetzt werden.[1]

Besonders Digitalisierungsinitiativen wie Google Books, Early English Books Online (EEBO) oder dem Verzeichnis der im deutschen Sprachraum erschienenen Drucke des 16. und 17. Jahrhunderts (VD16/17) ist die mittlerweile sehr gute Abdeckung an frei verfügbaren Bildfaksimiles frühneuzeitlicher alchemischer Drucke zu verdanken. Das Vorhandensein einer solchen Datengrundlage erinnert an die Forderung Principes und Newmans, dem relativ schlechten Erschließungszustand der alchemischen Überlieferung zu begegnen.[2] Um diese Alchemica allerdings im Zuge von (digitalen) Editionen verfügbar zu machen, werden digitale Transkripte dieser Datengrundlage benötigt. Doch händische Transkription ist sehr zeitaufwändig. Eine Software zur (halb-)automatischen Anfertigung von Rohtranskriptionen ist *Transkribus*, nach eigener Definition eine „Plattform für die Digitalisierung, Texterkennung mithilfe künstlicher Intelligenz, Transkription und das Durchsuchen von historischen Dokumenten", entwickelt an und im Umfeld der Universität Innsbruck.[3]

In diesem Beitrag soll daher illustriert werden, welche Möglichkeiten die Software Transkribus zur Corpuserzeugung bietet. Zuerst wird auf das Arbeiten mit Handschriften eingegangen[4] und in weiterer Folge auf Drucke, für die das NOSCEMUS-Modell

1. Martinón-Torres 2011, 233

2. Principe und Newman betonen: „[…] rigorous historical attention to the issues of textual purity and authorial biography should be one important focus for alchemical studies over the next decades. Critical editions of important individual works are needed and more comprehensively, editions of the complete opera of important figures, containing careful discrimination between the strata of authentic, interpolated, and spurious works." In Principe und Newman 2001, 419.

3. Hervorgegangen ist Transkribus aus dem READ-Projekt (https://readcoop.eu/de/transkribus), das im Rahmen von Horizon 2020 von der EU gefördert wurde (https://ec.europa.eu/programmes/horizon2020/en); Zugriff: 09.05.2021.

4. Ein alchemisches Beispiel für eine mit Transkribus transkribierte Handschrift ist *Alchymistische Kunst-Stücke in gutter Ordnungk* (= ÖNB Cod. 11450, 1596): Camen 2018. Siehe dazu auch den Beitrag von Rudolf Werner Soukup in diesem Band.

out-of-the-box sehr gute Ergebnisse erzielt.[5] Ferner wird kurz auf das NOSCEMUS-Projekt selbst eingegangen, im Zuge dessen eine große Menge neulateinischer wissenschaftlicher Literatur mithilfe der Transkribus-Software automatisiert transkribiert wurde, darunter auch Texte zu Alchemie und Chymie, wie beispielsweise das *Lexicon Alchemiae Rulandi* (1612) oder Michael Maiers *Atalanta fugiens* (1617/18).[6]

Funktionalitäten der Transkribus-Software

Nach Anlegen eines Benutzerkontos und dem Download über genannte Homepage ist die graphische Benutzeroberfläche von Transkribus in der Vollversion auf dem lokalen Rechner verfügbar. Gleichzeitig besteht eine Verbindung zur Server-Infrastruktur, weswegen eine aktive Internetverbindung für das Arbeiten mit Transkribus unbedingt erforderlich ist. Die Vollversion – der sogenannte eXpert-Client – ermöglicht dem Prinzip nach alle komplexen Transkriptionsworkflows ohne Einschränkungen, einschließlich händischer Korrekturen neben bzw. nach den automatischen Transkriptionen. Eine etwas vereinfachte und web-basierte graphische Benutzeroberfläche für den Internet-Browser steht mit dem Web-Tool *Transkribus Lite* ebenfalls zur Verfügung[7]. In diesem Fall kann auf eine lokale Installation auf dem eigenen Rechner verzichtet werden. Der Vorgang der Registrierung und des Log-Ins bleibt dabei weiterhin erforderlich. Die Web-Version von Transkribus wurde seitens ihrer Entwickler:innen zunächst bewusst als reduzierte Variante verstanden, wobei Benutzer:innenfreundlichkeit im Vordergrund stehen sollten; man reagiert damit auf Wünsche der User:innen nach einfacherer Anwendung, wonach die Bedienung der derzeitigen Vollversion punktuell umständlich und zu wenig intuitiv sei. Langfristig soll allerdings durch die allmähliche Integrierung aller Funktionen des eXpert-Clients die Weboberfläche zur eigentlichen Arbeitsplattform für den „gewöhnlichen" Transkribusnutzer werden. Ein erster Schritt in diese Richtung war die Veröffentlichung der neuen, vollständig überarbeiteten und in ihrem Funktionsumfang stark ausgebauten Weboberfläche im Sommer 2023.[8]

