hosted by Hamburg University Press

apropos

[Perspektiven auf die Romania]

Sprache/Literatur/Kultur/Geschichte/Ideen/Politik/Gesellschaft

Digital, global, transdisziplinär

Impulse für die Romanistik

hrsg. von
José Calvo Tello
Nanette Rißler-Pipka
Jan Rohden

9
Winter
2022

www.apropos-romania.de

ISSN: 2627-3446

Impressum

apropos [Perspektiven auf die Romania] 2022, Nr. 9
ISSN: 2627-3446
DOI: https://doi.org/10.15460/apropos.9

Herausgeber*innen

Dossier-Herausgeber*innen für dieses Heft

José Calvo Tello, Nanette Rißler-Pipka & Jan Rohden

Autor*innen dieser Ausgabe

José Calvo Tello, Alejandro Cienfuegos Pérez, Beatrice Colcuc, Annette Gerstenberg, Ulrike Henny-Krahmer, Laura Hernández-Lorenzo, Georg A. Kaiser, Anne Klee, Christoph Müller, Cord Pagenstecher, Nanette Rißler-Pipka, Anna Rodella, Jan Rohden, Julia Röttgermann, Markus Trapp, Johannes von Vacano, Anja Weingart, Florian Zacherl

Wissenschaftlicher Beirat

Lektorat, Gestaltung, Satz

Christoph Behrens, Beate Kern, María Teresa Laorden, Joris Lehnert

Bildrechte

Soweit nicht anders vermerkt, liegen die Bildrechte bei den Autor*innen selbst oder es handelt sich um gemeinfreie Bilder.

Coverbilder

Computergeneriertes Bild (DreamStudio) nach Claude Monet: Prompt "computers and literatures" & nach Amedeo Modigliani: Prompt "a crowd speaking in the alps mountains" (CC0 1.0) © José Calvo Tello

Copyright

Kontakt

www.apropos-romania.de – redaktion@apropos-romania.de

Digital, global, transdisziplinär: Impulse für die Romanistik

9

hosted by Hamburg University Press

Dossier

Digital, global, transdiziplinär: Impulse für die Romanistik

Computerlinguistik und Sprachdaten

Premiers travaux

2022, n°9
pp. 5-12
doi: https://doi.org/10.15460/apropos.9.1933

Nanette Rißler-Pipka

Einleitung
Für eine transdisziplinäre digitale Romanistik – Herausforderungen durch Multilingualität und Digitalisierung

Nanette Rißler-Pipka
ist habilitierte Romanistin und National Coordinator für Deutschland im europäischen Infrastruktur-konsortium für die Geistes-wissenschaften DARIAH.
nanette.rissler-pipka@gwdg.de

Keywords

Digitalisierung – Forschungsdaten – Romanistik – Infrastruktur – Metadaten – FAIR-Prinzipien – CARE-Prinzipien

Die Auswirkungen der Digitalisierung auf das romanistische Selbstverständnis und die Einzelphilologien

Die disziplinäre Aufteilung der Romanistik überwindet die traditionellen Einzel- und Nationalphilologien und schafft einerseits einen zentralen Fachverband wie den Deutschen Romanisten Verband (DRV)[1] und gleichzeitig besteht die Notwendigkeit der einzelphilologisch orientieren Verbände (Italianisten, Hispanisten, Franko-romanisten, Lusitanisten, Balkanromanisten),[2] die sich wiederum alle gemeinsam in der AG-ROM organisieren. Das historische Verständnis der Romanistik im Sinne einer vergleichenden Literaturwissenschaft und des gemeinsamen sprach-wissenschaftlichen Interesses an historisch-vergleichender Grammatik (Kalkhoff 2010) führt letztlich zu einer vielfachen Ausdifferenzierung, die wir auch an Fachtagungen wie dem Romanistentag sehen können. Während die Schaffung neuer fachlicher Stränge wie der Kulturwissenschaft oder der Fachdidaktik möglich waren, konnte der Antrag der AG Digitale Romanistik,[3] eine transversale Sektion einzurichten, die sich keiner der genannten Sub-Disziplinen unterordnet, nicht befürwortet werden.

[1] <https://www.deutscher-romanistenverband.de/>
[2] <https://www.deutscher-romanistenverband.de/interesse-an-romanistik/romanistische-fachverbaende/>
[3] <https://www.deutscher-romanistenverband.de/ag-digitale-romanistik/>

Dennoch waren wir zufrieden, mit dem Sektionsvorschlag „Digital, global, transdisziplinär: Impulse für eine transdisziplinäre Digitale Romanistik“ angenommen worden zu sein. Die letztliche Zuordnung zur Kulturwissenschaft war daraufhin eher der Not geschuldet, denn sowohl der Call for Papers als auch die endgültige Zusammensetzung der Beiträge zeugen von literatur- und sprachwissenschaftlichen, aber auch kulturwissenschaftlichen und landeskundlichen Beiträgen sowie zu einem wichtigen Anteil auch infrastrukturellen Belangen.

Die Zusammenführung von fachwissenschaftlich orientierten Beispielen, die zumeist aus dem Projektkontext oder aus Qualifikationsarbeiten stammen, auf der einen Seite und informationswissenschaftlich orientierten Beiträgen der Infrastrukturanbieter auf der anderen Seite, ist ein großer Gewinn dieser Sektion und des vorliegenden Dossiers.

Die Einblicke in die konkrete Projektarbeit, wie in den Beiträgen von Klee/ Röttgermann zu Topic Modeling, Gerstenberg/Pagenstecher zu Oral History, auch von Hernández-Lorenzo, Henny-Krahmer und Rohden zu konkreten stilometrischen und gattungsstilistischen Analysen der spanischen, lateinamerikanischen und italienischen Literatur, zu Wörterbüchern als Ressource (Zacherl), zu linguistischer Annotationssoftware (Weingart/Kaiser) oder zu dem regionalwissenschaftlichen und regionalsprachlichen Projekt VerbAlpina (Colcuc/Rodella) bilden jeweils neben den Ergebnissen auch die Frage der FAIRen Bereitstellung von (Meta-)Daten, Code, Visualisierungen und Analysen ab. Wie in den allermeisten geisteswissenschaftlichen Disziplinen wurden Forschungsdaten in der Romanistik bisher zumeist in Form von wissenschaftlichen Artikeln geteilt oder in größeren Veröffentlichungen auch als Wörterbücher, Editionen und in anderen Druckformaten zur Verfügung gestellt. Während diese Formen der Verbreitung weiter bestehen bleiben, sind sie jedoch nicht mehr in der Lage, das vorhandene und wachsende Spektrum romanistischer Forschungsdaten abzubilden und würden die Funktionalität und den Sinn zu einem großen Teil einschränken. Selbst in Forschungsprojekten, die weder an digitalem Forschungsmaterial arbeiten noch digitale Methoden verwenden, entstehen im Projektverlauf fast unvermeidlich digitale Forschungsdaten (Informationssammlungen, Tabellen, Präsentationen, Texte, Webseitenauftritte, Bibliographien, etc.), die als Vorstufen zur Druckpublikation wesentlich schneller veröffentlicht werden können und nach Projektende archiviert werden sollen. Daneben stellen maschinenlesbare Daten – und dazu gehören Korpora ebenso wie Programmiercode, Datenbanken, etc. – hohe Anforderungen an den Publikationsprozess. Diese Anforderungen haben sich in den FAIR-Prinzipien über alle Disziplinen hinweg niedergeschlagen (Wilkinson et al. 2016).

Welche fachspezifischen Probleme sich für die Romanistik daraus ergeben, verfolgen wir in der AG Digitale Romanistik in einer Blog-Reihe[4] und anderen Aktivitäten[5], die auch diese Sektion des Romanistentags 2021 vorbereitet haben. Multilingualität ist dabei nur ein fachspezifischer Aspekt, der sich sowohl auf die

[4] <https://blog.fid-romanistik.de/ag-digitale-romanistik/>
[5] <https://www.deutscher-romanistenverband.de/ag-digitale-romanistik/forschungsdaten/>

Daten selbst (Sprache und Codierung in Text- und Audiodaten) als auch auf die Metadaten bezieht. Daneben stellen internationale Beziehungen über Länder- und Infrastrukturgrenzen hinweg eine weitere Herausforderung für das gemeinsame Erstellen, Bearbeiten, Sichern, Publizieren und Teilen von Forschungsdaten dar.

Die FAIR und CARE Prinzipien – Forschungsdaten der Romanistik

Die für die romanistische Forschung ganz selbstverständliche internationale Zusammenarbeit, insbesondere mit Ländern des Global South (Lateinamerika, frankophones Afrika), bringt im Zusammenhang von Forschungsdaten und kulturellem Erbe sensible Themen auf, die leider noch zu wenig im Fokus von großen Infrastrukturanbietern und auch der Forschung selbst stehen. Auf interdisziplinärer Ebene wurden für diesen Bereich neben den FAIR-Prinzipien zusätzlich die CARE-Prinzipien (Carroll 2020, Imeri/Rizzolli 2022) erarbeitet. Gerade vor dem historischen Hintergrund von NS-Enteignung und NS-Raubkunst muss sich auch die deutsche Romanistik die Frage stellen: Wie gehen wir mit Forschungsdaten um, die entweder originär aus anderen Ländern stammen oder von deutschen Romanist*innen in diesen Ländern gewonnen wurden? Schon vor dem Zeitalter der Digitalisierung nutzten Forschende aus Deutschland die Archive und Museen anderer Länder, um das Material vor Ort zu untersuchen oder zeichneten Sprachforschende Dialekte und Sprachen auf, um diese zu sichern und zu analysieren. Das ist in erster Linie eine sehr wertvolle und wichtige Arbeit, die zur Erhaltung von Kulturgut und Minoritätensprachen beiträgt. Dennoch gilt es im Zuge der Massendigitalisierung und Veröffentlichung von Forschungsdaten, sensibel mit den Themen von Herkunft, Ursprung und Urheberrechte umzugehen.

Ein konkretes Beispiel, das noch nicht unmittelbar in unserer Sektion diskutiert werden konnte, ist die jüngste Datenveröffentlichung von „Transcripciones de Documentos Inéditos Recolectados Del Archivo de La Real Audiencia de Guadalajara“ von Sarah Albiez-Wieck im DARIAH-DE Repository. Die wertvolle Veröffentlichung der abfotografierten, bisher unveröffentlichten Dokumente aus dem kleinen Archiv in Guadalajara (Mexico) sowie deren Transkriptionen, die die Forscherin selbst anfertigte, wurde von ihr und dem Team des Maria Sibylla Merian Centre mit den mexikanischen Partnern ausgehandelt. Offenbar hat sich die Forscherin entschieden, nicht alle Dokumente, die Grundlage ihrer Arbeit waren, zu veröffentlichen: „The documents presented here are only a very small selection of documents analyzed within the project in a wide range of archives in Peru, Mexico and Spain.“ (Albiez-Wieck 2021) Auch wenn keine weiteren Gründe hier angegeben wurden, so kann doch anhand dieses Beispiels auf ein allgemeines Problem hingewiesen werden.

Die Bedingung der Fördergeber wie in diesem Fall der DFG (das entsprechende Projekt „Processes of construction of ‚the Ethnic‘ in Michoacan, Mexico, and Cajamarca, Peru. Translocational positionalities of indigenous migrants under colonial rule“ wird von der Autorin ebenfalls genannt), alle Projektergebnisse möglichst Open Access der weiteren Nutzung und der Forschung zur Verfügung zu

stellen und damit auch den FAIR-Prinzipien zu entsprechen, kann für die Arbeit mit indigenen Kulturgütern nicht ohne die gleichzeitige Beachtung der CARE-Prinzipien erfolgen. Es gilt für jedes einzelne Dokument nicht nur die rechtliche Lage zu klären, sondern auch in ethischer Hinsicht zu überprüfen und mit den Produzenten oder Besitzern der Daten unter Kenntnisnahme aller verfügbaren Kontextinformationen auszuhandeln, was veröffentlicht werden darf und was nicht.

Es ist auch in Open Access Repositorien wie dem DARIAH-DE Repository möglich, eine entsprechend einschränkende Creative Commons Lizenz zu vergeben (im obigen Beispiel ist dies: by-nc-sa/4.0/) oder zusätzlich wie in unserem Beispiel in der Readme-Datei zu schreiben: „Please do not reproduce the documents without contacting me first“ (Albiez-Wieck 2021). Es gibt hier demnach durchaus Abstufungen sowohl rechtlicher als auch ethischer Natur, die gegenüber einer vollkommen freien Wiederverwendung von romanistischen Forschungsdaten abgewogen werden müssen. Dies gilt für historische Daten des indigenen Kulturerbes wie in diesem Fall, aber fast noch deutlicher wird es für Daten noch lebender Personen und Zeitzeugen, wie sie im Projekt „Oral History. Digital“ (vgl. den Beitrag von Gesternberg/Pagelstecher in diesem Dossier) geschaffen und aufgezeichnet werden. Dabei handelt es sich ebenso wie bei der Sicherung von Forschungsdaten aus Archiven, deren Kulturgüter vom materiellen Zerfall bedroht sind, um Erinnerungskultur und weltweites Engagement und Zusammenarbeit. Gleichzeitig muss sehr genau auf die Persönlichkeitsrechte der aufgezeichneten oder auch nur genannten Menschen geachtet werden. Ebenso spielen die oft komplexen Urheber- und Besitzrechte eine große Rolle, die nicht nur das Schriftstück, die Fotografie, die Transkription oder den Text betreffen, sondern auch die aufbewahrende Institution im Land selbst, deren Bedeutung durchaus auch vom Alleinstellungsmerkmal und der Attraktivität des Materials abhängen kann. Durch den exklusiven Zugang zu eben jenem Kulturgut überleben die entsprechenden Institutionen. Es besteht daher gerade im romanischen Sprachraum nicht zwangsläufig ein unproblematisches Bekenntnis zum Open Access und zu der Schaffung und Öffnung digitaler Archive. Wird die Reise nach Lateinamerika der Forscherin künftig nicht mehr von der DFG oder anderen Fördergebern bezahlt, weil die Forschungsobjekte digitalisiert ortsunabhängig verfügbar sind? Werden umgekehrt Entdeckungen durch deutsche Romanist*innen in Archiven des Global South dort abfotografiert und in europäischen Repositorien veröffentlicht? Wer von den Akteuren erhält dann mehr Anerkennung, die Forschung am Objekt und die digitale Bereitstellung oder das lokale Archiv, das sich um den Erhalt des materiellen Kulturgutes bemüht?

Dies mag alles in Teilen zutreffen, am Ende wird aber hoffentlich trotz der ortsunabhängigen Verfügbarkeit von Forschungsdaten, die ja auch Forschung jenseits der Drittmittelförderung erlaubt, weiterhin die Forschung vor Ort nötig und wichtig sein. Außerdem werden in enger Zusammenarbeit zwischen Romanist*innen und lokalen Infrastruktur- und Kulturinstitutionen die Verantwortung und die Meriten offen und gleichwertig verteilt, wie dies auch obige Beispiele zeigen. Die Datenveröffentlichung begleitet im Regelfall nur eine Studie und muss zunächst als

eigener Wert geschätzt werden, der von vielen verschiedenen Akteuren gemeinsam hergestellt und ermöglicht wird. Eine mögliche Lösung bildet zumindest für Textdaten auch die Veröffentlichung von abgeleiteten Textformaten, die insbesondere im Falle rechtlicher Einschränkungen einen wichtigen Kompromiss darstellt (vgl. Schöch et al. 2020).

Zu Metadaten und Daten – Infrastruktur und Forschung

Um ein Problembewusstsein bezüglich FAIR und CARE auf der einen Seite und eine Kultur des offenen Teilens von Forschungsdaten im Sinne der Open Science auf der anderen Seite in der Romanistik zu schaffen, reicht die Arbeit der AG Digitale Romanistik allein nicht aus. Vielmehr sind wir auf feste Infrastrukturangebote beispielsweise der romanistischen Fachinformationsdienste bzw. FID (s. Trapp/ Vacano für den FID Romanistik und Müller für den FID Lateinamerika und Karibik in diesem Dossier) angewiesen, die Einzelberatungen übernehmen, Lösungen und Daten anbieten sowie die Verbindung zum Forschungsdatenmanagement auf professioneller Ebene herstellen können.

Gemeinsam mit den Anbietern von Forschungsdaten sollten wir überlegen, welche Metadaten wir als Forschende brauchen, um genau das Material zu finden (digital oder analog spielt hinsichtlich der Metadaten keine Rolle), das wir zur Beantwortung unserer Forschungsfrage benötigen. So unterschiedlich wie die Forschungsfragen werden dabei auch die Anforderungen an romanistisch-spezifische Metadaten sein. Dennoch kann man sich vermutlich auf einige elementare Informationen einigen:

1. Die Angabe der Sprache – ein Metadatum, das viele für selbstverständlich halten, aber oft fehlt,[6] fehlerhaft ist oder nicht standardisiert verwendet wird.[7] Außerdem sollte die Information über Sprache und Sprachdaten auch bei einzelnen Bildern hinterlegt sein (wenn es sich beispielsweise um Fotografien von Text/Bild-Kombinationen handelt). Besonders interessant für die Romanistik, aber auch andere Philologien, ist die Information, ob es sich bei einem Text um eine Übersetzung handelt, und wenn ja, in welcher Originalsprache dieser geschrieben wurde.

[6] Ein sehr schönes Beispiel ist eine der ältesten digitalen Bibliotheken in der nationalen Infrastrukturlandschaft: Die digitale Bibliothek in TextGrid (<https://textgrid.de/de/digitale-bibliothek>) zählt mit Werken der Weltliteratur von mehr als 600 Autor*innen zu einer wichtigen digitalen Ressource für die Komparatistik. Allerdings enthielten die vom Erstanbieter mitgelieferten Metadaten keine Informationen zur Sprache. Das lässt sich nachträglich nur noch mit sehr hohem Aufwand ändern und war für das vornehmlich germanistisch orientierte Zielpublikum des Repositoriums zum Zeitpunkt der Veröffentlichung auch nicht relevant. Bei neuen Veröffentlichungen wird die Angabe der Sprache allerdings nun zur Pflicht. Geplant ist aktuell ferner die Veröffentlichung des Korpus ELTeC mit Romanen aus zunächst 6 und später mehr als 20 verschiedenen Sprachen (vgl. <https://www.distant-reading.net/eltec/>) im TextGrid Repository.

[7] Allerdings kann es auch für Bibliotheken und Anbieter von Repositorien oder Archiven manchmal schwierig sein, eine einheitliche Verwendung von Sprachcodes über die Jahre hinweg beizubehalten. So gibt es selbst innerhalb des ISO-Standards 2- oder 3-Zeichen lange Varianten und unterschiedliche Empfehlungen innerhalb von bibliothekarischen oder Web-Standardisierungsorganisationen wie dem W3C (vgl. Ishida 2016).

2. Die Angabe des Dateiformats – ist oft nicht eindeutig für Nutzer*innen zu erkennen. Je nach Portal und Anbieter werden zu einer Quelle unterschiedliche Formate zur Verfügung gestellt (z.B. txt und XML oder PDF und IIIF).[8] Eine breite Auswahl an Formaten ist natürlich ein Vorteil und entspricht den unterschiedlichen Bedürfnissen der Forschenden. Für die Anbieter der Daten ist es umgekehrt oft schwer die Balance zwischen Service, Offenheit und den komplexen Lizenzmodellen zu finden (vgl. auch Lehmann 2022, 12-17).
3. Die Unterscheidung von Primär- und Sekundärliteratur – ist leider in Bibliothekskatalogen oder auch in Verlagsinformationen nicht einheitlich oder explizit gegeben, für die Forschung, Literaturrecherche und Suche nach Forschungsdaten allgemein jedoch sehr wichtig. Zu diesem Zweck gibt es zwar große Fachbibliographien (z.B. Klapp-Online), doch gerade in der Forschung mit bibliographischen Metadaten direkt am und mit dem Katalog ist dies problematisch, wie José Calvo Tello an einer beispielhaften Studie in diesem Dossier zeigt.

Die Unterscheidung und die unterschiedliche Behandlung von Metadaten und Daten wird zunehmend überflüssig. Zum einen, weil Daten ohne Metadaten kontext- und sinnfrei werden, da sie sich selten vollständig selbst erklären können: ein Text ohne Titel, Zeit und Autor mag zwar dennoch als poetisch oder fiktional identifiziert werden, aber im Moment der Publikation und damit beim Eintritt in die ‚Welt' erhält der Text unweigerlich einen Kontext und unabhängig von analog oder digital mindestens bibliographische Metadaten. Zum anderen fügen Forschende immer mehr Informationen hinzu (durch Annotation, etc.), die ebenfalls zum Metadatum werden können – dazu gehören nicht nur Auswertungen mit digitalen Werkzeugen, sondern auch Artikel und Bücher, also Sekundärliteratur, die wiederum selbst als Forschungsdaten mit eigenen Metadaten gesehen werden können. Werden Metadaten nicht nur zum Forschungsobjekt (z.B. die bibliografische Information zum Text), sondern darüber hinaus auch untereinander verknüpft (z.B. durch Verweise auf Normdaten und Linked Open Data), dann entsteht vernetztes Wissen und mit der entsprechenden technischen Funktionalität ein Knowledge Graph. Ganz praktisch würde dies unsere Arbeit erleichtern, wenn in digitalen Bibliotheken, Bibliothekskatalogen und darüber hinaus jeweils sämtliche Sekundärliteratur, Primärwerke und wenn möglich sogar dazu gehörende Forschungsdaten miteinander vernetzt wären. Diese scheinbar simple Aufgabe ist aber für Infrastruktureinrichtungen nicht allein lösbar, grenzt doch die Identifikation von Primärwerken an die Frage „Was ist Literatur?" oder noch weiter „Was ist ein primäres Werk oder Forschungsdatum?". Nicht nur an dieser Stelle ist die Zusammenarbeit zwischen Forschung und Infrastruktur gefragt, wie sie auch in der Sektion und in diesem Dossier erprobt wurde.

[8] Siehe Beispiele hier: <https://gdz.sub.uni-goettingen.de/id/PPN659351714?tify={%22pages%22:[6],%22view%22:%22export%22}> oder im TextGrid Repository.

Ausblick

Auf politischer Ebene ist zur Vorbereitung und Implementierung der Nationalen Forschungsdateninfrastruktur in Gremien wie der Gemeinsamen Wissenschaftskonferenz (GWK)[9], dem Rat für Informationsinfrastrukturen (RfII)[10] oder communitybesetzten Gremien wie dem Geisteswissenschaftlichen Forum NFDI[11] sowie innerhalb der einzelnen Konsortien die Rede von den Herausforderungen der Digitalisierung, aber auch von den Chancen und der Notwendigkeit der datengetriebenen, interdisziplinären Wissenschaft auf der einen Seite und mangelndem Personal und Kompetenzen auf der anderen Seite (RfII 2019, 7-16).

Wenn Junkerjürgen eine Reform-Romanistik fordert und die „romanistische Literaturwissenschaft wäre demnach in eine romanistische Medienwissenschaft zu überführen" (Junkerjürgen 2021, 105), so muss dies nicht bei einer Öffnung bezüglich des Gegenstands (von der Literatur zu allen auch digitalen Medien) enden, sondern kann auch die Vermittlung von digitaler Kompetenz und Methoden sowie die Beteiligung an einer disziplinübergreifenden Debatte um datengetriebene Wissenschaft und dazu erforderliche Infrastrukturen einbeziehen.

Literatur

ALBIEZ-WIECK, Sarah. 2021. „Transcripciones de Documentos Inéditos Recolectados Del Archivo de La Real Audiencia de Guadalajara." DARIAH-DE, 2021. <https://doi.org/10.20375/0000-000E-556A-C>.

BECKER, Lidia et al. 2020. *Fachbewusstsein der Romanistik Romanistisches Kolloquium XXXII*. Tübingen: Narr Francke Attempto Verlag.

CARROLL, Stephanie Russo et al. 2020. „The CARE Principles for Indigenous Data Governance." *Data Science Journal* 19 (1) (4. November 2020): 43. <https://doi.org/10.5334/dsj-2020-043>.

IMERI, Sabine und Michaela Rizzolli. 2022. „CARE Principles for Indigenous Data Governance: Eine Leitlinie für ethische Fragen im Umgang mit Forschungsdaten?" *o-bib. Das offene Bibliotheksjournal* / Herausgeber VDB 9, Nr. 2 (14. Juni): 1–14. <https://doi.org/10.5282/o-bib/5815>.

ISHIDA, Richard. 2016. „Zweibuchstabige oder dreibuchstabige Sprachcodes." W3C. Internationalisierung. Making the World Wide Web worldwide, 25.05. <https://www.w3.org/International/questions/qa-lang-2or3.de.html>.

JUNKERJÜRGEN, Ralf. 2021. „Reform-Romanistik. Ein Plädoyer." *apropos [Perspektiven auf die Romania]* 7, 102–106. <https://doi.org/10.15460/apropos.7.1842>.

KALKHOFF, Alexander M. 2010. *Romanische Philologie im 19. und frühen 20. Jahrhundert: Institutionengeschichtliche Perpektiven*. Romanica Monacensia, Band 78. Tübingen: G. Narr.

LEHMANN, Jörg. 2022. „The Tragedy of the Cultural Commons. Research Report and Data Publication." Zenodo, 3. Mai. <https://doi.org/10.5281/zenodo.6513596>.

[9] <https://www.gwk-bonn.de/>
[10] <https://rfii.de/de/>
[11] <https://nfdi.hypotheses.org/>

RfII – Rat für Informationsinfrastrukturen. 2019. „Digitale Kompetenzen - dringend gesucht! Empfehlungen zu Berufs- und Ausbildungsperspektiven für den Arbeitsmarkt Wissenschaft“, Göttingen, 56 S. <https://rfii.de/?p=3883>.

SCHÖCH, Christof et al. 2020. „Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen.“ *Zeitschrift für digitale Geisteswissenschaften*. Wolfenbüttel. text/html Format. <DOI: 10.17175/2020_006>.

WILKINSON, Mark D. et al. 2016. „The FAIR Guiding Principles for Scientific Data Management and Stewardship.“ *Scientific Data* 3 (15. März): 160018. <https://doi.org/10.1038/sdata.2016.18>.

Zusammenfassung

Digitale Ansätze restrukturieren den Wissenschaftsbetrieb und stärken seinen globalen Charakter. Sie können ferner neue Synergien bilden, sowohl zwischen unterschiedlichen Fächern (transdisziplinär), als auch innerhalb der Teildisziplinen eines Faches (intradisziplinär). Für die Romanistik konnten Auswirkungen und Potential der Digitalisierung bereits für unterschiedliche Teildisziplinen aufgezeigt werden. In welcher Weise die Digitalisierung darüber hinaus ein Bindeglied für disziplinübergreifende Forschung romanistischer Prägung bilden kann, zeigen die Beiträge der Sektion "Impulse für eine transdisziplinäre digitale Romanistik" des Romanistentags 2021, die hier versammelt werden. Einleitend werden die Besonderheiten des Faches Romanistik im Zusammenhang mit Multilingualität und Digitalisierung ebenso betrachtet wie das Verhältnis zu den FAIR und CARE Prinzipien. Damit soll die Romanistik auch angeregt werden, sich in aktuellen wissenschaftspolitischen und infrastrukturellen Bewegungen wie beispielsweise der NFDI (Nationale Forschungsdateninfrastruktur) zu positionieren.

Abstract

The disciplinary concept of Romance Philology in Germany has per se a transdisciplinary character which is more and more difficult to maintain in everyday practice. Today, challenges regarding digitisation and multilingualism ask us to have again a closer look at the concept of our discipline. The working group Digital Romance Studies hosted a section on “Digital, global, transdisciplinary: Impulses for transdisciplinary Digital Romance Studies” at the Romanistentag 2021. The proceedings of the section are published here in apropos. The introduction highlights the necessary collaboration between sub-disciplines, researchers, infrastructure providers and politics (funding organisations) in order to be able to share openly research data according to the FAIR and CARE principles.

Dossier

Digital, global, transdisziplinär: Impulse für die Romanistik

Teil 1 Computational Literary Studies

Bild: Computergeneriertes Bild (DreamStudio) nach Diego Rivera: Prompt "novels and latin american identities" (CC0 1.0)

apropos [Perspektiven auf die Romania]

Winter 2022

9

www.apropos-romania.de

ISSN: 2627-3446

2022, n°9
pp. 14-36
doi: https://doi.org/10.15460/apropos.9.1893

Ulrike Henny-Krahmer

Novelas originales y americanas

A Digital Analysis of References to Identity in Subtitles of Spanish American 19th Century Novels

Ulrike Henny-Krahmer
is junior academy professor for Digital Humanities at the University of Rostock.
ulrike.henny-krahmer@uni-rostock.de

Keywords

genre – identity – Spanish America – 19th century novel – metadata analysis

1. Introduction

1.1. Literary texts, identity constitution, and genre

Literature is one of the central media through which identities are represented and constructed. "Identity" is a status that is achieved by means of identification processes and it can concern the personal identity of individuals as well as the collective identity of groups. To constitute identity means to integrate disparate experiences and conceptions of the self and the world, differing expectations and cultural role models into a relatively static and harmonic whole (cf. Horatschek 2013, 323). One of the areas in which the special role that literary texts play for the constitution of identities has been discussed is in cultural studies of memory. From that perspective, literary texts are a medium of collective memory and they fulfill specific functions in the culture of remembrance. They can form ideas about past worlds, convey images of history, serve to reflect on processes and problems of collective memory, and also shape concepts of identity. As "collective texts", they circulate in specific cultural contexts and contribute to generate, communicate, and provide perspectives on collective memory and identity (cf. Erll 2017, 167–190).

A characteristic of collective identities is that they are bound to the development of group-specific cultural forms. Collective identities can be related to different kinds of groups, for instance linguistic communities, political entities such as nation states, or cultural groups on local, regional, cross-regional, or supra-national levels. Feldman (2001), for example, examines narratives of American national identity as group narratives and discusses the relationship between identity stories and literary genres as specific cultural forms. In general, group-defining stories are

highly patterned and genres provide such patterns for literary texts. In the case of national narratives, they have been expected to relate to the romance genre, with a "superior hero" and a "high mimetic mode" as opposed to an "ordinary hero" and "comedy and modern realistic fiction" (Feldman 2001, 130, referring to Frye 1957). However, in her analysis, Feldman finds, that "[n]ational identity stories may have a distinctive genre, but which genre is chosen is bound to vary. We may find a tragedy in one place, a romance in another" (Feldman 2021, 130). So, although there is a relationship between literary genres as patterns for group-specific cultural forms which serve to represent and constitute collective identities, this relationship is not fixed and pre-established but can instead be considered the result of group-specific identification processes.

The special role that literary texts and genres play in these processes is set forth by Erll, who describes the characteristics of literary texts as one specific symbolic form of memory culture. Through processes of convergence, in literary texts, complex events are concentrated in specific topoi, narratives, places, or characters. Furthermore, collective memory is built through narrative processes which are also central for narrative literary genres. What is narrated is selected and combined from a wealth of impressions and data, and genre patterns can be understood as conventions of the codification of events. As fictional texts, literary texts have a restricted claim to be referential and objective. This provides them, however, with the privilege to construct realities which can contribute to the constitution of collective memories and identities (cf. Erll 2017, 167–172).

Here, these general considerations on the relationship between literary texts, identity constitution, and genre are taken into account as a basis for a digital analysis of 19th century Spanish American novels, of their subgenres, and their function in the formation of collective identities, starting from references to identity which were found in the subtitles of the novels.

1.2. Spanish American novels in the 19th century and questions of identity

In Spanish America, the 19th century was marked by the independence movements of the Spanish American colonies, which aspired and, in most cases, achieved to become independent from the mother country Spain in the course of the century. During the colonial period, access to novels had been limited, but in the 19th century, the genre became popular and spread in connection with the development of local literary markets. For Argentina, for instance, a comprehensive study of the novel's emerging reality in the 19th century cultural system has been undertaken by Molina, who compares the novels with mushrooms springing up (cf. Molina 2011).

In many cases, the novels served to address social, political, and historical issues, and they fulfilled important functions in the formation of distinct national identities. For Latin America as a whole, Sommer writes about "foundational fictions" and "national romances" (Sommer 1993), discussing the role that romantic novels played in the process of national consolidation. Also, Lindstrom

dedicates several chapters of her book on early Spanish American narrative to the interconnections between narratives and nationhood: "The Struggle for Nationhood and the Rise of Fiction", "The Mid-Nineteenth Century: Romanticism, Realism, and Nationalism", "Late-Nineteenth-Century Narratives of Social Commentary and National Self-Reflection" (Lindstrom 2004). The significance of narrative fiction and the novel genre in processes of political emancipation and the definition of national spaces and identities has also been addressed for individual Spanish American countries, for example by Hanway (2003) for Argentina, Brushwood (1966) for Mexico, or Ferrer (2018) for Cuba.

As the number of different studies on the topic shows, the question of national Spanish American identities and of decolonialization has been in the focus of research on 19th century Spanish American novels. However, there are other types of identity issues that are significant in relation to the Spanish American novels. Towards the end of the 19th and in the early 20th century, a genuine, supranational, and cultural Spanish American identity began to develop in the wake of the Modernist current. This was a current that aimed to provide the industrialized and modernized society with an equally modern literature, oriented towards contemporary French literary currents. Following this, there was again a turn to regional themes that foregrounded rural settings (Gálvez 1990, 148–194). This means that in addition to the question of a political, national identity, the novels also addressed aspects of cultural identity that related to spaces of a different scale and nature. As Julio Ortega puts it in the introduction to a volume dedicated to the search for a distinct and universal Hispanic American literature, although 19th and 20th century narratives are part of collective memories of nation-building, they cannot be reduced to the definition of a single and unambiguous identity, but represent and shape it in a process of productive creativity:

> Si la narrativa es una memoria de la formación nacional, cuya fábula o romance es un proceso de autoidentificación, la identidad deja de ser un catálogo de deudas impagables y se construye, más bien, como alteridad y pluralidad; al punto que esta narrativa demuestra, más bien, la identidad como proceso abierto, que en lugar de definirse como carencia, se define como exceso de filiaciones, alianzas y consensos. La literatura parece decirnos que en vez de un "problema de identidad" (de legitimidad, autoconocimiento, pertenencia, comunidad), tenemos una re-solución de identidades. Así, el relato de la identidad, donde se despliega y pone a prueba, no sólo sostiene al "yo" heroico sino a su interlocutor, en el espacio de concurrencia donde forjan su libertad mutua (Ortega 2011, 17).

This observation about the function of Spanish American narratives for cultural and national identities is in line with the considerations from memory research referred to above, which describe that literary texts can open up different perspectives on collective identities, that they can themselves contribute to the constitution of identity in a variety of ways, and that even genre patterns used to define group identities are applied in different ways in specific historical and cultural contexts.

1.3. Identity references in subtitles of novels

The different aspects of identity constitution by means of narratives, that have been raised in the previous section on 19th century Spanish American novels, shall be examined in this article from a specific perspective, that of identity references

in subtitles of novels. In order to explain how this particular investigation came about, it is necessary to go into its more general background, which is the study *Genre Analysis and Corpus Design: 19th Century Spanish American Novels (1830-1910)*, a dissertation which has been produced by the author of this article in the context of the project *Computational Literary Genre Stylistics* (CLiGS) at the University of Würzburg[1] and to be published soon. The goal of the dissertation was to analyze thematic subgenres of the novels and literary currents on the basis of stylistic features and distributions of topics in the texts and to investigate how well they can be classified with different quantitative methods. Another central question was to analyze which textual features are distinctive for the different subgenres and literary currents in question. To be able to conduct the empirical study, a corpus of 256 novels from Argentina, Mexico, and Cuba, which had been published between 1830 and 1910, was compiled and prepared in digital formats.[2] The corpus has been published under the name of *Corpus de novelas hispanoamericanas del siglo XIX* (*Conha19*, cf. Henny-Krahmer 2021). Besides the corpus, a more comprehensive collection of bibliographical data about 829 novels from the same historical context was developed. It is called *Bib-ACMé: Bibliografía digital de novelas argentinas, cubanas y mexicanas (1830-1910)* and has been published as data (cf. Henny-Krahmer 2017a) as well as in the form of a web application (cf. Henny-Krahmer 2017b).

Both the corpus and the bibliography contain detailed metadata about the subgenres of the novels and the literary currents that they have been attributed to. The information about the subgenres and currents was collected from literary histories, monographs, and research articles on Spanish American novels, but also from title pages of the novels' historical editions. All the genre labels that were found were collected and classified regarding their source – literary historical or contemporary – and their type. The following types of subgenre labels were identified in the corpus and the bibliography: (1) thematic labels such as "novela histórica" or "novela sentimental", (2) those referring to literary currents, as for example "novela romántica" or "novela naturalista", (3) identity labels connected to linguistic, cultural, geographical, regional, or national identity as for instance "novela original", "novela mexicana", or "novela habanera", (4) labels related to the mode of the narration in terms of its relationship to reality, its form of representation, its medium, attitude, or intention, for example "cuadros", "estudio", "memorias", "novela satírica", or "novela de propaganda". The groups of genre labels were developed starting from semiotic models of genre, especially those formulated by Raible (1980) and Schaeffer (1983), which emphasize the role of genre names as complex linguistic signs that point to different levels of meaning. Whenever a literary work is associated with such a genre name, the generic signal

[1] The project ran between 2015 and 2020 and was funded by the German Ministry for Education and Research (BMBF). The primary goal of the research group was "to provide a methodological linkage between new techniques of quantitative analysis of literary texts and the fundamental issues of literary studies in the domain of genre theory and stylistics [...] on the basis of several large text collections that consist of French dramas of the classical period and the Enlightenment as well as French and Spanish novels of the 19th century" (CLiGS n.d.).

[2] To be able to focus on specific literary historical contexts, only novels from three selected Spanish American countries were chosen.

opens up an interpretive framework and specific genre conventions, against the background of which the text can be analyzed and understood.

The dissertation focused on the first two levels of genre labels, i.e., the thematic subgenres and literary currents, both of which have been at the centre of literary historical research. Here, the third group of genre labels, which relate to different forms of identity, is addressed. During the preparation of the corpus and the bibliography, it became clear that this type of genre label frequently occurs in the subtitles of the novels as they appear on the covers of historical editions of the texts, for example "Esqueletos sociales. Novela original de J. Rivera y Rio" (Mexico, edition of 1873), "Otilia. Novela americana" (written by the author Ventura Aguilar, who was probably Argentine, edition of 1895), or "La campana de la tarde, o vivir muriendo. Novela cubana por Julio Rosas" (Cuba, edition of 1873), to mention some individual cases.

However, these historical references to identity which occur in the subtitles of the novels have not been analyzed systematically so far, even though questions of identity are a prominent subject in literary historical research on 19th century Spanish American novels. A reason for this might be that most – though not all – studies on Spanish American novels concentrate on a relatively small number of canonical works and that it has not yet been noticed how often these references to some kind of identity occur. A second reason is probably that historical genre labels, that is, labels that have been added to titles of works either by authors or editors, must in no way be systematic or especially meaningful. They may point to textual characteristics of the novels in question, but they can as well express extra-textual functions of the novels or indicate that certain designations for the novels were in vogue and were used to sell the books well and attract readers. Moreover, since they may be the result of both authorial decision and editor influence, it is not clear to what extent they correspond to an authorial will.

Some discussions of this kind of genre labels can however be found. Botrel, who discusses the Spanish novel between 1830 and 1930 as an editorial genre, mentions the designation "novela original" and remarks:

> Las normas/formas tipográficas bibliográficas permiten también observar cómo después de un período en el que se precisa el origen de la novela («novela escrita en francés por Mr.» o «Madama…», «en inglés por Mistress…» o «Sr…» y «traducida al castellano por…» iniciales) la preeminencia del título unida con la hispanización casi sistemática de los nombres de los autores traducidos (Javier de Montepín, Pablo Feval, etc.) y la importancia numérica de las traducciones, con la desaparición de la mención del traductor, al menos en las referencias bibliográficas, hace que el género novela venga disociado de una por otra parte deseada hispanidad y asociado con una patronímica y toponimia extranjerizante, como producto extranjero o, más probablemente, asimilado. La mención «novela original» o «española» introducirá durante cierto tiempo una distinción poco decisiva, estadísticamente al menos (Botrel, 2001, paragraph 12, footnotes omitted).

Botrel thus analyzes the designation "novela original" as a marker indicating that the novels were originally written in Spanish and are not translations, a label which became necessary to distinguish these novels from others originally written by foreign authors in foreign languages. Also, Molina mentions the label "novela original" in her discussion of subtitles of Argentine novels published between 1838

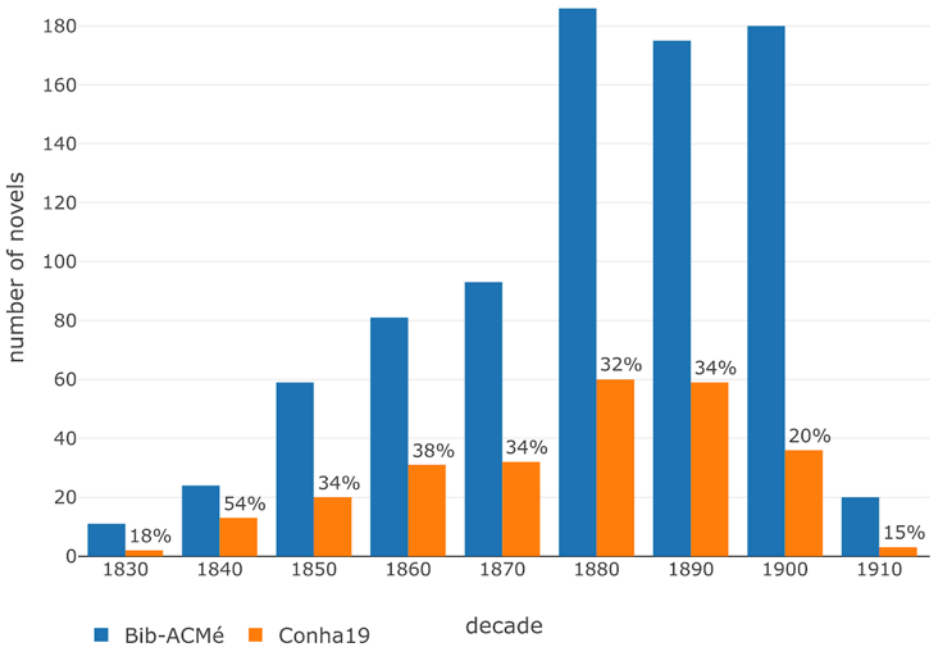
und 1872 and interprets it in the same way as Botrel, as a marker of local authorship: "Hacia mediados de la década de 1850, el apelativo 'original' empieza a leerse en la portada de libros editados en Buenos Aires; y con él quiere significarse que se trata de un texto escrito por autor local, que no es una traducción" (Molina 2011, 54) and "El adjetivo 'original' – según se explicó más arriba – señala las obras producidas por autores locales; aparece en un 21 % de nuestro corpus" (Molina 2011, 231). Both Botrel and Molina point out that a large proportion of the novels sold in Spain and Argentina in the early nineteenth century came from abroad (see the quote from Botrel above and Molina 2011, 25–26), so it seems particularly necessary to mark novels of local origin.