Nach Anlegen eines Benutzerkontos gilt es in einem weiteren Schritt, hochauflösende digitale Fotos oder Faksimile-Scans von der Quelle, die man bearbeitet, für Transkribus verfügbar zu machen, also hochzuladen, wodurch sie am Server des Anbieters gespeichert werden. Die unterstützen Dateiformate sind PDF, JPEG, PNG

5. Ein englischsprachiges Tutorial zur Verwendung von Transkribus für Drucke unter Benutzung des NOSCEMUS-Modells anhand eines alchemischen Beispiels findet sich in Lang 2019.

6. Ruland 1612; Maier 1617/18; NOSCEMUS „Semantic Drilldown": https://wiki.uibk.ac.at/noscemus/Special:BrowseData/Works; sowie der Disziplin „Alchemie/Chemie" zugeordnete Werke: https://wiki.uibk.ac.at/noscemus/Special:BrowseData/Works?Discipline%2FContent=Alchemy%2FChemistry.

7. https://transkribus.eu/lite/.

8. Vgl. https://readcoop.eu/de/coming-soon-new-transkribus-web-app/.

und TIFF.[9] Im Internet verfügbare Digitalisate können gegebenenfalls auch über einen DFG-Viewer- oder METS-Link direkt hochgeladen werden. Bei großen Dateien bietet sich zudem der Upload über FTP an. Pro hochgeladener Quelle legt man eine Sammlung (*collection*) an, die um beliebig viele Quellen erweiterbar ist. Jede Sammlung besteht zumindest aus den hochgeladenen Bildern einer Quelle. Allenfalls kommen manuell oder semiautomatisch angefertigte Transkriptionen hinzu. Über die Zugänglichkeit jeder Sammlung kann individuell entschieden werden (sicht- und/oder auch bearbeitbar nur für jene Person selbst, die die Sammlung anlegt, oder auch für weitere, ausgewählte auf Transkribus registrierte Benutzer:innen). Der Grundeinstellung nach sind hochgeladene Digitalisate und Transkriptionen grundsätzlich privaten Charakters (nicht öffentlich einsehbar) und können zu Beginn nur von jenen Personen bearbeitet werden, unter deren Account die Sammlung angelegt wird.

Anwendung zur Schrifterkennung

Nach dem Upload des Bildmaterials gibt es zwei Möglichkeiten, eine maschinelle Transkription anfertigen zu lassen: Entweder man gewöhnt („trainiert“) Transkribus an das individuelle Schriftbild der Quelle, die man bearbeiten möchte, und fertigt so ein individuelles *Modell* auf Basis der Quelle selbst an oder man verwendet ein bereits vorhandenes, das dem der eigenen Quelle möglichst ähnelt.

Mittlerweile ist eine beachtliche Anzahl (132, Stand November 2023) an verschiedensten Modellen für eine Viezahl von Sprachen, Handschriften und Schrifttypen vom Mittelalter bis in die Gegenwart für alle Benutzer frei zugänglich. Die Transkribus-Website bietet eine gut aufbereitete Liste der öffentlichen Modelle, in welcher mittels einer Reihe von Filtern spezifisch nach für den zu transkribierenden Text möglicherweise in Frage kommenden Modellen gesucht werden kann.[10] alternativ können aber auch eigene Modelle mit anderen Nutzer:innen geteilt werden.

Das Training der Software an ein bisher nicht vorhandenes Schriftbild erfolgt mittels manueller Transkription in Transkribus. Seit der Version 1.15.1 ist sowohl die Trainingsfunktion für Schriftbilder (PyLaia HTR) wie auch die Option für das Training von Layout-Modellen (P2PaLA) für alle Benutzer:innen frei zugänglich. Das Schriftbild der im Vorhinein manuell transkribierten Original-Seiten fungiert im Trainingsprozess als *Modell*, um damit weitere semiautomatische Transkriptionen der Quelle mithilfe von Transkribus durchführen zu können. In der Regel sollte der Umfang

9. Vgl. https://readcoop.eu/de/transkribus/anleitungen/transkribus-in-10-schritten/ (Pkt. 5).
10. Eine gut aufbereitete Liste mit allen öffentlich verfügbaren Modellen bietet die Transkribus-Homepage unter: https://readcoop.eu/de/transkribus/oeffentliche-modelle/.

des bereits als Gold-Standard-Transkript vorliegenden Textes, auf dessen Basis das Modell trainiert werden soll, zwischen 25 und 75 Seiten betragen, entsprechend einer Wortanzahl zwischen 5.000 und 15.000. Für das Training eines Modells sollten die ausgewählten Original-Seiten zudem möglichst frei von sekundären Hinzufügungen, Hand- oder Schriftstilwechseln sein, da diese während des Trainings nicht berücksichtigt werden können und später wieder manuell gekennzeichnet bzw. hinzugefügt werden müssen. Während des Trainingsprozesses werden einige Originalseiten, zu denen manuell angefertigte Transkriptionen existieren, von der Maschine zur Seite gelegt, in der Regel ein bis zwei à 50 bis 100 Seiten der Quelle. Diese Seiten fließen selbst nicht in den Trainingsprozess mit ein, sondern dienen Transkribus als Unbekannte zu Testzwecken seiner selbst: Nach Abschluss des Trainingsprozesses versucht Transkribus, sie automatisch zu transkribieren und vergleicht das Resultat mit der manuell angefertigten, „korrekten" Transkription. Aus den Abweichungen zwischen automatischer und manueller Transkription ergibt sich die zeichenbezogene Fehlerquote (*character error rate* – CER) einschließlich falsch transkribierter Leerzeichen, fehlender oder falsch gesetzter Satzzeichen, zusätzlicher Zeichen und Fehler bei der Groß- und Kleinschreibung.

Die CER gilt als Richtwert für die Genauigkeit, die bei der automatischen Transkription weiterer Seiten zu erwarten ist. Allgemein gesprochen liegt die zeichenbezogene Treffergenauigkeit von Transkribus bei Handschriften momentan im Bereich von etwa 95 Prozent, je nach Umfang und Qualität des zur Verfügung gestellten Trainingsmaterials. In der Regel beträgt die CER bei einigermaßen gleichmäßigen Handschriften, die weitestgehend frei von Hinzufügungen späterer Hände oder ähnlicher Zusätze sind, demnach etwa fünf bis maximal zehn Prozent, das heißt: bei einem Aufkommen von zwanzig Zeichen sind davon ein bis zwei falsch. Damit kann aus handschriftlichen Quellen derzeit eine gut lesbare Rohtranskription generiert werden, ein editionstauglicher Text jedoch nur sehr bedingt. Bei Drucken stellt sich die Situation deutlich besser dar. Die zeichenbezogene Fehlerquote liegt bei großen Modellen (100.000 Wörter und mehr) regelmäßig bei 0,5 Prozent und weniger. Erneutes Training von Transkribus mit erweiterter Ausgangsdatenbasis und die Anwendung des verbesserten Modells auf die Quelle kann außerdem die Treffergenauigkeit verbessern. Vor der eigentlichen Texterkennung oder händischen Transkription muss die Quelle einer Layout-Analyse unterzogen werden. Diese kann wiederum automatisch oder – sehr arbeitsintensiv— manuell erfolgen. Bei der automatischen Layout-Analyse gilt es allerdings zu beachten, dass bei komplexen Dokumenten (z. B. Tabellen) die Layout-Erkennung noch recht fehleranfällig ist, was eine zeitintensive Nachbearbeitung bzw. Korrektur per Hand notwendig machen kann.[11]

11. Als Beta stehen in der neuen Weboberfläche zwei neue Funktionen – „Field Models" und „Table Models" – für das Training von Layout-Modellen zur Verfügung, die, genügend Trainingsdaten vor-

Aus einer Auswahl verfügbarer Schriftmodelle kann für die automatische Transkription einer beliebigen Anzahl an Seiten derzeit nur jeweils *ein* Modell pro Transkriptions*vorgang* (*job*) ausgewählt werden, der mehrere Seiten umfassen kann. Das bedeutet: Auch wenn auf einer Seite mehrere (und als solche definierte!) Textabschnitte (*textregions*) vorhanden sind, die im Schriftstil voneinander abweichen, kann die gesamte Seite stets nur mit *einem* Schriftmodell von der Maschine gelesen und transkribiert werden. Je nach Anzahl der Seiten, die Transkribus pro Vorgang bearbeitet, kann dessen Dauer variieren. Im Falle *einer* Seite pro Transkriptionsvorgang beträgt diese Dauer nur wenige Sekunden.