From the perspective of historically oriented, empirical genre research, the identity references that appear in the subtitles of the novels are certainly of interest. In a broader context, it would be instructive to investigate whether identity references in subtitles of novels in the nineteenth century occurred predominantly in Spanish-speaking countries, or, for example, in French- or English-speaking countries as well. Such an approach, however, goes far beyond the present study, in which the focus is on the corpus of Spanish American novels and in which the following questions are posed: Which kind of identity labels occur and how frequent and numerous are they? Are they connected to extra-textual features such as the nationalities or cultural identities of the authors or the period in which they were published? And do they relate to other levels of genre, for example the thematic subgenres or literary currents? How about their relationship to the content and style of the texts, is there a pattern and a correspondence between certain identity labels and the kind of novels to which they were attached? Or are the labels used rather randomly by authors or editors? Finally, what can be learned about the concepts of identity that the 19th century Spanish American novels represented and constituted? In the following, these questions are examined with a digital analysis of the identity references that occur in the corpus *Conha19* and the bibliography *Bib-ACMé*.

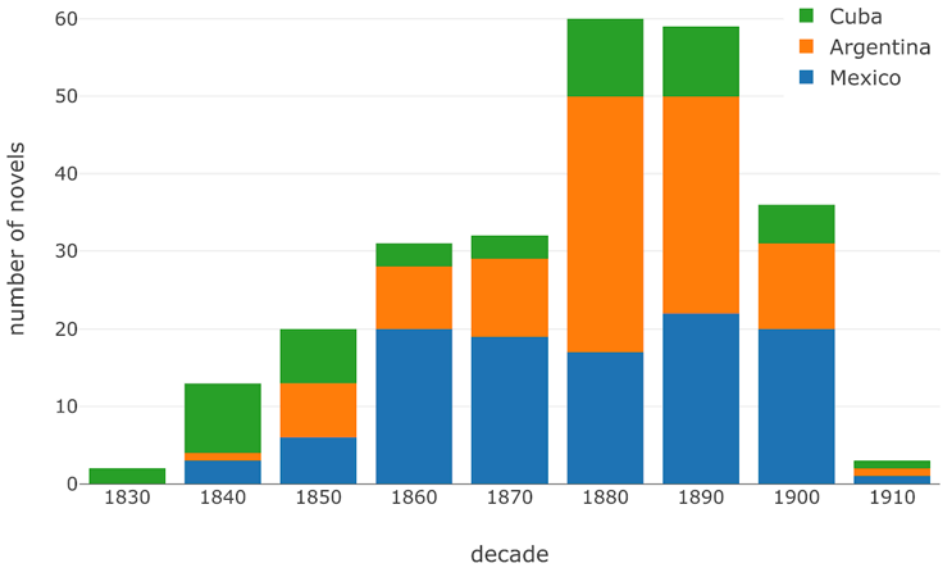
2. Digital analysis of identity references in *Bib-ACMé* and *Conha19*

2.1. Starting point: data, data modeling, and methods

The digital analysis of identity references is based on the two resources *Bib-ACMé* and *Conha19*, which comprise 829 and 256 novels each. In both cases, novels were chosen that were published between 1830 and 1910 and that were either written by Argentine, Mexican, and Cuban authors or had first been published in the respective countries. Only novels that were originally written in Spanish are considered. Fig. 1 and 2 illustrate the distribution of novels per decade, to give an impression of the contents of the bibliography and the corpus:



1 | Number of novels per decade in *Bib-ACMé* and *Conha19*



2 | Number of novels per decade and by country in *Conha19*

In the first figure, for each decade, the number of novels in *Bib-ACMé* is compared to *Conha19*, showing that the corpus contains approximately one third of the novels that are part of the bibliography. The main difference between these two resources is that the bibliography consists only of metadata about the novels – in particular about their authors, editions, and subgenre labels – while the corpus also contains the digital full texts of the novels. This means that the bibliography can be used in its entirety for an analysis of the metadata, but only the corpus is suitable for text analysis. In this contribution, both aspects are combined, which means that the metadata analysis with the bigger data set is complemented by textual analyses

of the corpus to be able to take into account text-internal features, as well. Fig. 2 illustrates the distribution of novels in the corpus by country and decade. *Conha19* contains 108 novels from Mexico, 99 novels from Argentina, and 49 novels from Cuba. Both figures show that the number of novels is highest in the 1880s and 1890s. This reflects the fact that from the middle of the 19th century more and more novels were published. On the other hand, the number of novels from 1900 onwards is again lower in the digital resources, because some of the texts in the 20th century are still subject to copyright and are therefore not as accessible as the earlier texts.

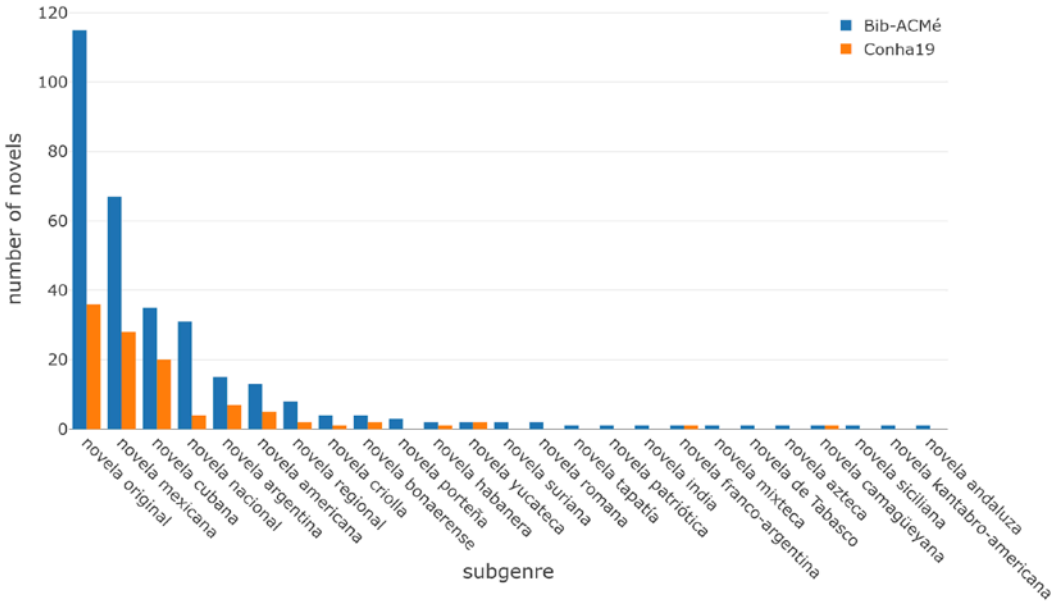
Before the results of the analyses can be presented, three important methodological issues need to be addressed. First, how were the genre labels referring to different kinds of identity collected and treated? Second, which other extra- and intratextual aspects of the texts are considered and linked to the identity references, and how were these data obtained? Third, which digital methods are used to analyze the identity references in the subtitles of the novels in connection with the other metadata and textual data?

The first methodological issue concerns the identity references in the subtitles of the novels themselves. In *Bib-ACMé*, for each novel, all the editions that could be found and that were published between 1830 and 1910 were registered. This means that each novel as an individual work of art can be represented by several different editions. The titles and subtitles of the works are found on those historical editions, which amount to 1,220 instances in *Bib-ACMé*. For each edition, it was checked whether the subtitle contained some kind of reference to identity, such as "novela original", "novela americana", "novela cubana", etc. These references were then collected for each novel and it is on the level of the individual novels as works of art that the identity references are analyzed here. The decision to derive the references from the editions but to analyze them on the level of the novels as literary works was made because the references to identity were treated as genre labels here and genre is considered an aspect that is attributed to the work of art, not to individual editions of the work. This methodological choice has several consequences. It prevents certain novels that have been published very often from dominating over those for which there is only a single edition. If the results were recorded at the level of editions, individual novels could distort them. However, the disadvantage of this modeling is that subtle differences of subtitle assignment that may exist between different editions are not taken into account.

To be able to analyze references to identity quantitatively, they were normalized, which means that grammatical and orthographic differences in the forms and different reference words of the identity attributes were homogenized. However, the original forms were also retained in the metadata so that the standardization process can be tracked. For example, labels such as "novela histórica original", "novela de costumbres mexicanas", "cuento camagüeyano" were normalized to "novela original", "novela mexicana", and "novela camagüeyana" for the purpose of identity analysis. A further step was taken by grouping the identity references to geographical and cultural units at a higher level. The goal of this grouping was to create more homogeneous groups that could be analyzed quantitatively, since

references to local, historical, or indigenous identities in particular were very sporadic. As in the case of the linguistic normalization, also here the original forms were kept to ensure a transparent methodology. As an example for grouping, references such as “novela cubana”, “novela camagüeyana”, or “novela habanera” are all considered “novelas cubanas”. On an even higher level, “novela cubana”, “novela mexicana”, “novela argentina”, “novela americana”, or “novela criolla” are all considered part of the group “novela americana”, as opposed to novels attributed to non-American identities or to no kind of identity at all. These groupings provide additional levels for quantitative analysis, but of course they introduce a form of interpretation and abstraction, which has to be kept in mind. All the information about identity references in the subtitles of the novels is part of the metadata that has been collected for *Bib-ACMé*, and it is retained in the digital bibliography, where it is encoded in XML, following the standard of the *Text Encoding Initiative* (TEI Consortium 2022).

The first question that can be answered by analyzing this metadata is which kind of identity labels occur in the bibliography and corpus and how frequent and numerous they are. In *Bib-ACMé*, 33 % of the novels carry some kind of identity label and in *Conha19*, 39 % of the novels have such a label. An overview of which labels occur how often in each resource is given in Fig. 3:



3 | Number of novels with identity references in their subtitles, in *Bib-ACMé* and *Conha19*

In the following, the evaluation focuses on the occurrences of the labels in the bibliography (how often these occur in the corpus in comparison can be seen in Fig. 3). All in all, there are 25 different subgenre labels related to the linguistic, geographical and socio-cultural identity. The most important identity label is the general term “novela original”, carried by 113 (14 %) of the works in the bibliography. It is followed by the labels related to the three selected countries (“novela mexicana”, “novela cubana”, “novela argentina”), by other general labels (“novela nacional”, “novela regional”) and by labels referring to the American continent (“novela americana”, “novela criolla”, “novela india”). Among the various

identity labels of minor importance, there are several related to the countries' capitals ("novela bonaerense", "novela porteña", "novela habanera")[3], to specific regions or cities in Mexico or Cuba ("novela yucateca", "novela suriana",[4] "novela tapatía",[5] "novela de Tabasco", "novela camagüeyana"), and to Mexican indigenous people ("novela mixteca", "novela azteca"). Also, there are references to European regions and culture ("novela romana", "novela franco-argentina", "novela siciliana", "novela kantabro-americana", "novela andaluza"). How these different labels are grouped for further analysis is summarized in Tab. 1 below.

group	labels
novela americana	novela americana, novela argentina, novela azteca, novela bonaerense, novela camagüeyana, novela criolla, novela cubana, novela de Tabasco, novela franco-argentina, novela habanera, novela india, novela kantabro-americana, novela mexicana, novela mixteca, novela porteña, novela suriana, novela tapatía, novela yucateca
novela argentina	novela argentina, novela bonaerense, novela franco-argentina, novela porteña
novela cubana	novela camagüeyana, novela cubana, novela habanera
novela mexicana	novela azteca, novela de Tabasco, novela mexicana, novela mixteca, novela suriana, novela tapatía, novela yucateca

Tab. 1 | Groupings of identity labels

Besides the treatment of the identity references found in the subtitles on historical editions of the novels, the second methodological aspect that needs to be clarified is which other extra- and intratextual aspects of the texts are taken into account in the metadata and text analysis of the identity references and how this data was gathered. The following features that are external to the texts are analyzed:

- the years of publication of the novels, for which the year of the novels' first edition is decisive; this allows to check if references to identity occurred primarily in certain subperiods of the 19th century;
- the country that the novels are associated with; which is either the nationality of the author or the country in which the novel was first published (if the author is not Argentine, Mexican, or Cuban); with that, it can be analyzed if references to identity were used more often in one of the Spanish American countries than in another;
- the primary thematic subgenre of the novels as indicated by literary historians or explicitly given on historical editions; the question here is if there are thematic subgenres that were more often used in combination with identity references than others;

[3] In the case of the "novela mexicana" it cannot be decided if it refers to the country or the capital.
[4] According to the "Diccionario de la lengua española" of the Spanish Royal Academy, "suriana" means "coming from the south of Mexico" (Real Academia Española 2021a).
[5] "Coming from Guadalajara" (Real Academia Española 2021b).

- the literary current that the novels were associated with by literary historians; also, here it is checked whether there are correlations with the identity references[6].

All this information is available for all the novels in *Bib-ACMé*. There are further text-internal characteristics that are only available for the subset of novels that is part of *Conha19*, because access to the full texts was needed to be able to determine these features that are related to the content, plot, and style of the texts themselves:

- the continent of the setting (America or Europe)[7];
- the time period of the setting (past, recent past, or contemporary)[8];
- the narrative perspective (first person or third person)[9];
- words preferred and avoided (in comparison to novels not carrying the same kind of identity reference).

Because there can be shifts in the course of the text, for the categories "continent of the setting", "time period of the setting", and "narrative perspective", the modes that were predominant throughout the texts were chosen. To look for correlations between the narrative perspective and identity references is of interest because it reveals ways in which identity is communicated in the novels – through interior views and individual perspectives of characters or from a more neutral and panoramic perspective that is independent of single characters.

For all the kinds of extra-textual and text-internal features listed above, the goal of the analysis is to look for correlations between them and the identity references that are attributed to the novels in their subtitles. The different features cover contextual aspects (period, authorship, publication place, thematic genre, literary current) as well as textual ones (aspects of the setting, narrative perspective, and stylistic features).

[6] Of course the thematic subgenres and literary currents that the novels have been associated with by literary historians and by contemporaries are not only text-external characteristics, in the sense of generic conventions applied to the texts, but relate to text-internal features. They are listed here as aspects that are external to the texts because the genre labels were collected in that way – not by analyzing the contents and style of the novels, but by gathering genre attributions that have been made by others.

[7] "Setting" refers to the geographic location or time (including the historical period) of a narrative and can also be designated as the "story world".

[8] The time period of the setting is determined in relation to the publication years of the novels' first edition. "Past" means that the narrated time is more than 60 years before the publication date, "recent past" that it is between 30 and 60 years before the publication date, and "contemporary" that the narrated time is within 30 years before or after the publication date or that it is not marked at all.

[9] The term "narrative perspective" refers to the point of view that the narrator of a story has. This can be, for example, an autodiogetic perspective, which means that the narrator is at the same time the protagonist and tells from her or his perspective, or a homodiegetic perspective, which means that the narrator holds the view of a minor character, or a heterodigetic perspective, which corresponds to an omniscient narrator. For the purpose of this analysis, autodiegetic and homodiegetic narrators are subsumed under the category "first person" and heterodiegetic narrators under "third person" because the difference in linguistic style (first person vs. third person form) mattered most for the corpus analysis.

With this range of features, various aspects of the novels are covered and related to the question of identity construction. The content and style of the novels are not analyzed here directly in terms of identity, though, but in relation to the explicit identity references found in subtitles with which the novels were marked on the historical editions. These are understood as signals or indicators of identity functions that the novels had.

The third aspect that needs to be explained regarding the methodology of the analysis is how correlations between identity references and other features of the novels are determined and interpreted. For all the information that is encoded as metadata – publication years, countries, thematic subgenres, literary currents, continent and time period of the setting, and narrative perspective – it is examined whether certain identity references occur relatively more frequent for some of the metadata categories than for others. The kinds of questions asked are for example: Is the share of "novelas originales" in the bibliography higher in some decades of the 19th century than in others? Do "novelas mexicanas" have a higher proportion of a setting in the recent past than novels that do not bear this kind of identity label? The results of these comparisons of the proportion of novels associated with certain identity labels and textual categories can only be interpreted as tendencies. As a rule of thumb, each time that there is a difference of 5 % or more in the proportion of novels with a certain identity label, it is considered for the results of the analysis here. The differences can however not be expected to be statistically significant in all cases because the number of novels in *Bib-ACMé* and *Conha19* is not very high in statistical terms, especially if only one third of the novels carry identity references and subgroups of them are analyzed.

For the stylistic analyses, which identify words that are preferred or avoided by novels carrying certain identity references, contrastive analyses are performed with the tool *stylo* (Eder et al., 2016). In a contrastive analysis, the corpus is divided into two partitions to determine the differences between the two subsets of the corpus. In this way, stylistic features of the sub-corpora can be worked out, in the form of words that are preferred or avoided in one sub-corpus over the other, which means that certain words that are overrepresented can be interpreted as distinctive or characteristic for the sub-corpus in question. A contrastive measure that is implemented in *stylo* and used for the analyses here is Craig's Zeta (Craig & Kinney, 2009).[10]

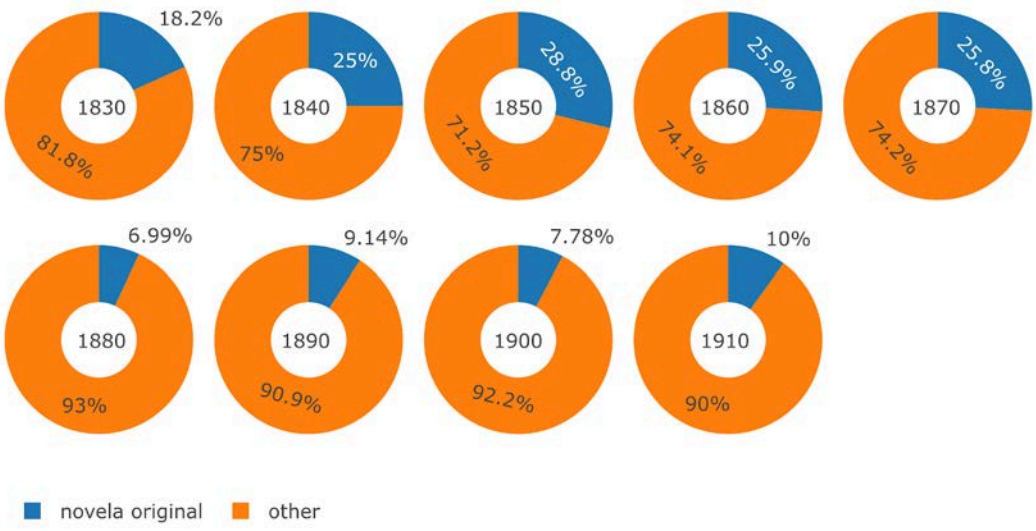
In the following, the analysis focuses on those identity references that occur most frequently or form the largest groups. First, the most frequent type of identity reference in *Bib-ACMé* and *Conha19*, the "novelas originales", are examined, followed by all the references that can be grouped as "novelas americanas". Finally, the three groups of references referring to the three countries, "novelas mexicanas", "novelas argentinas", and "novelas cubanas", are analyzed. Thus,

[10] Craig's Zeta is a further development of the contrastive measure Burrow's Zeta (cf. Burrows 2007). Both measures provide interpretative lists of words that are distinctive for sub-corpora because they focus on words from the middle frequency spectrum (neither the most common nor the rarest words). Craig's Zeta is used here because it provides both words preferred and avoided in the sub-corpus that is contrasted with another one.

emphasis is placed on linguistic identity, on the question of a trans- and supranational Spanish American identity, and on national identities.[11]

2.2. Novelas originales

The first aspect of the novels that carried the label "novela original" that is examined here is their distribution over time. How many percent of the novels published in each decade between 1830 and 1910 had this kind of reference in their subtitles? The results are shown in Fig. 4 below:



4 | "Novelas originales" vs. other novels by decade

Up to the 1870s, the proportion of "novelas originales" ranges between approximately one fifth and one fourth of the novels. In contrast, from the 1880s onwards, only 10 % or less of the novels carry that kind of identity reference. The "novela original" clearly is a phenomenon of the earlier decades in the 19th century. Apparently, the need to mark the linguistic originality of the novels, i.e., that they were originally written in Spanish and are not translations of foreign works, was not as strong anymore when more and more novels were published in the Spanish American countries around 1880.

If we look at the shares of novels with the label "novela original" in the novels from the three countries Mexico, Argentina and Cuba, we notice that they are lower for Argentina (6 % below average) than for the other two countries.[12] This observation is probably related to the fact that there are more Mexican and Cuban novels in the early decades of the 19th century and because the number of novels from Argentina only increases later in the century, when the label "novela original" was not used so much anymore.

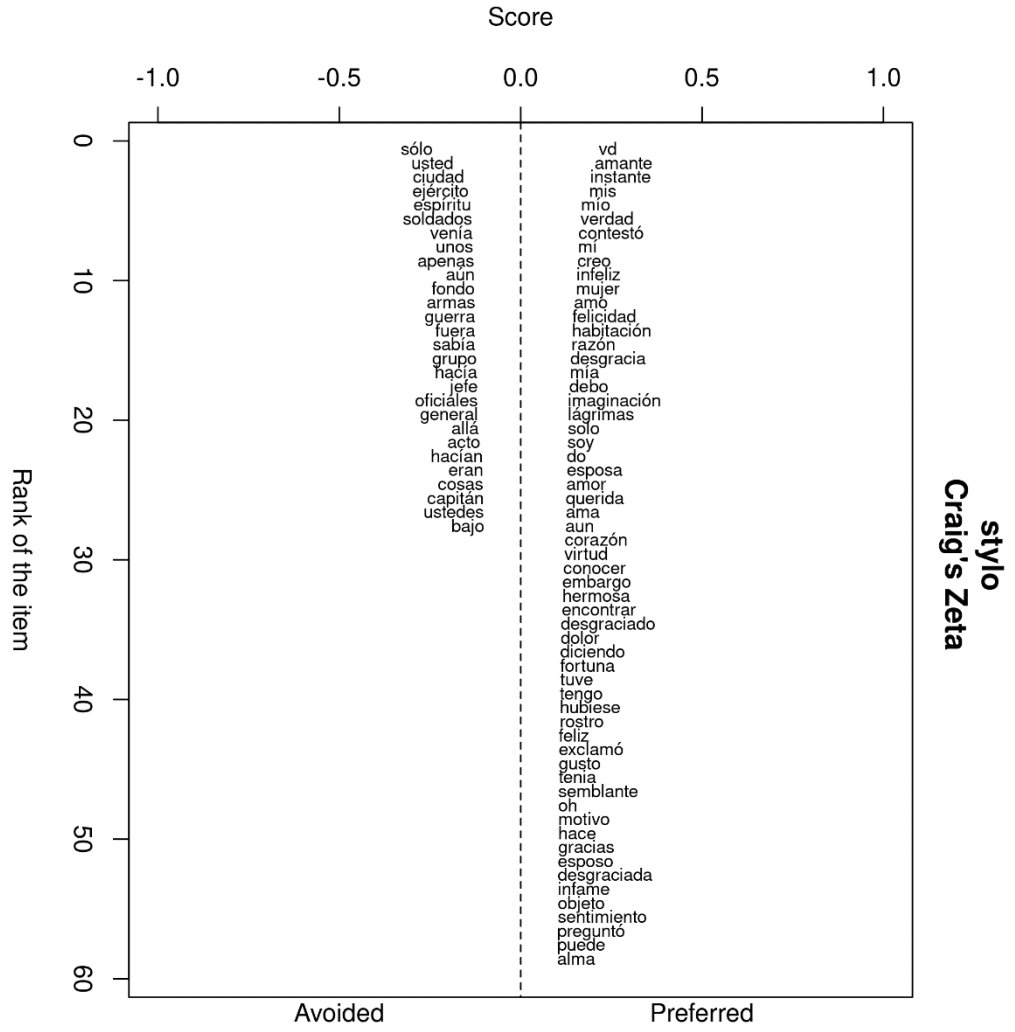
[11] The scripts that were used to perform the analyses and their results are published on GitHub (cf. Henny-Krahmer 2022).

[12] Charts which show the numbers mentioned in this article are available at <https://github.com/hennyu/original_american_romtag21/tree/main/images> 27.2.2022. All the percentages that are given in this article are rounded.

Turning to the thematic subgenres, "novelas originales" are overrepresented in sentimental novels (7 % above average) and underrepresented in novels of customs ("novelas de costumbres", with -5 %), with the other thematic subgenres ranging in between these values. Both sentimental novels and novels of customs can be found in the whole 19th century, so the different proportion of "novelas originales" in the novels attributed to these two thematic subgenres is an interesting finding. One hypothesis to explain this result is that the sentimental novel as a subgenre was also very common in non-Spanish speaking countries, so here it was necessary to signal the local origin, while the "novelas de costumbres" were a subgenre typical of novels from Spain and Spanish America, so perhaps in the case of these novels it was not so urgent to indicate the own, linguistic identity. Regarding the literary currents, the biggest proportion of "novelas originales" is found in the group of romantic novels (+6 %), while the proportions are much smaller for realist novels (-6 %) and naturalistic novels (-12 %). This again confirms that the "novelas originales" occur primarily in the early 19th century because the realist and naturalistic currents begin to dominate around 1880.

The next step is to analyze the results for those text features that are only available for the novels in the *Conha19* corpus, that is, for a subset of all novels. For the continent of the setting, it can be stated that the "novelas originales" are clearly overrepresented in the group of novels with a European setting (+28 %), which suggests that the custom to label novels as "novelas originales" has a European origin. This assumption is supported by the observations made by Botrel (2001, paragraph 12) and could be tested by further empirical investigation of the occurrence of this subtitle in nineteenth century novels from Spain. An examination of the temporal setting also reveals interesting differences, because "novelas originales" with a setting in the past are underrepresented (-11 %) in comparison to novels with a setting in contemporary times or the recent past. This may be correlated to the number of sentimental novels with that label, as these are usually not set in the past, in contrast to historical novels. Finally, for the narrative perspective, there is no difference between novels written in first person or in third person with regards to the share of "novelas originales".

In addition to the various metadata related to text-external and internal factors, textual features themselves will now be analyzed in terms of their relationship to the identity reference "novela original." Fig. 5 illustrates the words preferred and avoided by novels carrying the label "novela original", when compared to all other novels in the corpus.



5 | Words avoided and preferred in “novelas originales”

The words that are underrepresented in “novelas originales” can be related to historical themes, as there are several words from the word fields of fight and army: “ejército”, “soldados”, “armas”, “guerra”, “jefe”, “oficiales”, “general”, and “capitán”. Furthermore, there are several words in perfect tense: “venía”, “sabía”, “hacía”, “hacían”, “eran”. The words that are overrepresented in “novelas originales” can be associated with a sentimental theme: “amante”, “instante”, “infeliz”, “amo”, “felicidad”, “habitación”, “desgracia”, “lágrimas”, “esposa”, “amor”, etc. The stylistic analysis of the “novelas originales” confirms the tendency of the novels to belong to the sentimental subgenre and not to the historical one.

2.3. Novelas americanas

After the novels with the label “novela original”, we will now look at those that have an identity reference that points to the American continent, either in a general form as in “novela americana” or “novela criolla”, or in a form that is specific for an individual Spanish American country, region, city, or a certain cultural or indigenous group, for instance “novela mexicana”, “novela tapatía”, or “novela azteca”. In this section, these novels are however analyzed from a quantitative point of view and

therefore the different individual labels are grouped by country or for the whole continent.[13]

Looking at the development of the American labels over time, there is no clear tendency as for the "novelas originales". Only novels which carry an explicit label referring to the Argentine context tend to have been published in the later decades of the 19th century. The first novels with such an identity reference appeared in the 1860s and in the 1900s there were eight novels of this kind. So this trend must be interpreted with caution because the overall number of novels carrying a label from the group "novela argentina" is low. Besides that, it was already mentioned that the number of Argentine novels is in general lower in the first half of the 19th century when compared to Mexican and Cuban novels.

To examine the "novelas americanas" by country makes most sense for the group as a whole because obviously "novelas mexicanas" were always Mexican, "novelas argentinas" always Argentine, and "novelas cubanas" Cuban. Still, the proportion of novels that carried a national identity label may vary for each country. Of the Cuban novels, 33 % (14 % above average) had an American identity reference and 30 % a particular Cuban one. On the other hand, only 12 % (7 % below average) of the Argentine novels had a reference as "novela americana", of which 7 % referred specifically to the Argentine context. The novels from Mexico have average proportions in this regard. Apparently, in the case of the Cuban novels it was most important to signal the American and especially Cuban identity by means of explicit references in subtitles of the novels, which might be related to the long struggle for independence of the Cuban colony, which lasted until the end of the 19th century. In contrast, when the number of novels from Argentina increased, the country had been independent for more than half a century. In quantitative terms, the Mexican novel developed earlier than the Argentine one. At the same time, Mexico became independent much earlier than Cuba. Both aspects might explain the average number of American (and Mexican) references to identity in the subtitles of Mexican novels.

When the thematic subgenres are concerned, "novelas de costumbres" are most overrepresented for novels with a label of the type "novelas americanas" (+29 %), while political (-11 %) and sentimental novels (-10 %) are most underrepresented in the same group of novels. This is in contrast to the "novelas originales". So, an explicit indication of Spanish American identity seems to be related to the depiction of local customs in the novels, as opposed to novels with a political or a "classic" sentimental theme of European origin. If one considers the identity labels of the individual countries in connection with thematic subgenres and how they differ from one country to another, the "novelas mexicanas" have a comparatively high share of historical novels (+11 %), while the "novelas argentinas" and "novelas cubanas" do not occur in political novels (0 %, each) and less frequently in historical ones (-2 % and -3 %, respectively). The numbers of explicit "novelas argentinas" and

[13] As for the "novelas originales", all the results can be found in the GitHub repository mentioned above (Henny-Krahmer 2022).

"novelas cubanas" are however quite low, so that it is hard to speak of any trends in these cases.

As to the literary currents, the highest proportion of novels with explicit references to an American identity are realist novels (+11 %). The same holds for the "novelas mexicanas" (+9 %), the "novelas argentinas" (+5 % for the realist and equally +5 % for the naturalistic novels), but not for the "novelas cubanas", for which the realist current is slightly underrepresented (-3 %). So unlike the "novelas originales", the novels with references to American identities tend to belong to the realist current.

With regard to those text features that were only collected for *Conha19*, it can be stated that the continent of the setting is American for all the novels whose subtitles include identity references to the American continent or to individual Spanish American countries. Among the different time periods of the setting, the "novelas americanas" are overrepresented in the group of novels set in the recent past (+9 %), but this is only due to the "novelas mexicanas" and "novelas cubanas", because there are no "novelas argentinas" with that kind of temporal setting. This can be interpreted to mean that in Spanish American novels, coming to terms with one's recent past was an important aspect in the process of identity constitution.

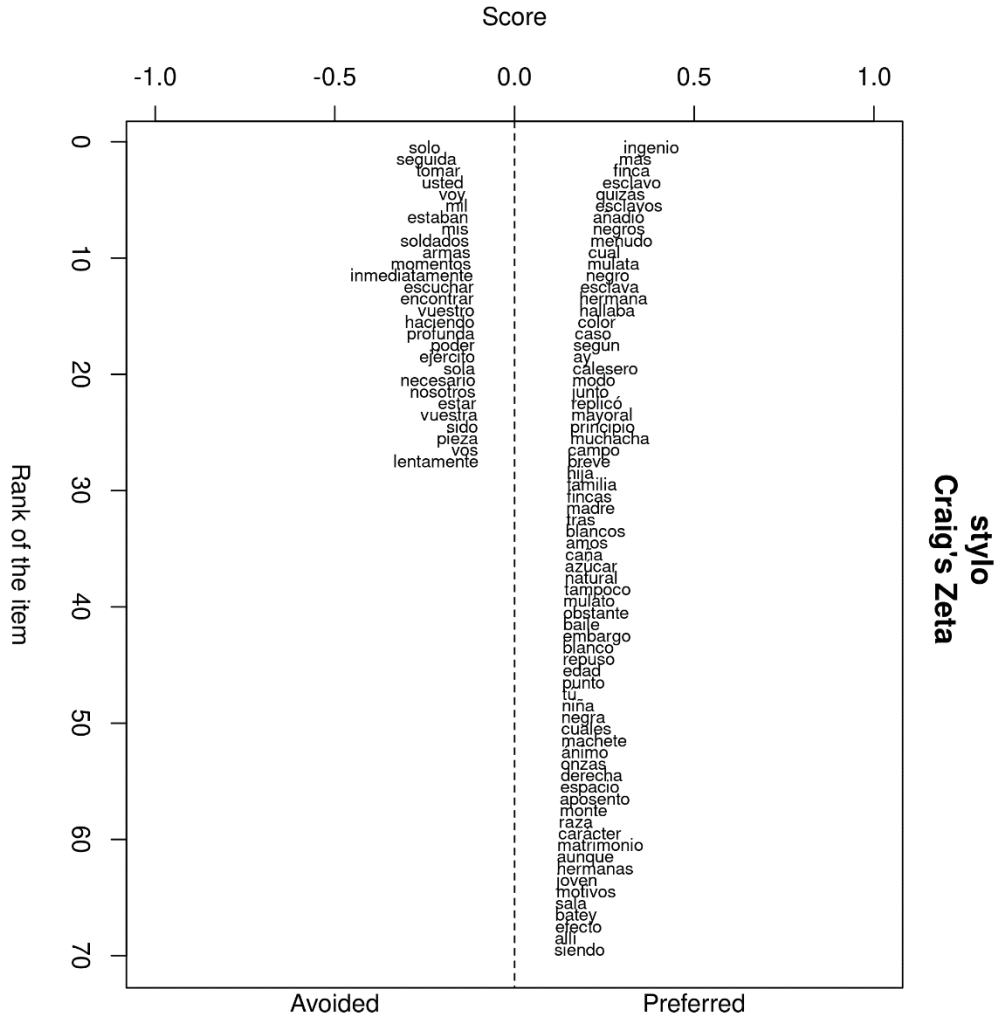
For the connection between identity references that refer to the Spanish-American space and the narrative perspective of the novels, the following observations can be made: for the "novelas americanas", narrations in first person are under-represented (-9 %), which is also the case for "novelas mexicanas" (- 10 %) and "novelas cubanas" (- 7 %), but not for the "novelas argentinas". So, except for the "novelas argentinas", for which no trend deviating from the average is visible, the novels with American identity references tend to avoid first person narratives in favor of narrations in third person. The representation and constitution of identity in these novels is thus more strongly connected to a perspective that is external to the view of individual characters, which fits well with their tendency to be part of the realist current and to depict local customs as opposed to the romantic and sentimental novels in which narrations in first person are not so rare.

Also, for the "novelas americanas" and the novels with national identity references, the words that are preferred and avoided in them when compared to the group of other novels in contrastive analyses were analyzed. For the group of "novelas americanas" as a whole, the words preferred are mainly forms of address and specific words designating people: "señor", "usted", "señora", "muchacha", "niña", "caballero", "hermana", and "padre". Furthermore, there are some adverbs and an interjection that are preferred by the "novelas americanas" ("conque", "luego", and "ay"), as well as the verb forms "vamos" and "comenzó". The only words avoided are "solo" and "hacia", but this result may be related to orthographic issues in the full texts of the novels, because the form "sólo" with accent appears in the list of words preferred and "hacia" may occur as a form where the accent is missing because it was not used in historical spelling of the verb form "hacía". These results can be interpreted to mean that the novels with explicit references to American identity are stylistically distinguished from the other novels primarily by linguistic characteristics, especially by specific forms of address, that do not occur in this way in the group of other novels. On the other hand, the "novelas americanas" basically

do not show any words that are underrepresented, which means that beyond the overrepresented words, they operate with the same vocabulary as the other novels in the corpus. However, there is no thematic vocabulary that all the "novelas americanas" have in common, in contrast to the "novelas originales". This means that there is not the one common American theme associated with these novels, so that one has to look for characteristic themes in the individual national novels.

The lists of words preferred and avoided by the "novelas mexicanas" shows that terms that can be associated with sentimental novels because they are used to describe physical features of characters ("expresión", "labios", "semblante", "belleza", "espíritu", "bella", "sonrisa"), dialogue and interaction ("añadió", "mirar", "repuso", "miró"), or a sentimental theme ("dolor"), are under-represented. On the other side, terms that can be related to the depiction in particular of rural customs are overrepresented in the "novelas mexicanas" ("negocio", "sombrero", "pesos", "caballero", "pueblo", "hacienda", "multitud", "justicia", "lance"), as are some specific verb forms (e.g. "comenzó", "vaya", "voy", "gritó"). These results correspond to the observation that "novelas mexicanas" are more often associated with the subgenre "novelas de costumbres" than with sentimental novels.

Among the words preferred by "novelas argentinas", above all distinctive forms of address and nouns for people stick out: "tío", "tía", "misia", "viejo", "madre", "hermana", "marido", "joven", "hermano", "familia", "niña", "niño", "mamá", "hijita", "jóvenes", "padre", "papá", "pareja", "vieja", "doctor", and "padrino". The list of words avoided, on the other hand, is not so easy to interpret and is made up of verbs, adjectives, adverbs, nouns, and conjunctions. Nouns that are under-represented are for example "pueblo", "justicia", "crimen", "poder", and "dinero". So, the novels with Argentine identity references seem to have a quite specific linguistic style – in so far as a whole list of words from different grammatical categories is avoided – and they also avoid some specific content words. On the other side there are distinctive forms of address and mentions of people in them, but the words preferred do not reveal any specific themes. Finally, the words preferred and avoided by "novelas cubanas" are shown in Fig. 6 below:



6 | Words avoided and preferred in “novelas cubanas”

In the list of words preferred by “novelas cubanas”, some terms turn up that can be associated with novels of customs and in particular description of plantation settings and slavery, for instance “ingenio”, “finca”, “esclavo”, “negros”, “mulata”, “calesero”, “mayoral”, “azúcar”, “machete”, “monte”, “batey”, so a specific Cuban theme becomes visible here. Some of the words avoided can be attributed to a military theme, which would be typical of historical novels: “soldados”, “armas”, “poder”, “ejército”. As was seen before, the “novelas cubanas” are underrepresented in the subgenre of historical novels. In addition, there are also some other verb forms, adverbs, and pronouns which are avoided in the novels with Cuban identity references.

2.4. Synthesis: identity types of novels

In Tab. 2 depicted below, the results of the comparison of proportions of novels by period, country, thematic subgenre, literary current, continent and time period of the setting, narrative perspective as well as words preferred and avoided are summarized for all the different kinds of identity references that were analyzed, i.e. the “novelas originales”, “novelas americanas”, “novelas argentinas”, “novelas

mexicanas", and "novelas cubanas". The "+" sign is used to mark all the aspects that are overrepresented in the novels with a certain identity reference when compared to their average proportion in the bibliography and corpus. Correspondingly, the "-" sign serves to indicate that the identity references are underrepresented for the extra-textual or text-internal feature in question. An "=" sign means that there is no difference in one direction or the other. For the three kinds of national novels, the "novelas mexicanas", "novelas argentinas", and "novelas cubanas", those values that deviate from the "novelas americanas" in general are highlighted in red to stress the differences between these three types of national identity references.

identity reference / property	*novela original*	*novela americana*	*novela mexicana*	*novela argentina*	*novela cubana*
period	+early	=	=	+late	=
country	-Argentina	+Cuba =Mexico -Argentina	+Mexico	+Argentina	+Cuba
thematic subgenre	+sentimental -*costumbres*	+*costumbres* -political -sentimental	+*costumbres* +historical -sentimental	+*costumbres* -political -historical	+*costumbres* -political -historical
literary current	+romantic -realist -naturalistic	+realist	+realist	+realist +naturalistic	-realist
continent of setting	+Europe	+America -Europe	+America -Europe	+America -Europe	+America -Europe
time period of setting	-past	+recent past	+recent past	-recent past	+recent past
narrative perspective	=	-first person	-first person	=	-first person +third person
words preferred / avoided	+sentimental -historical	+forms of address	+*costumbres* -sentimental	+people (forms of address) -mixed	+*costumbres* (plantation) -military

Tab. 2 | Overview of identity types of novels

In summary, the "novelas originales" can be characterized as novels which tend to have been published early in the 19th century, and which tend to be Mexican and Cuban novels with a sentimental theme. They belong mostly to the romantic current, tend to have a European setting and to treat events that are contemporary or set in the recent past. No preference for a specific narrative perspective can be found in them, and the words preferred by these novels also point in the direction of sentimental novels.

Compared to that, the "novelas americanas" occur in the whole 19th century, are inclined towards novels of customs and the realist current. They obviously all have an American setting and the narrated events belong preferably to the recent past. In these novels, first person narrators tend to be avoided and stylistically, they are marked by specific forms of address. In several regards, they are quite the opposite as the "novelas originales".

Looking closer at the national variants, the “novelas mexicanas” have a tendency towards historical themes, the “novelas argentinas” occur rather lately, and are connected to the naturalistic current. Futhermore, they are not so often set in the recent past, rather in comtemporary times or a more remote past. First person narratives are not avoided by the “novelas argentinas”. The “novelas cubanas”, on the other hand, tend to avoid historical themes and are less inclined towards the realist current. The words preferred by these novels show specific Cuban themes.