Arbeiten mit bestehenden Modellen

Vor der Initiierung eines einzelnen Transkriptionsvorgangs wählt man ein zur Verfügung stehendes Modell aus bzw. ist ein Modell bereits voreingestellt. Eine wachsende Anzahl öffentlich einsehbarer Schriftmodelle steht zur Verfügung, bei denen man ausprobieren kann, ob sie für die eigene Quelle geeignet sind. Auf ein Modell für die Transkription von frühneuzeitlichen Antiqua-Drucken, das für alle Nutzer:innen frei zugänglich ist, wird im Folgenden noch näher eingegangen.[12] Alle anderen Modelle muss man im Rahmen des jeweiligen User-Accounts selbst trainieren und anlegen. Die automatische Erkennung eines Textquantums von bis zu 500 Seiten ist dabei kostenlos, darüber hinaus ist der Erwerb von *Credits* erforderlich. Diese können auf Anfrage – *on demand* – oder als Abonnement bezogen werden. Quelle und Modell sind voneinander prinzipiell unabhängig. Das heißt: Man kann grundsätzlich jedes zur Verfügung stehende Modell auf jede Quelle anwenden, was jedoch nur Sinn macht, wenn zwischen der Schrift der Quelle und jener des Modells hinreichend Ähnlichkeit besteht. In Zukunft könnte es möglich sein, pro definiertem Textabschnitt jeweils eigene Modelle auszuwählen, sodass Textregionen einer Seite auch unabhängig voneinander, selektiv und ihrem Schriftbild entsprechend transkribiert werden können.

---

ausgesetzt, auch bei sehr komplexen Layouts und Tabellen sehr gute Ergebnisse lieferen. Vgl. https://readcoop.eu/de/introducing-field-models-trainable-layout-ai-in-transkribus/ und https://readcoop.eu/de/introducing-table-models-trainable-layout-ai-in-transkribus/.

12. Das Modell orientiert sich eher am Schriftbild der Drucke als an deren Sprache. Transkribus kennt zwar im Grunde keine Sprachen, doch lassen sich die Modelle mithilfe der Language Models, die entweder aus einem externen Wörterbuch gespeist oder auch mit den Transkriptions-Trainingsdaten kreiert werden können, einzelnen Sprachen zuordnen. Letzteres liefert meist deutlich bessere Ergebnisse. Die Fehlerquote in einem lateinischen Text aus dem 17. Jahrhundert ist beispielsweise beim NOSCEMUS-Modell deutlich geringer als bei einem italienischen Text aus derselben Zeit. Zudem verlangt Transkribus im Rahmen des Trainings die Angabe der Sprache beziehungsweise der Sprachen, die sich in den Trainingsdatensätzen finden, nicht aber die Angabe des Schrifttyps.

Transkribus als Editonswerkzeug

Verhältnismäßig niedrige Zeichenfehlerquoten und durchaus vielversprechende Weiterentwicklungen in näherer Zukunft lassen die Frage berechtigt erscheinen, ob und inwiefern Transkribus im Rahmen der Erstellung historisch-kritischer Editionen zum Einsatz kommen sollte. Trotz entscheidender Verbesserungen gilt nach wie vor: „Transkribus […] generates diplomatic transcriptions, not edited text."[13] Transkribus eignet sich allenfalls für die Anfertigung von *Basis*transkriptionen, die in weiterer Folge (innerhalb oder außerhalb des Programms) manuell korrigiert werden. Ob der Weg hin zur *de facto* fehlerfreien Basistranskription schneller beschritten werden kann, wenn man in Transkribus korrigiert oder wenn man *a priori* ohne Transkribus arbeitet, muss von Fall zu Fall getestet werden. Für die Anbringung bestimmter editorischer Anmerkungen an die Transkription ist Transkribus derzeit indes noch nicht zur Standardpraxis geworden. Kritik seitens der DH-Community bestand außerdem lange Zeit an der Tatsache, dass es sich beim XML-Output von Transkribus um PAGE-XML handelte. Dies wurde allerdings in jüngerer Zeit durch die Möglichkeit, auch TEI-XML herunterzuladen, deutlich verbessert.

Für den Gebrauch innerhalb von Editionsunternehmen, bei denen *de facto* fehlerfreie Transkriptionen benötigt werden, empfiehlt es sich, die in Transkribus generierten Daten zu exportieren, in andere Dateiformate umzuwandeln bzw. zu kopieren und am ehesten *außerhalb* von Transkribus weiter zu bearbeiten, sobald editorische Anmerkungen angebracht werden müssen, z. B. zu interlinearen Hinzufügungen durch spätere Hände oder Spezifika handschriftlicher Korrespondenz. Im Rahmen eines Editionsunternehmens ist die parallele Anwendung von Transkribus *neben* traditionellen Methoden zu erwägen. Zeitgleich mit der Erstellung eines Editionstextes außerhalb der Software können digital angefertigte Rohtranskriptionen (teils unter Nutzung der genannten Exportfunktion) mehreren Zwecken dienen: Sie können mittels digitaler Eingabe und Suche von Stichworten gezielt nach unterschiedlichen Phänomenen mit sprachlichem Niederschlag im Text durchsucht werden, noch *bevor* der zu Grunde liegende Text der Quelle vollständig ediert ist. Somit könnten historische Forschungsergebnisse zum Inhalt der Quelle, wenngleich Transkribus die Erstellung des Editionstextes selbst nicht wesentlich beschleunigt, schon während des Editionsprozesses schneller generiert werden.

Vorbemerkung: Neulateinische Editionsphilologie

Wirft man einen Blick auf das weite Feld der neulateinischen Literatur – Literatur im weitesten Sinne – so sieht man sich einerseits der Tatsache gegenübergestellt,

13. Petrolini und Wallnig 2019, 246

dass neulateinische Texte nicht so sehr von klassischen Philolog:innen (Latinist:innen) gelesen werden, sondern dass sie eher von Vertretern anderer Fachrichtungen herangezogen und als Quellen nutzbar gemacht werden, darunter Theolog:innen, Jurist:innen, Neuphilolog:innen und natürlich Historiker:innen aller Couleur. Andererseits muss beachtet werden, dass sich in der neulateinischen Philologie anders als bei der Editionstätigkeit an Texten des lateinischen und griechischen Altertums noch kein allgemein anerkannter Standard für die Editionsarbeit hat durchsetzen können.[14] Die Empfehlungen dazu, wie man einen neulateinischen Text präsentieren soll, reichen von der Forderung nach größtmöglicher Authentizität mit strikter Beibehaltung aller Eigenarten des zu edierenden Texts in Orthographie, Interpunktion und Lautstand über zahlreiche vermittelnde Versuche[15] , die die Notwendigkeit betonen, stets von Fall zu Fall abzuwägen, bis hin zur Anmahnung einer konsequenten Normierung unter Anlehnung an moderne Klassikerausgaben mit dem Ziel der leichteren Lesbarkeit.

Das Textcorpus

Im Kontext des ERC-geförderten Projektes NOSCEMUS („Nova Scientia: Early Modern Scientific Literature and Latin"), das sich der Untersuchung der Rolle von Latein als Sprache der (Natur-) Wissenschaften in der frühen Neuzeit widmet, ist im Laufe der letzten Jahre ein Modell trainiert worden, das auf die Transkription lateinischer (Wissenschafts-)Texte hin ausgelegt ist.[16] Um die formale Vielheit und inhaltliche Breite der lateinischen Wissenschaftsliteratur fassen und zu einem Gesamtbild zusammenführen zu können, galt es zu Beginn, ein möglichst repräsentatives Textkorpus hinsichtlich literarischer Form, wissenschaftlicher Disziplin und zeitlichem Rahmen in einer auf MediaWiki basierenden Datenbank zusammenzutragen, die um die Semantic-MediaWiki-Erweiterung ergänzt ist.[17]

In dieser Datenbank sind nach Stand November 2023 knapp 1000 Werke aus sechs Jahrhunderten verzeichnet – vom Beginn des Buchdruckes im 15. Jahrhundert bis zur Marginalisierung von Latein als Wissenschaftssprache im (späten) 19. Jahrhundert. Diesem großen zeitlichen Rahmen steht die inhaltliche und formale Spanne in nichts nach. Die insgesamt 21 literarischen Genera reichen von der Biographie über die Dissertation und die akademische Rede bis hin zum Lehrgedicht. Nicht weniger breit ist auch das inhaltliche Feld der insgesamt neun im NOSCEMUS-Corpus vertretenen

14. Vgl. dazu die Ausführungen bei Ijsewijn/ Sacré 1992–1998, hier Bd. 2, 434–501.
15. Ein prominenter Vertreter einer solchen im deutschen Sprachraum sehr populären Mittelposition ist Mundt 1992.
16. https://www.uibk.ac.at/projects/noscemus/.
17. https://wiki.uibk.ac.at/noscemus/Main_Page.

Wissenschaftsdisziplinen. Es erstreckt sich von Alchemie und Chemie[18] über Biologie und Medizin bis zur Mathematik, zu Astrologie und Astronomie. Obwohl in die Datenbank grundsätzlich nur lateinische Werke aufgenommen werden, findet sich darin doch eine nicht unbeträchtliche Zahl von Texten, die zum Teil mehrsprachig sind (v. a. Latein und Griechisch) oder in denen wenigstens bald längere bald kürzere Abschnitte in einer anderen Sprache gehalten sind (neben der dritten alten Sprache Hebräisch in erster Linie die großen europäischen Volkssprachen Deutsch, Englisch, Italienisch und Französisch).