3. Conclusions

The digital analysis of identity references in subtitles of Spanish American 19th century novels revealed that different types of novels show up when the various types of identity labels are analyzed in connection with a range of extra-textual and text-internal features. The “novela original” tends to be more European and in the tradition of romances than the novels with American identity references. Furthermore, it was found out that each type of national novel, that is, novels with explicit references to a national context in their subtitle, has its own characteristics. There is not the one common genre, theme, or style that serves to represent, express, and form national identity, but there are forms that are specific for each case, a result which corresponds with the general considerations from memory studies. Several tendencies of the novels with identity references could be worked out, whereby one can assume that these references were not just arbitrarily assigned by authors or editors. Still, the varying proportions and distinctive words cannot be understood as statistically significant differences, just as quantitative inclinations or trends, because in the whole, the share of novels with explicit identity references is rather low.

A meaningful extension or variation of the analysis presented here would be to analyze the identity references on the level of individual editions instead of literary works as more abstract categories. Like that, also the places of publications of individual editions and the publishing houses and editors behind them could be taken into account. This would, however, require to model the data about identity references on the level of the individual editions and also to identify all the publishers and editors of the novels, a work which still has to be done.

Beyond a quantitative analysis of the most frequent identity references and the types of novels that they are linked to, it could also be fruitful to take a look at the whole spectrum of different identity labels and to analyze individual novels by close reading. The comprehensive overview of identity references in the bibliography and corpus given here can be seen as a good starting point for such a more in-depth, qualitative analysis that can take into account more and other kinds of Spanish American novels than the ones most often analyzed in terms of linguistic, cultural, or political identity.

References

BOTREL, Jean-François. 2001. “La novela, género editorial (España, 1830-1930).” In *La novela en España (siglos XIX-XX), ed. Aubert, Paul,* 35–51.

Madrid: Casa de Velázquez.
<http://books.openedition.org/cvz/2631>.
BRUSHWOOD, John S. 1966. *Mexico in its Novel. A Nation's Search for Identity.* Austin: University of Texas Press.
BURROWS, John. 2007. "All the Way Through: Testing for Authorship in Different Frequency Strata." *Literary and Linguistic Computing* 22 (1): 27–47.
<https://dx.doi.org/10.1093/llc/fqi067>.
CLiGS. n.d. *CLiGS – Computational Literary Genre Stylistics.*
<https://cligs.hypotheses.org/sprachen/english> 27.2.2022.
CRAIG, Hugh & Arthur F. Kinney. 2009. *Shakespeare, Computers, and the Mystery of Authorship.* New York: Cambridge University Press.
EDER, Maciej et al. 2016. "Stylometry with R: a package for computational text analysis." *R Journal* 8 (1): 107–121.
ERLL, Astrid. 2017. *Kollektives Gedächtnis und Erinnerungskulturen: Eine Einführung.* Stuttgart: J. B. Metzler.
FERRER, José Luis. 2018. *La invención de Cuba: Novela y nación (1837-1846).* Madrid: Editorial Verbum.
FLEISHER Feldman, Carol. 2001. "Narratives of national identity as group narratives: Patterns of interpretive cognition." In *Narrative and Identity. Studies in Autobiography, Self and Culture*, ed. Brockmeier, Jens & Donal Carbaugh, 129–144, Amsterdam, Philadelphia: John Benjamins.
GÁLVEZ, Marina. 1990. *La novela hispanoamericana (hasta 1940).* Madrid: Taurus.
HANWAY, Nancy. 2003. *Embodying Argentina: Body, Space and Nation in 19th Century Narrative.* Jefferson, North Carolina, London: McFarland.
HENNY-KRAHMER, Ulrike (ed.). 2017a. *Bib-ACMé. Bibliografía digital de novelas argentinas, cubanas y mexicanas (1830-1910).* Versión 1.2. Github.com.
<https://github.com/cligs/bibacme> 27.2.2022.
HENNY-KRAHMER, Ulrike (ed.). 2017b. *Bib-ACMé. Bibliografía digital de novelas argentinas, cubanas y mexicanas (1830-1910).*
<http://bibacme.cligs.digital-humanities.de> 27.2.2022.
HENNY-KRAHMER, Ulrike (ed.). 2021. *Corpus de novelas hispanoamericanas del siglo XIX (conha19)*. Versión 1.0.1. Github.com.
<https://github.com/cligs/conha19> 27.2.2022.
HENNY-KRAHMER, Ulrike. 2022. *Novelas originales y americanas. Data and Scripts.* Github.com.
<https://github.com/hennyu/original_american_romtag21> 27.2.2022.
HORATSCHEK, Anna-Margaretha. 2013. „Identität, kollektive." In *Metzler Lexikon Literatur- und Kulturtheorie: Ansätze – Personen – Grundbegriffe*, ed. Nünning, Ansgar, 323–324, Stuttgart, Weimar: J. B. Metzler.
LINDSTROM, Naomi. 2004. *Early Spanish American Narrative.* Austin: University of Texas Press.
MOLINA, Hebe Beatriz. 2011. *Como crecen los hongos. La novela argentina entre 1838 y 1872.* Buenos Aires: Teseo.
ORTEGA, Julio. 2011. "Introducción." In *La búsqueda perpetua: lo propio y lo universal de la cultura latinoamericana, vol. 3*, *La literatura hispanoamericana,* ed. Vega Armijo, Mercedes de, 17–21. México: Secretaría de Relaciones Exteriores, Dirección General del Acervo Histórico Diplomático.
RAIBLE, Wolfgang. 1980. „Was sind Gattungen? Eine Antwort aus semiotischer und textlinguistischer Sicht." *Poetica 12*, 320–349.
REAL ACADEMIA ESPAÑOLA (ed.). 2021a. "suriano, na." *Diccionario de la lengua española (DLe).*
<https://dle.rae.es/suriano> 27.2.2022.
REAL ACADEMIA ESPAÑOLA (ed.). 2021b. "tapatío, a." *Diccionario de la lengua*

española (DLe).
<https://dle.rae.es/tapatío> 27.2.2022.
SCHAEFFER, Jean-Marie. 1983. *Qu'est-ce qu'un genre littéraire?* Paris: Seuil.
SOMMER, Doris 1993. *Foundational Fictions. The National Romances of Latin America.* Berkeley, Los Angeles: University of California Press.
TEI CONSORTIUM (ed.). 2022. *Guidelines for Electronic Text Encoding and Interchange.*
<http://www.tei-c.org/P5/> 27.2.2022.

Abstract

Relationships between literary texts, identity constitution, and genre are explored in this digital analysis of 19th century Spanish American novels from Mexico, Argentina, and Cuba, of their subgenres, and their function in the formation of collective identities, starting from references to identity which were found in the subtitles of the novels. In particular, the label "novela original", as well as identity references that can be subsumed under the terms "novela americana", "novela mexicana", "novela argentina", and "novela cubana" are analyzed. It is found that each type of identity novel, that is, novels with explicit references to a linguistic, cultural, or national context in their subtitle, has its own characteristics. There is not the one genre, theme, or style that serves to represent and constitute identity, but there are forms that are specific for each case, a result which corresponds with findings from memory studies on group-defining stories.

Zusammenfassung

In der vorliegenden digitalen Analyse hispanoamerikanischer Romane des 19. Jahrhunderts aus Mexiko, Argentinien und Kuba werden Zusammenhänge zwischen literarischen Texten, Identitätskonstitution und Gattungen untersucht. Berücksichtigt werden die Untergattungen der Texte sowie ihre Funktion für die Herausbildung kollektiver Identitäten. Hierzu werden Identitätsbezüge in den Untertiteln der Romane als Ausgangspunkt herangezogen. Insbesondere werden die Bezeichnung als "novela original" sowie Identitätsbezüge durch die Begriffe "novela americana", "novela mexicana", "novela argentina" und "novela cubana" analysiert. Es lässt sich feststellen, dass jede Art von Identitätsroman, d. h. Romane, die sich im Untertitel explizit auf einen sprachlichen, kulturellen oder nationalen Kontext beziehen, spezifische Eigenschaften aufweist. Es gibt keine einheitliche Gattung, Thema oder Stil, um Identität zu repräsentieren oder zu konstituieren; vielmehr weist jeder Fall spezifische Formen auf. Dieses Ergebnis deckt sich mit Ergebnissen aus der Gedächtnisforschung (*memory studies*) zu identitätsstiftenden Erzählungen.

2022, n°9
pp. 37-56
doi: https://doi.org/10.15460/apropos.9.1875

Laura Hernández-Lorenzo

La prosa de Gustavo Adolfo Bécquer en los límites de la poesía : análisis estilométrico[1]

Laura Hernández-Lorenzo
disfruta de un contrato postdoctoral Margarita Salas en la Universidad de Sevilla.
lhernandez1@us.es

Palabras clave

Gustavo Adolfo Bécquer – Estilometría – géneros literarios – prosa – poesía

> Sucede […] lo que con otras muchas cosas del mundo, en que todo es cuestión de la distancia a que se miran, y la mayor parte de las veces, cuando se llega a ellos, la poesía se convierte en prosa.
>
> G. A. Bécquer

1. Introducción

Una de las disciplinas de las Humanidades Digitales que pueden ser de sumo interés para establecer una Romanística digital transdisciplinaria es la Estilometría. Aunque desde sus orígenes ha predominado su aplicación a problemas de atribución de autoría, en los últimos años han aumentado los trabajos que muestran su potencial para el análisis de otras cuestiones estilísticas, en lo que se ha llamado la Estilometría más allá de la autoría (*Stylometry Beyond Authorship Attribution*). Dentro de esta última línea, destacan los estudios estilométricos de géneros literarios en diferentes tradiciones literarias, en los que se han obtenido resultados de gran interés, que suponen un aliciente para estudios posteriores (Jannidis & Lauer 2014; Schöch 2018). En el caso de la literatura española, este tipo de estudios son menos comunes y solo se han aplicado hasta el momento en la evaluación y detección de géneros literarios de la novela de la Edad de Plata (Calvo Tello 2021) y a la poesía del Siglo de Oro (Navarro 2018, Hernández-Lorenzo 2022).

[1] Para la realización de este trabajo he contado con financiación del proyecto *Large-Scale Text Analysis and Methodological Foundations of Computational Stylistics*, financiado por el *National Science Center of Poland* (SONATA-BIS 2017/26/E/HS2/01019). Además, este estudio le debe mucho a mi Trabajo de Fin de Máster, dirigido por María Victoria Utrera, a quien agradezco sus consejos y comentarios. Me gustaría expresar también mi agradecimiento a los editores de este número por sus valiosas sugerencias para mejorar mi trabajo.

En el presente estudio, estos métodos se aplican al análisis del género literario en las obras literarias de un único autor: Gustavo Adolfo Bécquer (1836-1870), siguiendo la estela de estudios anteriores (Jannidis & Lauer 2014; Calvo Tello 2019). Además de ser uno de los poetas y narradores más influyentes de todos los tiempos, Bécquer es considerado el gran precursor del poema en prosa en la literatura española, el cual cultivó, junto a la prosa poética, en algunas de sus *Leyendas*, al tiempo que acometía la renovación de la prosa.[2] Aunque por lo general la crítica ha aceptado la clasificación de obras becquerianas en prosa poética y poema en prosa que propone Luis Cernuda (1975), muchos añaden otros textos en los que creen ver o bien prosa poética o bien poema en prosa. La aplicación de métodos estilométricos a la obra becqueriana tiene como objetivo principal determinar qué textos se encuentran más cerca del género lírico (medido por la proximidad o similitud estilística con respecto a la poesía en verso escrita por el propio Bécquer) y cuáles se encuentran más cerca del género narrativo. Al mismo tiempo, se pretende explorar si es posible realizar mediante estas técnicas una gradación de los textos en prosa becquerianos hacia lo lírico. Por último, el presente trabajo constituye una muestra de cómo puede contribuir la estilometría al análisis de géneros literarios limítrofes, como es el caso del poema en prosa, confirmando o refutando con métodos objetivos las clasificaciones realizadas por la crítica literaria.

La estructura de este artículo es la siguiente: después de esta introducción (sección 1), se presenta el estado de la cuestión del papel de Bécquer como precursor del poema en prosa (sección 2), seguido de una descripción del corpus utilizado y de su preparación (sección 3). A continuación, se expone la metodología empleada (sección 4) y se detallan los resultados obtenidos (sección 5). Finalmente, se presentan las conclusiones del estudio (sección 6).

2. Gustavo Adolfo Bécquer como precursor del poema en prosa

2.1. Prosa poética y poema en prosa

El poema en prosa y la prosa poética son considerados modalidades limítrofes entre la narrativa y la lírica, caracterizados por la ruptura entre las fronteras tradicionales del verso y la prosa mediante la inclusión de elementos poéticos en el formato prosístico. Si acudimos a la bibliografía sobre el poema en prosa, ambos términos aparecen con frecuencia, y se utiliza prosa poética para un eslabón o modalidad intermedia entre la prosa narrativa y el poema en prosa, en la que el componente narrativo va diluyéndose hasta llegar al género del poema en prosa, sin suponer la desaparición de la misma:

> […] debe asociarse la prosa poética a un determinado "modo de escritura" especificado por la figuración del lenguaje (repeticiones, imágenes, formas rítmicas, etc.), susceptible de realizarse en géneros diversos (novela, cuento, libro de viajes, ensayo, etc.), y de afectar tanto a obras enteras como a párrafos, por lo que caracteres como la "unidad" son ajenos a

[2] Para una distinción entre prosa poética y poema en prosa, véase la sección 2.1 de este trabajo.

> su dominio. El poema en prosa, en cambio, además de constituir un "género literario" independiente y diferenciado, participa de una particular morfología y, por qué no, de una "arquitectura" intrínseca (Agudo 2004, 223).[3]

Atendiendo a esta delimitación y a cómo se aplica concretamente al caso de Bécquer, se puede distinguir la prosa poética por ser una modalidad en la que lo lírico aparece desde el principio supeditado a lo narrativo; por tanto, lo poético actúa como un adorno que embellece y sublima la narración. Los recursos retóricos aparecen diseminados a lo largo del texto, de modo que la concentración de los mismos es menor. En cambio, el poema en prosa se caracteriza por su brevedad y mayor concentración de recursos retóricos. Además, tiene un carácter más fragmentario, pues aparece inserto en una narración en la que supone un corte.

2.2. Los orígenes del poema en prosa

Para una parte importante de la crítica, los orígenes del género del poema en prosa se encuentran en el movimiento romántico y en la revolución que este supuso (Cernuda 1975; Aullón de Haro 1979; Utrera 1999; Agudo Ramírez 2004). [4] En este sentido, el romanticismo puso de manifiesto la fragilidad de los géneros literarios como casillas estancas y, en su búsqueda de un nuevo lenguaje, inició un camino de experimentación que continúa hasta nuestros días. Esto unido a la importancia adquirida por la poesía lírica, y al auge y la renovación de la prosa,[5] favoreció la creación de nuevos géneros como el poema en prosa, la prosa poética o el verso libre. En el caso del poema en prosa, se han señalado sus orígenes en las tradiciones literarias germánica y francesa. De acuerdo con algunos estudiosos, la aparición del poema en prosa en Alemania es anterior, siendo los *Himnos a la noche* (1800) de Novalis la obra inaugural del género[6] (Aullón de Haro 1979; Utrera 1999). Aunque posterior, es fundamental la aparición en Francia de traducciones en prosa de textos originariamente compuestos en verso, normalmente adaptaciones de cantos folclóricos extranjeros, que mantienen el aliento lírico de la composición

[3] En términos similares se expresa Aullón de Haro, que resalta la extensión y unidad del texto, junto a la intencionalidad del autor como factores diferenciadores de ambas modalidades: "El proceso de conformación del poema en prosa no es en modo alguno desconectable de la llamada prosa poética, eslabón intermedio que lo une y aleja a un tiempo de otros tipos de prosa. Quiero decir, el uno es impensable sin la existencia de la otra. Ahora bien, qué cosa sea prosa poética y qué cosa poema en prosa no parece susceptible de delimitación sino refiriéndonos a extensión del texto y unidad del mismo. Otro factor es el que implica la intencionalidad del autor, ya que, en efecto, dentro de una narración poética existe la posibilidad de aislar fragmentos que puedan plenamente considerarse como poemas en prosa" (Aullón de Haro 1979, 109).

[4] Algunos autores prefieren situar el nacimiento de este género en una época más tardía, pues consideran que el movimiento modernista es esencial para la aparición del género (Jiménez Arribas 2009). Sin pretensión de menospreciar la relevancia del Modernismo para la construcción del poema en prosa, estoy de acuerdo con los otros estudiosos en que los intentos de los autores del Romanticismo español, aunque no constituyen poemas en prosa tan consolidados, son claros antecedentes del género.

[5] La poesía lírica se convirtió en el género predilecto de los románticos por permitir la expresión de los sentimientos y de la interioridad, y actúa como lámpara que alumbra las profundidades del ser (Meyer Abrams 1975). Entre los textos que promueven este género, destacan la *Defensa de la poesía* de Shelley (1999) y las opiniones de autores como Novalis, Hegel o los Schegel (Novalis 1994). Al mismo tiempo, la prosa experimenta un auge como expresión de lo verdadero y por permitir una mayor libertad frente a las normas constreñidas del verso. De acuerdo con la crítica, este auge no puede separarse del proceso de renovación de la misma con el movimiento romántico (Berenguer 1974).

[6] También se destaca la importancia de las canciones, baladas y *lieds* (Utrera 1999).

original y que generaron una moda por imitar este género también en la prosa[7]. Para Aullón de Haro destaca la aparición en 1842 del *Gaspard de la Nuit* de Aloysius Bertrand, que supone la «aparición del poema en prosa francés como género» (Aullón de Haro 1979, 112), consagrado posteriormente por Charles Baudelaire (*Petits Poèmes en Prose*, 1869), Arthur Rimbaud (*Les Illuminations*, 1886) y Stéphane Mallarmé en distintos poemas en prosa en sus obras.

2.3. Bécquer y el poema en prosa

En España, destaca el nombre de Gustavo Adolfo Bécquer como precursor clave del género, cuyo legado influiría en obras emblemáticas del poema en prosa español, como *Diario de un poeta recién casado* de Juan Ramón Jiménez u *Ocnos* de Luis Cernuda (Utrera 1999). De este modo, Bécquer, autor destacado del Romanticismo tardío por considerarse el comienzo de la poesía moderna en español con sus *Rimas*, aparece como un antecedente fundamental del proceso de renovación de la prosa en España gracias a sus *Leyendas*. Berenguer considera que llevó a cabo «uno de los intentos españoles más curiosos de prosa poética, nuevo o casi nuevo en el medio literario de su tiempo y patria» (Berenguer 1974, 3), y para Cernuda «es Gustavo Adolfo Bécquer quien adivina en España la necesidad de la poesía en prosa[8] y quien responde a ella y le da forma en sus *Leyendas*»[9] (Cernuda 1975, 984). Estos intentos becquerianos[10] están relacionados con su papel de renovador del verso y de la prosa (Cernuda 1975) y por su gusto por la experimentación, consecuencia de su pasión de explorar[11] (Pageard 1995). Para otros autores, otros

[7] A algunos de estos emergentes poemas en prosa se les da el nombre de *fragments*, que remite a la brevedad y fragmentación propia del género lírico. Además, puede señalarse una tradición de prosa poética en Francia en las obras de Jean-Jacques Rousseau o François-René de Chateaubriand (Utrera 1999).

[8] Para Cernuda, Bécquer tuvo conocimiento de la moda francesa de las traducciones y decidió crear algo similar en español (Cernuda 1975). Efectivamente, de las investigaciones de Robert Pageard ha resultado que Bécquer aprendió francés en el Colegio de San Telmo (Pageard 1990), a lo cual debemos sumar que su madrina, Manuela Monnehay, era francesa. La crítica destaca especialmente la posible influencia del *Smarra ou les Démons de la Nuit* (1821) de Charles Nodier como antecedente del género (Cernuda 1975). Además, pudo verse influido por la tradición germana del poema en prosa por tres vías diferentes: la primera, la aparición de numerosas traducciones al español de cuentos y composiciones de los Grimm, los Heine o Hoffmann con anterioridad a 1860 en publicaciones periódicas; la segunda, su amistad con Augusto Ferrán, quien probablemente entró en contacto con el poema en prosa durante su estancia en Alemania y se encargó de difundirlo a su vuelta a España mediante traducciones en prosa de Heine; y la tercera, la aparición de traducciones de autores alemanes en prosa en *El Museo Universal*, publicación con la que Bécquer estuvo vinculado y colaboró a menudo, llegando a ser director de la sección literaria de la revista en 1866 (Aullón de Haro 1979; Pageard 1990). Por último, se ha señalado la influencia de la tradición de la leyenda en el Romanticismo español.

[9] Las *Leyendas* son relatos fantásticos escritos por Bécquer como resultado de sus viajes por toda España recuperando las tradiciones y leyendas populares. En ellas tiene un papel clave el elemento sobrenatural.

[10] Para Cernuda y otros autores, se trata de intentos de prosa poética y poemas en prosa, pues la narración predomina sobre lo lírico (Cernuda 1975; Utrera 1999).

[11] Para Cernuda, la renovación de la prosa que emprende Bécquer está relacionada con la que lleva a cabo en el verso: «Eso concierta por otra parte con lo que ya sabemos era su intención al escribir verso: atentar contra la conformación clásica o clasicista del mismo, haciéndolo más flexible, convertirlo en instrumento adecuado de lo que quería hacer y decir con él. Así, paralelamente a como aproxima el verso a la prosa, trata también de acercar la prosa al verso, no para escribir una prosa poética, sino para hacer de la prosa instrumento efectivo de la poesía» (Cernuda 1975, 987). Y Pageard describe en los siguientes términos cómo la continua experimentación y la búsqueda de una forma de expresión adecuada (también conocido como el problema de la inefabilidad becqueriana) marcan la obra de Bécquer: «El primero es la necesidad en Bécquer de

factores como la búsqueda de lo ideal o la importancia del elemento sobrenatural también influyen en la abundancia de lo lírico y en la conformación del poema en prosa becqueriano (López Castro 2002; Foss 2010).

2.4. Clasificaciones de poemas en prosa y prosa poética en la obra de Bécquer

Los textos en prosa de Bécquer más cercanos a lo lírico han sido clasificados en dos categorías: poemas en prosa y prosa poética. La clasificación más aceptada y valorada por la crítica es la realizada por el poeta Luis Cernuda (1975), que puede resumirse en los siguientes términos:

1. Textos escritos en prosa poética. Para Cernuda, tres narraciones breves de Bécquer están escritas íntegramente en prosa poética. Se trata de *El caudillo de las manos rojas*, *La Creación* y *Creed en Dios*. Los subtítulos de estas *Leyendas*, además, las presentan como si fueran traducciones o adaptaciones de un texto extranjero, siguiendo la moda de las traducciones generada en Francia[12].
2. Textos escritos como poemas en prosa. Para Cernuda, se encuentran ensayos parciales de poemas en prosa en *La ajorca de oro*, *La corza blanca*, *El gnomo*, *Las hojas secas* y *El caudillo de las manos rojas*. En *La ajorca de oro*, el poema en prosa comprendería los tres primeros párrafos de la *Leyenda*, mientras que en *La corza blanca* se localizarían en los dos coros de voces. En el caso de *El gnomo*, el diálogo a cuatro voces entre el agua, el viento, Marta y Magdalena constituiría «si no un poema en prosa, un fragmento de poesía en prosa: una prosa libre» (Cernuda 1974, 993). Los poemas en prosa insertos en *El caudillo de las manos rojas* se encontrarían en los dos cantares de Siannah, el primero, épico (que queda interrumpido) y el segundo, «La vuelta del combate», lírico. Por último, Cernuda nos dice que en *Las hojas secas* no es posible aislar un fragmento como poema en prosa, pues el tono poético flota a lo largo de toda la composición, creando una atmósfera lírica.

Aunque la mayoría de los autores aceptan la clasificación de Cernuda, como es el caso de Aullón de Haro (1979) o León Felipe (1999), muchos añaden otros textos en los que creen ver o bien prosa poética o bien poema en prosa. Dentro de las

someterse a una fuerza violenta, penosa, y, por consiguiente, breve, para dar forma social (empleaba la palabra "vestir") a las creaciones de su imaginación; designaba esos potentes esbozos, más o menos complejos, por el vocablo de "ideas". El segundo factor [...] era una tendencia constante a la novedad y variedad a partir tanto de la experiencia personal como de las impresiones procedentes de la cultura escrita y oral; hasta el fin de su corta vida [...] siempre buscó Bécquer nuevas formas de expresión. [...] De la combinación de ambos factores –necesaria brevedad y búsqueda de algo nuevo- nació un arte fragmentario, animadísimo, de fugacidad en todos los dominios, muy sorprendente para los lectores y observadores de los años 1860 y 1870» (Pageard 1995, 17-18).

[12] Los subtítulos de estas *Leyendas* son: *El caudillo de las manos rojas (Leyenda India)*, *La Creación (Poema Indio)* y *Creed en Dios (Cantiga provenzal)*. Como puede comprobarse, además de aparecer como traducciones y resaltar el componente exótico, en algunos de los subtítulos se incide en la cercanía a lo lírico de los textos (*La Creación* es presentada como poema indio y *Creed en Dios* como cantiga provenzal).

Leyendas, los cuatro primeros párrafos de *La promesa* y la segunda descripción del concierto musical en *Maese Pérez el organista* han sido señalados como fragmentos poéticos en prosa por Pascual Izquierdo, mientras que Jesús Rubio califica a *La venta de los gatos* como texto cercano al poema en prosa[13] (Izquierdo 1995; Rubio Jiménez 2006).

3. Preparación del corpus

Para llevar a cabo este estudio, los textos de Bécquer han sido recopilados de diversas fuentes, entre las que destacan la Biblioteca Virtual Miguel de Cervantes[14] y algunas webs educativas[15]. A la hora de preparar el corpus se han tenido en cuenta las siguientes cuestiones:

1. Puesto que el presente estudio se centra en textos literarios, se ha recogido la poesía en verso de Bécquer, contenida en las *Rimas*, y sus narraciones breves, entre las que se encuentran las *Leyendas*. Por esta misma razón, no se incluyen textos programáticos y de teoría literaria, como las *Cartas literarias a una mujer*, ni textos histórico-arqueológicos, como la *Historia de los templos de España*. Tampoco se ha incluido el *Apólogo* debido a sus enormes similitudes temáticas con *La Creación*, leyenda con la que comparte incluso los mismos personajes, lo cual habría influido en los resultados.
2. Los fragmentos considerados poemas en prosa por Cernuda se han separado de las *Leyendas* de las que procedían, agrupándose en un único texto («Poemas en prosa»). En consecuencia, para evitar ruido, estas *Leyendas* no se incluyen en el corpus (*La corza blanca*, *La ajorca de oro* y *El gnomo*).
3. Debido a su breve extensión, los poemas en verso se aglutinan y reparten en dos archivos («Rimas1» y «Rimas2»).

El corpus resultante está constituido por 24 textos becquerianos, cuyos datos se encuentran recogidos en la Tabla 1. Los textos pueden también consultarse en el repositorio GitHub del estudio: <https://github.com/lamusadecima/Becquer-corpus>.

[13] Otros críticos también han señalado posibles poemas en prosa o prosa poética en el *Apólogo*, la *Historia de los templos de España*, la serie de *Pensamientos* o las *Cartas desde mi celda* (Montalvo 1995; Pageard 1995; López Castro 2002).

[14] Web: <http://www.cervantesvirtual.com/portales/gustavo_adolfo_becquer/>. [consulta: 25/12/2021].

[15] Especialmente se ha utilizado la web Lit2Go del *Florida Center for Instructional Technology*: <https://etc.usf.edu/lit2go/49/obras-de-gustavo-adolfo-becquer-tomo-primero/> [consulta: 25/12/2021].

Título	Colección	Clasificación según Cernuda	Extensión (*tokens*)
Creed en Dios	Narración breve	Prosa poética	3662
Desde mi celda	Cartas que incluyen narraciones	-	35790
El beso	Narración breve	-	4762
El caudillo de las manos rojas	Narración breve	Prosa poética	10506
El Cristo de la calavera	Narración breve	-	3885
El maestro Herold	Narración breve	-	1648
El miserere	Narración breve	-	3336
El monte de las ánimas	Narración breve	-	2740
El rayo de luna	Narración breve	-	3564
Es raro	Narración breve	-	3361
La Creación	Narración breve	Prosa poética	2274
La cruz del diablo	Narración breve	-	6161
La cueva de la mora	Narración breve	-	2024
La mujer de piedra	Narración breve	-	3273
La promesa	Narración breve	-	3440
La rosa de pasión	Narración breve	-	3076
La venta de los gatos	Narración breve	-	4128
Las hojas secas	Narración breve	Constituye un poema en prosa en su totalidad	1293
Los ojos verdes	Narración breve	-	2684
Maese Pérez	Narración breve	-	5211
Poemas en prosa ("Poema_en_prosa-fragmentos")	Fragmentos de narraciones	Ensayos de poemas en prosa. Fragmentos que proceden de las leyendas *La corza blanca*, *La ajorca de oro, El gnomo* y *El caudillo de las manos rojas*	1508
*Rimas*1	Poesía	-	4032
*Rimas*2	Poesía	-	3083
Tres Fechas	Narración breve	-	6489

Tab. 1 | Lista de las obras literarias de Bécquer que constituyen el corpus de estudio.

4. Metodología

Para abordar el poema en prosa y la prosa poética en la obra literaria becqueriana, en este estudio se emplean las metodologías de los enfoques cuantitativos en Humanidades Digitales (Hoover 2008). Estos se han visto reforzados en los últimos años gracias a la emergencia de las teorías de la lectura distante (Moretti 2013) y

el macroanálisis (Jockers 2013), y en el último *Companion* se ubican parcialmente dentro de la línea de investigación de minería de textos (Jockers y Underwood 2016). Dentro de estas metodologías, se emplean específicamente la Estilometría y el Análisis de redes.

Desde sus orígenes, la principal aplicación de la Estilometría ha sido dilucidar la autoría de obras de atribución dudosa o anónimas, con notables éxitos, entre los que destacan el caso de los *Federalist papers* (Mosteller y Wallace 1964) o el caso Rowling (Juola 2015). Sin embargo, contamos con algunos precedentes de estudios estilométricos que analizan cuestiones estilísticas no relacionadas con la atribución de autoría, como el trabajo pionero sobre la cronología de las obras de Platón realizado por Lutosławski (1897). En los últimos años han aumentado las aplicaciones estilométricas al estudio de corrientes literarias y artísticas (Burrows 2003), la influencia del sexo del autor (Rybicki 2016) o el género literario. En el caso de este último, destacan los estudios estilométricos de las obras de los románticos alemanes (Jannidis y Lauer 2014), el análisis computacional de novelas de diferentes géneros (Underwood 2016), y los trabajos realizados dentro del proyecto CLiGS, con aplicaciones de la Estilometría al género literario en el teatro clásico francés (Schöch 2018) y a la narrativa española e hispanoamericana de finales del siglo XIX y principios del XX (Calvo Tello 2019; Calvo Tello 2017; Henny-Krahmer 2018). Los buenos resultados obtenidos por todos estos estudios suponen un aliciente para el análisis estilométrico del género literario en otras obras y autores, y especialmente para el caso de la literatura española, en la que las aplicaciones estilométricas son aún escasas.

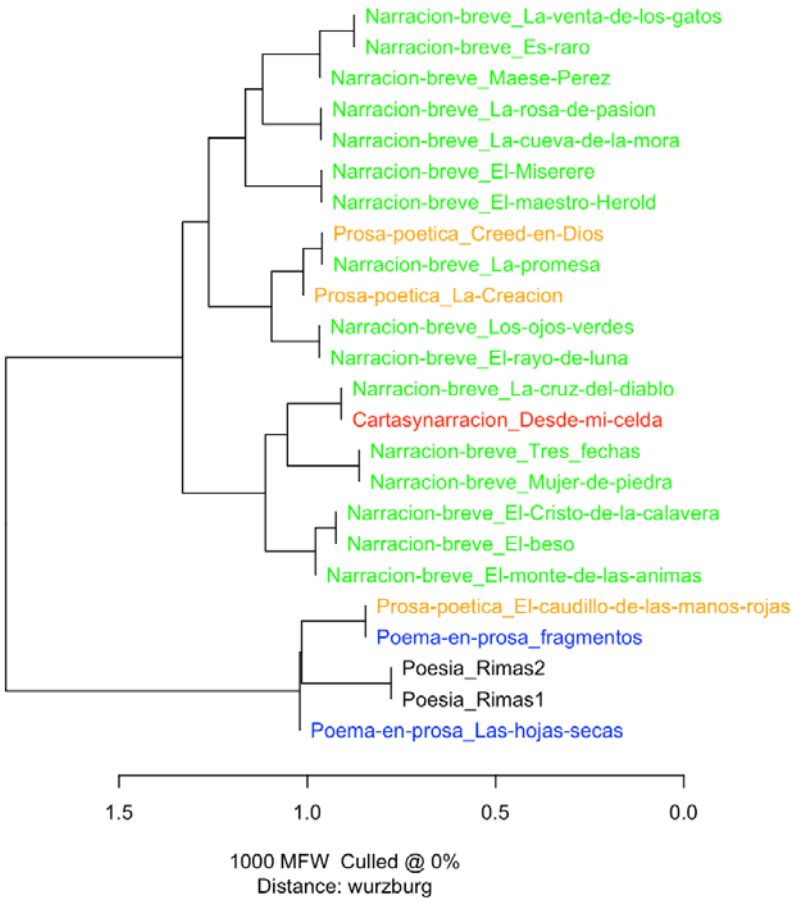
Con el objetivo de realizar un primer acercamiento y exploración de lo que pueden aportar los métodos computacionales al estudio de la progresión de los textos hacia lo lírico en la obra becqueriana, en el presente estudio se emplean métodos estilométricos no supervisados. Estos nos permitirán observar qué textos se agrupan más cerca de la poesía lírica en verso (nuestro referente del máximo grado de presencia de lo lírico) sin predefinir qué textos están escritos en prosa poética y cuáles son poemas en prosa o narraciones con un menor componente lírico, como un primer resultado que pueda ser corroborado en estudios posteriores mediante métodos supervisados. Más concretamente, se utilizan algunos de los métodos empleados en la bibliografía estilométrica mencionada con buenos resultados (Jannidis y Lauer 2014), como los análisis de grupos o *clustering* (sección 5.1) y los árboles de consenso (sección 5.2). En cuanto a las medidas de distancia y parámetros elegidos, se ha aplicado Cosine Delta (Smith y Aldridge 2011) sobre las 100 palabras más frecuentes en el caso del análisis de grupos, y sobre las 100 a 1000 palabras más frecuentes, en el caso del árbol de consenso. Se ha elegido esta medida por ser la que produce mejores resultados a nivel general e independientemente del corpus utilizado (Evert et al. 2017; Ochab et al. 2019). Los valores de parámetros elegidos tienen la ventaja de que el análisis estará restringido fundamentalmente a las palabras función e incluyen también, en el caso del árbol de consenso, una cantidad suficiente de iteraciones, por lo que se asegura la estabilidad de las conexiones textuales obtenidas. Estos análisis han sido llevados a cabo mediante el paquete *stylo* de R (Eder et al. 2016).

Además, se presenta una red estilométrica (sección 5.3), obtenida a partir del *output* del árbol de consenso siguiendo la implementación de Eder (2017) a través del programa Gephi (Bastian et al. 2009). Este método se utiliza para visualizar y analizar las relaciones estilísticas y estilométricas entre los textos, puesto que la red se apoya en más conexiones que el árbol de consenso original, lo cual permite apreciar matices y conexiones estilísticas no relacionadas con la autoría. Con el objetivo de distinguir cuáles son los textos que presentan mayores similitudes estilísticas, se aplica como procedimiento de detección de comunidades el cálculo de la Modularidad siguiendo el algoritmo de Lovain (Blondel et al. 2008). Se trata de uno de los más utilizados con este fin (Newman 2010) y, al encontrarse implementado en Gephi, los resultados pueden ser reproducidos fácilmente por otros investigadores.

5. Resultados

5.1. Análisis de grupos

El análisis de grupos (*clustering*) obtenido del corpus becqueriano se encuentra recogido en la Figura 1. Como puede observarse en la misma, se distinguen dos grupos principales: un primer grupo en la parte superior que contiene casi todos los relatos y leyendas, y un segundo grupo en la parte inferior que incluye la poesía lírica en verso (las *Rimas*), *Las hojas secas*, los ensayos de poemas en prosa según Cernuda y la leyenda de *El caudillo de las manos rojas*.



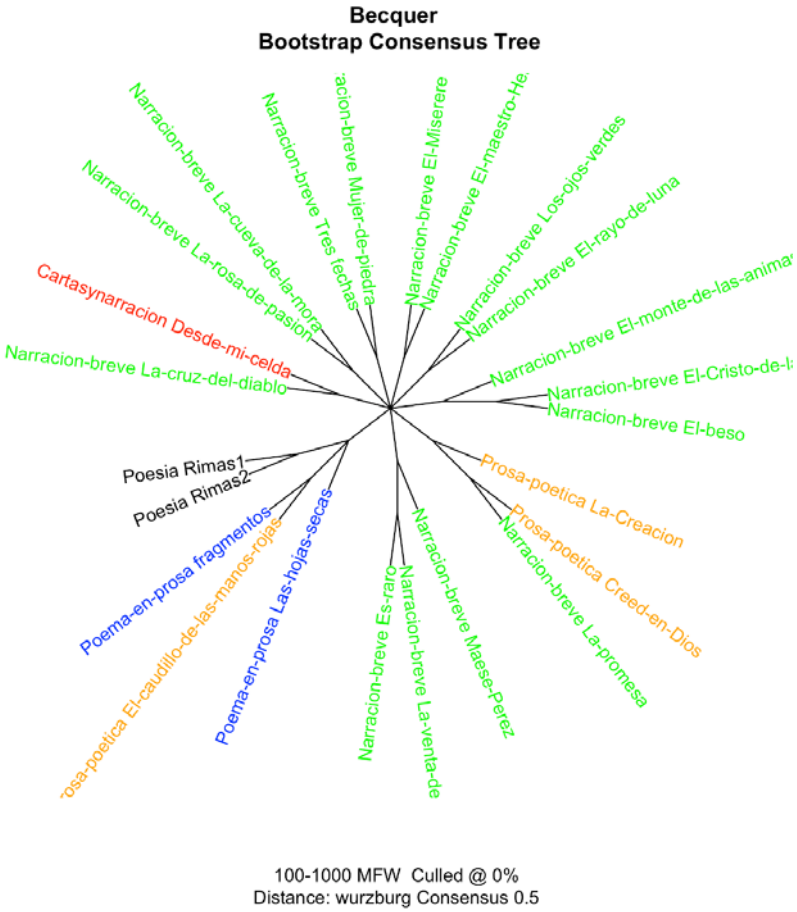
1 | Análisis de grupos del corpus becqueriano recogido en este estudio. Se ha realizado mediante el paquete *stylo* aplicando Cosine Delta sobre las 1000 palabras más frecuentes.

En el primer grupo, destaca la aparición de un subgrupo o *cluster* que contiene las leyendas de *La promesa*, *Creed en Dios*, *La Creación*, *Los ojos verdes* y *El rayo de luna*. Pero resulta más llamativo el segundo grupo, ya que concuerda con la opinión de Cernuda sobre la proximidad de estos textos al género lírico. Para realizar una aproximación a los mecanismos estilísticos que están detrás de la diferenciación de los textos en poema en prosa frente al resto de relatos becquerianos, se ha examinado la tabla de frecuencias empleada en este análisis[16]. En esta se han podido distinguir mayores frecuencias relativas de nombres en los textos líricos, los poemas en prosa y en la leyenda de *El caudillo de las manos rojas* (que forman el segundo grupo de la figura anterior) frente a las narraciones becquerianas. Por contrario, las obras más cercanas a lo lírico presentan menores frecuencias relativas de formas verbales en comparación con el resto de los textos becquerianos. Esto sugiere que los textos más líricos favorecen el uso de sustantivos, mientras que los verbos son menos frecuentes.

5.2. Árbol de consenso

A continuación, se presenta el árbol de consenso (Eder 2013) obtenido (Figura 2). Como puede observarse, y en línea con lo que han señalado otros investigadores (Jannidis y Lauer 2014), el resultado parece menos interesante que el análisis de grupo para explorar cuestiones estilísticas más allá de la atribución de autoría. De todos modos, puede destacarse la aparición de una rama con la poesía lírica (las *Rimas*), *El caudillo de las manos rojas*, *Las hojas secas* y los fragmentos que según Cernuda constituyen poemas en prosa. Esta rama coincide con uno de los *clusters* detectados en el análisis de grupos, y puede interpretarse como la existencia de un gran componente lírico en estos relatos y en los supuestos poemas en prosa. Igualmente, puede resaltarse la aparición de otra rama con las leyendas *La Creación*, *La promesa* y *Creed en Dios*, que pone de manifiesto que existe una conexión muy estable entre estos textos.

[16] Esta puede consultarse en el repositorio GitHub de este estudio: <https://github.com/lamusadecima/Becquer-corpus>.