Das Transkribus-Modell „Noscemus GM“

Auf all die oben skizzierten Umstände galt es bei der Entwicklung des projekteigenen Transkribus-Modells Rücksicht zu nehmen. Das Modell sollte möglichst ein „generelles“ sein, das auf alle in der Datenbank zusammengetragenen Drucke anwendbar ist. Ein besonderes Problem stellte hierbei der Umstand dar, dass in der Zeitspanne, die NOSCEMUS abzudecken sucht, viele unterschiedliche Schriftarten in den Pressen zum Einsatz gekommen sind, neben zahlreichen Antiquatypen, die vom 15. bis ins 19. Jahrhundert eine starke Wandlung durchgemacht haben, verschiedene griechische und eine ganze Reihe von gebrochenen Schriften (v. a. Fraktur- und Schwabachervarianten).

Um ein Modell zu trainieren, das auf all die genannten Punkte adäquat Antwort geben kann, wurden zentrale Transkriptionsrichtlinien erarbeitet, wobei im Hintergrund neben dem Gedanken einer ökonomischen Arbeitsweise bei der Erstellung einheitlicher Trainingsdaten stets das Ziel der leichten Lesbarkeit und Handhabung in der Volltextsuche – bei schmutziger, also nicht händisch nachbearbeiteter, Optical Character Recognition (OCR) – mitschwang. Die wichtigsten sollen hier kurz angeführt werden: Ligaturen (z. B. *Æ* und *æ*, *Œ* und *œ*, *ſt*) und Standardabkürzungen (z. B. *-q;* = *-que,* -⁹ = *-us,* -**R**᷒ = *-rum, …m̃… = …mm…, …ñ… = …nn…*) wurden kommentarlos aufgelöst, Lang-s (ſ) wurde als (gewöhnliches) s transkribiert, Kapitälchen wurden als (gewöhnliche) Großbuchstaben transkribiert, Diakritika und markante Sonderzeichen (z. B. *&, ë, ï, ę* oder alchemische Symbole) wurden hingegen nach Möglichkeit beibehalten.

Das NOSCEMUS-Transkribus-Modell setzt sich in seiner aktuellen Version (Noscemus GM 6) aus insgesamt sechs Einzelmodellen (15./16. Jahrhundert, 17. Jahrhundert, 18. Jahrhundert, 19. Jahrhundert, Latein/Fraktur und Latein/Griechisch) zusammen,

18. Nach Stand November 2023 sind in der Datenbank 82 alchemische und chemische Schriften verzeichnet. https://wiki.uibk.ac.at/noscemus/Special:BrowseData/Works?Discipline%2FContent=Alchemy%2FChemistry.

von denen jedes rund 800 Seiten beinhaltet, welche aus durchschnittlich zehn verschiedenen Drucken stammen. Die Auswahl der Drucke erfolgte systematisch, v. a. unter den Gesichtspunkten Druckjahr – es sollten möglichst immer Drucke von Anfang, Mitte und Ende eines Jahrhunderts in gleicher Zahl vertreten sein – und verwendete Typen.

Bekannte Probleme und Schwächen in der aktuellen Version

Obwohl das Modell in seiner derzeitigen Form recht breit aufgestellt ist und bei „gewöhnlichen" Drucken – d. h. Drucken, die kein besonderes Layout aufweisen oder mit wenigen Sonderzeichen auskommen – sehr gute Ergebnisse liefert (CER 0,80 %), krankt es dennoch an einigen Schwächen. Bei der Transkription von mathematischen Zahlenverbindungen (und Formeln) ebenso wie bei Diakritika und seltenen Abkürzungen oder Ligaturen in Inkunabeln ist die Fehlerrate noch recht hoch. Dasselbe gilt, wenn auch in deutlich abgeschwächter Form, für frühe (humanistische) griechische Drucke.

Das auffälligste Problem ist derzeit aber wohl noch die fehleranfällige Erkennung von Sonderzeichen, v. a. im alchemischen Bereich. Für den Großteil der in frühneuzeitlichen alchemischen Drucken verwendeten Sonderzeichen und Symbole gibt es eine Unicode-Nummer bzw. ein Unicode-Zeichen.[19] Der Aufbau von Noscemus GM, in dessen Rahmen – wie bereits erwähnt – primär das Ziel verfolgt wurde, ein Modell zu trainieren, das möglichst breit aufgestellt ist, hat es aber mit sich gebracht, dass in die Trainingsdaten nur vergleichsweise wenige Seiten aus alchemischen bzw. chemischen Drucken gelangt sind, welche die für diese Disziplinen typischen Sonderzeichen in für das Training des Modells genügender Anzahl enthalten. Dies schlägt sich in einer entsprechend hohen Fehlerrate nieder, wenn es gilt, solche Symbole zu erkennen und richtig zu transkribieren. Mit Version 6 ist das NOSCEMUS-Modell aktuell an einen Punkt gelangt, an dem eine Erweiterung der Trainingsdaten nach den bisherigen Kriterien keine spürbare Verbesserung mehr bringen wird.