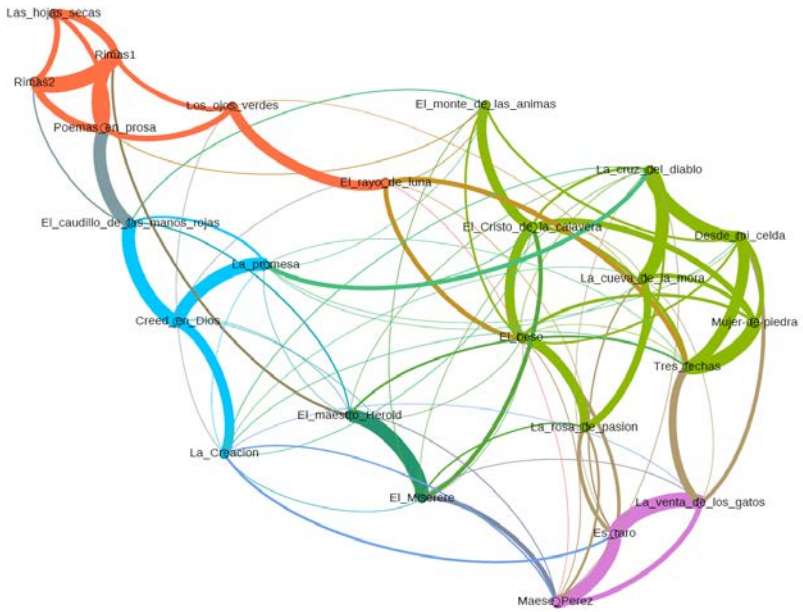
2 | Árbol de consenso del corpus becqueriano. Se ha realizado mediante el paquete stylo aplicando Cosine Delta sobre las 100 a 1000 palabras más frecuentes.

5.3. Red estilométrica

Seguidamente, las Figuras 3 y 4 contienen la red estilométrica obtenida tras importar el *output* del árbol de consenso en Gephi y aplicar el algoritmo Force Atlas 2 (Jacomy et al. 2014), tal y como recomienda Eder en su trabajo sobre esta implementación de Estilometría y Análisis de redes (Eder 2017). En la Figura 3 se ha aplicado el cálculo de la modularidad con una resolución de 1, lo cual ha dado lugar a cinco comunidades diferentes. La comunidad de color naranja contiene la poesía lírica (las *Rimas*), el texto de *Las hojas secas*, las leyendas *El rayo de luna* y *Los ojos verdes*, y los ensayos de poemas en prosa, que serían los textos más cercanos al género lírico. Seguidamente, la comunidad de color celeste incluye los textos que para Cernuda están escritos en prosa poética (*El caudillo de las manos rojas*, *Creed en Dios* y *La Creación*) junto a la leyenda *La promesa*. La tercera comunidad, en color verde oscuro, es la más pequeña y contiene solo dos textos, *El maestro Herold* y *El Miserere*, leyendas becquerianas donde la música tiene un papel protagonista[17]. Se aprecia otra pequeña comunidad en color violeta, que

[17] Resulta llamativo que *Maese Pérez el organista*, leyenda muy relacionada también con la música, no se encuentre en esta comunidad, si bien parece estar fuertemente conectado con *El Miserere*, como se aprecia en los nodos de la red.

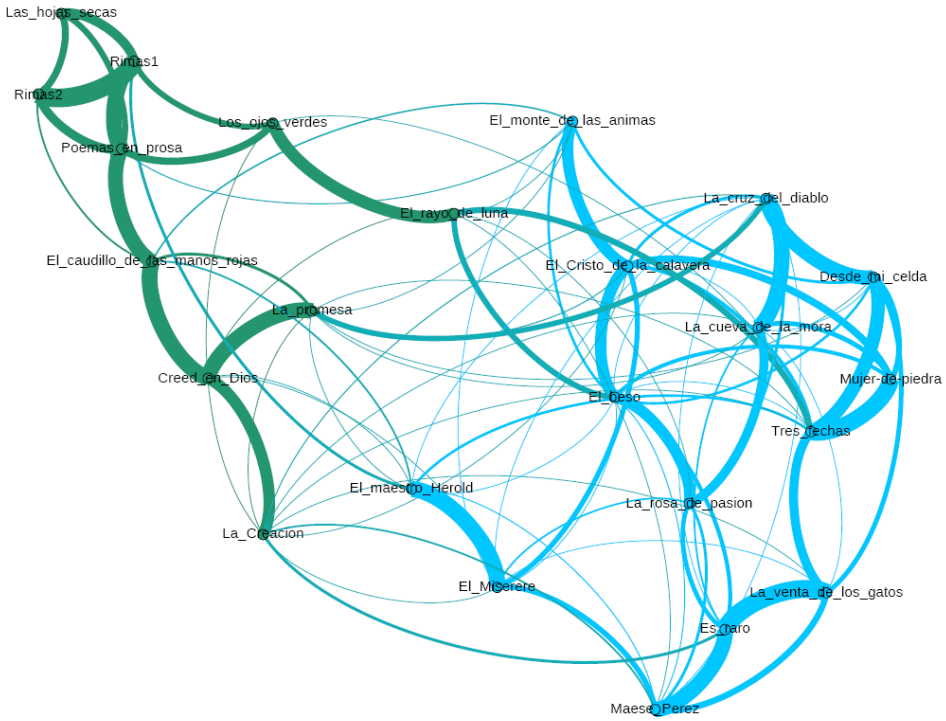
incluye la leyenda de *Maese Pérez* y las narraciones *Es raro* y *La venta de los gatos*, textos que destacan por contener un importante elemento costumbrista[18]. Por último, la comunidad de color verde claro abarca el resto de las leyendas y textos becquerianos, que parecen ser los más narrativos y alejados de lo lírico. Si se compara este resultado con los cuatro grupos (dos principales y uno de ellos que puede dividirse en tres) obtenidos en el análisis de grupos (Figura 1), se observa que en la red hay también un grupo o comunidad de poesía lírica junto a poemas en prosa (la comunidad naranja), otro de prosa poética (la comunidad celeste) y el resto corresponden a las narraciones breves menos cercanas a lo lírico.



3 | Red estilométrica del corpus literario becqueriano realizada mediante Gephi a partir del output del árbol de consenso de la Figura 2, y coloreada de acuerdo a las comunidades obtenidas en el cálculo de la Modularidad (con una resolución de 1).

Sin embargo, la distinción entre textos predominantemente líricos o narrativos es más clara al realizar el cálculo de la modularidad con una resolución de 2, para obtener menos comunidades. De esta forma, como puede observarse en la Figura 4, solo se obtienen dos: una comunidad de color verde con la poesía en verso (las *Rimas*), los supuestos ensayos de poemas en prosa, las leyendas en prosa poética según Cernuda (*El caudillo de las manos rojas*, *La Creación* y *Creed en Dios*) y los textos más cercanos al género lírico; y una comunidad de color celeste con el resto de los relatos y leyendas, que estarían más lejanos de lo lírico y contarían con un mayor componente narrativo.

[18] En el caso de *Maese Pérez*, destaca el parlamento de una señora en la iglesia con su vecina, repleto de expresiones populares.



4 | Red estilométrica del corpus literario becqueriano realizada mediante Gephi a partir del output del árbol de consenso de la Figura 2, y coloreada de acuerdo a las comunidades obtenidas en el cálculo de la Modularidad (con una resolución de 2).

Finalmente, la Tabla 2 recoge en forma esquemática la clasificación de cada texto según Cernuda y los resultados obtenidos en este estudio (principalmente, análisis de grupos y red estilométrica).

Título	Clasificación siguiendo a Cernuda	Análisis de grupos	Red estilométrica
Creed en Dios	Prosa poética	Grupo narrativo. Subgrupo de prosa poética	Resolución 1: comunidad de prosa poética
			Resolución 2: comunidad poética
Desde mi celda	Cartas que incluyen narraciones	Grupo narrativo. Tercer subgrupo narrativo	Resolución 1: comunidad narrativa general
			Resolución 2: comunidad narrativa
El beso	Narración	Grupo narrativo. Tercer subgrupo narrativo	Resolución 1: comunidad narrativa general
			Resolución 2: comunidad narrativa
El caudillo de las manos rojas	Prosa poética	Grupo poético	Resolución 1: comunidad de prosa poética
			Resolución 2: comunidad poética

Título	Clasificación siguiendo a Cernuda	Análisis de grupos	Red estilométrica
El Cristo de la calavera	Narración	Grupo narrativo. Tercer subgrupo narrativo	Resolución 1: comunidad narrativa general
			Resolución 2: comunidad narrativa
El maestro Herold	Narración	Grupo narrativo. Primer subgrupo narrativo	Resolución 1: comunidad narrativa relacionada con la música
			Resolución 2: comunidad narrativa
El miserere	Narración	Grupo narrativo. Primer subgrupo narrativo	Resolución 1: comunidad narrativa relacionada con la música
			Resolución 2: comunidad narrativa
El monte de las ánimas	Narración	Grupo narrativo. Tercer subgrupo narrativo	Resolución 1: comunidad narrativa general
			Resolución 2: comunidad narrativa
El rayo de luna	Narración	Grupo narrativo. Subgrupo de prosa poética	Resolución 1: comunidad más lírica
			Resolución 2: comunidad poética
Es raro	Narración	Grupo narrativo. Primer subgrupo narrativo	Resolución 1: comunidad narrativa costumbrista
			Resolución 2: comunidad narrativa
La Creación	Prosa poética	Grupo narrativo. Subgrupo de prosa poética	Resolución 1: comunidad de prosa poética
			Resolución 2: comunidad poética
La cruz del diablo	Narración	Grupo narrativo. Tercer subgrupo narrativo	Resolución 1: comunidad narrativa general
			Resolución 2: comunidad narrativa
La cueva de la mora	Narración	Grupo narrativo. Primer subgrupo narrativo	Resolución 1: comunidad narrativa general
			Resolución 2: comunidad narrativa
La mujer de piedra	Narración	Grupo narrativo. Tercer subgrupo narrativo	Resolución 1: comunidad narrativa general
			Resolución 2: comunidad narrativa

Título	Clasificación siguiendo a Cernuda	Análisis de grupos	Red estilométrica
La promesa	Narración	Grupo narrativo. Subgrupo de prosa poética	Resolución 1: comunidad de prosa poética
			Resolución 2: comunidad poética
La rosa de pasión	Narración	Grupo narrativo. Primer subgrupo narrativo	Resolución 1: comunidad narrativa general
			Resolución 2: comunidad narrativa
La venta de los gatos	Narración	Grupo narrativo. Primer subgrupo narrativo	Resolución 1: comunidad narrativa costumbrista
			Resolución 2: comunidad narrativa
Las hojas secas	Constituye un poema en prosa en su totalidad	Grupo poético	Resolución 1: comunidad más lírica
			Resolución 2: comunidad poética
Los ojos verdes	Narración	Grupo narrativo. Subgrupo de prosa poética	Resolución 1: comunidad más lírica
			Resolución 2: comunidad poética
Maese Pérez	Narración	Grupo narrativo. Primer subgrupo narrativo	Resolución 1: comunidad narrativa costumbrista
			Resolución 2: comunidad narrativa
Poemas en prosa ("Poema_en_prosa-fragmentos")	Ensayos de poemas en prosa. Fragmentos que proceden de las leyendas *La corza blanca*, *La ajorca de oro, El gnomo* y *El caudillo de las manos rojas*	Grupo poético	Resolución 1: comunidad más lírica
			Resolución 2: comunidad poética
Tres Fechas	Narración	Grupo narrativo. Tercer subgrupo narrativo	Resolución 1: comunidad narrativa general
			Resolución 2: comunidad narrativa

Tab. 2 | Clasificaciones de las *Leyendas* y textos narrativos becquerianos siguiendo a Cernuda y los resultados obtenidos en los análisis de este estudio (principalmente, del análisis de grupos y la red estilométrica).

6. Conclusiones y futuras líneas de investigación

En este estudio se han aplicado metodologías estilométricas y de análisis de redes a la obra literaria de Gustavo Adolfo Bécquer con el objetivo de determinar cuáles de sus textos se encuentran más cerca del género lírico (desde un punto de vista estilométrico) y si los resultados coinciden con la clasificación de poemas en prosa y prosa poética generalmente aceptada por la crítica. En este sentido, la clasificación de Cernuda queda reforzada, ya que los resultados obtenidos coinciden con esta a grandes rasgos:

En primer lugar, los supuestos ensayos de poemas en prosa insertos en algunas *Leyendas* becquerianas se encuentran siempre agrupados con su poesía en verso, las *Rimas*, y en la misma comunidad que estas, en el caso del análisis de redes. Lo mismo ocurre con *Las hojas secas*, considerado también por Cernuda como un poema en prosa en su totalidad. Estos textos representarían, por tanto, el mayor grado de proximidad al género lírico.

En segundo lugar, y en otro nivel, dos de las leyendas consideradas como escritas en prosa poética por Cernuda, *La Creación* y *Creed en Dios*, siempre se encuentran agrupadas juntas y en la misma comunidad. La tercera, *El caudillo de las manos rojas* aparece en el análisis de grupos y en el árbol de consenso en el mismo grupo y rama que las *Rimas*, muy posiblemente debido a que dos de los ensayos de poemas en prosa proceden de esta leyenda, pero en la red cae en la misma comunidad que las dos anteriores. Esto parece indicar que efectivamente se trata de textos cercanos al género lírico, pero en un grado menor que los poemas en prosa.

Además, en este estudio se realizan nuevas aportaciones, entre las que destaca cómo la leyenda *La promesa*, que ha sido señalada por Izquierdo como un posible texto escrito en prosa poética, se ubica en la misma comunidad que *La Creación*, *Creed en Dios* y *El caudillo de las manos rojas*, lo cual apoya la postura de este estudioso. En el caso de otros textos considerados por algunos estudiosos como poemas en prosa o prosa poética, como *Maese Pérez el organista* y *La venta de los gatos*, estos muestran algunas particularidades, junto a *Es raro*, frente a otros textos narrativos de Bécquer, cuya razón parece ser la inclusión de un lenguaje costumbrista y popular. Sin embargo, los tres textos se agrupan y aparecen con el resto de las *Leyendas* y narraciones cuando se reduce el número de comunidades, lo cual parece indicar que se encuentran más distantes del género lírico. Otra aportación de interés es el caso de *Los ojos verdes* y *El rayo de luna*, que parecen tener un mayor componente lírico que otras leyendas, lo cual abre nuevas vías de investigación para futuros trabajos. Desde el punto de vista de las metodologías utilizadas, los buenos resultados obtenidos ponen de manifiesto y refuerzan el interés de la Estilometría para el análisis de géneros literarios limítrofes, y concretamente, del poema en prosa en la literatura española.

Ahora bien, este es un estudio inicial de una investigación en marcha, por lo que serán necesarias nuevas investigaciones antes de llegar a conclusiones definitivas. Por cuestiones de espacio, en futuros trabajos se examinará con más detenimiento la señal lírica obtenida, se profundizará en los mecanismos estilísticos que están

detrás de los géneros del poema en prosa y la prosa poética (de los cuales se adelanta en este trabajo el mayor empleo de sustantivos en textos líricos), y se aplicarán métodos estilométricos supervisados y de clasificación al análisis de los textos literarios del poeta de las *Rimas*.

Bibliografia

ABRAMS, Meyer H. 1975. El espejo y la lámpara: teoría romántica y tradición crítica. Barcelona: Barral.

AGUDO RAMÍREZ, Marta. 2004. «La poética romántica de los géneros literarios: el poema en prosa y el fragmento. Situación europea y su especificación en España». Alicante: Universidad de Alicante. <http://hdl.handle.net/10045/10558>.

AULLÓN DE HARO, Pedro. 1979. «Ensayo sobre la aparición y desarrollo del poema en prosa en la literatura española». Analecta Malacitana 2 (1): 109-36.

BASTIAN, Mathieu, Sebastien Heymann, y Mathieu Jacomy. 2009. «Gephi: An Open Source Software for Exploring and Manipulating Networks». Third International AAAI Conference on Weblogs and Social Media, 361-62. <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154%5Cnpapers2://publication/uuid/CCEBC82E-0D18-4FFC-91EC-6E4A7F1A1972>.

BERENGUER CARISOMO, Arturo. 1974. La prosa de Bécquer. Sevilla: Secretariado de Publicaciones de la Universidad de Sevilla.

BLONDEL, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, y Etienne Lefebvre. 2008. «Fast unfolding of communities in large networks». Journal of Statistical Mechanics: Theory and Experiment 2008 (10). IOP Publishing: P10008. doi:10.1088/1742-5468/2008/10/p10008.

BURROWS, John. 2003. «Questions of Authorship Attribution and Beyond: A Lecture Delivered on the Occasion of the Roberto Busa Award ACH-ALLC 2001, New York». Computers and the Humanities 37 (1). Springer: 5-32. <http://www.jstor.org/stable/30204877>.

CALVO TELLO, José. 2017. «What does Delta see inside the Author?: Evaluating Stylometric Clusters with Literary Metadata». En III Congreso de la Sociedad Internacional Humanidades Digitales Hispánicas Sociedades, políticas, saberes: 153-61. Málaga. <http://hdh2017.es/wp-content/uploads/2017/10/Actas-HDH2017.pdf>.

CALVO TELLO, José. 2019. «Delta Inside Valle-Inclán: Stylometric Classification of Periods and Groups of His Novels». Romanische Studien. Theorien von Autorschaft und Stil in Bewegung: Stilistik und Stilometrie in der Romania, n.º 6: 151-64. <http://romanischestudien.de/index.php/rst/article/view/625/1320>.

CALVO TELLO, José. 2021. The Novel in the Spanish Silver Age. A Digital Analysis of Genre using Machine Learning. Bielefeld: Bielefeld University Press.

CERNUDA, Luis. 1975. «Bécquer y el poema en prosa español». En Prosa completa, 984-93. Barcelona: Barral.

EDER, Maciej. 2013. «Computational stylistics and Biblical translation: how reliable can a dendrogram be?» En The Translator and the Computer, editado por T. Piotrowski y L. Grabowski: 155-70. Wroclaw.

EDER, Maciej. 2017. «Visualization in stylometry: Cluster analysis using networks». Digital Scholarship in the Humanities 32 (1): 50-64.

EDER, Maciej, Jan Rybicki, y Mike Kestemont. 2016. «Stylometry with R: A Package for Computational Text Analysis». The R Journal 8 (1): 107-21. <https://journal.r-project.org/archive/2016/RJ-2016-007/index.html>.

EVERT, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström,

Christof Schöch, y Thorsten Vitt. 2017. «Understanding and explaining Delta measures for authorship attribution». Digital Scholarship in the Humanities 32 (June 2017): ii4-16.

Foss, Emily. 2010. «Síntesis de poesía y prosa, sueño y realidad: "la novela lírica" de Bécquer y Nerval». Espéculo. Revista de estudios literarios, n.º 44.
<https://webs.ucm.es/info/especulo/numero44/nlirica.html>.

Henny-Krahmer, Ulrike. 2018. «Exploration of Sentiments and Genre in Spanish American Novels». En Digital Humanities 2018 Puentes-Bridges, 399-403.
<https://web.archive.org/web/20190204111506/https://dh2018.adho.org/wp-content/uploads/2018/06/dh2018_abstracts.pdf>.

Hernández-Lorenzo, Laura. 2022. «Stylistic Change in Early Modern Spanish Poetry Through Network Analysis (with an Especial Focus on Fernando de Herrera's Role)». Neophilologus 106: 397-417.
doi:10.1007/s11061-021-09717-2.

Hoover, David L. 2008. «Quantitative Analysis and Literary Studies». En A Companion to Digital Literary Studies, editado por Ray Siemens y Susan Schreibman. Oxford: Blackwell.
<http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-6-9&toc.depth=1&toc.id=ss1-6-9&brand=9781405148641_brand>.

Izquierdo, Pascual. 1995. «Presencia de lo lírico, atmosférico y maravilloso en las Leyendas de Bécquer». En Bécquer. Origen y estética de la modernidad. Actas del VII Congreso de Literatura Contemporánea, Universidad de Málaga, 9, 10, 11 y 12 de noviembre de 1993, editado por Cristóbal Cuevas, 33-61. Málaga: Publicaciones del Congreso de Literatura Española Contemporánea.

Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann, y Mathieu Bastian. 2014. «ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software». PLoS ONE 9 (6): e98679.
doi:10.1371/journal.pone.0098679.

Jannidis, Fotis, y Gerhard Lauer. 2014. «Burrows's Delta and Its Use in German Literary History» En Distant Readings. Topologies of German Culture in the Long Nineteenth Century, editado por Matt Erlin y Lynne Tatlock, 29-54. Rochester: Camden House.

Jiménez Arribas, Carlos. 2009. «Apuntes para un breve estudio panorámico del poema en prosa en España e Hispanoamérica». Zurgai: Euskal herriko olerkiaren aldizkaria: Poetas por su pueblo, n.º 7: 6-9.

Jockers, Matthew L. 2013. Macroanalysis: Digital Methods and Literary History. Urbana: University of Illinois Press.

Jockers, Matthew L., y Ted Underwood. 2016. «Text-Mining the Humanities». En A New Companion to Digital Humanities, editado por Susan Schreibman, Ray Siemens, y John Unsworth, 291-306. Wiley-Blackwell.

Juola, Patrick. 2015. «The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions». Digital Scholarship in the Humanities 30 (1): 100-113.

León Felipe, Benigno. 1999. «El poema en prosa en España (1940-1990)». Tenerife: Universidad de La Laguna.

López Castro, Armando. 2002. «Bécquer y la poesía en prosa». Salina: revista de lletres, n.º 16: 143-52.

Lutosławski, Wincenty. 1897. The origin and growth of Plato's logic: with an account of Plato's style and of the chronology of his writings. London: Longmans, Green & Co.

Montalvo, Yolanda. 1995. «Sublimación e irrisión en las narraciones

orientales de Bécquer». En Bécquer. Origen y estética de la modernidad. Actas del VII Congreso de Literatura Contemporánea, Universidad de Málaga, 9, 10, 11 y 12 de noviembre de 1993, editado por Cristóbal Cuevas, 241-49. Málaga: Publicaciones del Congreso de Literatura Española Contemporánea.

MORETTI, Franco. 2013. Distant Reading. London: Verso.

MOSTELLER, F., y D. Wallace. 1964. Inference and disputed authorship: The Federalist. Addison-Wesley.

NAVARRO-COLORADO, Borja. 2018. «On Poetic Topic Modeling: extracting themes and motifs from a corpus of Spanish poetry». Frontiers in Digital Humanities 5 (15).

NEWMAN, M. E. J. 2010. Networks. An Introduction. Oxford: Oxford University Press.

NOVALIS, Friedrich, Friedrich Schiller, August Wilhelm Schlegel, Heinrich von Kleist, y Friedich Hölderlin. 1994. Fragmentos para una teoría romántica del arte. Editado por Javier Arnaldo. Madrid: Tecnos.

OCHAB, Jeremi K., Joanna Byszuk, Steffen Pielström, y Maciej Eder. 2019. «Identifying Similarities in Text Analysis: Hierarchical Clustering (Linkage) versus Network Clustering (Community Detection)». En Digital Humanities 2019: Utrecht University. <https://dev.clariah.nl/files/dh2019/boa/0981.html>.

PAGEARD, Robert. 1995. «Bécquer o la peligrosa pasión de explorar: poesía, poemas en prosa, crítica y variedad periodística». En Bécquer. Origen y estética de la modernidad. Actas del VII Congreso de Literatura Contemporánea, Universidad de Málaga, 9, 10, 11 y 12 de noviembre de 1993, editado por Cristóbal Cuevas, 13-32. Málaga: Publicaciones del Congreso de Literatura Española Contemporánea.

RUBIO JIMÉNEZ, Jesús. 2006. Pintura y literatura en Gustavo Adolfo Bécquer. Sevilla: Fundación José Manuel Lara.

RYBICKI, Jan. 2016. «Vive la différence: Tracing the (authorial) gender signal by multivariate analysis of word frequencies». Digital Scholarship in the Humanities 31 (4): 746-61. <http://dx.doi.org/10.1093/llc/fqv023>.

SCHÖCH, Christof. 2018. «Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie». En Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven, editado por Toni Bernhart, Sandra Richter, Marcus Lepper, Marcus Willand, y Andrea Albrecht, 77-94. Berlin: De Gruyter.

SMITH, Peter W.H., y W. Aldridge. 2011. «Improving authorship attribution: Optimizing Burrows' Delta method». Journal of Quantitative Linguistics 18 (1): 63-88.

UNDERWOOD, Ted. 2016. «The Life Cycles of Genres». Cultural Analytics 2 (2). <https://doi.org/10.22148/16.005>.

UTRERA TORREMOCHA, Victoria. 1999. Teoría del poema en prosa. Sevilla: Secretariado de Publicaciones de la Universidad de Sevilla.

Resumen

Además de ser uno de los poetas y narradores más influyentes de todos los tiempos, Gustavo Adolfo Bécquer (1836-1870) es considerado el gran precursor del poema en prosa en la literatura española, el cual cultivó, junto a la prosa poética, en algunas de sus *Leyendas*. Aunque por lo general la crítica ha aceptado la clasificación de obras becquerianas en prosa poética y poema en prosa propuesta por Luis Cernuda, muchos añaden otros textos en los que creen ver una de estas

dos modalidades. En este estudio se aplican métodos estilométricos con el objetivo de determinar qué textos se encuentran más cerca del género lírico. Al mismo tiempo, se pretende explorar si es posible realizar mediante estas técnicas una gradación de los textos en prosa becquerianos hacia lo lírico. Los resultados ayudarán a arrojar nueva luz sobre la obra lírica en prosa becqueriana, además de constituir una muestra de cómo puede contribuir la Estilometría al análisis de géneros literarios limítrofes, como es el caso del poema en prosa.

Abstract

In addition to being one of the greatest poets and writers of all time, Gustavo Adolfo Bécquer (1836-1870) is considered the great precursor of the prose poem in Spanish literature, which he cultivated, together with poetic prose, in some of his *Legends*. Although scholars have generally accepted the prose poem and poetic prose classification of Bécquer's works proposed by Luis Cernuda, many add other texts in which they believe one of these two modalities to be present. In this study, stylometric methods are applied in order to determine which texts are closer to the lyric genre. At the same time, the aim is to explore whether it is possible, by means of these techniques, to make a gradation of Bécquer's prose texts towards the lyric genre. The results will help to shed new light on Bécquer's lyric work in prose, and will provide an example of how Stylometry may contribute to the analysis of borderline literary genres, such as the prose poem.

2022, n°9
pp. 57-86
doi: https://doi.org/10.15460/apropos.9.1888

Anne Klee & Julia Röttgermann

„Nuit, correspondance, sentiment"

Topic Modeling auf einem Korpus von französischen Romanen 1750-1800

Anne Klee
ist wissenschaftliche Mitarbeiterin am *Trier Center for Digital Humanities* der Universität Trier.
klee@uni-trier.de

Julia Röttgermann
ist wissenschaftliche Mitarbeiterin am *Trier Center for Digital Humanities* der Universität Trier.
roettger@uni-trier.de

Keywords

18. Jahrhundert – Französische Literatur – Topic Modeling – Linked Open Data – Wikidata

1. Topic Modeling

Wie lassen sich große Korpora hinsichtlich ihrer literarischen Themen explorativ digital erforschen? Das Verfahren Topic Modeling, das in der hier vorgestellten Implementation auf der statistischen Methode *Latent Dirichlet Allocation* basiert (Blei, Ng, und Jordan 2003), kann zur Analyse verschiedener Arten von Daten eingesetzt werden: Neben beispielsweise Bildern, genetischen Daten oder Zeitungsarchiven (Blei 2011) lässt sich der Algorithmus mit entsprechenden Parametereinstellungen auch auf literarische Texte anwenden, wie zahlreiche Studien zeigen konnten (Underwood 2012; Rhody 2013; Jockers 2014). Die Methode, die von einer möglichst großen Datenmenge profitiert, berücksichtigt vor allem die Kookkurrenz von Wörtern und generiert auf dieser Grundlage Gruppen von semantisch verwandten Wörtern, die als Topics bezeichnet werden.

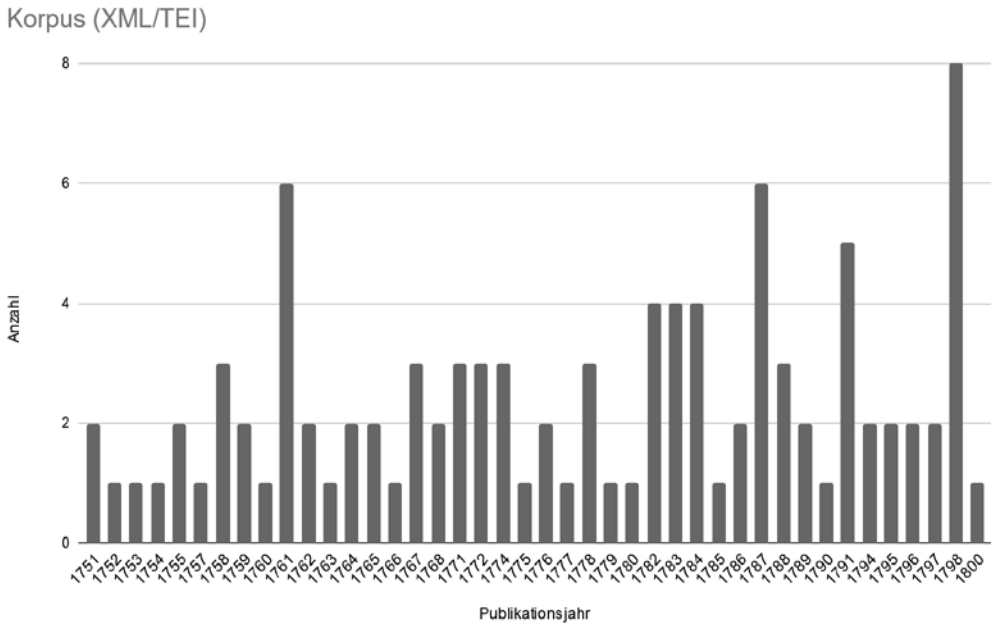
Beispiele für den erfolgreichen Einsatz der Methode Topic Modeling auf literarischen Texten sind Christof Schöchs Studie zu Topic Modeling auf französischen Dramen-Texten der Klassik und Aufklärung, die Topic Modeling Ergebnisse und Gattungen korreliert, oder auch Katherine Bodes Topic Modeling

Arbeiten auf fiktionalen Texten in australischen Zeitungen, in denen sie beispielsweise als zusätzlichen Parameter Gender analysiert (Bode 2018, 156-198; Schöch 2017).

Französische Texte aus dem 18. Jahrhundert haben Charakteristika, die in der Topic Modeling Pipeline durch entsprechende Preprocessing-Schritte berücksichtigt werden müssen. Forschungsprojekte wie ARTFL (American and French Research on the Treasury of the French Language) haben Topic Modeling bereits auf Texten des 18. Jahrhundert erprobt und die Artikel der *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers* mithilfe von Topic Modeling analysiert. Dabei unterstreichen sie vor allem die Möglichkeit, über Topic Modeling Diskurse abzubilden (Roe, Gladstone & Morrissey 2016). Elizabeth Andrews Bond und Robert M. Bond haben unlängst eine Studie zu Topic Modeling in vorrevolutionären französischen Presseartikeln vorgelegt (Bond & Bond 2020), wobei sie ihre Analyse mit dem stm-Package von R durchgeführt haben.

Im Kontext des Verbundprojekts „Mining and Modeling Text" (MiMoText), das vom *Trier Center for Digital Humanities* koordiniert wird, wurde Topic Modeling mit MALLET in Python auf das im Projekt entstehende Romankorpus „Collection de romans français du dix-huitième siècle (1750-1800)" in der Version 0.1.0 (kurz: roman18) angewendet. Die Topic Modeling-Pipeline basiert auf Schöch (Schöch 2020) und wurde an die Anforderungen des Projektes angepasst (Klee & Röttgermann 2020).

2. Datengrundlage: Romankorpus MiMoText



1 | MiMoText Romankorpus (XML/TEI): Werke pro Erstpublikationsdatum (Stand 13.10.2021)

Datengrundlage des Topic Modeling ist das roman18-Korpus an 80 französischen Romanen 1751-1800 (Röttgermann et al. 2020), das sich aus mehreren Quellen speist: eigene Volltextdigitalisierung per Double-Keying-Verfahren, eigene Volltext-digitalisierung mithilfe der Software OCR4all (Reul u. a. 2019), Wandlung von EPUB-Dateien aus den Quellen *Wikisource*[1], *Ebooks libres et gratuits*[2], *GoogleBooks*[3], *Frantext*[4] *und Rousseau Online*[5]. Alle Input-Dateien wurden in TEI-konformes XML nach den Richtlinien der *Text Encoding Initiative* (Burnard 2014) nach dem Schema der *European Literary Text Collection* (*ELTeC*) kodiert (Burnard & Odebrecht 2019). Mit Hilfe eines Python-Skripts[6], das historische Verbformen und Schaft-S in den Texten umwandelt, wurden die Texte teilmodernisiert, normalisiert und als Plaintext extrahiert.

Ein Vergleichshorizont der Daten ist zudem eine nahezu vollständige Dokumentation der literarischen Produktion französischer fiktionaler Werke 1751-1800 in Form einer Bibliographie (Martin, Mylne & Frautschi 1977), die uns dank wissenschaftlicher Vorarbeiten (Lüschow 2019) digitalisiert in Form eines RDF-Graphen vorliegt.

Aufgrund der statistischen Merkmale dieser bibliographischen Daten konnten wir uns hinsichtlich der Ausgewogenheit unseres Romankorpus den Proportionen in diesen Metadaten annähern und Merkmale wie Genderverteilung oder Erst-publikationsdaten approximativ im Romankorpus abbilden. Die Angabe zum Geschlecht der Autor:innen ist nicht explizit in den Metadaten enthalten, konnte jedoch über einen Abgleich mit VIAF ermittelt werden.

[1] *Wikisource* (ein Schwesterprojekt von Wikipedia) vereint Primärtexte in über 70 Sprachen. In einer auf Crowdsourcing basierenden Transkriptionsumgebung werden Faksimile und per Optical Character Recognition erkannter Text nebeneinander gestellt und von der Crowd korrigiert. Die Dateien durchlaufen verschiedene Qualitätsstufen. Wir haben uns dazu entschieden Texte aufzunehmen, die mindestens durch ein gelbes oder grünes Label ausgezeichnet sind, also vollständig transkribiert und von mindestens zwei verschiedenen Personen korrigiert wurden.

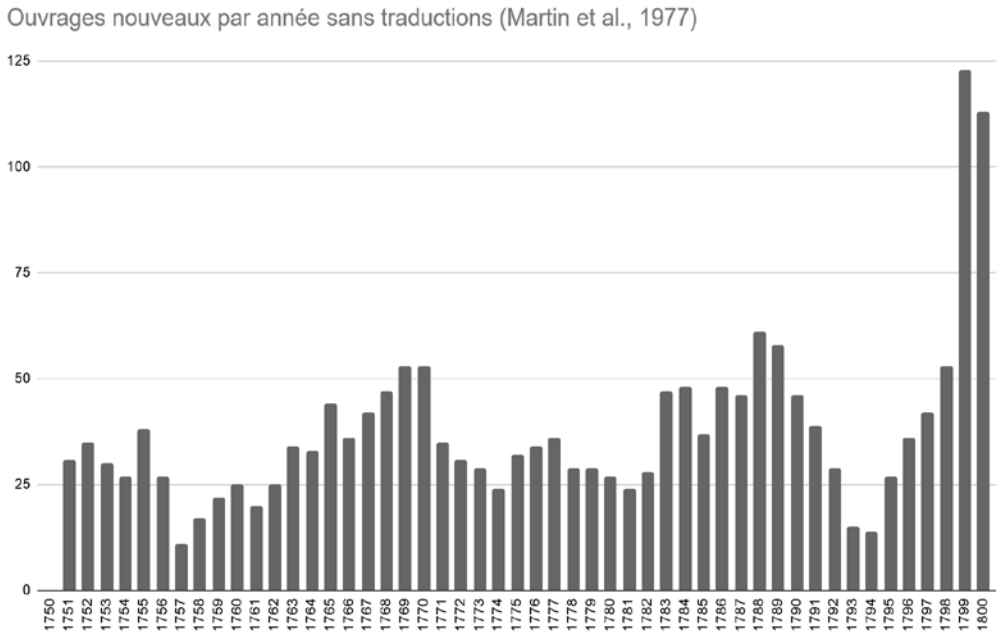
[2] Auf der Website *Ebooks libres et gratuits* konnten wir vor allem Werke kanonischer Autor:innen finden. Sie wurden als EPUB heruntergeladen und in TEI/XML konvertiert. Ein Nachteil der Plattform ist, dass sich die Provenienz der analogen Datenquelle (Print) leider nicht einsehen lässt.

[3] Die Datenqualität der französischen Romane, die wir im entsprechenden Publikationsdatum frei verfügbar als EPUB auf *GoogleBooks* finden konnten war sehr divers. Wir haben Werke nach der verfügbaren Qualität im Hinblick auf eine möglichst geringe OCR-Fehlerquote ausgewählt.

[4] *Frantext* enthält Primärtexte verschiedener Gattungen vom IX. bis XXI. Jahrhundert. Wir haben Texte verwendet, die den Kriterien Publikationsdatum 1750-1800, Gattung: "roman" und Lizenz "licence libre" entsprechen. Das auf Frantext vorliegende TEI/XML enthält eine Vielzahl an linguistischen Annotationen, die mit Hilfe eines Skripts entfernt wurden : <https://github.com/MiMoText/roman18/blob/master/Python-Scripts/transformation_frantext/Umwandlung_Frantext-Werke_in_TEI.py>, 28.01.2022.

[5] *rousseauonline.ch* bietet Zugang zu allen Werken von Jean-Jacques Rousseau (1712-1778) in ihrer ersten Referenzausgabe. Die Texte sind online zum Lesen zugänglich und stehen zum kostenlosen Download als PDF oder EPUB zur Verfügung. <https://www.rousseauonline.ch/>, 28.01.2022.

[6] <https://github.com/MiMoText/roman18/tree/master/Python-Scripts/modernization_and_transformation_to_plaintext>, 28.01.2022.



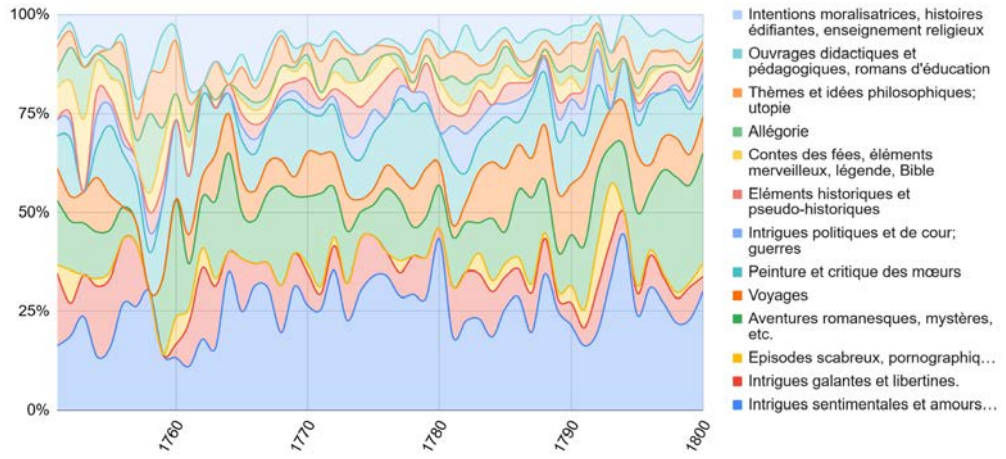
2 | Auswertung der Bibliographie du genre romanesque français 1751-1800 (Martin, Mylne & Frautschi 1977) hinsichtlich des Erstpublikationsdatums ohne Berücksichtigung der Übersetzungen.

Das Romankorpus wird fortlaufend um weitere erschlossene Quellen erweitert und wächst somit kontinuierlich. Der aktuelle Stand ist auf GitHub einsehbar. Die Grundlage des hier besprochenen Topic Modelings ist unter „Topic Model of roman18 corpus (November 2020)" auf Zenodo archiviert (Klee und Röttgermann 2020).

Die Besonderheit, dass die Grundgesamtheit[7] der literarischen Produktion der entsprechenden Dekaden innerhalb der Bibliographie (Martin, Mylne & Frautschi 1977) sorgfältig dokumentiert wurde und dass Schlagworte in Form thematischer Inhalte der Romane analysiert wurden, stellt für das Projekt einen reichen Datenschatz dar, der gerade im Vergleich mit den Ergebnissen des unsupervised Machine Learnings des Topic Modelings eine interessante Kontrastfolie bietet.

Das Korpus der bibliographischen Einträge 1751-1800 wurde von Seiten der Bibliograph:innen bezüglich der literarischen Themen und Handlung („thèmes et intrigues") in folgende Kategorien eingeteilt (s. Abb. 3).

[7] Zu möglicherweise fehlenden Werken cf. Dawson 1978, 497–508.



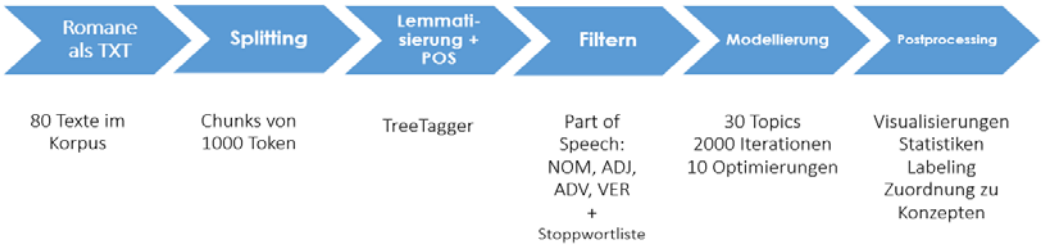
3 | Kategorien der Themen und Handlungselemente („thèmes et intrigues") des französischen Romans 1751-1800 im Zeitverlauf, Daten: Martin et al., 1977, S. xlviii.

Wir können in der Betrachtung der Themenkategorien im Zeitverlauf (Abb. 3) erkennen, dass religiöse und moralisierende Themen im französischen Roman der zweiten Hälfte des 18. Jahrhunderts eine abnehmende Tendenz aufweisen. Die Themen Erziehung („éducation") oder Reise („voyage") hingegen nehmen beispielsweise in den Dekaden ab 1780 prozentual zu.

Vor dem Hintergrund dieser 1977 publizierten Themenwerte für den französischen Roman der Zeit von 1751 bis 1800 in den bibliographischen Metadaten wollen wir nun betrachten, welche Topics im Gegenzug der Topic Modeling Algorithmus mit MALLET auf den Volltexten (Röttgermann et al. 2020) generiert.

3. Topic Modeling Workflow

Für die automatische Gewinnung von thematischen Aussagen aus dem MiMoText-Romankorpus wurde ein Topic Modeling-Workflow entwickelt, der mit Hilfe von Pythonskripten durchgeführt wird. Unsere Topic Modelling-Pipeline beinhaltet neben der Modellierung und verschiedenen Nachbearbeitungsschritten im Postprocessing zunächst einige Vorverarbeitungsschritte, die — unter der Berücksichtigung von Spezifika literarischer Prosatexte — dazu dienen, die Texte auf den Prozess des Topic Modelings vorzubereiten.



4 | Topic Modeling-Pipeline im Projekt „Mining and Modeling Text".

3.1 Input

Die Texte werden als Plaintext in die Pipeline eingespeist. Dabei sind diese bereits im Zuge der Korpuserstellung modernisiert und normiert worden. Dies beinhaltet die Anpassung historischer Wort- und Flexionsformen an die moderne Sprache sowie die Ersetzung historischer Schriftzeichen wie beispielsweise das Schaft-S.[8]

Da der Topic Modeling-Algorithmus Wörter in ihrem Kontext betrachtet, sollte dieser für die eingespeisten Texte nicht zu groß ausfallen und in einer einheitlichen Größe vorliegen. Die Romane stellen jedoch eine vergleichsweise umfangreiche Textsorte dar. Zusätzlich variieren die verschiedenen Texte im Korpus stark in ihrem Umfang. Aus diesem Grund werden die Inputdateien zu Beginn in Textstücke mit einer Wortlänge von 1000 Token gesplittet.

Den Textdateien beigefügt wird eine Metadatentabelle[9], welche später im Post-processingschritt bei der statistischen Auswertung und Visualisierung der Topic Modeling-Ergebnisse verwendet wird. Diese enthält Informationskategorien wie das Autorengeschlecht, das Publikationsjahr und die Erzählform.

3.2 Preprocessing: POS-Tagging und Filtern

Vor der eigentlichen Modellbildung werden die eingespeisten Textdateien mit üblichen Preprocessingschritten vorverarbeitet. In einem ersten Schritt werden die Texte lemmatisiert. Das bedeutet, die einzelnen Wortformen werden auf ihre Grundform zurückgeführt, sodass flektierte Formen eines Lemmas bei der Modellierung als gleiche Wörter behandelt werden. Für diese Wortformen wird zusätzlich ein *part of speech*-Tagging durchgeführt, bei dem sie mit ihrer Wortart annotiert werden. Dazu wird das Tool TreeTagger (Schmid 1994) verwendet, das Modelle für die Anwendung auf über 40 verschiedene Sprachen zur Verfügung stellt. Für die Wahl des geeigneten Modells wurden zwei Ausführungen erprobt: das von TreeTagger bereitgestellte Modell für die moderne, französische Sprache sowie das Presto-Modell (PRESTO 2014)[10], welches speziell für das Französisch des 16. und 17. Jahrhunderts auf der Grundlage historischer Texte trainiert wurde (Souvay & Pierrel 2009). Eine exemplarische Überprüfung der Tagging-Ergebnisse hat gezeigt, dass Presto zwar bessere Ergebnisse bei der Erkennung von historischen Wortformen liefert, sich jedoch im Vergleich zum TreeTagger-Modell durch eine deutlich unzuverlässigere POS-Klassifizierung auszeichnet (cf. Abb. 5). Zudem weist das Modell für die moderne Sprache eine bessere Named-Entity-Recognition auf, was beim Herausfiltern von Eigennamen eine wichtige Rolle spielt. Insgesamt überwiegen die Stärken des vom TreeTagger bereitgestellten Modells, weshalb dieses für die Vorverarbeitung in der Topic Modeling-Pipeline verwendet

[8] Dies erfolgt mit Hilfe eines Pythonskriptes und einer Modernisierungsliste (<https://github.com/MiMoText/roman18/tree/master/Python-Scripts/modernization_and _transformation_to_plaintext>). Für die Liste wurden mit einem Spellcheck und der Pythonbibliothek enchant (<http://pythonhosted.org/pyenchant/>) die im Korpus vorkommenden historischen Wortformen ermittelt.

[9] <https://github.com/MiMoText/roman18/blob/master/XML-TEI/xml-tei_metadata.tsv>, 28.01.2022.

[10] <http://presto.ens-lyon.fr/>.

wird. Seine Schwäche bei der Erkennung historischer Sprachformen wird durch den Modernisierungsschritt im Vorfeld aufgefangen.

	TreeTagger fr-Modell		Presto	
	POS	Lemma	POS	Lemma
étoit	Substantiv	étoit	Konjugiertes Verb (être & avoir)	être
errois	Adjektiv	errois	Konjugiertes Verb	errer
sçavois	Verb im Subjonctiv imperfect	sçavois	Konjugiertes Verb	savoir
connoissance	Substantiv	connoissance	Substantiv	connaissance
quinze	Numeral	quinze	Substantiv	quinze
vingt-quatre	Numeral	vingt-quatre	Adjektiv	vingt-quatre
dernier	Adjektiv	dernier	Substantiv	dernier

5 | TreeTagger und Presto in einem exemplarischen Vergleich. Die erste Spalte beinhaltet Wortformen, wie sie im Romantext auftreten, die Spalte „POS" beinhaltet die Wortklasse, die das Modell der jeweiligen Form zuordnet, und „Lemma" die Grundform, auf die die Wortform zurückgeführt wird.

Es konnte nachgewiesen werden, dass durch die Beschränkung auf ausgewählte Wortarten kohärentere Topics erzielt werden (cf. Uglanova & Gius 2020). Um semantisch aussagekräftige Topics zu generieren, werden die Texte deshalb im Vorfeld nach Wortarten gefiltert. In die Modellierung eingespeist werden nur Lemmata der Wortarten Substantiv, Adjektiv, Adverb und Verb.

Mit Hilfe zweier Stoppwortlisten werden zusätzlich sowohl Hilfsverben als auch die für die Textsorte Roman charakteristischen Figurennamen, die beim POS-Tagging nicht als Eigennamen erkannt wurden, herausgefiltert.[11] Beide Kategorien von Wörtern kommen in den Texten sehr häufig vor, haben aber wenig bis keinen semantischen Gehalt und tragen damit also nicht zur Gewinnung thematischer Muster bei.

3.3 Modellierung

Für die Durchführung des Modellierungsschrittes wurden zwei verschiedene Varianten getestet: die Pythonbibliothek Gensim (Rehurek & Sojka 2010)[12] und das Java-basierte Tool MALLET (McCallum 2002)[13].

[11] Die in den Texten vorkommenden Figurennamen wurden mit Hilfe von Named Entity Recognition mit SpaCy ermittelt.
[12] <https://pypi.org/project/gensim/>.
[13] <http://mallet.cs.umass.edu/topics.php>.

	Gensim	MALLET
0	cœur	cœur
1	voir	amour
2	point	point
3	aimer	aimer
4	amour	jamais
5	ami	voir
6	croire	rendre
7	tout	âme
8	jamais	sentiment
9	sentiment	bonheur
10	moins	ami

6 |Vergleich zwischen Gensim und MALLET bezogen auf die Verteilung der zehn wichtigsten Topicwörter des Topics amour_sentiment. Die Topics stammen aus vergleichbaren Modellen, die für 10 Topics mit 500 Iterationen vorgenommen wurden.

Da die Bibliothek von Gensim für die Anwendung auf sehr großen Textkorpora entwickelt wurde[14], war im Vorfeld zu erwarten, dass in unserem Fall eines vergleichsweise kleinen Korpus das Tool MALLET besser geeignet ist. Diese Annahme konnte im Abgleich der berechneten Modelle bestätigt werden. Unsere Untersuchungen haben gezeigt, dass die mit MALLET erstellten Modelle konsistentere — das heißt für die Ableitung von thematischen Aussagen besser interpretierbare — Topics[15] geliefert haben. Illustrieren lässt sich dies am Beispiel des Topics amour_sentiment. Wie in Abb. 6 zu sehen ist, finden sich in dem mit MALLET erstellten Topic unter den zehn wichtigsten Wörtern deutlich mehr solcher Wörter, die semantisch dem Themenfeld Liebe/Gefühl zuzuordnen sind.

Durch den Wrapper LdaMallet von Gensim[16] kann die Modellierung mit MALLET in die Python-Pipeline eingebunden werden. Nach Tests mit verschiedenen Modellgrößen wurde ein Topic Model mit 30 Topics trainiert. Diese Anzahl bietet bei der Korpusgröße von rund 80 Texten ein ausgewogenes Topic-Spektrum. Eine höhere Zahl Topics vergrößert zwar die Menge unterschiedlicher Topics, unter diesen finden sich jedoch in größerer Zahl generische Topics und es kommt vermehrt zu semantischen Überschneidungen. Des Weiteren handelt es sich zunehmend um sehr spezifische Topics, welche nur in einzelnen Werken vorkommen und welche wir ohnehin bei der Einspeisung in unser Wissensnetzwerk ignorieren. Eine kleinere Anzahl führt zu mehrdeutigen Topics und würde verhindern, dass manche Wortkookurrenzen überhaupt aufgedeckt werden.

[14] Cf. Rehurek & Sojka 2010, die Gensim auf ein Korpus von 61.293 Volltexten und insgesamt über 270 Mio. Token anwenden.

[15] Zu einem ähnlichen Ergebnis kommt auch Dipanjan Sarkar, der ein Korpus an wissenschaftlichen Artikeln mit Topic Modeling analysiert und die coherence scores von MALLET und Gensim miteinander vergleicht: "You can clearly see that the model from MALLET is much better (...) as compared to the default LDA model from Gensim." (Sarkar 2019, 402)

[16] <https://radimrehurek.com/gensim_3.8.3/models/wrappers/ldamallet.html>.

Die Anzahl der Iterationen und Optimierungen beim Machine Learning-Prozess sind zwei weitere festzulegende Parameter. Umso mehr Iterationen durchgeführt werden, desto stärker verlängert sich die Laufzeit, desto mehr steigert sich jedoch die Qualität des resultierenden Topic Models. Für das vorliegende Korpus wurden 2000 Iterationen und 10 Optimierungen für den Trainingsprozess als geeignete Größe ermittelt. Die Optimierungen führen zu einer differenzierteren Wahrscheinlichkeitsverteilung der Topics im berechneten Modell.[17]

3.4 Postprocessing

Das Postprocessing umfasst die Erzeugung verschiedener Statistiken und Visualisierungen der Topic Modeling-Ergebnisse, die der Analyse und Extraktion der thematischen Statements dienen, aber auch selbst den Nutzer:innen des Wissensnetzwerkes bereitgestellt werden.

Das berechnete Topic Model besteht aus einer zuvor definierten Anzahl von Topics, die aus einer Wahrscheinlichkeitsverteilung der eingespeisten Wörter bestehen, sowie einer Wahrscheinlichkeitsverteilung dieser Topics für jedes Textdokument des Korpus. Bestehend aus diesen Informationen werden verschiedene CSV-Dateien[18] erstellt. Auf der Basis der wahrscheinlichsten Wörter wird jedem Topic ein Label zugewiesen (vgl. 3.5). Zusammen mit dieser Information werden aus der Verteilung der Top-Topics pro eingespeistes Werk schließlich Themenaussagen abgeleitet. Dabei berücksichtigen wir die fünf wahrscheinlichsten Topics für jeden Roman, bei vorheriger Aussortierung aller Topics, die in weniger als 10 % und in mehr als 80% der Korpuswerke enthalten sind.[19] Es werden dadurch einerseits sehr seltene, zum Teil werkspezifische, und andererseits sehr häufige, in der Regel generische, Topics ausgeschlossen, da sie für einen werkübergreifenden Themenabgleich keinen Gewinn bringen. Es verbleiben damit 25 Topics, die bei der Generierung der Themenaussagen einbezogen werden.

Aufgrund der in der französischen Literatur des 18. Jahrhunderts vorherrschenden Themen ähneln sich viele Werke in Bezug auf die Top-Topics. Um dennoch Aussagen darüber treffen zu können, wie sich die einzelnen Werke thematisch voneinander unterscheiden, ist es hilfreich, die distinktiven Topics pro Werk zu ermitteln. Ein Topic ist distinktiv für ein Werk, wenn es in diesem überrepräsentiert

[17] Dazu passt der Algorithmus interne Hyperparameter fortlaufend so an, dass sowohl die Topics als auch die Wörter ihrer Zusammensetzung im resultierenden Modell stärker unterschiedlich gewichtet werden (cf. Steyvers & Griffiths 2007).

[18] *topicwords.csv* enthält die 50 Top-Wörter für jedes Topic, *wordprobs.csv* zeigt für jedes Lemma im Vokabular den Score für jedes Topic. *doc-topic-matrix.csv* enthält für jedes Textchunk eine Verteilung der Topics, welche in *chunkmatrix.csv* um relevante Metadaten pro Textstück ergänzt ist. Die Topicwahrscheinlichkeiten bezogen auf die Romane als Gesamttexte sind in der *mastermatrix.csv* zusammengefasst. *topicranking.csv* listet für jedes Werk die zehn wahrscheinlichsten Topics mit ihren Wahrscheinlichkeitswerten.

[19] Hier ist anzumerken, dass im Grunde jedes Topic in jedem Werk vorhanden ist. Es tritt jedoch erst ab einer bestimmten Wahrscheinlichkeit signifikant in Erscheinung, ab der wir vereinfacht davon sprechen, dass es in einem Werk vorkommt. Der Schwellwert ist von der Korpusgröße und Topicanzahl abhängig. Für das hier beschriebene Topic Model haben wir eine Wahrscheinlichkeit von 0.03 als Schwellwert angewandt. Mithilfe von diesem lässt sich berechnen, in wieviel Prozent der Texte jedes Topic vorkommt.

ist, also im Vergleich zum Gesamtkorpus eine überdurchschnittliche Wahrscheinlichkeit aufweist. Demnach definieren wir seltene Topics, die nur in wenigen Romanen des Korpus vorkommen und dadurch eine geringe Wahrscheinlichkeit bezogen auf das Gesamtkorpus aufweisen, für ein Einzelwerk distinktiv, sofern sie dort unter den Top-Topics auftreten.

MiMoText-ID	Top 5-Topics	Distinktive Topics = Seltene in Top 5
Senac_Emigre	Correspondance, Bonheur, Attraction_Personnalité, [temps paraître encore], Amour_Sentiment	
Maistre_ Voyage	Bonheur, Art, Nuit, Mort, Nature	Art
Sade_Aline	Philosophie, Interaction, Amour_Sentiment, Deuil, Migration_Voyage	
Sade_Justine	Terreur, Philospophie, Mort, Nuit, Interaction	
Bernadin_Paul	Nature, Bonheur, Art, Mort, Famille	Art
Laclos_Liaisons	Correspondance, Interaction, Amour_Sentiment, Bonheur, [temps paraître encore]	
Retif_Paysanne	Famille, Correspondance, Nuit, Interaction, Deuil	
Mercier_An	Philosophie, Monarchie, Art, Mort, Richesse	Monarchie, Art
Retif_AntiJustine	Nuit, Famille, [temps paraître encore], Relation amoureuse, Interaction	Relation amoureuse
Rousseau_Julie	[point voir moins], Amour_Sentiment, Éducation_Enfance, Bonheur, Deuil	
Voltaire_Candide	Alimentation_Sociabilité, Migration_Voyage, [point voir jour], Richesse, Deuil	Alimentation_ Sociabilité

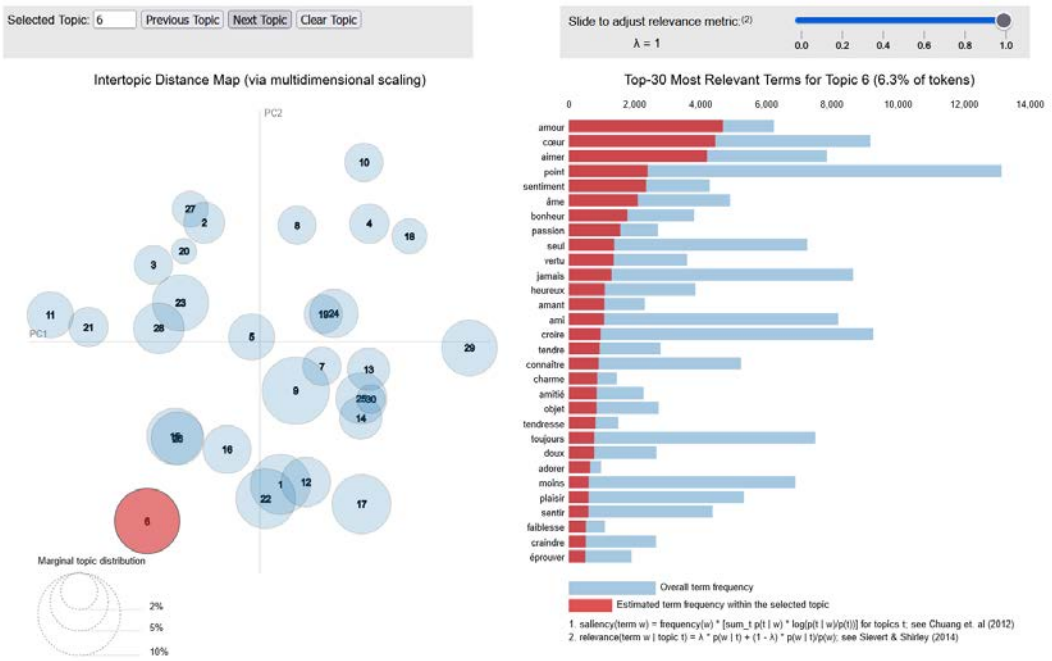
7 | Verteilung der distinktiven Topics bei den Pilotwerken (Ausschnitt aus dem gesamten Romankorpus). Topics, die keinem der Einträge des thematischen Vokabulars zugeordnet werden können, sind mit ihren drei wichtigsten Wörtern in eckigen Klammern gelabelt (vgl. Kapitel 3.5).

Mithilfe der Verteilung der Topics im Korpus können die seltenen Topics ermittelt werden. Als selten wurden hier Topics definiert, die nur in maximal 20% aller Romane im Korpus vorkommen. Die Schnittmenge mit den Top 5 Topics pro Werk zeigt schließlich, durch welche prävalenten Topics sich ein Werk von den übrigen abhebt.[20]

Auf der Grundlage der Statistiken werden zur Veranschaulichung der Topic Modeling-Ergebnisse verschiedene Visualisierungen erstellt.

[20] Zur Ermittlung von distinktiven Topics sind außerdem noch andere statistische Verfahren denkbar. Cf. dazu das Projekt "Zeta und Konsorten", welches verschiedene statistische Distinktivitätsmaße erforscht (cf. Schöch 2018, Du et al. 2021).

Mithilfe der Pythonbibliothek PyLDAVis[21] lässt sich eine interaktive HTML-Visualisierung[22] des Topic Models erstellen, mit welcher die Nutzer:innen die Verteilung und Zusammensetzung der Topics explorieren können.[23]



8 | Topic Modeling-Visualisierung in PyLDAVis. Auf der linken Seite sind die 30 Topics als Kreise zu sehen. Ihre Größe symbolisiert ihre Gewichtung im Korpus, ihre Lage zueinander bildet die Ähnlichkeit ihrer Wortverteilungen ab. Die Auswahl eines Topics links ermöglicht eine genauere Exploration seiner Wortverteilung auf der rechten Seite der interaktiven Visualisierung.

Einen Überblick über die häufigsten Wörter in jedem Topic bieten die mit der Pythonbibliothek wordcloud[24] erstellten Wortwolken. Die unterschiedlich großen Wahrscheinlichkeitswerte werden hier durch die Größe der Wörter abgebildet.

[21] <https://pyldavis.readthedocs.io/en/latest/#> [28.01.2022].

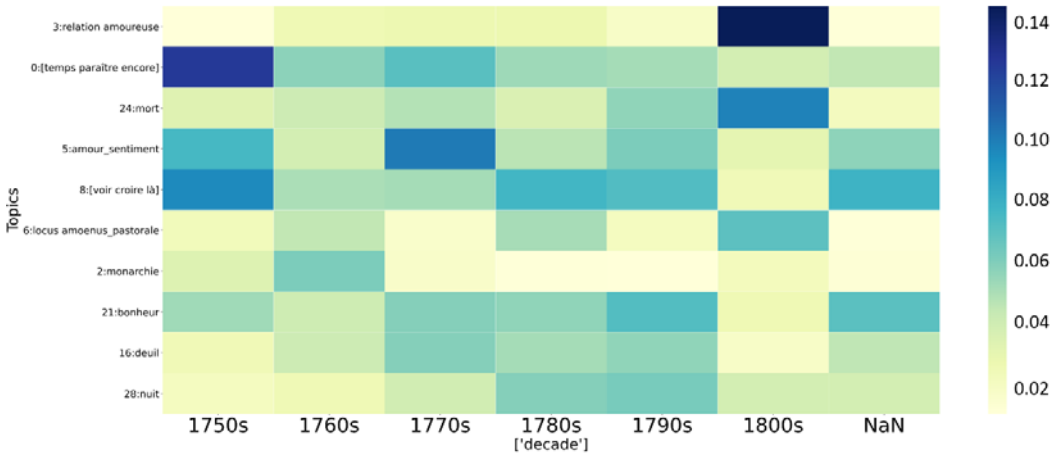
[22] Cf. die Visualisierung zum hier behandelten Topic Model: <https://github.com/MiMoText/mmt_2020-11-19_11-38/blob/main/results/mmt_2020-11-19_11-38/visualization.html> [21.10.21]. Um die Visualisierung wie in Abb. 8 anschauen zu können, ist es erforderlich, die verlinkte HTML-Datei zunächst lokal abzuspeichern und dann im Browser zu öffnen.

[23] Da PyLDAVis mit dem Output der Topic Modeling-Bibliothek von Gensim arbeitet, müssen die von MALLET erstellten Ergebnisse zunächst in das entsprechende Format transformiert werden: <https://github.com/MiMoText/topicmodeling/blob/master/scripts/make_overview.py> [28.01.2022].

[24] <https://amueller.github.io/word_cloud/> [28.01.2022].

9 | Wordcloud zum Topic „amour_sentiment".

Mithilfe der Bibliothek seaborn werden sogenannte Heatmaps zu einem bestimmten Metadaten-Parameter erstellt. Aktuell ist es möglich, Heatmaps zur Topicverteilung bezüglich der TextID, des Autor:innennamens, des Autor:innen-geschlechts, der Dekade oder der Erzählform zu generieren. Die Auswahl wird über die Parametereinstellungen im Script gesteuert.



10 | Heatmap mit der Verteilung von Topics im Romankorpus, dominante Topics pro Dekade.[25]

3.5 Labeling mit Hilfe eines thematischen Vokabulars

Um die aus dem Topic Modeling gewonnen Aussagen in das Wissensnetz einzuspeisen und mit den thematischen Informationen aus der Sekundärliteratur und der Bibliographie vergleichbar zu machen, ist es notwendig, die Topics zu labeln und zu normalisieren. Hierfür dient uns ein auf das Projekt zugeschnittenes kontrolliertes Themenvokabular. Bei der Erarbeitung waren bestimmte Eigenschaften von besonderer Wichtigkeit: Die Begriffe sollten die Themenkonzepte der französischen Aufklärung abdecken, ein gewisses Abstraktionslevel aufweisen, damit sie als kategorische Begriffe fungieren können, und ihre Zusammenstellung

[25] Die hier verwendete Anzahl an Romanen (80) der Version 0.1.0 des roman18-Korpus ist noch nicht groß genug, um robuste Aussagen hinsichtlich der Topics pro Dekade zu treffen. Wir sind dabei das Romankorpus im Laufe des Jahres 2022 auf bis zu 200 Volltexte zu erweitern. Die bibliographischen Metadaten enthalten derzeit ca. 2000 items, die in ca. 30.000 RDF-Tripeln resultieren und somit robustere Aussagen erlauben.

sollte transparent und nachvollziehbar sein. Eine erste Grundlage bildet das Themeninventar des *Dictionnaire européen des Lumières* (Delon 1997). Die Artikelstichwörter bieten eine gute Abdeckung an gesellschaftlich, politisch, ideengeschichtlich oder kulturell relevanten Themen der Epoche und stellen somit einen geeigneten Grundstock an möglichen Labeln für die in den Romanen vorkommenden Themen. Dennoch enthält die Ressource Begriffe, die entweder zu spezifisch (z.B. „pyrrhonisme") oder zu generisch (z.B. „fonction") sind, um durch sie literarische Themen zu beschreiben, weshalb diese für das Vokabular nicht berücksichtigt wurden. Ergänzt wurden die Begriffe um fehlende Konzepte zum Labeln der Topics, um thematische Schlagworte aus der Bibliographie (cf. Martin, Mylne & Frautschi 1977) sowie um Themenkonzepte, die in der Sekundärliteratur erwähnt werden, wenn diese anderweitig nicht repräsentiert waren. Das Vokabular ist nun konsolidiert, kann aber auch in Zukunft bei Bedarf erweitert werden.

Um die multilinguale Vergleichbarkeit zwischen französischsprachigen Primärtexten und deutschsprachiger Sekundärliteratur zu gewährleisten, und im Sinne der Anschlussfähigkeit an und Interoperabilität mit anderen Datenbeständen, werden die Themenkonzepte auf einen Normdatensatz (Wikidata) gemappt, wodurch das kontrollierte Vokabular konsolidiert und multilingual erfasst ist (siehe Abb. 11).[26]

théologie	theology	Theologie	Q34178	https://www.wikidata.org/wiki/Q34178	DEL
tolérance	toleration	Toleranz	Q183225	https://www.wikidata.org/wiki/Q183225	DEL
tradition	tradition	Tradition	Q82821	https://www.wikidata.org/wiki/Q82821	DEL
traduction	translation	Übersetzung	Q7553	https://www.wikidata.org/wiki/Q7553	DEL
tragédie	tragedy	Tragödie	Q80930	https://www.wikidata.org/wiki/Q80930	DEL
transport	transport	Transport	Q7590	https://www.wikidata.org/wiki/Q7590	DEL
travail	work	Arbeit	Q6958747	https://www.wikidata.org/wiki/Q6958747	DEL
troubadour	troubadour	Troubadour	Q186370	https://www.wikidata.org/wiki/Q186370	BGRF
tyrannie	tyranny	Tyrannei	Q22082330	https://www.wikidata.org/wiki/Q22082330	Seklit
universalisme	universalism	Universalismus	Q875797	https://www.wikidata.org/wiki/Q875797	Seklit
urbanisme	urbanism	Urbanistik	Q59950	https://www.wikidata.org/wiki/Q59950	DEL
utilitarisme	utilitarianism	Utilitarismus	Q160590	https://www.wikidata.org/wiki/Q160590	DEL
utopie	utopia	Utopie	Q131156	https://www.wikidata.org/wiki/Q131156	DEL
valeur	value	Wertvorstellung	Q194112	https://www.wikidata.org/wiki/Q194112	Seklit
vanité	vanity	Eitelkeit	Q1321250	https://www.wikidata.org/wiki/Q1321250	Seklit
vengeance	revenge	Rache	Q1712140	https://www.wikidata.org/wiki/Q1712140	Seklit
vérité	truth	Wahrheit	Q7949	https://www.wikidata.org/wiki/Q7949	Seklit
vertu	virtue	Tugend	Q157811	https://www.wikidata.org/wiki/Q157811	DEL
vie	life	Leben	Q3	https://www.wikidata.org/wiki/Q3	Seklit

11 |Ausschnitt aus dem kontrollierten Vokabular zur Extraktion thematischer Konzepte. „DEL" bezeichnet dabei Ressourcen aus dem *Dictionnaire européen des Lumières* (Delon 1997), „BGRF" Ressourcen aus den thematischen Schlagworten der *Bibliographie du genre romanesque français 1751-1800* (Martin, Mylne & Frautschi 1977) und „Seklit" Ressourcen aus der im Projekt ausgewerteten Sekundärliteratur.

Dieses aktuell 368 Einträge umfassende Vokabular der Themenbegriffe liefert die Konzept-Items für die Objektposition aller thematischen Statements. Zur Erstellung der Topic Modeling-Statements wird folglich jedem der Topics ein Label aus diesem Inventar zugewiesen. Dazu wurde zunächst ein automatischer Ansatz mit Einsatz von Word Embeddings (Mikolov u. a. 2013) und der Ermittlung von Topic-Zentroiden erprobt. Notwendig dafür ist ein Word Embedding Modell, welches auf

[26] Zur Dokumentation der Liste: <https://github.com/MiMoText/vocabularies>, letzter Zugriff 28.01.2022.

einer ausreichend großen Menge französischer Texte trainiert wurde. Hierfür finden sich bereits vortrainierte Modelle, die auf frei verfügbaren Texten im Web wie den Texten der französischen Wikipedia oder Nachrichtenkorpora basieren (cf. Fauconnier 2015) und für den Versuch des Labelings nachgenutzt werden können. Mithilfe des Modells wird für jedes Topic ein Vektor berechnet, der Topiczentroid, welcher sich aus dem Durchschnitt der Vektoren zu den wahrscheinlichsten Topicwörtern (z.B. die Top-20 Wörter) ergibt. Auch für jedes der Labelkandidaten, d.h. den Wörtern aus dem vorher definierten Labelvokabular, kann ein Vektor bestimmt werden. Über einen Distanzabgleich der Topicvektoren und Labelvektoren kann nun für jedes Topic ein zugehöriges Label ermittelt werden: Die Paarungen mit den geringsten Distanzen zeigen potentiell geeignete Label.

Ein weiterer Ansatz besteht in der Abbildung der Top-Topicwörter auf einen gemeinsamen generischen Begriff unter Einsatz eines lexikalisch-semantischen Netzes wie WordNet (<https://wordnet.princeton.edu/>). Mit dem automatischen Ansatz sind verschiedene Schwierigkeiten und Probleme verbunden: Klassische Word Embedding-Modelle enthalten nur Vektoren für Einzelwörter. Somit können mit dem dargelegten Verfahren keine Mehrwortbegriffe als Label zugeordnet werden. Die vortrainierten Modelle enthalten darüber hinaus in der Regel keine sehr speziellen Begriffe, die allerdings für die französische Literatur des 18. Jahrhunderts relevant sein könnten. Aus diesen Gründen wäre das Trainieren eines eigenen Modells anzudenken, welches einerseits Mehrwort-Lexeme beinhaltet und andererseits auch Texte zu Themen der Literatur der Aufklärung berücksichtigt. Allerdings stellte sich der Word Embedding-Ansatz noch aus einem anderen Grund als weniger gut geeignet heraus: Bei den Wörtern mit der geringsten Distanz zu den jeweiligen Topiczentroiden handelt es sich häufig um sehr spezifische Begriffe, die weniger gut geeignet sind, um ein Topic in seiner Ganzheit zu labeln. Beispielsweise würde „livre" als Label für ein Topic ausgewählt werden, das nicht rein literarische Aspekte enthält, sondern auch weitere Bereiche der Kunst wie die Musik und die Malerei abdeckt. Ähnlich gestaltet es sich beim WordNet-Ansatz schwierig, einen Oberbegriff für die Topicwörter mit einem geeigneten Abstraktionsniveau zu finden.

Aus diesen Gründen fiel die Entscheidung für eine manuelle Vergabe der Label. Die Zusammensetzung der Topicwörter einiger Topics lässt die Zuweisung eines eindeutigen Labels nicht zu, da hier verschiedene Konzepte repräsentiert sind. In diesen Fällen werden Doppellabel vergeben wie beim Topic „alimentation_sociabilité" in Abb. 12: Wörter wie „table", „manger", „boire" und „vin" deuten auf das Thema der Ernährung hin, daneben die vorkommenden Personen und Lebewesen „homme", „femme", „maître", „hôte" und „chien" auf eine gesellige Zusammenkunft. Das Topic wird daher durch die Verknüpfung zweier Begriffe des thematischen Vokabulars repräsentiert.

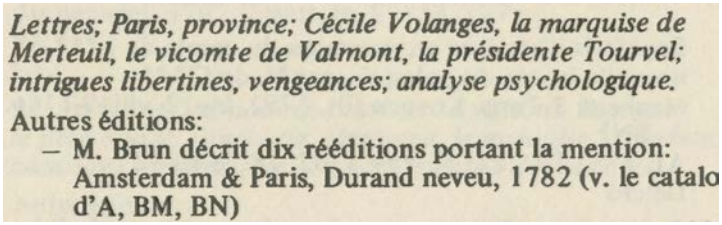
12 | Wordle des Topics „alimentation_sociabilité".

Da manche der Topics sehr generisch sind und nur eine geringe semantische Kohärenz aufweisen, ist es nicht immer möglich, ein passendes Label zu vergeben. In diesen Fällen werden die ersten drei Topicwörter als Label vergeben wie zum Beispiel „[point voir jour]".

4. Auswertung der Ergebnisse des Topic Modelings anhand von zwei Fallbeispielen

4.1 *Les Liaisons Dangereuses* von Choderlos de Laclos (1782)

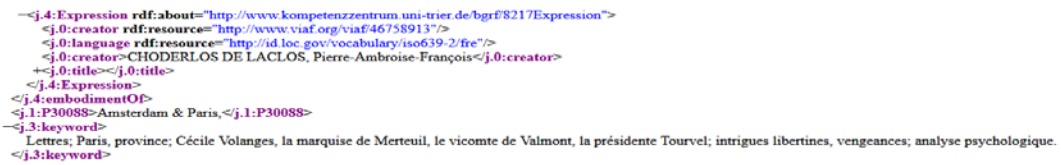
Anhand mehrerer Fallbeispiele seien hier die Ergebnisse des Topic Modelings analysiert und illustriert. Wir wählen als erstes Fallbeispiel ein Schlüsselwerk des 18. Jahrhunderts aus, um zu vergleichen, wie sich die Informationsextraktion aus der Bibliographie und aus dem Topic Modeling des Primärtextes hinsichtlich der extrahierten Themen verhalten: *Les Liaisons Dangereuses* von Choderlos de Laclos aus dem Jahr 1782. In der *Bibliographie du genre romanesque français 1751-1800* (Martin, Mylne & Frautschi 1977) werden von Seiten der Bibliograph:innen folgende Spezifikationen hinsichtlich der literarischen Themen des Werks vorgenommen:


Lettres; Paris, province; Cécile Volanges, la marquise de Merteuil, le vicomte de Valmont, la présidente Tourvel; intrigues libertines, vengeances; analyse psychologique.

Autres éditions:
– M. Brun décrit dix rééditions portant la mention: Amsterdam & Paris, Durand neveu, 1782 (v. le catalo d'A, BM, BN)

13 |Die Schlagworte zur Gattung, Handlungsort, Protagonisten und zu den Themen des Romans *Les Liaisons Dangereuses*, Screenshot aus der *Bibliographie du genre romanesque français 1751-1800* (Martin, Mylne & Frautschi 1977, 246).

Die Einträge der *Bibliographie du genre romanesque français 1751-1800* (Martin et al., 1977) wurden in einen RDF-Graphen überführt (Lüschow 2019), als Beispiel hier der Eintrag zu dem analysierten Werk:



14 | Der bibliographische Eintrag zu *Les Liaisons Dangereuses*, modelliert in RDF (Lüschow 2019)

Ausgehend von den Keywords innerhalb des RDF-Graphen wurden diese extrahiert und mit einem Python-Skript sortiert, mit dem Ziel die Einteilung der Bibliographen in die Kategorien Erzählform (hier „Lettres"), Handlungsort (hier: „Paris, province"), Protagonisten (hier „Cécile Volanges, la marquise de Merteuil, le vicomte de Valmont") Schlagworte der Handlung (hier: „intrigues libertines, vengeances") und Stil/Haltung/Tonalität (hier: „analyse psychologique") nachzuvollziehen.

ID	author	Label	narrative perspective_ string	Narrative location_string	characters_ string	plot_theme	style_attitude_ tonality
82.17	CHODERLOS DE LACLOS, Pierre-Ambroise-François	Les liaisons dangereuses ou lettres recueillies dans une société & publiées pour l'instruction de quelques autres, par M. C.....de L . . .	Lettres	Paris, province	Cécile Volanges, la marquise de Merteuil, le vicomte de Valmont, la présidente Tourvel	intrigues libertines, vengeances	analyse psychologique

Tab 1: Ergebnis der automatischen Keyword-Sortierung[27] für den Eintrag zu *Les Liaisons Dangereuses* (Laclos, 1782), die Metadaten stammen aus dem Eintrag 82.17 (Martin, Mylne & Frautschi 1977, 246).

Anhand dieser Kategorisierung innerhalb der XML-Dateien lassen sich die thematischen Schlagworte extrahieren. Relevant ist das Keyword-Feld „Themen und Handlung".

Nach einer Vereinheitlichung der Keywords mittels des kontrollierten Vokabulars[28] ergibt sich damit aus den bibliographischen Daten folgende Aussage zu literarischen Themen des Werks, hier formuliert in der Struktur eines RDF-Triples (Resource Description Framework Triples):

[27] Cf. <https://github.com/MiMoText/KeywordExtractor>.
[28] Cf. <https://github.com/MiMoText/vocabularies>.

Laclos_Liaisons	about	libertinage	http://zora.uni-trier.de:11000/wiki/Item:Q1
Laclos_Liaisons	about	vengeance	http://zora.uni-trier.de:11000/wiki/Item:Q1

Wie verhalten sich die extrahierten Aussagen aus dem Topic Modeling der Primärtexte dazu? Folgende Topics sind prävalent in *Les Liaisons Dangereuses*:

Laclos_Liaisons	about	correspondance	https://doi.org/10.5281/zenodo.4493224
Laclos_Liaisons	about	interaction	https://doi.org/10.5281/zenodo.4493224
Laclos_Liaisons	about	amour_sentiment	https://doi.org/10.5281/zenodo.4493224
Laclos_Liaisons	about	bonheur	https://doi.org/10.5281/zenodo.4493224
Laclos_Liaisons	about	[temps paraître encore]	https://doi.org/10.5281/zenodo.4493224

Das Topic „correspondance" umfasst Topicwörter wie Brief („lettre"), schreiben („écrire"), Antwort („réponse"), empfangen („recevoir") etc. und enthält einen Hinweis auf die Gattung von *Les Liaisons Dangereuses*: Es handelt sich um einen Briefroman. Die Bibliograph:innen der *Bibliographie du genre romanesque français 1751-1800* haben dies ebenfalls mit der Gattungszuschreibung „Lettres" analysiert (Martin, Mylne & Frautschi 1977, 246).

Topic „correspondance" und Topic „interaction"

15 |Wordles der Topics „correspondance" und „interaction", <https://doi.org/10.5281/zenodo.4493224>.

Das Topic, das sich im Ranking der Gewichtungen als nächstes für Choderlos Laclos' Werk als Ergebnis zeigt, ist das Topic „interaction". Es umfasst Topicwörter wie verlangen („demander"), verweigern („refuser"), geben („donner"), Mittel („moyen"), vorschlagen („proposer") etc. Das Label interaction fasst dieses semantische Cluster zusammen.

Es verweist auf Interaktionen und Prozesse des Aushandelns, die im Kontext des Romans auf psychologische Verhandlungen hindeuten. Die Bibliograph:innen

haben dies mit den Schlagworten Intrigen („intrigues") und Rache („vengeance") terminologisch stärker auf Plotmuster hin formuliert. Hier zeigt sich eine Eigenschaft des Topic Modeling Algorithmus: Er bewegt sich gewissermaßen an der Oberfläche der Wörter und bildet Verteilungen und Kookurrenzen ab, hat jedoch nicht die interpretatorische Kompetenz des menschlichen Lesers oder der Leserin, die beim Lesen der Briefe psychologische Zusammenhänge, die nicht explizit benannt werden, erschließen (das Konzept der Rache, der Intrige).

Topic „sentiment" und Topic „bonheur"

16 |Wordle der Topics „sentiment", „bonheur" (Klee & Röttgermann 2020).

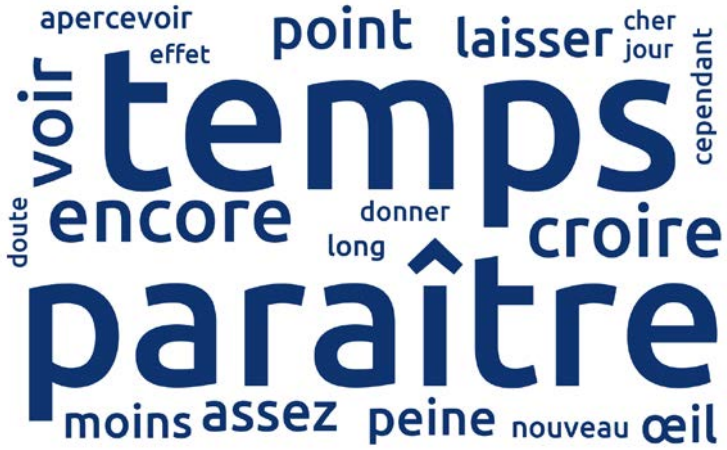
Das dritthäufigste Topic laut Topic Modeling Durchgang aus November 2020[29] für das Werk *Les Liaisons Dangereuses* ist das Topic „amour_sentiment", das Topicwörter wie Gefühl („sentiment"), Liebe („amour"), Herz („cœur"), Seele („âme") umfasst. In den statistischen Auswertungen der gesamten literarischen Produktion 1751-1800 von Seiten der Bibliograph:innen zeigt sich, dass für das Gesamtkorpus der bibliographischen Daten als vorherrschendes Thema mit 25,2 % das Thema der „intrigues sentimentales" den größten thematischen Raum im Korpus der Bibliographiedaten einnimmt (cf. dazu auch Abb. 3). Die „intrigues sentimentales" sind laut Daten aus der Bibliographie ein häufiger Topos des Romans des 18. Jahrhunderts.