Für die Zukunft bleibt zu hoffen, dass es gelingt, auch nach Auslaufen des NOSCEMUS-Projektes im April 2023 durch die Beisteuerung Datenmaterial, welches z. B. alchemistische und chemische Sonderzeichen in genügender Fülle enthält, das NOSCEMUS-Modell auch an solcherlei Fronten stärker aufzustellen.

19. Dank der Bemühungen um William Newmans Projekt *The Chymistry of Isaac Newton* sind die wichtigsten alchemischen Symbole mittlerweile im Unicode-Standard enthalten sowie durch die Schriftart *Newton Sans* abbildbar, vgl. W. R. Newman, Walsh et al. 2009; W. R. Newman 2015.

Conclusio

Transkribus bietet mit den frei verfügbaren, direkt nachnutzbaren und mitunter sehr performativen Modellen eine sehr gute Möglichkeit zur automatisierten Erstellung digitaler Transkripte von Texten, die bereits als digitale Faksimiles vorliegen, wie dies bei einer Vielzahl von Alchemica der Fall ist. Besonders für Druckwerke lateinischer Sprache eignet sich das Noscemus General Model, das im Zuge des Innsbrucker NOSCEMUS-Projekts erstellt wurde.

Damit lädt die Software Forschende ein, der Forderung Principes und Newmans, dass mehr und mehr alchemische Grundlagentexte ediert und kommentiert werden sollten, nachzukommen. Über Möglichkeiten der Verfügbarmachung derart transkribierter Texte in Form digitaler Editionen ist in Zukunft weiter nachzudenken. Transkribus bietet dafür eine hausgemachte Lösung an, die allerdings bisher für die meisten kleineren Projekte der Alchemieforschung unerschwinglich sein dürfte. Doch ist es wichtig, dass Grundlagentexte der Alchemiegeschichte so schnell als möglich für ein größeres Publikum verfügbar gemacht werden. Dies könnte beispielsweise im Sinne von Minimal Digital Editions umgesetzt werden. So könnten die Texte möglichst zeitnah von der Community nachgenutzt und mit der Zeit tiefenerschlossen werden.

Das Vorliegen digitaler Transkripte ermöglicht allerdings nicht nur die (digitale) Edition alchemischer Texte, sondern auch die Anwendung einer Vielzahl an Methoden aus dem Repertoire der Digital Humanities auf diese, so beispielsweise die der quantitativen Textanalyse. Somit hilft die Transkribus-Software – und im Speziellen das Noscemus General Model – dabei, den Weg für Edition, Kommentierung und Tiefenerschließung alchemischer Texte zu ebnen.

Michael Fröstl Jahrgang 1984, aufgewachsen in Niederösterreich, Studium in Wien (lateinische Philologie, historische Hilfswissenschaften; Teilstudium der kathol. Theologie; nebenbei Ausflüge in die Bereiche alte Geschichte und antike Philosophie); flüchtige Begegnungen mit Alchemie sowie mit Digital Humanities als Projektmitarbeiter der Universität Graz, am Institut für Österreichische Geschichtsforschung und am Austrian Centre for Digital Humanities der Österreichischen Akademie der Wissenschaften (linguistische Annotation, Transkriptionssoftware); einige Zeit AHS-Lehrer für Latein an niederösterreichischen Gymnasien (Tulln & Klosterneuburg); Mitarbeit an der Neuauflage des lateinisch-deutschen Schulwörterbuchs Stowasser; zwischen 2020 und 2023 als wissenschaftlicher Archivar des Kantons St. Gallen ansässig in der Schweiz: Erschließung von Urkunden und Akten ab dem 15. Jahrhundert in der dortigen Unesco-Weltkulturerbestätte des Stiftsbezirks St. Gallen (Stiftsarchiv) mit Forschungs- und Publikationstätigkeit zur barocken Heiligenverehrung im Bodenseeraum. Seit Dez. 2023 Archivdirektor der niederösterreichischen Landeshauptstadt St. Pölten. Aktuelle Forschungsinteressen: Mitteleuropäische Stadt- und Landesgeschichte, mittelalterliche Hagiographie; Klosterkultur und Klosterchronistik der Frühen Neuzeit, speziell von Frauenklöstern.

Stefan Zathammer studierte klassische Philologie und Rechtswissenschaften an der Universität Innsbruck. 2020 erfolgte mit einer Arbeit zum Tiroler Kirchenhistoriker und neulateinischen Dramatiker Joseph Resch (1716–1782) die Promotion zum Dr. phil. Von 2017 bis 2022 arbeitete er am ERC-Projekt NOSCEMUS („Nova Scientia: Early Modern Scientific Literature and Latin"), innerhalb dessen er den Bereich der Digital Humanities betreute, mit und war von 2019 bis 2022 auch wissenschaftlicher Mitarbeiter am Ludwig Boltzmann Institut für Neulateinische Studien. Seit Anfang 2023 ist er wissenschaftlicher Koordinator für den Standort Innsbruck der School of Medieval and Neo-Latin Studies (Kooperationsprojekt der Universitäten Freiburg i. Br., Innsbruck und Zürich). Neben dem neulateinischen Schultheater gehören die antike und humanistische Historiographie sowie die antike und frühneuzeitliche Rechtsgeschichte zu seinen besonderen Forschungsinteressen.