[29] Topic Model des roman18 Korpus (Nov 2020), Release v0.1.0. Trier: Trier Center for Digital Humanities 2021. URL: <https://github.com/MiMoText/mmt_2020-11-19_11-38>. DOI: 10.5281/zenodo.4493224. Die hier dokumentierte Datengrundlage enthält 92 Dateien. Eingespeist in unser Wissensnetzwerk wurden jedoch nur Statements zu 79 Werken (Romanen). Die Differenz ergibt sich daraus, dass einige Files in Form von Einzelbänden vorlagen und in einem späteren Bearbeitungsschritt zu einer Datei fusioniert wurden.

Zu einem ähnlichen Schluss kommt man auch bei genauer Betrachtung der Auswertung der Topic Modeling Ergebnisse. Das Topic „amour_sentiment" hat sich bei der Ermittlung der distinktiven Topics[30] pro Werk als nicht distinktiv gezeigt. Das bedeutet, dass es ein Topic ist, welches in einer Vielzahl an Werken des Korpus prominent vorkommt. Bibliographische Daten und Topic Modeling kommen hier zu einem ähnlichen Ergebnis: Die „intrigues sentimentales" sind als literarisches Thema breit im Korpus der Bibliographie vertreten und das Topic „sentiment" ist im gesamten Romankorpus präsent.

Das Topic amour_sentiment beinhaltet Top-Topicwörter wie Liebe („amour"), Leidenschaft („passion") und Liebhaber („amant"), die semantische Überschneidungen mit dem Themenwert „libertinage", der von den Bibliograph:innen dem Werk zugeordnet wurde, aufweisen.

Topic [temps - paraître - encore]



17 | Wordcloud zum Topic [temps - paraître - encore], <https://doi.org/10.5281/zenodo.4493224>.

Die Problematik, dass einige Topics nicht ganz eindeutig zu labeln sind, zeigt sich am Beispiel des Topics [temps - paraître - encore] (Abb. 13) gut. Innerhalb der Liste der thematischen Labelkandidaten findet sich nicht nur keine Entsprechung, das gesamte Topic weist zudem eine geringe Kohärenz[31] auf. Einige Literaturwissenschaftler:innen wie zum Beispiel Ted Underwood sehen in schwer eindeutig zu labelnden, heterogenen Topics dennoch eine literaturwissenschaftlich interessante Ressource (Underwood 2012). Für das Wissensnetzwerk werden diese Topics nicht aussortiert, sondern mit einem Label, welches aus den Top-Topicwörtern besteht, eingespeist.

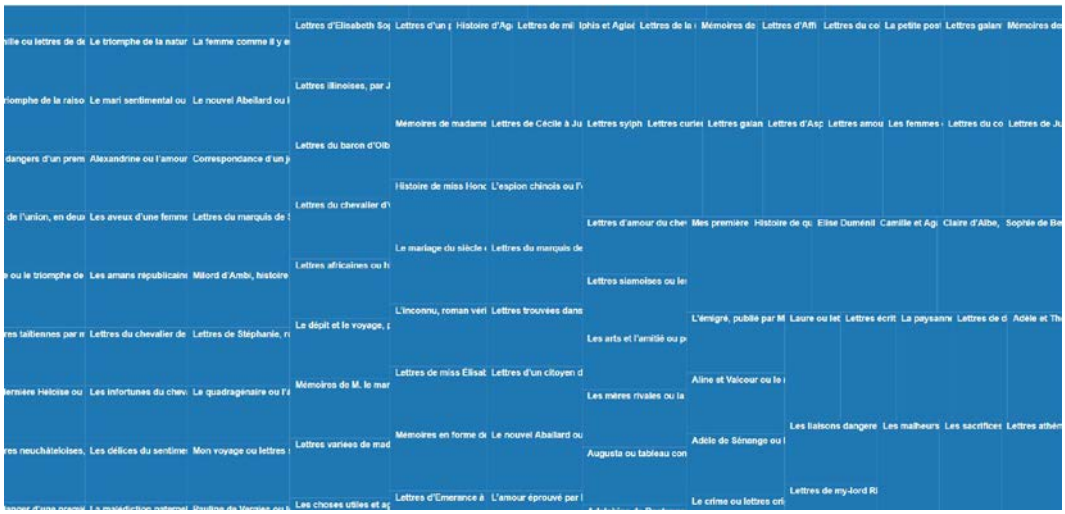
Im Abgleich der Daten zeigen sich für *Les Liaisons Dangereuses* teilweise semantische Überschneidungen (amour_sentiment/libertinage) zwischen den Ergebnissen des Topic Modelings und den bibliographischen Daten, teilweise zeigen sich aber auch komplementäre thematische Konzepte. Unterstreicht die

[30] Zur Ermittlung der distinktiven Topics vgl. Kapitel 3.4.3
[31] Zur Frage der menschlichen Bewertung der Kohärenz von Topics cf. (Chang et al. 2009).

Bibliographie die psychologischen Themen des Romans wie das Thema der Rache oder der Intrigen, extrahiert der Algorithmus des Topic Modelings in diesem Fall als weitere Topics „bonheur" und „interaction" und somit zwei Topics, die erkennen lassen, dass das doppelte Spiel der Protagonist:innen, Zynismus und Rache, nicht an der Oberfläche der Wörter erkennbar ist, sondern sich aus dem Handlungszusammenhang der Briefe und den Schlüssen des Lesenden ergeben.

Das Topic „correspondance" verweist hier auf die Gattung/Erzählform des Romans, eine Information, die auch über die bibliographischen Metadaten im MiMoText Wissensnetzwerk gespeichert ist. Interessiert sich ein Forschender für weitere Briefromane aus dem 18. Jahrhundert, lässt sich ein Überblick über die narrativen Formen im Korpus mit folgender SPARQL-Abfrage[32] im MiMoText-Wissensnetzwerk einsehen:

```
#defaultView:TreeMap
SELECT
   ?narrativeformLabel ?item ?itemLabel
WHERE
{?item wdt:P54 ?narrativeform.
 ?item wdt:P54 wd:Q3718. #?item has narrative perspective "epistolary novel"
  SERVICE wikibase:label { bd:serviceParam wikibase:language "fr". }
  }
```



18 | SPARQL-Abfrage zu narrativen Formen mit defaultView „Treemap". In der treemap werden (anklickbar: <https://tinyurl.com/2oxxeyt2>) weitere Briefromane ausgegeben. Diese Informationen basieren auf Daten aus der *Bibliographie du genre romanesque français, 1751-1800* (Martin, Mylne & Frautschi 1977).

Die Analyse des Einzelromans *Les Liaisons Dangereuses* hinsichtlich der Topics zeigt insgesamt, dass sich sinnstiftende und plausible thematische Konzepte maschinell extrahieren lassen, die jedoch in diesem Beispieltext teilweise abweichend von den von Menschenhand extrahierten thematischen Konzepten des Romans sind.

[32] Link zum projektinternen SPARQL-Endpoint: <https://query.mimotext.uni-trier.de/>Weitere Informationen zu SPARQL und ein Tutorial findet man hier : <https://mimotext.github.io/MiMoTextBase_Tutorial/>.

4.2 *Voyage autour de ma chambre* von Xavier de Maistre (1794)

> Charmant pays de l'imagination, toi que l'Être
> bienfaisant par excellence a livré aux hommes
> pour les consoler de la réalité
> (De Maistre 1794, 104)

Voyage autour de ma chambre (1794) von Xavier de Maistre ist ein Werk, das im Zuge der weltweiten Coronapandemie, die für viele Menschen im Lockdown oder „confinement" damit einherging, sich vermehrt auf die eigenen Privaträume zurückzuziehen, als Roman des 18. Jahrhunderts eine Brücke zu den Lesenden des 21. Jahrhundert schlagen kann (Laurentin 2020; Villa Ramirez & Gartner Restrepo 2020). Das zentrale Thema des Romans ist ein 42-tägiger Hausarrest, den das erzählende Ich allein in seinem Zimmer verbringt.

Schon Blaise Pascal hatte ein Jahrhundert zuvor in seinen *Pensées* bemerkt, dass alles Unglück des Menschen davon herrühre, dass er nicht ruhig in einem Zimmer zu bleiben vermöge: „Tout le malheur des hommes vient de ne savoir pas demeurer en repos, dans une chambre." (Pascal 1670, 200) Jedwede Ablenkung, jedwedes „divertissement" wie Jagd oder Tänze diene dazu, den Menschen von seiner eigenen Sterblichkeit abzulenken. Daher, so Blaise Pascal, sei auch das Gefängnis eine Bestrafung, da es den Menschen dazu verdamme, alleine ohne Ablenkung in seinem Zimmer zu bleiben (Pascal 1670, 200).

In *Voyage autour de ma chambre (1794)* verlässt der Protagonist sein Zimmer nicht, lässt jedoch in Parodie auf das im 18. Jahrhundert beliebte Genre des Reiseromans seine Phantasie schweifen und reist so imaginär aus seinem Zimmer heraus in ferne Länder. Die Struktur des Romans in 42 Kapiteln spiegelt die Dauer der Quarantäne[33] (42 Tage) wider. Auch De Maistre selbst saß, als er den Roman schrieb, aufgrund eines Duells in einem 42-tägigen Hausarrest in Turin fest.

Die Bibliograph:innen haben den Plot des Werks folgendermaßen resümiert: „Récit fantaisiste: les objets que l'auteur voit dans sa chambre évoquent des souvenirs, inspirent des réflexions." (Martin, Mylne & Frautschi 1977, 376) Zu welchem Schluss kommt der Topic Modeling Algorithmus? Die dominanten Topics werden in folgenden Statements deutlich:

[33] Im 18. Jahrhundert bezeichnet die Quarantäne noch gemäß der etymologischen Herkunft eine Dauer von 40 Tagen, analog zur religiösen 40-tägigen Einteilung der Fastenzeit. Der Eintrag QUARANTAINE der *Enyclopédie* nennt einerseits die 40-tägige Quarantäne von Schiffen zur Verhinderungen der Übertragung von Seuchen, andererseits die Quarantäne in der Rechtsprechung (D'Alembert und Diderot 1751, 658), <https://fr.wikisource.org/wiki/L%E2%80%99Encyclop%C3%A9die/1re_%C3%A9dition/QUARANTAINE>.

Maistre_Voyage	about	bonheur	https://doi.org/10.5281/zenodo.4493224
Maistre_Voyage	about	art	https://doi.org/10.5281/zenodo.4493224
Maistre_Voyage	about	nuit	https://doi.org/10.5281/zenodo.4493224
Maistre_Voyage	about	mort	https://doi.org/10.5281/zenodo.4493224
Maistre_Voyage	about	nature	https://doi.org/10.5281/zenodo.4493224

Tab.2 |Struktur der thematischen Statements zu Werken im MiMoText Knowledgegraph; zur Erläuterung der Modellierung vgl. „5. Modellierung der Themenaussagen in RDF".

Betrachten wir das Topic „art" genauer am Beispiel der Visualisierung als Wordcloud. Die Größe der Schriftart spiegelt die Gewichtung der Topicwörter im Topic wider.

19 | Wordle-Visualisierung der Topics „art", <https://doi.org/10.5281/zenodo.4493224>.

Das Topic „art" setzt sich aus Top-Topicwörtern wie Werk („ouvrage"), Buch („livre"), Gemälde („tableau"), Musik („musique"), Theater („théâtre"), Geschmack („goût") etc. zusammen. Wertet man das Werk im Close Reading aus, so bewahrheitet sich, dass es in den philosophischen Betrachtungen und Reflektionen um Glück und Kunst geht, auch um das Verhältnis der verschiedenen Künste zueinander. So berichtet das erzählende Ich beispielsweise in Kapitel XXV, dass Mme de Hautcastel Überlegungen anstellt, dass sie die Musik von Cherubini (s. im Anhang Abb. 15) und Cimarosa im Vergleich zur „alten Musik" tief berühre und dass die bildende Kunst nur von einer sehr kleinen Klasse goutiert werde, während die Musik jedwedes Lebewesen verzaubere:

> Mais que m'importe à moi, me dit un jour Mme de Hautcastel, que la musique de Cherubini ou de Cimarosa diffère de celle de leurs prédécesseurs? Que m'importe que l'ancienne musique me fasse rire, pourvu que la nouvelle m'attendrisse délicieusement? Est-il donc nécessaire à mon bonheur que mes plaisirs ressemblent à ceux de ma trisaïeule? Que me parlez-vous de peinture, d'un art qui n'est goûté que par une classe très-peu nombreuse de personnes, tandis que la musique enchante tout ce qui respire? (De Maistre 1794, 71)

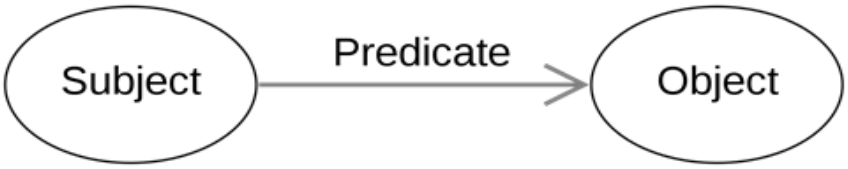
Das Close Reading bestätigt eine Diskussion des Kunstbegriffs, die sich im Topic „art" andeutete. Weit davon entfernt, nur Topics an der Oberfläche des Textes zu erkennen, konnte im diskutierten Werk mit Topic Modeling auch die psychologische und philosophische Dimension des Textes in Topics („mort", „nature", „art", „bonheur") abgebildet werden.

Am Ende von *Voyage autour de ma chambre* verlässt das erzählende Ich den geschützten Raum („la chambre") seiner Phantasiereisen und begibt sich in die Außenwelt. Denn – und hier ähnelt sich der Befund von Blaise Pascal und Xavier de Maistre – „la solitude ressemble à la mort." (De Maistre 1794, 105)

5. Modellierung der Themenaussagen in RDF

Wie in Kapitel 3 und 4 beschrieben haben wir ein Korpus an 80 Romanen der zweiten Hälfte des 18. Jahrhunderts mithilfe von Topic Modeling mit MALLET analysiert. Im Ergebnis erhalten wir 30 Topics, die mithilfe eines kontrollierten Vokabulars[34] gelabelt wurden und somit vergleichbar mit einer weiteren Informationsquelle, den bibliographischen Metadaten, sind. Die Topics wurden unter Berücksichtigung der distinktiven Topics in Relation zu den Werken zunächst in einer Liste erfasst, die alle Zuordnungen aus Top Topics und Werken enthält. Als Cutoff der berücksichtigten Topics – denn theoretisch ist jedes Topic in einer bestimmten Wahrscheinlichkeit in jedem Werk enthalten – haben wir die ersten fünf Top Topics gewählt. Die Entscheidung beruht darauf, auch hier eine Vergleichbarkeit zu den weiteren Quellen des Textminings zu erreichen (die Metadaten nennen auch bis zu fünf Themen pro Werk).

Die Ergebnisse modellieren wir sodann in Form von RDF (Resource Description Framework)[35]-Tripeln, deren Struktur sich aus drei Elementen zusammensetzt: Subjekt, Prädikat und Objekt.



20 | Grundlegendes RDF-Diagram (Wikicommons, Basic RDF Graph-en.svg, User: cmplstofB, Lizenz: WTFPL).

Übertragen auf die Modellierung unserer Themenwerte für Romane hier ein Beispiel zur Veranschaulichung:

Subjekt = Voyage autour de ma chambre

Prädikat = About

Objekt = art

[34] <https://github.com/MiMoText/vocabularies>, 24.9.2021.
[35] <https://www.w3.org/RDF/>, 24.9.2021.

In folgender Beispieltabelle (ergänzt um weitere Aussagen) bildet jede Zeile ein Tripel:

Subjekt	Prädikat	Objekt
Maistre_Voyage	about	bonheur
Maistre_Voyage	about	art

Tab. 3 | Jede Zeile bildet ein RDF-Tripel (hier in menschenlesbarer Form angedeutet).

Bezüglich des Prädikats (hier: „About") nutzen wir nach Möglichkeit bereits vorhandene Ontologien nach. „About" wurde als Prädikat aus dem Vokabular schema.org nachgenutzt.[36]

Das Projekt „Mining and Modeling Text" folgt dem Prinzip von Linked Open Data (Berners-Lee u. a. 2006), das davon ausgeht analog zu Hyperlinks in Dokumenten nun Datenobjekte bzw. Entitäten miteinander zu vernetzen. Dazu müssen Daten so ausgezeichnet werden, dass sie anschlussfähig an die Linked Open Data Cloud sind. Daher wurden im Projekt alle Themenwerte mit Identifiern aus dem Linked Open Data Hub Wikidata verknüpft. Dies ermöglicht eine Disambiguierung und bietet die Möglichkeit bei Bedarf über diese einheitliche ID automatisiert weitere Informationen aus der Linked Open Data Cloud abzurufen, beispielsweise das Label des Themenkonzepts in einer anderen Sprache (hier: Englisch).

Subjekt	Prädikat	Objekt	Wikidata ID	Label: engl
Maistre_Voyage	about	bonheur	https://www.wikidata.org/wiki/Q8	-> happiness
Maistre_Voyage	about	art	https://www.wikidata.org/wiki/Q735	-> art

Tab.4 |Über Wikidata-Identifier erreichen wir eine Disambiguierung und können weitere Informationen zu den Themenwerten abrufen, beispielsweise Label in anderen Sprachen.

Die RDF-Tripel, die auf den Ergebnissen des Topic Modelings basieren, werden aggregiert und in unsere lokale Instanz der offenen und freien Software Wikibase importiert. Dort liegen Sie nach dem Import in einer wikifizierten Form vor und können über den projekteigenen SPARQL-Endpoint abgefragt werden.

[36] <https://schema.org/about>, 28.01.2022.

21 | Projektinterne Wikibase-Instanz <http://data.mimotext.uni-trier.de>, hier der Eintrag zu *Voyage autour de ma chambre* (1794).

Durch die Aggregation tausender RDF-Tripel[37] zu den französischen Romanen 1751-1800 bildet sich ein Graph, der sich via SPARQL abfragen lässt. Es ließe sich zum Beispiel fragen, welche Romane der Zeit das Thema „famille" enthalten oder ob sich das Thema „éducation" im Zeitverlauf 1751 bis 1800 verändert, beispielsweise vor und nach 1789. Auch Kombinationen an Abfragen sind möglich: Enthalten Romane der Kategorie „Brief" vorrangig ein bestimmtes Thema? Zeige mir Romane, die von Frauen geschrieben wurden und das Topic „nature" enthalten etc. Das übergeordnete Ziel der hier vorgestellten Extraktion von Topics aus französischen Volltexten ist es, diese im Zusammenspiel mit aus weiteren Quellen (insbesondere Sekundärliteratur und bibliographische Metadaten) extrahierten Aussagen im Sinne einer "data-rich literary history" (Bode 2018, 37–57) als Wissensgraphen zu modellieren.

Die Annäherung an den Wissensgraphen über die thematischen RDF-Tripel ermöglichte es uns, unseren technischen Workflow zu etablieren und einen Grundstock an relevanten Aussagen zu den Primärtexten in unser Netzwerk einzuspeisen. Insbesondere der Vergleich der Ergebnisse aus unüberwachtem maschinellem Lernen (Topic Modeling) mit den Ergebnissen aus der Analyse der bibliographischen Metadaten ist aufschlussreich, da sie einen Mensch-Maschine-Vergleich ermöglicht.

Nachdem die erste Projektphase von MiMoText vom Korpusaufbau und dem Einspeisen von Tripeln zu Erzählformen, Textlänge, Themen, Publikationsdatum und Autor:innen getragen war, planen wir als nächste Schritte weitere RDF-Tripel

[37] Aktuell sind ca. 30.000 Tripel zu Autor:innen und Werken der französischen Prosa 1751-1800 eingespeist (Stand 28.01.2022).

zu Figuren, Handlungsorten und zur Tonalität der Romane zu erheben und auf diese Weise das Wissensnetzwerk kontinuierlich weiter anzureichern.

Hinweise

Die verwendeten Daten, Zwischenergebnisse und Visualisierungen sind auf GitHub verfügbar.

ROMANKORPUS
<https://github.com/MiMoText/roman18/>
DOI: 10.5281/ZENODO.4061904.
TOPIC MODELING WORKFLOW
<https://github.com/MiMoText/topicmodeling>
doi:10.5281/zenodo.4493223.

Das Projekt „Mining and Modeling Text" wird an der Universität Trier durchgeführt und durch die Forschungsinitiative des Landes Rheinland-Pfalz 2019-2023 gefördert.

Bibliographie

Software

KELLY, Ryan. 2004-2011. "PyEnchant".
<https://github.com/pyenchant/pyenchant>, 24.8.2021.

MCCALLUM, Andrew Kachites. 2002. "MALLET: A Machine Learning for Language Toolkit."
<http://mallet.cs.umass.edu/topics.php>, 3.12.2020.

PRESTO TEAM. 2014. "PRESTO. Projet ANR/DFG: L'évolution du système prépositionnel du français."
<http://presto.ens-lyon.fr/?page_id=197>, 14.1.2020.

ŘEHŮŘEK, Radim, und Petr Sojka. 2010. "Gensim: topic modelling for humans."
<https://pypi.org/project/gensim/>, 3.12.2020.

ŘEHŮŘEK, Radim, und Petr Sojka. 2010. "Gensim KeyedVectors."
<https://radimrehurek.com/gensim/models/keyedvectors.html>, 17.12.2020.

ŘEHŮŘEK, Radim, und Petr Sojka. 2010. "Gensim mallet wrapper."
<https://radimrehurek.com/gensim/models/wrappers/ldamallet.html >, 3.12.2020.

REUL, Christian et al. 2019. "OCR4all — An open-source tool providing a (semi-) automatic OCR workflow for historical printings." *Applied Sciences* 9 (22).
<https://github.com/OCR4all/getting_started>, 4.12.2020.

SCHMID, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees." *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK: Association for Computational Linguistics.
<https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>, 3.12.2020.

SCHÖCH, Christof. 2020. *Simple Topic Modeling pipeline using TextBlob and gensim.*
<https://github.com/dh-trier/topicmodeling/>, 21.01.2022

Datensätze

FAUCONNIER, Jean-Philippe. 2015. *French Word Embedding Models*.
<https://fauconnier.github.io/>.

KLEE, Anne. & Röttgermann, Julia. 2020. *Doing topic modeling on French 18th century novels in the context of MiMoText project [Data set]*.

<https://github.com/MiMoText/topicmodeling>.
<https://doi.org/10.5281/ZENODO.4493223>.

LÜSCHOW, Andreas. 2019. *Bibliographie du genre romanesque français 1751-1800: RDF model [Data set]* [French]. Trier University.
<http://doi.org/10.5281/zenodo.3401428>.

RÖTTGERMANN, Julia (ed.), contributors: Dudar, J., Klee, A., Konstanciak, J., A., Ondraszek S., Probst, A., Schöch, C. 2020. *Collection de romans français du dix-huitième siècle (1750-1800) / Eighteenth-Century French Novels (1750-1800) [Data set]*, Release v0.1.0. Trier: TCDH, 2020.
URL: <https://github.com/mimotext/roman18>.
DOI: <https://doi.org/10.5281/zenodo.4061903>.

Referenzen

BERNERS-LEE, Tim et al. 2006. „A Framework for Web Science". *Foundations and Trends in Web Science* 1, Nr. 1, 1–130.
<https://doi.org/10.1561/1800000001>.

BLEI, D. M. 2011. "Introduction to Probabilistic Topic Models". *Communications of the ACM*, 1–16.

BLEI, D. M. et al. 2003. "Latent dirichlet allocation." *The Journal of Machine Learning Research*, 3, 993–1022.

BODE, Katherine. 2018. „`Man people woman life´/`Creek sheep cattle horses´: Influence, Distinction, and Literary Traditions". In *A World of Fiction: Digital Collections and the Future of Literary History*, 157–98. University of Michigan Press.
<https://www.jstor.org/stable/j.ctvdtpj1d.10>.

BODE, Katherine. 2018. *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press.

BOND, Elizabeth Andrews & Robert M. Bond. 2020. „Topic Modelling the French Pre-Revolutionary Press". In *Digitizing Enlightenment : Digital Humanities and the Transformation of Eighteenth-Century Studies*, ed. Simon Burrows und Glenn Roe, 247–76. Oxford: Liverpool Univ. Press.

BURNARD, Lou. 2014. "What Is the Text Encoding Initiative? : How to Add Intelligent Markup to Digital Resources." *Encyclopédie Numérique*. Marseille: OpenEdition Press.
<http://books.openedition.org/oep/426>.

BURNARD, Lou & Carolin Odebrecht. 2019. „COST-ELTeC/Schemas: level0 and level1 release." *Zenodo*.
<https://doi.org/10.5281/zenodo.3490758>.

CHANG, Jonathan et al. 2009. „Reading Tea Leaves: How Humans Interpret Topic Models." In *NIPS*.
<http://books.nips.cc/papers/files/nips22/NIPS2009_0125.pdf>.

D'ALEMBERT, Jean Le Rond & Denis Diderot. 1751. *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers*. Paris: Briasson.
<https://fr.wikisource.org/wiki/Encyclop%C3%A9die,_ou_Dictionnaire_raisonn%C3%A9_des_sciences,_des_arts_et_des_m%C3%A9tiers>.

DAWSON, Robert L. 1978. „Review :The Martin, Mylne, Frautschi Bibliographie du genre romanesque français." Eighteenth-Century Studies 11, Nr. 4, 497–508.
<https://doi.org/10.2307/2737969>.

DELON, Michel. 1997. *Dictionnaire européen des Lumières*. Paris: PUF.

DE MAISTRE, Xavier. 1794. *Voyage autour de ma chambre*. Paris: Firmin-Didot et Cie.
<https://fr.wikisource.org/wiki/Voyage_autour_de_ma_chambre>.

DU, Keli et al. 2021. „Zeta & Eta: An Exploration and Evaluation of two Dispersion-based Measures of Distinctiveness." In *Proceedings of the*

Conference on Computational Humanities Research 2021. Amsterdam. <http://ceur-ws.org/Vol-2989/short_paper11.pdf>.

JOCKERS, Matthew L. 2014. „Topic Modeling." In *Text Analysis with R for Students of Literature*, ed. Matthew L. Jockers, 135–59. Quantitative Methods in the Humanities and Social Sciences. Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-319-03164-4_13>.

LAURENTIN, Emmanuel. 2020. „Lire ‚Voyage autour de ma chambre', un texte ô combien d'actualité !" *France Culture,* 22.03. <https://www.franceculture.fr/litterature/lire-voyage-autour-de-ma-chambre-un-texte-o-combien-dactualite>.

MARTIN, A., V. Mylne & R. L. Frautschi. 1977. *Bibliographie du genre romanesque français, 1751-1800*. London: Mansell.

MIKOLOV, T. et al. 2013. *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781 [cs], 6.09. <http://arxiv.org/abs/1301.3781>.

PASCAL, Blaise. 1670. *Pensées de M. Pascal sur la religion et sur quelques autres sujets, qui ont esté trouvées après sa mort parmy ses papiers*. Paris: Guillaume Desprez. <https://fr.wikisource.org/wiki/Livre:Pascal_-_Pens%C3%A9es,_%C3%A9dition_de_Port-Royal,_1670.djvu>.

ŘEHŮŘEK, Radim & Petr Sojka. 2010. "Software framework for topic modelling with large corpora.", In *Proceedings of the LREC 2010 Workshop on new Challenges for NLP Frameworks*, 45-50. <https://is.muni.cz/publication/884893/lrec2010-rehurek-sojka.pdf>.

REUL, Christian et al. 2019. OCR4all -- An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. arXiv:1909.04032 [cs], September. <http://arxiv.org/abs/1909.04032>.

RHODY, Lisa M. 2013. "Topic Modeling and Figurative Language." *Journal of Digital Humanities*. 7 April. <http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody>.

ROE, Glenn, Clovis Gladstone & Robert Morrissey. 2016. „Discourses and Disciplines in the Enlightenment: Topic Modeling the French Encyclopédie." *Frontiers in Digital Humanities 2*, Lausanne: Frontiers Media SA. <https://doi.org/10.3389/fdigh.2015.00008>.

SARKAR, Dipanjan. 2019. *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*. Second edition. New York: Apress.

SCHÖCH, Christof. 2017. „Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html >.

SOUVAY, G. & J.-M Pierre. 2009. "LGeRM: Lemmatisation de mots en moyen français." *Traitement Automatique des Langues*, 50 (2). <https://halshs.archives-ouvertes.fr/halshs-00396452/document>.

STEYVERS, M. & T. Griffiths. 2007. „Probabilistic topic models." In *Handbook of latent semantic analysis*, ed. Landauer, T. K. et al., 427–448, Mahwah: Lawrence Erlbaum Associates Publishers.

UNDERWOOD, Ted. 2012. „What kinds of "topics" does topic modeling actually produce?" *The Stone and the Shell* (blog). <http://tedunderwood.com/2012/04/01/what-kinds-of-topics-does-topic-modeling-actually-produce/>.

UGLANOVA, Inna & Evelyn Gius. 2020. "The Order of Things. A Study on Topic Modelling of Literary Texts." *CHR 2020: Workshop on Computational Humanities Research*, November 18–20, 2020, Amsterdam, The

Netherlands.
<http://ceur-ws.org/Vol-2723/long7.pdf>.

VILLA RAMIREZ, Oscar Jhony & Carolina Fernanda Gartner Restrepo. 2020. „Voyage autour de ma chambre dans le temps du covid-19: confrontant la réalité." *Miguilim - Revista Eletrônica do Netlli* 9, Nr. 3, 331–41.
<https://doi.org/10.47295/mgren.v9i3.2582>, 7.12.2020.

Zusammenfassung

Wie lassen sich romanistische Korpora hinsichtlich ihrer literarischen Themen mit digitalen Methoden explorativ erforschen? Im Kontext des Verbundprojekts „Mining and Modeling Text" wurde Topic Modeling mit MALLET (McCallum 2002) auf ein Korpus von 80 französischen Romanen aus der Zeit von 1750 bis 1800 (Röttgermann et al. 2020) angewandt. Ziel des Topic Modeling-Ansatzes ist es dabei, Aussagen über die Themen von Werken zu treffen, die in Form von RDF-Tripeln in ein auf Wikibase basierendes Wissensnetzwerk einfließen. Die übergeordnete, interdisziplinäre und neuartige Idee ist es dabei, datenbasierte Literaturgeschichtsschreibung zu betreiben. Neben der Informationsextraktion aus Primärtexten speist sich das Wissensnetzwerk auch aus bereits digitalisierten bibliographischen Daten (Martin, Mylne & Frautschi 1977; Lüschow 2019). Im Zusammenspiel dieser beiden Informationsflüsse lässt sich über ein gemeinsames kontrolliertes Vokabular ein aufschlussreicher Datenabgleich vollziehen: Welche Themen der Werke wurden durch die Bibliograph:innen identifiziert und welche Topics treten durch den Topic Modeling-Algorithmus zutage? Zwei Fallstudien zu Choderlos de Laclos und Xavier de Maistre exemplifizieren die Vorgehensweise und das Potential dieses Ansatzes.

Abstract

How can Romance corpora be digitally researched exploratively with regard to their literary topics? In the context of the project "Mining and Modeling Text", topic modeling with MALLET (McCallum 2002) was applied to a corpus of 80 French novels 1750-1800 (Röttgermann et al. 2020). The aim of the topic modeling approach is to generate statements about the topics of the novels, which then are imported into a knowledge graph based on Wikibase. The overriding, interdisciplinary and novel idea is to practice data-based literary historiography. In addition to information extraction on primary texts, the knowledge network is also fed from digitized bibliographic data (Martin, Mylne & Frautschi 1977; Lüschow 2019). In the interplay of these two data types, a comparison can be carried out: Which "topics" of the novels were identified by the bibliographers and which topics are revealed by the algorithm? Two case studies on Choderlos de Laclos and Xavier de Maistre exemplify this.

2022, n°9
pp. 87-117
doi: https://doi.org/10.15460/apropos.9.1878

Jan Rohden

Stilometrische Annäherungen an den italienischen Petrarkismus

Jan Rohden
ist Mitarbeiter bei der Deutschen Forschungsgemeinschaft (DFG).
janrohden@gmail.com

Keywords

Petrarkismus – Stilometrie – Kontrastive Analyse – Kookurrenzanalyse – Netzwerkanalyse

1. Francesco Petrarca als Begründer des europäischen Petrarkismus

Francesco Petrarca (1304–74) kann als einer der einflussreichsten Autoren der europäischen Literatur angesehen werden. Einer der Hauptgründe dafür ist seine viel beachtete Gedichtsammlung in italienischer Sprache namens *Canzoniere*, die nicht nur die literarische Landschaft seiner Zeit, sondern auch die Werke vieler späterer Autoren beeinflusste.

Im Zentrum des *Canzoniere* steht die unerwiderte Liebe des Sprechers zu der verheirateten Laura. Obgleich die Sammlung aus thematischer Perspektive somit wenig ungewöhnlich erscheint, weist sie im Vergleich zu früherer Liebeslyrik einige Besonderheiten auf, beginnend damit, dass die Liebe des Sprechers auch nach dem Tod der Geliebten fortbesteht und poetisch verbalisiert wird. Die gemäß einer planvollen Struktur angeordneten 366 Gedichte der Sammlung greifen Motive aus der lateinischen, provenzalischen und italienischen Literatur auf. Dabei schildert der *Canzoniere* detailliert die oft widersprüchlichen Gefühle des Sprechers für Laura, der häufig hin- und hergerissen ist zwischen dem Glücksgefühl seiner Liebe und dem Schmerz, der aus seinem unerfüllten Verlangen resultiert. Diese von der Forschung oft als „dolendi voluptas" bezeichnete, konträre Gefühlslage des Sprechers ist ein illustratives Beispiel für die Gegensätzlichkeit, die Form und Inhalt des *Canzoniere* widerspiegeln (cf. Friedrich 1964, 217–219; Forster 1969).

Petrarcas Gedichtsammlung wurde nicht nur in Italien, sondern in ganz Europa stark rezipiert. Sein *Canzoniere* avancierte für Jahrhunderte zu einem poetischen Modell, an dem sich die Gedichtsammlungen zahlreicher Autoren orientierten. Dies verdeutlicht eine Vielzahl an Untersuchungen, die Elemente der Dichtung Petrarcas

in den Texten anderer europäischer Autoren nachweisen konnten. Zur Beschreibung jener virulenten Auswirkungen Petrarcas auf die europäische Literatur hat sich in der literaturwissenschaftlichen Forschung der Begriff „Petrarkismus“ etabliert. Obwohl Petrarcas literarischer Einfluss somit gut dokumentiert ist, existiert ein lebhafter Forschungsdiskurs über die Definition des Petrarkismus, wie nicht zuletzt eine einschlägige Bibliographie darlegt (cf. Hempfer, Regn & Scheffel 2005). Während ein großer Teil der Forschungsliteratur der Wirkung von Petrarcas poetischem Stil auf einzelne Autoren oder Texte nachgeht (cf. etwa Pyritz 1963; Regn 1987; Morales Saravia 1998; Schiffer 2000; Marnoto 2015), zielen einige Beiträge darauf ab, Petrarcas Einfluss auf die spätere literarische Landschaft systematisch zu beschreiben (cf. zum Beispiel Baldacci 1957; Forster 1969, 61–83; Hoffmeister 1973; Nardone 1998; Bernsen 2011; Regn 2013). Von den zuletzt genannten Untersuchungen erweisen sich die deutschsprachigen Arbeiten als besonders variantenreich. Unter ihnen stechen drei Ansätze hervor: Der erste von ihnen definiert den Petrarkismus in Anlehnung an die Systemtheorie als ein literarisches System von unterschiedlichen Elementen (cf. Hempfer 1987; Hempfer 1991; Regn 1987; Regn 1993).[1] Das entscheidende Merkmal des Petrarkismus sind jenem Konzept zufolge somit die im Text vorkommenden sprachlichen Aspekte: Nur ein Werk, das eine ausreichende Menge an für den Petrarkismus charakteristischen Elementen aufweist, ist als petrarkistisch zu betrachten.

Ausgehend von Michail Bachtins Konzeption der Dialogizität versteht der zweite Definitionsansatz von Rainer Warning (cf. Warning 1987) den Petrarkismus als eine Art der literarischen Aneignung miteinander konkurrierender Arten des lyrischen Sprechens, die besonders an der Gedichtform des Sonetts deutlich wird. Dem Verfasser zufolge ist ein Werk als petrarkistisch zu erachten, sofern die petrarkistische Sprechweise im Vergleich zu allen anderen im Text erkennbaren Diskursen dominiert.[2]

Einen dritten Ansatz zur Definition des Petrarkismus präsentiert Gerhard Regn in einem Lexikonartikel (cf. Regn 2013), der als Vermittlung der ersten zwei Forschungspositionen angesehen werden kann. Darin entwickelt Regn ein topographisch orientiertes Modell des Petrarkismus, das zwischen einem definitorischen Zentrum und einer Peripherie unterscheidet. Ein Werk ist im definitorischen Zentrum des Petrarkismus zu verorten, sofern es die von Petrarca begründete Liebeskonzeption mit Hilfe von spezifisch petrarkistischen Sprachelementen zum Ausdruck bringt. Werden hingegen lediglich einzelne Aspekte der Dichtung

[1] Aus systemtheoretischer Perspektive typische Merkmale des italienischen Petrarkismus des 16. Jahrhunderts sind beispielsweise eine gegensätzlich-paradoxale Affektstruktur, eine narrative Dimension der Lyrik mit autobiographischen Zügen und ein Zusammenspiel zwischen affektgeleiteten und affektübergreifenden, normgesteuerten Befindlichkeiten (cf. Regn 1987, 21–48).

[2] Warning stellt die Unterscheidbarkeit eines petrarkistischen Systems insofern grundsätzlich in Frage, als die dafür konstitutiven Elemente zwar eine Ähnlichkeit zu Petrarcas Werk aufzeigen können, jedoch nicht die für Warning entscheidende „Distinktivität eines ‚petrarkistischen Systems‘ gegenüber Petrarca selbst“ (Warning 1987, 338). Infolge dessen stehen sprachliche Elemente weniger im Zentrum von Warnings Definition, sondern stattdessen die Kombination unterschiedlicher Sprechweisen zu einem „Polylog, der die Sprache Petrarcas mit apetrarkischen Sprachen in je neue Beziehungen bringt“ (Warning 1987, 339) und besonders in der Gattung des Sonetts zur Sprache kommt.

Petrarcas aufgegriffen und mit fremden Elementen kombiniert, so distanziert sich ein Text vom petrarkistischen Mittelpunkt und nähert sich der Peripherie. Wird die Entfernung zum petrarkistischen Zentrum schließlich zu groß, so ist der betreffende Text nicht mehr dem Petrarkismus zuzuordnen. Poetisch wird eine Annäherung an das Zentrum durch dichterische Imitatio, eine Entfernung in Richtung Peripherie mittels Variation bzw. aemulatio erreicht.

Die vorgestellten Herangehensweisen zur Definition des Petrarkismus erweisen sich insbesondere in der Zusammenschau als erhellend, da sie unterschiedliche Aspekte als Gradmesser zur Klassifikation von Texten als petrarkistisch bzw. nicht petrarkistisch einsetzen. Während systemtheoretisch orientierte Ansätze vor allem auf sprachliche Elemente zur Bestimmung petrarkistischer Texte zurückgreifen, identifiziert Warnings Konzeption lyrische Diskursstrukturen als distinktiv für den Petrarkismus. Regn wiederum versteht den Petrarkismus als graduelles topographisches Modell, bei dem die Relation von petrarkistischen und nicht petrarkistischen Aspekten über den Grad an Virulenz des Petrarkismus im jeweiligen Text Auskunft gibt. Zur Evaluation der drei genannten Definitionen, aber auch zum besseren Verständnis des Petrarkismus als literaturwissenschaftliches Phänomen, erscheint die Analyse eines quantitativ umfassenden Korpus aus petrarkistischen Texten vor diesem Hintergrund aufschlussreich. Dank digitaler Verfahren sind quantitative Untersuchungen in den letzten Jahren nicht nur möglich, sondern zunehmend populärer geworden, auch wenn die digitale Modellierung literaturwissenschaftlich relevanter Informationen (noch) nicht für jede Forschungsfrage problemlos möglich und zudem selbst Gegenstand der Forschung ist.[3] In den Digital Humanities existieren unter anderem drei Methoden, mit deren Hilfe die drei vorgestellten Definitionsansätze auf quantitative Weise auf größere Textkorpora übertragen werden können. Es handelt sich dabei um kontrastive Analysen,[4] Kookkurrenzanalysen[5] und Netzwerkanalysen[6].

[3] Ein Projekt, das die Modellierung von literaturgeschichtlich relevanten Informationen in den Blick nimmt, ist beispielsweise *MiMoText* (cf. MiMoText 2022).

[4] Kontrastive Analysen zielen auf eine vergleichende Untersuchung mehrerer Texte oder Textgruppen ab. Zu diesem Zweck wird die zu analysierende Textsammlung nach einschlägigen Kriterien in zwei Partitionen (also Teilgruppen) aufgeteilt, um durch Vergleich auffällige Elemente zu ermitteln. Cf. zu kontrastiven Analysen umfassend Schöch 2018; Schöch et al. 2018a; Schöch et al. 2018b.

[5] Kookkurrenzen können definiert werden als „Gruppen von Wörtern, die häufiger zusammen auftreten, als dass es rein zufällig sein könnte." (Bubenhofer 2006–2022). Cf. zu Kookkurrenzanalysen in der Literaturwissenschaft etwa Suzuki et al. 2012.

[6] Die Netzwerkanalyse ist eine digitale Methode zur Visualisierung von Relationen innerhalb von Datenstrukturen, die seit Jahrzehnten in unterschiedlichen Kontexten eingesetzt wird. Obwohl die Vorzüge des Verfahrens am Beispiel von menschlichen Beziehungen, etwa im soziologischen Bereich, besonders ersichtlich sind, hat sich die Netzwerkanalyse auch aus literaturwissenschaftlicher Perspektive als aufschlussreich erwiesen (cf. Trilcke 2013; Jannidis 2017). In den letzten Jahren wurden netzwerkanalytische Ansätze beispielsweise zur Untersuchung stilometrischer Fragestellungen herangezogen, wobei unter anderem Vorteile der Visualisierung in Form von Netzwerken vermerkt wurden (cf. Eder 2017; Rotari, Jander & Rybicki 2021).