Sarah Lang studierte Klassische Philologie, Archäologie, Geschichte und Philosophie in Graz und Montpellier. Abschluss der Dissertation im Fach Digital Humanities zur Anwendung von Machine Reasoning auf die *Decknamen* des Druckkorpus des Iatrochymikers Michael Maier (1568–1622) erfolgte 2021. Für ihre Arbeit erhielt sie 2021 den Bader Preis für die Geschichte der Naturwissenschaften von der Österreichischen Akademie der Wissenschaften. Sie trug als Hauptorganisatorin im Organisationsteam zur Tagung „Alchemistische Labore / Alchemical Laboratories" (Wien, 19.–21.02.2020) bei. Als wissenschaftliche

Mitarbeiterin war sie 2016–2021 am Zentrum für Informationsmodellierung in Graz tätig, wo sie 2021–2027 als PostDoc arbeiten wird. Forschungsschwerpunkte: Neulateinische Alchemica, chymische Prozesse sowie die Anwendung des Methodenarsenals der Digital und Computational Humanities für deren Aufarbeitung. Seit 2023 ist sie im Vorstand des Verbands *Digital Humanities im deutschsprachigen Raum* (DHd).

Literaturverzeichnis

[1] Camen, Birte. 2018. „‘Alchymistische Kunst-Stücke in gutter Ordnungk’. Transkription und Beurteilung der Handschrift Artificia Alchimica der ÖNB (Cod. 11450) von 1596.“ Diplomarbeit, Universität Wien. URL: http://othes.univie.ac.at/52356/1/55310.pdf

[2] Ijsewijn, Jozef, Dirk Sacré. 1992–1998. *Companion to Neo-Latin Studies,* 2 Bde., 2. Auflage. Löwen.

[3] Lang, Sarah. 2019. „How to historical text recognition: A Transkribus Quickstart Guide.“ *LaTeX Ninja’ing and the Digital Humanities* Blog (10.11.2019). URL: https://latex-ninja.com/2019/11/10/how-to-historical-text-recognition-a-transkribus-quickstart-guide/.

[4] Maier, Michael. 1617/18. *Atalanta fugiens.* Oppenheim. URL: https://wiki.uibk.ac.at/noscemus/Atalanta_fugiens.

[5] Martinón-Torres, Marcos. 2011. „Some recent developments in the historiography of alchemy.“ *Ambix* 58/3: 215–37.

[6] Mundt, Lothar. 1992. „Empfehlungen zur Edition neulateinischer Texte.“ In *Probleme der Edition von Texten der Frühen Neuzeit. Beiträge zur Arbeitstagung der Kommission für die Edition von Texten der Frühen Neuzeit,* herausgegeben von Lothar Mundt, 186–192. Tübingen.

[7] Newman, William R., John A. Walsh et al. 2009. „Proposal for Alchemical Symbols in Unicode.“ In *The Chymistry of Isaac Newton,* herausgegeben von William R. Newman. URL: http://webapp1.dlib.indiana.edu/newton/fonts/Alchemy%5C%20Unicode%5C%20Proposal---March%5C%2031%5C%202009.pdf.

[8] Newman, William R. 2015. „Erläuterungen zur und Download der Schriftart Newton Sans.“ In *The Chymistry of Isaac Newton,* herausgegeben von William R. Newman. URL: http://webapp1.dlib.indiana.edu/newton/reference/font.do.

[9] Petrolini, Chiara, Thomas Wallnig. 2019. „Handwritten Text Recognition: *Transkribus* and Learned correspondence (with contributions by Günter Mühlberger).“ In *Reassembling the Republic of Letters in the Digital Age,* herausgegeben von Howard Hotson und Thomas Wallnig, 244–251. Göttingen.

[10] Principe, Lawrence, William R. Newman. 2001. „Some Problems with the Historiography of Alchemy.“ In *Secrets of Nature: Astrology and Alchemy in Early Modern Europe,* herausgegeben von William R. Newman und Anthony Grafton, 385–432. Cambridge, MA: MIT Press.

[11] Ruland, Martin. 1612. *Lexicon Alchemiae.* Frankfurt am Main. URL: https://wiki.uibk.ac.at/noscemus/Lexicon_Alchemiae.