2. Methodischer Ansatz

Trotz aller Unterschiede beschreiben alle Definitionen des Petrarkismus einen literarischen Einfluss des *Canzoniere* auf die Werke verschiedener Autoren. Daraus lassen sich zwei implizite Prämissen ableiten: Zum einen ist davon auszugehen, dass eine erkennbare Ähnlichkeit zwischen der Dichtung Petrarcas und den Texten seiner Rezipienten existiert. Dies bedeutet zum anderen, dass petrarkistische Werke offenbar charakteristische Merkmale aufweisen, die sie von vor-petrarkistischen und anderen nicht petrarkistischen Texten unterscheiden.

Digitale Verfahren können dabei helfen, derartige Unterscheidungsmerkmale zu finden. Je nach Methode können verschiedene Arten von Elementen in den Blick genommen werden, wodurch unterschiedlichen Erklärungsansätzen zum Petrarkismus Rechnung getragen werden kann, wie etwa den in Kapitel 1 vorgestellten Definitionen: So können beispielsweise kontrastive Analysen distinktive sprachliche Elemente ermitteln, die unter anderem für systemtheoretische Betrachtungen des Petrarkismus relevant sind. Auf diskursive Strukturen des lyrischen Sprechens, die für Warnings Konzept des Petrarkismus maßgeblich sind und sich mitunter in der Verknüpfung verschiedener Wort- und Motivfelder manifestieren, können Kookkurrenzanalysen Hinweise liefern. Topographische Konzeptionen des Petrarkismus im Sinne Regns lassen sich mittels Netzwerkanalysen veranschaulichen.

Die drei genannten digitalen Analyseverfahren sollen im Folgenden zur Untersuchung des italienischen Petrarkismus eingesetzt werden. Zu diesem Zweck soll ein umfassendes Textkorpus zuerst kontrastiv analysiert werden. Die dadurch eruierten distinktiven Elemente sollen danach Kookkurrenzanalysen unterzogen werden, um den Kontext jener Unterscheidungsmerkmale näher zu beleuchten und potentielle Verbindungen zwischen semantischen Feldern aufzudecken. Schließlich soll das untersuchte Korpus als Netzwerk[7] visualisiert werden, damit so Relationen und Entfernungsverhältnisse zwischen den Texten sichtbar werden.

Mit den drei genannten Methoden gilt es das nachfolgende Korpus zu untersuchen.

3. Korpus

Das Korpus besteht aus insgesamt 55 Gedichtsammlungen in italienischer Sprache, die vorwiegend aus dem 14., 15. und 16. Jahrhundert stammen.[8] Es ist in zwei Partitionen unterteilt: Die Zielpartition umfasst 51 Gedichtsammlungen, in denen die Forschung Anzeichen für den literarischen Einfluss Petrarcas aufgezeigt hat. Die Vergleichspartition hingegen beinhaltet in vier Sammlungen einen wesentlichen Teil der italienischen Liebesdichtung aus der Zeit vor dem Erscheinen von Petrarcas

[7] Die Netzwerkvisualisierungen beruhen jeweils auf den Häufigkeiten der gemäß den kontrastiven Analysen präferierten und somit distinktiven Wörter. Die Anzahl der für die Visualisierung herangezogenen Wörter variiert je nach Visualisierung zwischen mindestens 100 und maximal 1000. Für die Darstellung wurde der für kleinere Netzwerke (1-1000 Knoten) geeignete Layoutalgorithmus Fruchterman Reingold gewählt (cf. Schumacher 2020, § 75).

[8] Cf. für das gesamte Korpus sowie eine vollständige Liste der darin enthaltenen Sammlungen Rohden 2021a.

Canzoniere. Der unterschiedliche Umfang der beiden Partitionen des Korpus ist insofern der literaturhistorischen Überlieferungssituation geschuldet, als die Anzahl der seit Petrarcas *Canzoniere* veröffentlichten Sammlungen von Liebesdichtung die Zahl der zuvor erschienenen Sammlungen bei weitem übersteigt.[9] Bis auf eine Ausnahme (cf. Zaccagnini & Parducci 1915) basieren alle Texte des Korpus auf wissenschaftlichen Ausgaben, ediert von renommierten Herausgebern und Verlagen. Die digitalen Versionen jener Editionen entstammen der virtuellen Bibliothek *Biblioteca Italiana* (cf. Quondam, Alfonzetti & Asperti 2003).

In allen Texten des Korpus wurden die folgenden Elemente entfernt: Satzzeichen, Seiten- und Zeilennummern, Fußnoten, Endnoten, Überschriften, Titelblätter, Vorworte, Nachworte und sonstige Kommentare. Darüber hinaus wurden Passagen getilgt, die ausdrücklich als Prosatexte gekennzeichnet sind (etwa Widmungen, Kommentare und Einführungen der Verfasser). Ferner wurden alle Majuskeln in Minuskeln umgewandelt. Ohne jegliche weitere sprachliche Harmonisierung der Texte wurde jede einzelne Gedichtsammlung als Textdatei im txt-Format mit UTF-8-Kodierung gespeichert. Daraus ergibt sich das in Tabelle 1 im Überblick dargestellte Textkorpus:

Partition	*Art der Dichtung*	*Anzahl der Sammlungen*	*Zeichenanzahl (mit Leerzeichen)*
Zielpartition	Petrarkistisch	51	10.212.741
Vergleichspartition	Vorpetrarkistisch	4	680.707
Insgesamt		*55*	*10.893.448*

Tab. 1 | Korpus im Überblick

4. Kontrastive Analyse des Korpus

Für die Durchführung der kontrastiven Analysen werden die petrarkistische Zielpartition und die vorpetrarkistische Vergleichspartition des Korpus mit Hilfe der R-Bibliothek *stylo* (cf. Eder, Rybicki & Kestemont 2016) vergleichend analysiert. Als Distanzmaß zur mathematischen Ermittlung distinktiver Elemente dient dabei eine eigens entwickelte Variante des Dispersionsmaßes „Zeta“[10] namens „Ederlog-Zeta“. Der Grund für die Auswahl dieser spezifischen Zeta-Variante ist, dass Ederlog-Zeta bei Evaluationen durch Klassifikationstests nach dem Modell von Schöch (cf. Schöch et al. 2018a; Schöch et al. 2018b) am Beispiel petrarkistischer

[9] Dem Größenunterschied zwischen Ziel- und Vergleichspartition wurde durch die gezielte Wahl der Parameter der Analyse Rechnung getragen. So wurden bei der kontrastiven Analyse alle Gedichtsammlungen in vergleichsweise kleine Segmente (3.500 Wörter) eingeteilt, um sicherzustellen, dass von jeder Sammlung mindestens zwei Textteile in die Analyse eingehen. Bei der Kookkurrenzanalyse und der Netzwerkanalyse wurden die Schwerpunkte in erster Linie auf normalisierte bzw. gewichtete Werte gelegt. Für die Netzwerkvisualisierungen wurde überdies bewusst jeweils eine relativ geringe Anzahl an präferierten Wörtern berücksichtigt (100–1000).

[10] Cf. für Erläuterungen zu den mathematischen Hintergründen von Zeta Schöch 2018.

Lyrik in deutscher, englischer und italienischer Sprache eine vergleichsweise hohe Leistungsfähigkeit zeigt.[11]

Aus der kontrastiven Analyse des Korpus auf Basis von Ederlog-Zeta ergibt sich eine Liste aus 1.675 vergleichsweise präferierten und somit distinktiven sprachlichen Elementen (cf. Rohden 2021c),[12] von denen zahlreiche eine semantische Bedeutung und insofern eine recht hohe Interpretierbarkeit aufweisen.[13] Einige der Wörter aus der Liste lassen sich anhand ihrer semantischen Bedeutung gruppieren und einschlägigen semantischen Feldern zuordnen, wie Tabelle 2 veranschaulicht:

Semantisches Feld	*Wörter*
Alleinsein und Einsamkeit	sola, soli, deserto, disperso, perdi, perduti, persa, persi, perso
Alter und Antike	antica, antiche, antichi, antico, età, etate, greco, istoria, latini, passata, passati, secol, vecchi, vecchio
Architektur	casa, colonne, contrade, dimoranza, laberinto, marmo, mura, sassi, sasso, stanza, tempio, tetto, villa
Charaktereigenschaften	
- Bescheidenheit	umil, umili
- Ehre	honore, onora, onorata, onorati, onorato
- Ehrlichkeit	onesta, onestà, onestate, oneste, onesto
- Elend	miser, misera, miseri, misero
- Engherzigkeit	meschin
- Frömmigkeit	pia, pio
- Gelehrsamkeit	dotta
- Gnade	clemenza, pietade, pietanza, pietoso
- Grazie und Anmut	gratia, grazia, grazioso, leggiadre, leggiadri, leggiadro
- Grausamkeit	crudeltà, dispietato
- Gutmütigkeit	benvoglienza, benigna, benigno, bontà
- Höflichkeit	gradita, gradito
- Leidenschaft	passion, passione
- Mut	coragio
- Reinheit	casta, casto
- Seelen(Adel)	gentili
- Tugend	virtù, virtude, virtute, virtuti
- Verstand	ingegno, intelletti

[11] Die Evaluation dieser und weiterer Zeta-Varianten wird in einem bei der Zeitschrift *Digital Scholarship in the Humanities* zur Veröffentlichung eingereichten Aufsatz eingehend beschrieben. Cf. für den Code zur Implementierung von Ederlog-Zeta in *stylo* Rohden 2021b.

[12] Um die Wortliste auf ein überschaubares Maß zu reduzieren und ein zu seltenes Vorkommen der ermittelten Elemente im Korpus zu verhindern, beinhaltet diese Liste nur Wörter, die mindestens 15-mal im Korpus auftreten.

[13] Dass Varianten von Zeta insbesondere verglichen mit quantitativ ermittelten Metriken wie den am häufigsten auftretenden Wörtern eine hohe Interpretierbarkeit haben, konnte bereits an anderer Stelle nachgewiesen werden (cf. Schöch et al. 2018a).

- Verwirrung und Wahnsinn - Vorsicht - Würde - Wissen - Zahmheit - Zufriedenheit	confusa, confuso, turba, turbate, turbato cauto, prudente degni sapea, sapendo, saper, sappi, sappia, saprei mansueta contenti
Dichtung und Sprache - Dante - Gattungen der Dichtung - Lobpreisung - Sprache - Muse - Papier und Schrift - Reime und Verse - Singen - Stil	 beatrice, nuova, nuovi, nuovo[14] oda, ode adora, adorar, adori, adoro, ammiro, lode, lodo, lodata, lodato lingue musa carta, carte, libri, nota, penna, penne, pennel, pennello, scritte rime cantan, cantar, cantava, cante stil
Fehler	erra, errai, erranti, erranza, error, fallimento
Feuer	arda, ardea, ardenti, arder, ardisca, ardor, ardore, favilla, faville, fiamma, fiamme, infiamma, infiammi
Gedanken und Erinnerung	pensi, pensier, pensiero, memoria, rimembrar, rimirar, smarrita, smarrito, storia
Gefühle und Gemütszustände - Belastung und Beklemmung - Bedrohung und Gefahr - Eifersucht - Faulheit - Angst - Freundschaft - Freude - Empörung - Kummer und Sorge - Glückseligkeit	 gravosi minaccia, perigli, periglio gelosa pigro orrore, teme, temer, temi, timore amica gaudente, gaudio, giocondo, lieta, lieto, piacimento, piacque sdegna, sdegnando, sdegni, sdegno affanna, affanni, lamenti alegranza, beata, beate, beato, serena, serene, sereno

[14] Formen des Adjektivs „nuovo“ werden in der italienischen Dichtung – insbesondere jener des Mittelalters und der Renaissance – häufig als Anspielung auf die Lyrik des Dolce stil novo gedeutet. Zu den Texten, die die Forschung dieser literarischen Bewegung zuordnet, zählt unter anderem Dantes Prosimetrum *Vita Nova*, auf das auch die Bezeichnung Dolce stil novo zurückgeht. Obwohl Ausdrücke aus der semantischen Sphäre des Neuen bereits in der Dichtung vor Dante von Belang sind, etwa in der Lyrik der Dichterschule Scuola siciliana, erreicht die Verwendung von Wörtern aus jenem semantischen Feld bei Dante eine neue Dimension. Diesen Umstand dokumentiert beispielsweise die Häufigkeit des Auftretens von Wortformen von „nuovo“ in der Dichtung der Scuola siciliana im Vergleich zu der Lyrik Dantes. Während in der Sammlung der Dichtung der Scuola siciliana aus dem Vergleichskorpus in 338.544 Zeichen insgesamt 35 Wörter aus dem semantischen Feld „nuovo“ vorkommen, sind es in Dantes Lyrik bei nur 58.320 Zeichen 20 Ausdrücke. Anders formuliert: Wörter aus der semantischen Sphäre „nuovo“ kommen in Dantes Dichtung mehr als dreimal so häufig vor wie in der Sammlung der Gedichte der Scuola siciliana.

- Hoffnung	speme, sperar
- Kränkung	offenda, offende, offender, offesa, offese, offeso
- Langeweile	noie, noiose
- Leiden	soffersi, sofferto
- Liebe und Zuneigung	affetto, amarla, amò, amori, amorose, inamora,
- Müdigkeit	stanca, stanchi, stanco
- Neid	invia, invidia, invidio, invidiosa, piaga, piagato, piagente
- Schmerz	dolersi, duol, duole, duolo, pesanza, travagli, travagliato
- Ruhe	tranquilla
- Reue	pente, pentir, pento
- Schwäche	debile
- Sehnsucht	brama, brami, desia, desiderio, desio, desir, desire, desiri, disïanza
- Trost	consola
- Traurigkeit und Trauer	cruccioso, lacrime, lagrimando, lagrime, lagrimosa, lutto, malenanza, pianga, piangete, piangi, piangon, sconsolato, triste, tristi, tristo
- Vertrauen	fida, fidata, fidato, fide, fidele, fidi, fido
- Verzweiflung	disperata
- Wahnsinn	furore, vaneggia
- Wut und Hass	irata, irato, odia, odio
Gegensatz	contrari, contraria
Göttlichkeit, Religion, Transzendenz	
- Beten	prega, pregar, pregate, preghi, pregi, pregiate
- Christentum	caritate, confesso, croce, innocente, maria, martìr, peccar, perdona, salvar, salvo, spirti, spirto, vergin, vergine, vizi, vizii, voto
- Engel	ange, angeliche, angelico
- Glauben	credi
- Gott	diva, divin, divina, divino, iddio, idio
- Heiligkeit	benedetta, benedetto, benenanza, santa, santi, santo
- Himmel	ciel, cieli, cielo
- Kirche	chiesa
- Transzendenz und Ewigkeit	eterna, infinita, sempiterna, sempiterno, superna, superno, vaga, vaghezza, vaghi, vago
- Wunder	mirabil, miracol, miracoli
Herrschaft	conte, don, imperi, impero, libero, libertà, libertate, maestà, maestro, nobil, re, regina, regni, servi, servitute, tiranno, trono
Kälte und Eis	fredda, freddo, gela, gelata, gelato, gelide, gelo, giaccio
Körperlichkeit und Menschlichkeit	
- Blut	sangue, vena
- Essen	cibi, cibo, fame

- Familie	dote, figlia, figlio, figliuol, fratello, madre, moglie, padre, padri, patre, sorella, sposa, vedova
- Geburt	nacque, nasca, nascesti, nati
- Kleidung	falda, gonna, piega, vesta, veste, vesti, vestito
- Krankheit	febbre, infermo, medico, rimedio, risana, sani, velen, venen, veneno, venenoso
- Körperteile und Körperlichkeit	braccia, capei, capelli, carne, chioma, corpi, criatura, cuore, denti, dito, fianco, fiato, fronte, grembo, morso, occhio, ochi, orecchia, ossa, pelle, petti, petto, piè, piede, piedi, seno, ventre
- Menschlichkeit	fanciulla, fanciullo, pargoletto rido, riede, uman, umani, umano, uomini, viva, vivete, vivi
- Nacktheit	ignuda, ignudo, nude, nudo
- Trinken	ber, bere
Krieg	
- Angriff und Kampf	assalto, combattuto, contesa, contese, ferir, guerriera
- Befestigung und Verteidigung	difese, fortezza, schiera
- Flucht und Entkommen	scampa, scampar, scampo
- Feind	nimico
- Gewalt	menar, percossa, percosse, percosso, percuote, strugge, struggi
- Geleit	scorta, scorte
- Konflikt	liti, rapina
- Rebellen	rubella, rubello
- Sieger und Verlierer	vinca, vinci, vincitore, vinta, vinti, preda
- Waffen	archi, arco, armati, arme, armi, faretra, spada, strali
Kunst	copia, dipinti, disegno, imagine, opra, opre, pittura, publico
Mythologie	diana, dido, europa, minerva
Natur	
- Berge, Hügel, Gestein	alpe, alpestri, alpestro, colli, monte, monti, scioglie
- Erde	fango, pianeta, terre, terren, terrena, terreno
- Fauna	animal, aquila, artiglio, cani, caval, colomba, corno, farfalla, serpe, serpenti, nido, orso, ratto
- Feld und Wiese	campo, erba, erbe, pascendo, pasco, prati
- Flora	arbor, arbore, arbori, fioretti, fiori, fiorita, fiorito, flora, gigli, pino, rose, scorza, seme
- Früchte	mel, mele, noce, pomo
- Garten	orti
- Gewässer	fiume, fiumi, fonte, fonti, laghi, rio, rivo
- Luft	aere, aria
- Mond und Sterne	luna, stelle
- Tal	valle, valli
- Wald	bosco, selva
- Wasser	acqua, acque
- Wetter und Jahreszeiten	estate, lampo, nebbia, nubi, orizzonte, pioggia, piova, tempeste, tempestoso, verno

Recht	furto, giudice, giudizio, giuro, giusta, giusto, governa, iudicio, iusto, ladri, norma, oltraggio, parlamento
Schicksal und Ruhm	coronato, destino, fama, lauro, provedenza, sorte, sventura
Schmuck und Zierde	adorni, adorno, aureo, diamanti, ornamento, ornate, ornati, ornato, oro, perla, perle, pietra, pietre, preziosa, prezioso, smalto, tesoro, valenza, valimento
Schönheit	begli, bei, bell, belle, bellezza, beltà, beltade, bieltà, bieltate
Seele	alma, animo, aura
Sinnliche Wahrnehmung	
- Dunkelheit	oscura, oscuro, tenebre, tenebroso
- Farben	bianche, bianchi, bianco, bionde, bruna, bruno, colori, nera, nere, nero, rosso, tinto, verde, verdi, vermiglia, vermiglie, vermiglio, vïole
- Hören und Klang	ascolta, ascolti, ascolto, gridando, grido, mute, sentii, sorda, suon, suona, udì, udia, udir, udite
- Geruch	aulente, odor, odore
- Geschmack	amari, acerba, acerbi, acerbo
- Licht und Glanz	lucenti, lume, riluce, raggi, raggio, risplende, splende, splendor
- Schatten	ombra
- Sehen	chiare, chiari, cieca, cieco, discerne, mirando, mirar, mirate, mirava, orbo, vederai, vederla, vedermi, vedervi, vedessi, vedo, vedrai, vedrete, vedrò, veggia, veggiam, vider, visti, visto
- Sinne und Sinnestäuschung	fallanza, sensi, senso
- Stimme	gridando, grido, tacendo, voce, voci
Subjekt	soggetta, soggetto, subietto, suggetto
Süße	dolcezza, dolci, soave, soavemente, soavi, suavi
Tod	dipartita, muor, muore, mortal, mortali, salma, tomba, uccise
Vergessen	oblio
Wärme	caldo
Zusammensein und Gemeinschaft	compagno, inseme, insieme, nsieme, unita

Tab. 2 | Übersicht auffälliger semantischer Felder in der Liste präferierter Wörter.

Die in Tabelle 2 aufgelisteten semantischen Felder verweisen teils auf sprachliche Elemente, die von Forschenden als charakteristisch für die Dichtung des Petrarkismus betrachtet werden, beispielsweise von systemtheoretischen Beschreibungen (cf. Hempfer 1987; Hempfer 1991; Regn 1987; Regn 1993). Die semantischen Felder können somit dabei helfen, Elemente zu ermitteln, die aus systemtheoretischer Sicht für das petrarkistische System als konstitutiv gelten können.

Ein solches Element stellt beispielsweise die der petrarkistischen Lyrik zugeschriebene Gegensätzlichkeit dar, die nicht nur als eigenes Wortfeld vorkommt,

sondern von der auch die Virulenz konträrer semantischer Sphären zeugt.[15] Ein anderes von der Petrarkismusforschung thematisiertes Motiv, die Versinnbildlichung der Liebe als Krieg (cf. auch Hoffmeister 1973, 25), manifestiert sich im Korpus durch eine Vielzahl von Ausdrücken aus dem semantischen Feld des Krieges.[16] Auch weitere bemerkenswerte Aspekte des Petrarkismus lassen sich an den semantischen Feldern ablesen, wie die hochgradig subjektive Dimension der Lyrik oder die von dem Schwanken des Liebenden zwischen Lusterfüllung (Freude, Hoffnung, Liebe und Zuneigung) und Liebesschmerz (Leiden, Schmerz, Traurigkeit und Trauer, Verzweiflung, Wut und Hass) geprägte Liebeskonzeption des dolendi voluptas. Ähnlich verhält es sich mit dem semantischen Feld Süße, das als Anspielung auf die vorpetrarkistische Lyrik des Dolce stil novo gedeutet werden kann.[17] Die Referenz auf den Dolce stil novo wird durch weitere Aspekte untermauert, insbesondere durch rhetorische Strategien zur Idealisierung der Geliebten. Eine solche Stilisierung erfährt das Liebesobjekt in der Dichtung des Dolce stil novo gemäß dem Konzept des Seelenadels („gentilezza"). Dieser bezeichnet eine besondere Befähigung zur Liebe, die keineswegs von der Herkunft oder dem sozialen Status, sondern allein von der Seele abhängt. Durch die Idealisierung im Lichte der gentilezza schreibt der Sprecher der Geliebten übermenschliche, engelsgleiche Eigenschaften zu.

Im Vergleich zu den bisherigen kontrastiven Analysen petrarkistischer Lyrik (cf. Rohden 2022) deuten die semantischen Felder in Tabelle 2 ferner auf weitere für den Petrarkismus relevante Elemente hin. So erinnert das semantische Feld des Fehlers an das Einleitungssonett aus Petrarcas *Canzoniere*, das als programmatisch für die gesamte Sammlung interpretiert werden kann:[18]

> Voi ch'ascoltate in rime sparse il suono
> di quei sospiri ond'io nudriva 'l core
> in sul mio primo giovenile errore
> quand'era in parte altr'uom da quel ch'i' sono,
>
> del vario stile in ch'io piango et ragiono
> fra le vane speranze e 'l van dolore,
> ove sia chi per prova intenda amore,

[15] Ein Beispiel hierfür ist die Dichotomie von Alleinsein und Einsamkeit auf der einen sowie Zusammensein und Gemeinschaft auf der anderen Seite.

[16] Wie einige andere der aus den ermittelten semantischen Feldern hervorgehenden Motive, so tritt auch die Konzeption von Liebe als Krieg bereits in der Literatur der griechischen und römischen Antike auf, etwa bei Homer und Sappho (cf. Rissman 1983), aber auch bei Ovid (cf. zum Beispiel Cahoon 1988). Dass die Relevanz des Konzepts über die Sphäre der Literatur hinausgeht, wird exemplarisch anhand von Forschungsbeiträgen zur konzeptuellen Metaphorik deutlich, für die die Betrachtung von Liebe als Krieg als Paradebeispiel gilt (cf. Lakoff & Johnson 1980, 124). Der Umstand, dass Ausdrücke aus den semantischen Feldern Liebe und Krieg distinktive Elemente der untersuchten Zielpartition darstellen, liefert einen quantitativ fundierten Beleg dafür, dass die Konzeption der Liebe als Krieg für eine beträchtliche Anzahl von petrarkistischen Gedichtsammlungen von Bedeutung ist. Obgleich dieser Befund zunächst keine weiteren Informationen über die konkrete poetische Gestaltung des Motivs in den einzelnen Gedichtsammlungen liefert, illustriert er, dass die Versinnbildlichung von Liebe als Krieg für die italienische Dichtung des Petrarkismus in wesentlich höherem Maße charakteristisch ist als für die italienische Liebeslyrik vor Petrarca. Diese Beobachtung ist ein weiteres Indiz, das zu einem besseren Verständnis des Petrarkismus beitragen kann.

[17] Cf. für eine eingehende Darstellung des Dolce stil novo Pirovano 2014.

[18] Cf. für eine Analyse des Gedichts zum Beispiel Noyer-Weidner 1985.

spero trovar pietà, nonché perdono.

Ma ben veggio or sì come al popol tutto
favola fui gran tempo, onde sovente
di me medesmo meco mi vergogno;

et del mio vaneggiar vergogna è 'l frutto,
e 'l pentersi, e 'l conoscer chiaramente
che quanto piace al mondo è breve sogno.
(Petrarca 1964, 1)

Ein weiteres, zwar gleichsam im Eröffnungssonett des *Canzoniere* und in Tabelle 2 erkennbares, in der Zielpartition jedoch weiter gefasstes Element ist die Bezugnahme auf das semantische Feld der Dichtung. Während Dichtung im Eingangsgedicht des *Canzoniere* mit Blick auf formale („rime"), gattungsstilistische („del vario stile") und inhaltliche („in ch'io piango et ragiono") Gesichtspunkte thematisiert wird, verweist die Zielpartition auf die semantische Sphäre der Kunst allgemein, darunter auch andere mediale Formen der Kunst. Außerdem erweist sich die Referenz auf Dichtung in unterschiedlicher Hinsicht als spezifischer, zum einen durch die Präsenz des für weite Teile der frühitalienischen Dichtung signifikanten Aspekts des Lobes (cf. Seitscheck 2014) und zum anderen wegen Anklängen an die Liebeslyrik Dantes.[19]

Die kontrastive Analyse des Korpus offenbart demzufolge Merkmale, deren verstärktes Vorkommen in semantischen Feldern Rückschlüsse auf beachtenswerte Elemente der petrarkistischen Lyrik erlaubt. Diese distinktiven Merkmale können systemtheoretische Betrachtungen um eine quantitativ fundierte Perspektive bereichern. Über das Zusammenspiel jener Elemente, etwa ihre Kombination zu poetischen Diskursstrukturen, kann eine Betrachtung des Kontexts des jeweiligen sprachlichen Elements Informationen liefern – etwa mit Hilfe von Kookkurrenzanalysen.

5. Kookkurrenzanalyse des Korpus

Am Beispiel des Petrarkismus zeigt sich, wie Kookkurrenzanalysen die durch kontrastive Analysen gewonnenen Unterscheidungsmerkmale kontextualisieren und dadurch spezifizieren können. Auf diese Weise können Kookkurrenzanalysen durch semantische Felder verbalisierte Motive bestätigen wie auch konkretisieren. Dies verdeutlicht ein Blick auf die 18 Kookkurrenzen mit dem höchsten Kookkurrenzwert[20] zu „ghiacci.*"[21] in Tabelle 3, ermittelt mit der Anwendung *TXM* (Heiden, Magué & Pincemin 2010):

[19] Der substantielle Einfluss der Werke Dantes auf Pertrarca wurde von der literaturwissenschaftlichen Forschung wiederholt dokumentiert. Cf. zu der Rezeption Dantes durch Petrarca im Detail etwa Trovato 1979 und Santagata 1990.

[20] Der Kookkurrenzwert bemisst den Grad der Spezifik des gemeinsamen Auftretens von zwei Wörtern. Cf. für die mathematischen Hintergründe Lafon 1980.

[21] Bei der angegebenen Zeichenfolge handelt es sich um eine Suchanfrage nach Wörtern, die mit „fredd" beginnen und eine beliebige Anzahl von Zeichen haben können. Der Punkt („.") steht für ein beliebiges Zeichen mit Ausnahme des Leerzeichens; das Sternchen („*") signalisiert, dass das vorherige Zeichen beliebig

Nummer	*Kookkurrent*	*Frequenz*	*Co-Frequenz*	*Kookkurrenzwert*	*Mittlerer Abstand*
1	foco	1705	69	48	3,1
2	freddo	311	25	25	2,2
3	neve	400	27	25	2,4
4	nevi	59	11	15	2,5
5	sfaccio	21	8	14	6,1
6	il	32222	184	10	3,3
7	ghiaccio	296	13	10	5
8	braccio	155	9	8	5
9	fiamma	737	16	8	2,7
10	l	15106	96	7	4,1
11	pruine	9	4	7	2,2
12	caldo	392	12	7	3,2
13	fiamme	394	12	7	2,6
14	duro	597	14	7	1
15	fuoco	262	10	7	2,2
16	sole	1722	23	7	6,2
17	venti	383	11	6	3,5
18	verno	320	10	6	3,1

Tab. 3 | Die 18 Wörter mit den höchsten Kookkurrenzwerten zu „ghiacci.*“. Ausdrücke aus dem semantischen Feld des Feuers sind rot, jene aus der semantischen Sphäre der Kälte blau markiert.

Unter den 18 Ausdrücken mit dem höchsten Kookkurrenzwert lassen sich zwei Gruppen von Wörtern erkennen, die gegensätzlichen semantischen Feldern entstammen: auf der einen Seite vier Begriffe, deren semantische Bedeutung wie die von „ghiacci.*“ auf Kälte hindeutet; auf der anderen Seite sechs Wörter aus der semantischen Sphäre des Feuers bzw. der Wärme. Das Auftreten der beiden konträren Wortgruppen in den Kookkurrenzen zu „ghiacci.*“ zeugt von jener Gegensätzlichkeit, die auch die in Kapitel 4 herausgearbeiteten semantischen Felder aufzeigen.

oft oder auch gar nicht vorkommen kann. Für eine übersichtliche Einführung in Funktionen und Suchsyntax von *TXM*, cf. Schöch 2020.

Das gemeinsame Vorkommen kontrastiver Aspekte wird ebenfalls deutlich am Beispiel der beiden semantischen Felder Liebe und Schmerz, deren Kopräsenz auf das bereits thematisierte Motiv des dolendi voluptas hindeutet. Illustrativ dafür sind die in Tabelle 4 aufgelisteten Kookkurrenzen zu „dolo.*“:

Nummer	*Kookkurrent*	*Frequenz*	*Co-Frequenz*	*Kookkurrenzwert*	*Mittlerer Abstand*
1	mio	10666	566	104	2,4
2	pianto	1492	123	41	3,6
3	il	32222	829	20	3,6
4	core	2162	110	19	5,4
5	morte	3924	151	15	4,6
6	lagrime	318	34	15	4,1
7	sento	753	52	14	3,5
8	gioia	522	38	11	3,7
9	pietà	1352	66	11	4,4
10	tanto	3914	137	11	3,7
11	l’aspro	75	15	11	0,5
12	pena	969	50	9	3,7
13	lassa	270	24	9	4,5
14	mi	11226	300	9	4,2
15	magior	198	20	8	2,7
16	gran	3671	122	8	2,1
17	pene	491	32	8	4,3
18	pianti	344	26	8	2,6

Tab. 4 | Die 18 Wörter mit den höchsten Kookkurrenzwerten zu „dolo.*“. Wörter aus dem semantischen Feld des Schmerzes sind rot markiert, jene aus der semantischen Sphäre des Glücks blau.

Unter den Kookkurrenzen sind Begriffe vertreten, die auf Schmerz hindeuten, aber gleichsam auch positiv konnotierte Konzepte wie „core“ und „gioia“. Die daraus resultierende Opposition von Schmerz und Freude ist charakteristisch für die Konzeption des dolendi voluptas, wobei der Bezug zur Liebe symbolisch durch das Motiv des Herzens hergestellt wird.

Das gemeinsame Vorkommen von Wortgruppen aus zwei anderen konträren semantischen Feldern, der Göttlichkeit, Religion und Transzendenz einerseits und der Körperlichkeit bzw. Menschlichkeit andererseits, versinnbildlicht die nicht

zuletzt für die Dichtung des Dolce stil novo typische Idealisierung der Geliebten. Ein gutes Beispiel hierfür sind die Kookkurrenzen zu „divin.*“ in Tabelle 5:

Nummer	*Kookkurrent*	*Frequenz*	*Co-Frequenz*	*Kookkurrenzwert*	*Mittlerer Abstand*
1	spirto	817	47	16	2
2	bellezze	235	23	13	2,8
3	alte	250	23	12	1,5
4	virtù	1507	55	10	2,9
5	luce	1501	53	10	3,2
6	beltà	756	35	9	1,1
7	luci	441	26	9	1,8
8	mortal	903	37	8	3,7
9	contemplando	18	7	8	2,7
10	celeste	738	32	8	4,8
11	del	10388	200	7	3,9
12	angelica	96	11	7	2,1
13	umane	81	10	7	3,3
14	nel	5090	110	6	3,7
15	concetti	51	8	6	3,6
16	l’eterno	119	11	6	4
17	verbo	25	6	6	0,5
18	lume	1257	39	6	2,2

Tab. 5 | Die 18 Wörter mit den höchsten Kookkurrenzwerten zu „divin.*“. Ausdrücke aus der semantischen Sphäre der Körperlichkeit bzw. Menschlichkeit sind rot markiert, jene aus der semantischen Sphäre der Göttlichkeit, Religion und Transzendenz blau.

Bemerkenswert von den in Tabelle 5 dargestellten Ausdrücken ist das Wort „angelica“, welches das Motiv der „donna angelicata“ evoziert. Jenes für die Dichtung des Dolce stil novo, jedoch auch für Dantes Werk *Vita Nova* bedeutende Konzept (cf. Seitscheck 2014) kann gemäß der Petrarkismus-Konzeption Warnings als Ausdruck eines vorpetrarkistischen lyrischen Diskurses aufgefasst werden, der sich auf zwei Arten manifestiert: zum einen als semantisches Feld durch das Auftreten von Wörtern aus dem Bereich der Süße (cf. Kapitel 4), zum anderen in Form der Referenz auf das stilnovistische wie auch danteske Motiv der „donna angelicata“ in den Kookkurrenzen.

Die stilisierte Geliebte ist dabei zugleich Ausgangspunkt und Ziel des persönlichen Begehrens, das den Sprecher dazu veranlasst, die eigenen Gefühle zu verbalisieren. Die durch seine Liebe ausgelösten Gemütszustände sind in der Lyrik des Petrarkismus oftmals ebenso zwiespältig wie wechselhaft und schlagen sich im Text sowohl formal als auch inhaltlich nieder. Dies verdeutlichen neben den in der vorliegenden Untersuchung bereits herausgestellten Motiven der Gegensätzlichkeit die beiden dichotomischen semantischen Felder Alleinsein und Einsamkeit sowie Zusammensein und Gemeinschaft (cf. Kapitel 4).

Die subjektive Dimension des Petrarkismus zeigt sich allerdings auch anhand eines Elements, das in der literaturwissenschaftlichen Forschung bislang weniger stark fokussiert wurde: die sinnliche Wahrnehmung, deren Virulenz im untersuchten Korpus am Beispiel unterschiedlicher Aspekte zum Vorschein kommt. Im Besonderen gilt das für die visuelle Wahrnehmung, von der eines der wesentlichen Kernthemen der petrarkistischen Dichtung geprägt ist: die Liebe. Sichtbar machen dies die Kookkurrenzen zu „amor.*“ in Tabelle 6:

Nummer	*Kookkurrent*	*Frequenz*	*Co-Frequenz*	*Kookkurrenzwert*	*Mittlerer Abstand*
1	mi	11226	1334	75	3,8
2	core	2162	317	33	4,9
3	strali	236	72	26	3,4
4	d	640	125	24	2,6
5	cor	5018	544	22	4,9
6	pargoletti	34	25	21	0
7	strale	210	57	18	3,4
8	l’arco	243	60	17	3,4
9	m’ha	447	85	16	2,2
10	et	8790	826	15	4,4
11	sforza	229	55	15	3,8
12	m	433	81	14	3,6
13	me	8531	790	13	4,2
14	begli	739	111	13	4,7
15	occhi	3684	380	12	4,6
16	ove	2262	254	12	3,7
17	mio	10666	946	12	4,6
18	dolce	3926	397	12	4,5

Tab. 6 | Die 18 Wörter mit den höchsten Kookkurrenzwerten zu „amor.*“. Ausdrücke, die auf die Augen verweisen, sind rot markiert, Wörter aus dem semantischen Feld der Liebe blau und Begriffe aus der semantischen Sphäre Süße grün.

Abgesehen von dem Herz, das die Liebe symbolisiert, sind unter den Kookkurrenzen mit dem höchsten Kookkurrenzwert die Augen als das für die visuelle Wahrnehmung entscheidende Sinnesorgan zu finden. In Form des Adjektivs „dolce“ wird ersichtlich, dass der poetische Diskurs des Dolce stil novo auch die Wahrnehmung der Geliebten beeinflusst.[22] Ein weiteres Indiz dafür sind die Kookkurrenzen zu „vist.*“, abgebildet in Tabelle 7:

Nummer	*Kookkurrent*	*Frequenz*	*Co-Frequenz*	*Kookkurrenzwert*	*Mittlerer Abstand*
1	in	25194	503	20	3
2	dolce	3926	115	14	2,5
3	sua	3182	94	12	2,1
4	humana	82	14	11	1,5
5	l'ho	102	13	9	2
6	ho	1526	51	8	0,7
7	contrista	18	7	8	4
8	angelica	96	12	8	0,2
9	trista	228	17	8	4,5
10	altera	320	19	7	0,7
11	occhi	3684	86	6	5,1
12	umana	141	12	6	1,2
13	altrove	271	16	6	2,9
14	mirar	309	17	6	4,8
15	turbata	28	6	5	0,8
16	bella	2148	54	5	3,4
17	amata	99	9	5	0
18	lieta	620	23	5	3,3

Tab. 7 | Die 18 Wörter mit den höchsten Kookkurrenzwerten zu „vist.*“. Ausdrücke, die auf die Augen verweisen, sind rot markiert, Wörter aus dem semantischen Feld der Liebe blau und Begriffe aus der semantischen Sphäre Süße grün.

Die durch die Sicht repräsentierte visuelle Wahrnehmung steht insofern im Zeichen des Dolce stil novo, als mit „dolce“ ein Begriff aus dem Wortfeld der Süße ebenso unter den Kookkurrenzen vorkommt, wie das Wort „angelica“, das auf die

[22] Cf. zu weiteren Aspekten der Wahrnehmung der Geliebten durch den Sprecher in der Zielpartition auch Rohden 2022.

Idealisierung der schönen („bella“) Geliebten („amata“) verweist. Im Rahmen der visuellen Wahrnehmung vollzieht sich eine rhetorische Aneignung des stilnovistischen Diskurses durch die Art des lyrischen Sprechens des Petrarkismus, wie der Gegensatz zwischen menschlicher („humana“) und engelhafter („angelica“) Dimension der Geliebten ebenso skizziert, wie die Antinomie von Traurigkeit („contrista“, „trista“) und Fröhlichkeit („lieta“), die beim Sprecher Verwirrung hervorruft („turbata“). Dieses Zusammenspiel zwischen der petrarkistischen Art des Dichtens und dem poetischen Diskurs des Dolce stil novo zeigt sich ferner anhand der in Tabelle 8 aufgelisteten Kookkurrenzen zu „gentil.*“, einem bedeutenden Aspekt der Liebeskonzeption des Dolce stil novo:

Nummer	*Kookkurrent*	*Frequenz*	*Co-Frequenz*	*Kookkurrenzwert*	*Mittlerer Abstand*
1	donna	2133	151	45	1,6
2	alma	851	94	43	0,8
3	varchi	257	54	40	1,4
4	spirto	817	67	24	0,6
5	vile	390	47	24	5,7
6	pianta	439	44	19	1,1
7	bella	2148	99	16	3,7
8	anima	223	28	15	0,5
9	cortese	472	38	13	3,3
10	vaga	491	38	13	2,2
11	stile	464	36	12	5,8
12	cor	5018	162	12	2,1
13	cui	3653	121	10	3,5
14	sì	9034	234	8	3,9
15	animo	91	13	8	0,2
16	vago	837	40	7	3
17	umile	163	16	7	4,6
18	coppia	70	11	7	0,8

Tab. 8 | Die 18 Wörter mit den höchsten Kookkurrenzwerten zu „gentil.*“. Ausdrücke aus dem semantischen Feld der Körperlichkeit sind rot markiert, Wörter aus der semantischen Sphäre der Göttlichkeit, Religion, Transzendenz blau.

Die als schön („bella“) wahrgenommene Dame („donna“) wird von dem Sprecher zu einem Ideal stilisiert, das dank seiner übermenschlichen geistigen Merkmale

eine spirituelle Dimension erlangt („alma“, „spirto“, „anima“, „animo“). Die damit verbundene Lobpreisung der Dame prägt den poetischen Stil der Lyrik („stile“), der zwar in der Tradition der höfischen Dichtung („cortese“) steht, aber nicht mehr darauf beschränkt ist, da die entscheidenden Merkmale der Geliebten in der Wahrnehmung des Sprechers nicht aus deren Umfeld oder Herkunft resultieren, sondern aus deren geistiger Befähigung zur Liebe. Jene Auswirkungen der Wahrnehmung auf das poetische Schreibprogramm illustrieren die Kookkurrenzen zu einem der charakteristischen Elemente von Lyrik in Tabelle 9, dem Vers:

Nummer	*Kookkurrent*	*Frequenz*	*Co-Frequenz*	*Kookkurrenzwert*	*Mittlerer Abstand*
1	rime	436	50	33	2,2
2	miei	2057	80	20	1,7
3	i	11351	237	19	2,9
4	lagrime	318	29	17	2,5
5	prosa	27	12	16	2,3
6	pianto	1492	57	14	3,3
7	lacrime	233	21	12	2,2
8	sangue	780	37	12	3,3
9	versi	468	27	11	5,2
10	prose	37	10	11	1,8
11	tersi	19	8	10	4,8
12	mei	572	28	9	1,3
13	udite	61	10	8	4,1
14	fiume	458	22	7	4
15	occhi	3684	76	6	3,7
16	fiori	551	22	6	3,9
17	vena	141	11	6	3,3
18	dolci	1121	33	6	2,1

Tab. 9 | Die 18 Wörter mit den höchsten Kookkurrenzwerten zu „vers.*“

Die 18 Wörter mit den höchsten Kookkurrenzwerten lassen viele der bisher eruierten Elemente petrarkistischer Dichtung erkennen:

- Subjektivität („miei“, „mei“);
- Emotionalität („lagrime“, „pianto“);
- Körperlichkeit („sangue“, „vena“);

- visuelle Wahrnehmung („occhi“);
- das Aufgreifen des poetischen Diskurses des Dolce stil novo („dolci“).

Beachtenswert ist ferner die Referenz auf die Gattung der Prosa („prosa“, „prose“), die im Gegensatz steht zu der zumeist in Versen („versi“) verfassten Lyrik, wodurch das Stilmittel der Dualität veranschaulicht wird.

Unter den besagten Kookkurrenzen ist darüber hinaus ein weiteres semantisches Feld bemerkenswert. Es handelt sich dabei um die Natur, die in zwei Ausdrücken („fiume“, „fiori“) zur Sprache kommt. Die Bezugnahmen auf die Natur manifestieren sich zudem am Beispiel von anderen Konzepten, wie etwa „camp.*“, dessen Kookkurrenzen in Tabelle 10 aufgeführt sind:

Nummer	*Kookkurrent*	*Frequenz*	*Co-Frequenz*	*Kookkurrenzwert*	*Mittlerer Abstand*
1	i	11351	271	44	3,1
2	colli	323	42	34	3
3	boschi	282	40	34	2,6
4	selve	259	36	30	2,6
5	elisi	15	11	19	0
6	valli	166	21	17	3,4
7	le	13962	228	16	3,3
8	monti	419	29	16	3
9	fiumi	315	25	15	3,2
10	e	86639	992	15	4,2
11	rive	252	21	13	2,7
12	aprici	22	9	12	1,7
13	verdi	284	21	12	3
14	poggi	98	14	12	3,7
15	piagge	142	15	11	1,3
16	prati	102	13	11	2,5
17	alle	369	20	9	2,2
18	fioriti	31	8	9	1,1

Tab. 10 | Die 18 Wörter mit den höchsten Kookkurrenzwerten zu „camp.*“. Ausdrücke aus dem semantischen Feld der Natur sind grün markiert.

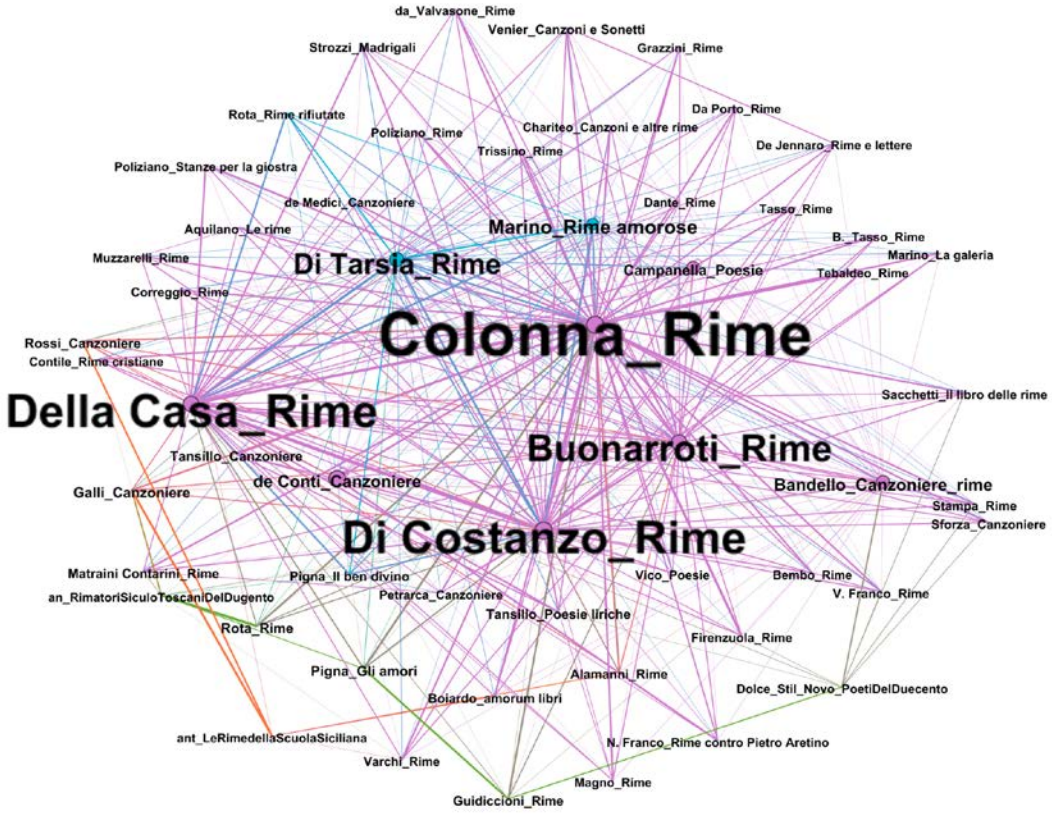
Beachtenswert ist die hohe Dichte an Wörtern aus der semantischen Sphäre der Natur, die als Kookkurrenzen zu „camp.*“ auftreten. Sie lässt erkennen, dass Anspielungen auf die Natur in der untersuchten Zielpartition nicht als singuläres Phänomen auftreten, sondern vielmehr in Form von Landschaftspanoramen, die sich aus der Kombination unterschiedlicher Begriffe aus dem semantischen Feld der Natur speisen.[23] Die poetische Darstellung und Verdichtung von Natur kann als Anknüpfen an eine Tradition verstanden werden, die Ernst Robert Curtius ausgehend von der Antike bis ins europäische Mittelalter nachgezeichnet hat (cf. Curtius 1993, 191–209). Bei Petrarca gewinnt die von dem Sprecher perzipierte Natur dann als Spiegel der eigenen Befindlichkeit im Lichte der Liebe zu Laura besondere Bedeutung, wie beispielsweise Hugo Friedrich (cf. Friedrich 1964, 210–214) oder Joachim Küpper (cf. Küpper 2002) argumentieren. Die Kookkurrenzanalysen deuten darauf hin, dass die Landschaft auch über Petrarcas *Canzoniere* hinaus zu einem relevanten Element petrarkistischer Dichtung wird. Diese Beobachtung deckt sich mit dem von literaturwissenschaftlichen Studien vermerkten Konnex zwischen dem Aufkommen der Landschaft als Denkfigur und der Herausbildung des modernen Subjekts (cf. etwa Stierle 1979; Ulmer 2010). Inwiefern die Adaption des Landschaftsmotivs im Petrarkismus jedoch als Aneignung einer poetischen Sprechweise im Sinne Warnings gedeutet werden kann, ist an dieser Stelle nicht abschließend zu klären. Dafür müsste sich die Darstellung der Landschaft auf einen einschlägigen poetischen Diskurs zurückführen lassen, was angesichts der langen Tradition des Motivs nicht ohne weiteres möglich ist.

6. Netzwerkanalyse des Korpus

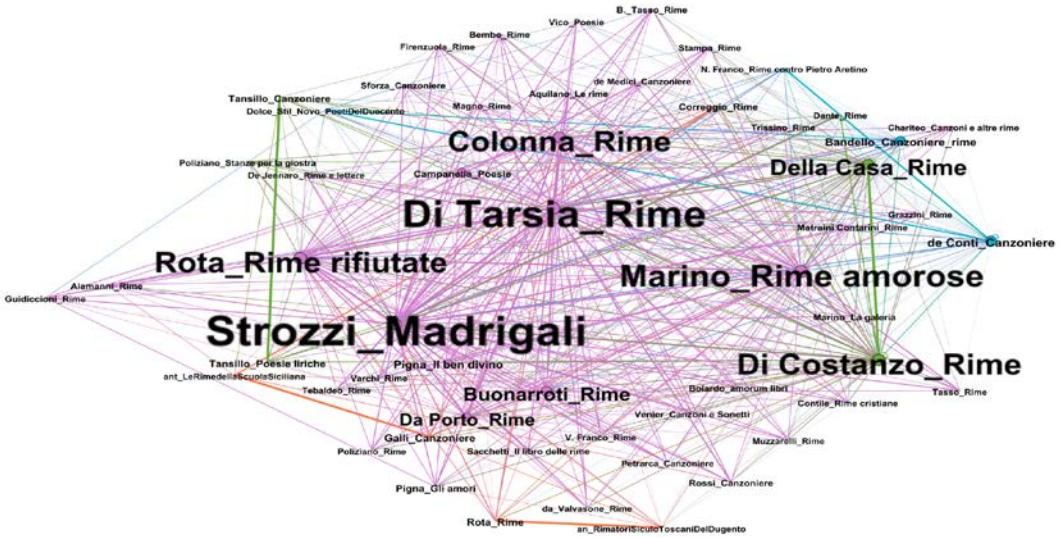
Kontrastive und Kookkurrenzanalysen können distinktive sprachliche Merkmale, beachtenswerte Motive sowie markante lyrische Sprechweisen aufzeigen und damit einen guten Überblick bieten. Diese Übersicht bleibt allerdings auf die Perspektive des Gesamtkorpus begrenzt und gibt nur wenig Aufschluss über Auffälligkeiten innerhalb des Korpus, wie beispielsweise Ähnlichkeiten oder Unterschiede zwischen den einzelnen Gedichtsammlungen. Durch Netzwerkanalysen lassen sich die Beziehungen von Elementen innerhalb einer Datensammlung visualisieren, wodurch Affinitäten und Distanzen sichtbar werden. Diesen Effekt kann man sich für die Untersuchung des Petrarkismus zunutze machen, indem man mit der stylo-Funktion des gleichnamigen r-Pakets stilometrische Analysen lanciert, die ausgehend von der Häufigkeit von sprachlichen Elementen die Distanz zwischen einzelnen Texten im untersuchten Korpus mathematisch berechnen.[24] Legt man dieser stilometrischen Analyse die durch die kontrastiven Analysen gewonnene Liste der präferierten Wörter zugrunde, lässt sich die Entfernung bzw. Nähe der einzelnen Sammlungen zueinander mit Blick auf die herausgearbeiteten distinktiven Elemente ermitteln. Das Resultat ist je eine csv-Datei mit Daten für Knoten

[23] Dies belegen auch die Kookkurrenzen zu anderen Wörtern, etwa „bosc.*“, „coll.*“ oder „fium.*“ (cf. Rohden 2021c).
[24] Cf. für eine anschauliche Einführung in *stylo* Horstmann 2019.

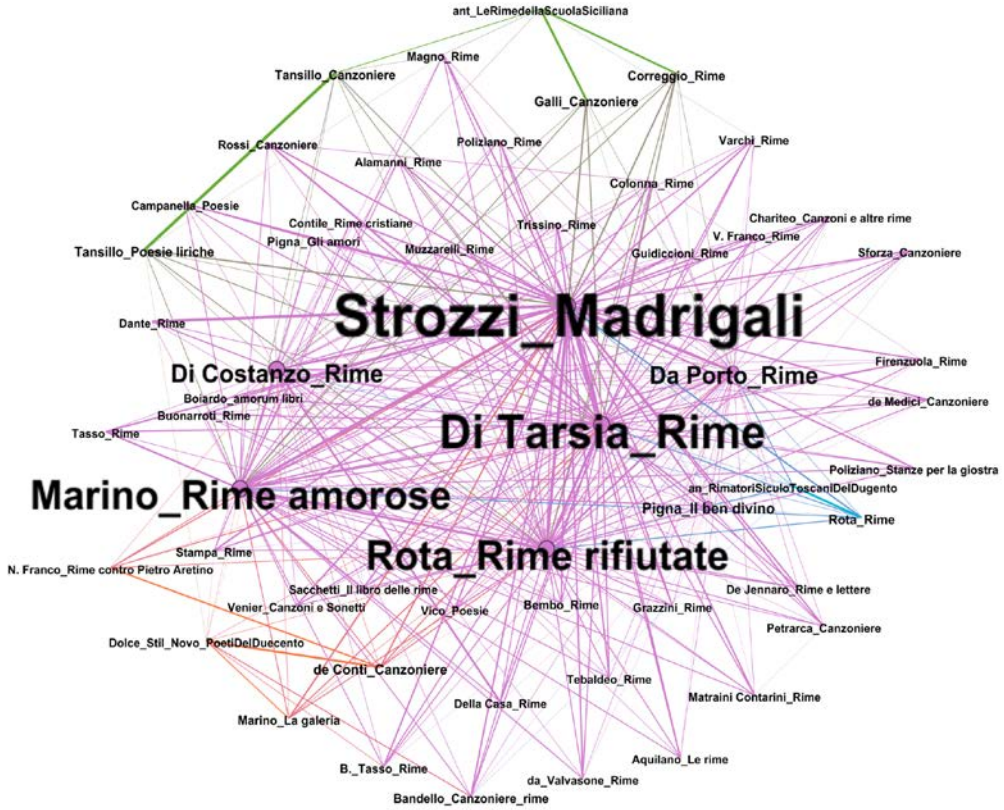
und Kanten eines Graphen. Importiert man beide Dateien in das Netzwerkanalysetool *Gephi* (cf. Bastian, Heymann & Jacomy 2009), lässt sich das untersuchte Korpus in Form eines Netzwerks veranschaulichen. Insbesondere der Vergleich mehrerer Visualisierungen mit unterschiedlichen Parametern (Abbildungen 1 bis 5) offenbart auffällige Beziehungen zwischen den einzelnen Gedichtsammlungen:



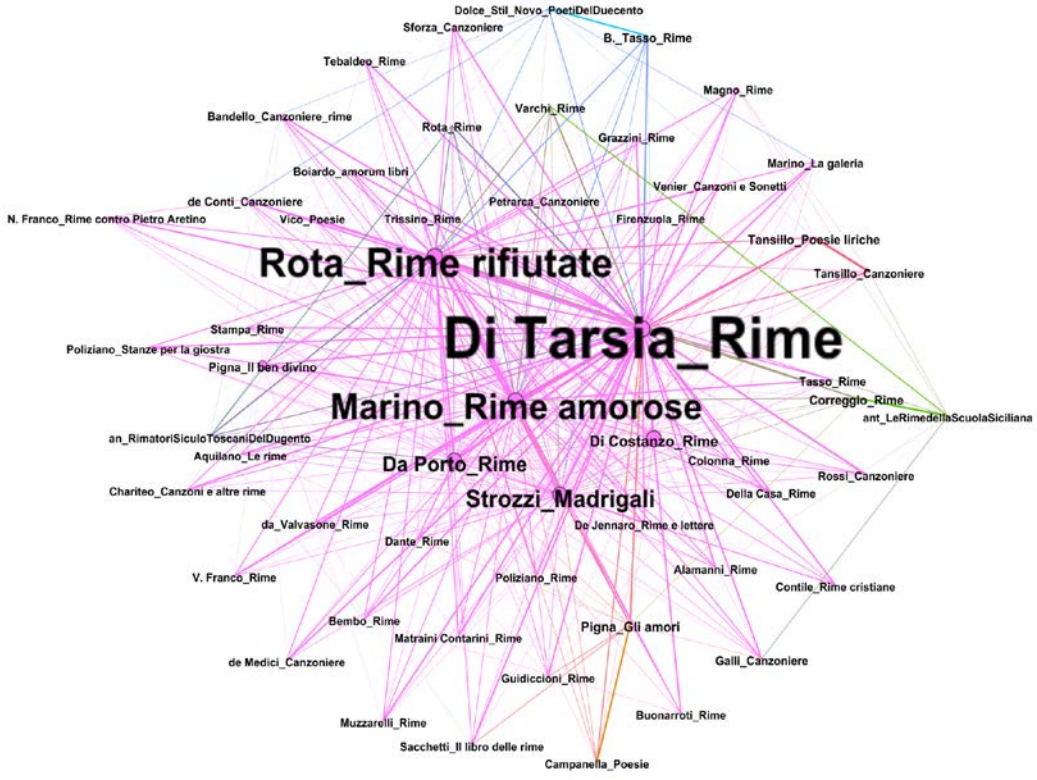
1 | Netzwerkvisualisierung des Korpus auf Basis der Häufigkeiten der 100-250 bevorzugten Ausdrücke, Layoutalgorithmus Fruchterman Reingold.



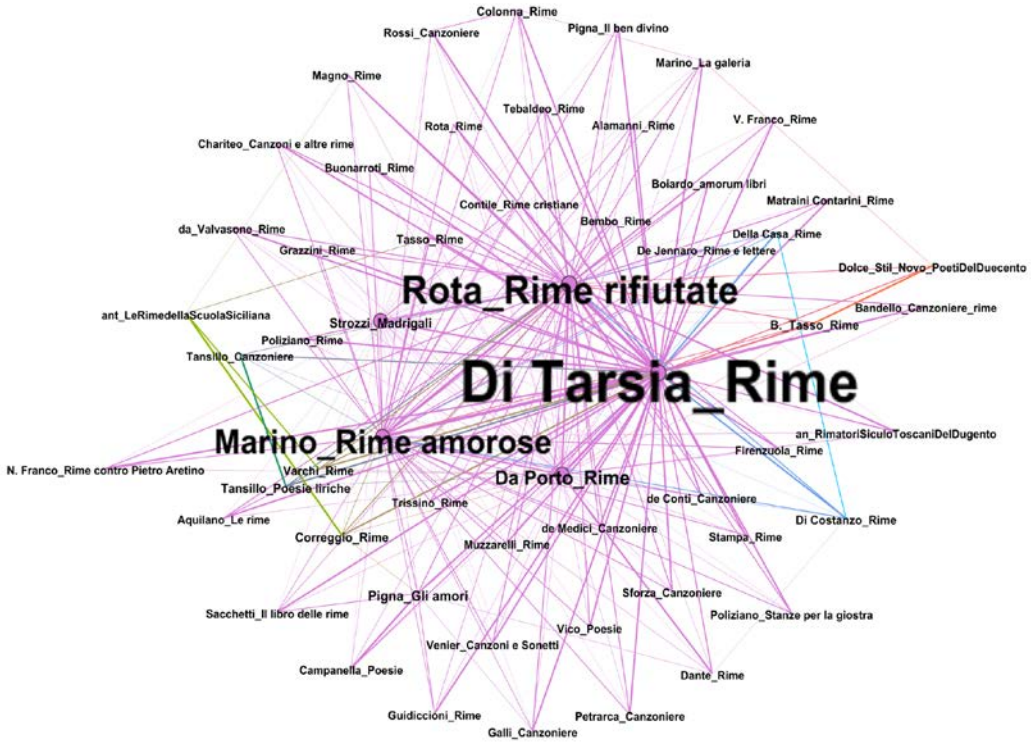
2 | Netzwerkvisualisierung des Korpus auf Basis der Häufigkeiten der 100-500 bevorzugten Ausdrücke, Layoutalgorithmus Fruchterman Reingold.



3 | Netzwerkvisualisierung des Korpus auf Basis der Häufigkeiten der 250-500 bevorzugten Ausdrücke, Layoutalgorithmus Fruchterman Reingold.



4 | Netzwerkvisualisierung des Korpus auf Basis der Häufigkeiten der 250-1000 bevorzugten Ausdrücke, Layoutalgorithmus Fruchterman Reingold.



5 | Netzwerkvisualisierung des Korpus auf Basis der Häufigkeiten der 500-1000 bevorzugten Ausdrücke, Layoutalgorithmus Fruchterman Reingold.

Die Abbildungen 1 bis 5 zeigen Netzwerkvisualisierungen des untersuchten Textkorpus. Die bunten Kanten kennzeichnen jeweils die mit der Modularitätsfunktion von *Gephi* ermittelten Gruppierungen innerhalb des Korpus.[25] Die Größe der Knoten orientiert sich an der Höhe des Werts der Betweenness-Zentralität,[26] die Schriftgröße der Beschriftungen an dem mittleren gewichteten Grad.[27] Die fünf Visualisierungen ziehen jeweils eine unterschiedliche Anzahl an distinktiven Wörtern zur Darstellung heran, beginnend mit 100 bis zu einem Maximum von 1000. Als Metrik zur mathematischen Beschreibung der Distanzen zwischen den einzelnen Texten in *stylo* wurde das Distanzmaß „Delta" (cf. Burrows 2002) genutzt, wobei das Clustering nach dem statistischen Verfahren des Consensus Tree[28] erfolgte, das die Ergebnisse mehrerer Clusterungsdurchläufe in einer gemeinsamen Darstellung zusammenführt. Im vorliegenden Fall wurde die Anzahl der für die

[25] Cf. für illustrative Einführungen in *Gephi* Grandjean 2015 und Schumacher 2020.

[26] Eine nachvollziehbare Definition von Betweenness liefert Mutschke: „Betweenness misst das Ausmaß, in dem ein Knoten auf kürzesten Pfaden zwischen anderen Knoten im Graphen positioniert ist." (Mutschke 2010, 370).

[27] Der „average weighted degree" kann definiert werden als „Average sum of weights of the edges of nodes" (Ayyappan, Nalini & Kumaravel 2017, 233), wobei „weight" eine Zahl bezeichnet, die einer Kante zwischen zwei Knoten zugewiesen wird.

[28] Horstmann liefert eine verständliche Erklärung zur Definition des Consensus Tree: „Der Consensus Tree (oder auch ‚Bootstrap Consensus Tree') ist eine runde Visualisierungsform, die mehrere Clusteranalysedurchgänge mit unterschiedlich vielen MFW bzw. Culling-Parametern in einer Ergebnisvisualisierung vereinigt. Dieses Verfahren wird daher auch ‚Bootstrap' genannt (was wörtlich mit ‚Stiefelriemen' übersetzt werden kann – wie die Schnürsenkel eines Schuhs werden die einzelnen Analysedurchläufe ineinander gewebt und am Ende festgezogen)." (Horstmann 2019, § 5).

Clusterung eingesetzten distinktiven Wörter jeweils in Zehnerschritten gesteigert, wobei die eingestellte „Consensus Strength" von 1 sicherstellte, dass nur Clusterverbindungen in den finalen Consensus Tree übernommen wurden, die ausnahmslos in allen Clustervorgängen auftraten.

Die auf diese Weise generierten Visualisierungen weisen sowohl Gemeinsamkeiten als auch Unterschiede auf. Interessant sind zunächst die im Zentrum des Netzwerks dargestellten Gedichtsammlungen, die Abweichungen zwischen einzelnen Visualisierungen zum Vorschein bringen, da mit steigender Anzahl der berücksichtigten distinktiven Wörter ein Wandel zu erkennen ist. So stechen in Abbildung 1 Colonnas *Rime*, Di Costanzos *Rime* und Della Casas *Rime* bezogen auf die Betweenness--Zentralität und insbesondere den mittleren gewichteten Grad heraus, dicht gefolgt von Buonarrotis[29] *Rime* und Di Tarsias *Rime*. Die herausragende Stellung der drei zuerst genannten Werke sowie der *Rime* Buonarrotis im Netzwerk nimmt jedoch mit jeder Visualisierung mehr und mehr ab. Stattdessen treten besonders Di Tarsias *Rime*, aber auch Marinos *Rime amorose* und Rotas *Rime rifiutate* immer stärker in den Vordergrund, wie deren zentrumsnahe Lage wie auch die jeweils hohen Werte für die Betweenness-Zentralität und den mittleren gewichteten Grad belegen.

Ähnlich ist in allen Abbildungen hingegen die Position von Petrarcas *Canzoniere*, die eher am Rand des Netzwerks liegt. Deutet man diese Lage im Lichte der topographischen Konzeption des Petrarkismus nach Regn so fällt auf, dass das für die Entstehung des Petrarkismus maßgebliche Werk hinsichtlich der ermittelten distinktiven Elemente nicht im definitorischen Zentrum, sondern vielmehr in der Peripherie zu verorten ist. Dieser Umstand ist insofern überraschend, als bereits die Bezeichnung Petrarkismus die Rolle Petrarcas als poetisches Leitbild hervorhebt. Dass nichtsdestotrotz eine hohe Anzahl an Texten innerhalb des Korpus existiert, die bezogen auf die eruierten Unterscheidungsmerkmale Parallelen zu Petrarcas *Canzoniere* aufweisen, verdeutlichen die Farben der Kanten. Diese machen in jeder der Abbildungen insgesamt 4 bis 6 Gruppen erkenntlich, bestehend aus je einer über 70% der Gedichtsammlungen umfassenden, violetten Hauptgruppe und mehreren deutlich kleineren Nebengruppen (maximal 11% der Texte). Beachtenswert ist die Zusammensetzung jener Gruppen: in fast allen Visualisierungen[30] Teil der Nebengruppen und des Weiteren in Randlage befindlich sind drei der vier vorpetrarkistischen Sammlungen, was deren kontrastiv analytisch festgestellte Divergenz zu den Werken der petrarkistischen Zielpartition visuell veranschaulicht. Eine Ausnahme bilden allerdings Dantes *Rime*, die in allen Abbildungen außer Visualisierung 2 der Hauptgruppe angehören. In Anbetracht der oben thematisierten Aspekte, die auf Dantes Lyrik (cf. Kapitel 4), aber auch auf die für seine frühe Dichtung relevante Bewegung des Dolce stil novo (cf. Pirovano 2014, 28–104) hindeuten (cf. Kapitel 4 und 5), werfen die vorliegenden Visualisierungen die Frage nach der Relation zwischen der Lyrik Dantes und jener des Petrarkismus auf.[31]

[29] Buonarroti ist vor allem als Maler besser unter seinem Vornamen Michelangelo bekannt.
[30] Abbildung 5 ist hier die Ausnahme.
[31] Zu dieser Frage existieren Forschungsbeiträge mit unterschiedlichen Stoßrichtungen (cf. zum Beispiel Marx 1998 und Pasquini 2017).

Die aus dem Vergleich der Abbildungen resultierenden Beobachtungen sind zusammenfassend zwar ein Indiz dafür, dass die Anzahl der der jeweiligen Netzwerkvisualisierung zugrunde gelegten Wörter einen deutlichen Einfluss auf die hervorstechenden Texte hat, zumindest im Hinblick auf die Eigenschaften Betweenness-Zentralität und mittlerer gewichteter Grad. Nichtsdestoweniger liefern die in den Netzwerken konstant auftretenden Elemente interessante Erkenntnisse, unter anderem den Umstand betreffend, dass nicht Petrarcas *Canzoniere* eine zentrale Position im Netzwerk einnimmt, sondern diese stattdessen am ehesten di Tarsias *Rime* zukommt. Mit Blick auf eine topographische Betrachtung des Petrarkismus im Sinne Regns stellt sich demnach die Frage, was das definitorische Zentrum der petrarkistischen Lyrik im Einzelnen konstituiert – wodurch gleichsam der Mehrwert von Netzwerkanalysen für literaturwissenschaftliche Fragestellungen illustriert wird.

7. Fazit

Ziel der vorliegenden Untersuchung war es, anknüpfend an existierende Definitionen durch digitale quantitative Analysen eines Korpus italienischsprachiger Gedichtsammlungen einen Beitrag zur Erforschung des Petrarkismus zu leisten. Zu diesem Zweck wurden mittels stilometrischer, Kookkurrenz- und Netzwerkanalysen signifikante sprachliche Elemente, auffällige poetische Diskurse und die Relationen zwischen den einzelnen Textsammlungen herausgearbeitet. Dadurch konnten nicht nur bereits von der literaturwissenschaftlichen Forschung beschriebene Merkmale des Petrarkismus bestätigt, sondern auch bislang weniger fokussierte Aspekte zum Vorschein gebracht werden. Hierzu gehören unter anderem ein erkennbarer Einfluss der Lyrik des Dolce stil novo sowie Dantes, die Virulenz der sinnlichen, insbesondere visuellen Wahrnehmung und die Kombination von Elementen der Natur zu poetischen Landschaftspanoramen. Darüber hinaus werfen einige der digital quantitativ ermittelten Resultate Fragen sowohl zur Abgrenzung des Petrarkismus als auch über die Beziehung petrarkistischer Werke zueinander auf. Dazu zählt etwa die Frage nach dem Verhältnis der Dichtung Dantes zum Petrarkismus oder die Frage danach, was das definitorische Zentrum des letzteren nach Regn konkret ausmacht.

Die zuletzt genannten Fragen verdeutlichen, dass eine Studie wie die vorliegende lediglich ein erster Schritt zur digitalen quantitativen Erforschung des Petrarkismus sein kann. Für eine umfassendere Untersuchung wären weitere Analysen aufschlussreich, etwa von petrarkistischen Korpora in anderen Sprachen oder im Hinblick auf intertextuelle Bezüge in petrarkistischen Texten. Erkenntnisse in diesen Bereichen wären nicht zuletzt deshalb wertvoll, weil bei allen Unterschieden zwischen den verschiedenen Definitionen des Petrarkismus weitgehender Konsens über die weitreichenden Auswirkungen von Petrarcas Lyrik auf die europäische Dichtung besteht. Zum besseren Verständnis jener Auswirkungen können digitale Methoden einen Beitrag leisten.

Bibliographie

AYYAPPAN, G., C. Nalini & A. Kumaravel. 2017. "A study on SNA: Measure average degree and average weighted degree of knowledge diffusion in GEPHI." *Indian Journal of Computer Science and Engineering (IJCSE)* 7 (6), 230–237.
<https://www.ijcse.com/docs/INDJCSE16-07-06-100.pdf>.

BALDACCI, Luigi. 1957. *Il Petrarchismo Italiano Nel Cinquecento.* Milano: Ricciardi.

BASTIAN, Mathieu, Sebastien Heymann & Mathieu Jacomy. 2009. "Gephi: An Open Source Software for Exploring and Manipulating Networks." In *Proceedings of the International AAAI Conference on Web and Social Media 3 (1)*, ed. AA.VV.,361–362.
<https://ojs.aaai.org/index.php/ICWSM/article/view/13937>.

BERNSEN, Michael. 2011. „Der Petrarkismus, eine lingua franca der europäischen Zivilisation." In *Der Petrarkismus – ein europäischer Gründungsmythos*, ed. Bernsen, Michael & Bernhard Huss, 15–30. Gründungsmythen Europas in Literatur, Musik und Kunst 4. Göttingen: V&R unipress.

BUBENHOFER, Noah. 2006-2022. *Einführung in die Korpuslinguistik: Praktische Grundlagen und Werkzeuge. Elektronische Ressource*.
<https://www.bubenhofer.com/korpuslinguistik/kurs/index.php?id=cosmas_client_kookk.html>.

BURROWS, John. 2002. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17, 267–287.
<https://doi.org/10.1093/llc/17.3.267>.

CAHOON, Leslie. 1988. "The Bed as Battlefield: Erotic Conquest and Military Metaphor in Ovid's Amores." *Transactions of the American Philological Association* 118, 293–307.
<https://doi.org/10.2307/284173>.

CURTIUS, Ernst Robert. 1993. *Europäische Literatur und lateinisches Mittelalter*. Elfte Auflage. Tübingen, Basel: Francke.

EDER, Maciej. 2017. "Visualization in Stylometry: Cluster Analysis Using Networks." *Digital Scholarship in the Humanities* 32 (1), 50–64.
<https://doi.org/10.1093/llc/fqv061>.

EDER, Maciej, Jan Rybicki & Mike Kestemont. 2016. "Stylometry with R: a package for computational text analysis." *R Journal* (8), 107–121.
<https://journal.r-project.org/archive/2016/RJ-2016-007/index.html>.

FORSTER, Leonard. 1969. *The Icy Fire: Five Studies in European Petrarchism.* Cambridge: Cambridge University Press.

FRIEDRICH, Hugo. 1964. *Epochen der italienischen Lyrik.* Frankfurt a. M.: Klostermann.

GRANDJEAN, Martin. 2015. *GEPHI – Introduction to network analysis and visualization*.
<http://www.martingrandjean.ch/gephi-introduction/>.

HEIDEN, Serge, Jean-Philippe Magué & Bénédicte Pincemin. 2010. «TXM: Une plateforme logicielle open-source pour la textométrie-conception et développement.» In *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010. Vol. 2*, ed. AA.VV., Milano: LED, 1021–1032.
<https://halshs.archives-ouvertes.fr/halshs-00549779/fr/>.

HEMPFER, Klaus W. 1987. „Probleme der Bestimmung des Petrarkismus. Überlegungen zum Forschungsstand." In *Die Pluralität der Welten: Aspekte der Renaissance in der Romania*, ed. Stempel, Wolf-Dieter & Karlheinz Stierle, 253–277. Romanistisches Kolloquium 4. München: Fink.

HEMPFER, Klaus W. 1991. „Intertextualität, Systemreferenz und Strukturwandel: Die Pluralisierung des erotischen Diskurses in der italienischen und französischen Renaissance-Lyrik (Ariost, Bembo, Du Bellay, Ronsard).“ In *Modelle des literarischen Strukturwandels*, ed. Titzmann, Michael, 7–44. Studien und Texte zur Sozialgeschichte der Literatur 33. Tübingen: Niemeyer.

HEMPFER, Klaus W., Gerhard Regn & Sunita Scheffel (ed.). 2005. *Petrarkismus-Bibliographie: 1972–2000.* Text und Kontext 22. Stuttgart: Steiner.

HOFFMEISTER, Gerhart. 1973. *Petrarkistische Lyrik.* Sammlung Metzler 119. Stuttgart: Metzler.

HORSTMANN, Jan. 2019. „Stilometrie mit Stylo.“ *forTEXT. Literatur digital erforschen*.
<https://fortext.net/routinen/lerneinheiten/stilometrie-mit-stylo>.

JANNIDIS, Fotis. 2017. „Netzwerke.“ In *Digital Humanities*, ed. Jannidis, Fotis, Hubertus Kohle & Malte Rehbein, 147–161. Stuttgart: Metzler.
<https://doi.org/10.1007/978-3-476-05446-3_10>.

KÜPPER, Joachim. 2002. „Mundus imago Laurae. Das Sonett ‚Per mezz'i boschi‘ und die ‚Modernität‘ des Canzoniere.“ In *Petrarca: das Schweigen der Veritas und die Worte des Dichters*, ed. Joachim Küpper, 54–88. Berlin: de Gruyter.

LAFON, Pierre. 1980. «Sur la variabilité de la fréquence des formes dans un corpus.» *Mots* 1 (1), 127–165.
<https://doi.org/10.3406/mots.1980.1008>.

LAKOFF, George & Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

MARNOTO, Rita. 2015. *O Petrarquismo Portugues do Cancioneiro Geral a Camões.* Manuais Universitariós. Lisboa: Imprensa Nacional-Casa da Moeda.

MARX, Barbara. 1998. „Petrarkismus im Zeichen von Dante: Pietro Bembo und die Asolani“. *Deutsches Dante-Jahrbuch* 73 (1), 9–50.
<https://doi.org/10.1515/dante-1998-0103>.

MIMOTEXT. 2022. „Projekt.“
<https://www.mimotext.uni-trier.de/projekt>, 23.2.22.

MORALES SARAVIA, José. 1998. «Vanitas y petrarquismo en el soneto XXIII de Garcilaso de la Vega.» *Iberoromania* (47), 47–71.

MUTSCHKE, Peter. 2010. „Zentralitäts- und Prestigemaße.“ In *Handbuch Netzwerkforschung*, ed. Stegbauer, Christian & Roger Häußling, 365–378. Wiesbaden: VS Verlag für Sozialwissenschaften.
<https://doi.org/10.1007/978-3-531-92575-2_33>.

NARDONE, Jean-Luc. 1998. *Pétrarque et le pétrarquisme.* Que sais-je? 3338. Paris: PUF.

NOYER-WEIDNER, Alfred. 1985. „Poetologisches Programm und ‚erhabener‘ Stil in Petrarcas Einleitungsgedicht zum ‚Canzoniere‘“. *Italienische Studien* 8, 5–26.

PASQUINI, Emilio. 2011. “Medieval Polarities: Dantism and Petrarchism.” In *Dante in Oxford: The Paget Toynbee Lectures 1995-2003*, ed. Kay, Tristan, Martin McLaughlin & Michelangelo Zaccarello, 167–180, Oxford: Legenda.
<https://www.taylorfrancis.com/chapters/edit/10.4324/9781315095141-9/medieval-polarities-dantism-petrarchism-emilio-pasquini>.

PETRARCA, Francesco. 1964. *Canzoniere*, ed. Contini, Giancarlo. Torino: Einaudi.
<http://www.letteraturaitaliana.net/pdf/Volume_2/t319.pdf>.

PIROVANO, Donato. 2014. *Il dolce stil novo.* Sestante 30. Roma: Salerno.

PYRITZ, Hans. 1963. *Paul Flemings Liebeslyrik: Zur Geschichte Des Petrarkismus.* Palaestra 234. Göttingen: V&R.

QUONDAM, Amedeo, Beatrice Alfonzetti & Stefano Asperti. 2003. *Biblioteca Italiana*. Roma: Sapienza Università di Roma. <http://www.bibliotecaitaliana.it/>.

REGN, Gerhard. 1987. *Torquato Tassos zyklische Liebeslyrik und die petrarkistische Tradition: Studien zur Parte Prima d. Rime (1591/1592).* Romanica Monacensia 25. Tübingen: Narr.

REGN, Gerhard. 1993. „Systemunterminierung und Systemtransgression. Zur Petrarkismus-Problematik in Marinos Rime amorose (1602)." In *Der Petrarkistische Diskurs: Spielräume und Grenzen; Akten des Kolloquiums an der Freien Universität Berlin, 23.10.–27.10.1991*, ed. Hempfer, Klaus W., 255–281. Text und Kontext 11. Stuttgart: Steiner.

REGN, Gerhard. 2013. „Petrarkismus." In *Historisches Wörterbuch Der Rhetorik Online*, ed. Ueding, Gert. Berlin, Boston: De Gruyter. <https://www.degruyter.com/view/HWRO/petrarkismus?pi=0&moduleId=common-word-wheel&dbJumpTo=Petra. Accessed 27.09.19>.

RISSMAN, Leah. 1983. *Love as war: Homeric allusion in the poetry of Sappho*. Beiträge zur klassischen Philologie 157. Königstein: Hain.

ROHDEN, Jan. 2021. *Corpus.* DARIAH-DE. <https://doi.org/10.20375/0000-000e-8add-e>.

ROHDEN, Jan. 2021. *Custom variants of Craig's Zeta*. DARIAH-DE. <https://doi.org/10.20375/0000-000e-8b18-b>.

ROHDEN, Jan. 2021. *DRT2021*. <https://github.com/ja-roh/DRT2021/>.

ROHDEN, Jan. 2022. "Petrarch's Poetic Style from a Computational Perspective: A Digital Quantitative Approach to Italian Petrarchism." In *Tackling the Toolkit: Plotting Poetry through Computational Literary Studies*, ed. Plecháč, Petr et al., 111–134. Prague: Institute of Czech Literature of the Czech Academy of Sciences. <https://doi.org/10.51305/ICL.CZ.9788076580336.08>.

ROTARI, Gabriela, Melina Jander & Jan Rybicki. 2021. "The Grimm Brothers: A Stylometric Network Analysis." *Digital Scholarship in the Humanities* 36 (1), 172–186. <https://doi.org/10.1093/llc/fqz088>.

SANTAGATA, Marco. 1990. *Per moderne carte: la biblioteca volgare di Petrarca*. Saggi 371. Bologna: Il Mulino.

SCHIFFER, James. 2000. "Shakespeare's Petrarchism." In *Shakespeare's Sonnets: Critical Essays*, ed. Schiffer, James, 163–183. Shakespeare Criticism 20. New York: Garland.

SCHÖCH, Christof et al. 2018. "Burrows' Zeta: Exploring and Evaluating Variants and Parameter." In *Mexico City: ADHO 2018*, ed. AA.VV. Mexico City. <https://dh2018.adho.org/burrows-zeta-exploring-and-evaluating-variants-and-parameters/>.

SCHÖCH, Christof et al. 2018. „Burrows Zeta: Varianten und Evaluation." In *DHd 2018: Kritik der digitalen Vernunft*, ed. AA.VV., 138–143. Köln. <http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf>.

SCHÖCH, Christof. 2018. „Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie." In *Quantitative Ansätze in den Literatur- und Geisteswissenschaften*, ed. Bernhart, Toni et. al., 77–94. Berlin, Boston: De Gruyter.

SCHÖCH, Christof. 2020. *TXM-Tutorial. Figurenbeschreibungsprojekt*. Zoom. <https://christofs.github.io/txm-tutorial/#/>.

SCHUMACHER, Mareike. 2020: „Netzwerkanalyse mit Gephi." *forTEXT. Literatur digital erforschen*. <https://fortext.net/routinen/lerneinheiten/netzwerkanalyse-mit-gephi>.

SUZUKI, Takafumi et al. 2012. "Co-Occurrence-Based Indicators for

Authorship Analysis." *Literary and Linguistic Computing* 27 (2), 197–214. <https://doi.org/10.1093/llc/fqs011>.

SEITSCHEK, Gisela. 2014. „Von der Donna angelicata zur gloriosa Beatrice. Stilo della loda oder Lobpreis der Herrin beim frühen Dante und den Stilnovisten." In *Das literarische Lob: Formen und Funktionen, Typen und Traditionen panegyrischer Texte*, ed. Franz, Norbert P., 55–84. Schriften zur Literaturwissenschaft 36. Berlin: Duncker & Humblot.

STIERLE, Karlheinz. 1979. *Petrarcas Landschaften: zur Geschichte ästhetischer Landschaftserfahrung*. Schriften und Vorträge des Petrarca-Instituts Köln 29. Krefeld: Scherpe Verlag.

TRILCKE, Peer. 2013. „Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft." In *Empirie in der Literatur-wissenschaft*, ed. Ajouri, Philip, Katja Mellmann & Christoph Rauen, 201–247. Poetogenesis – Studien zur empirischen Anthropologie der Literatur 8. Münster: Brill, mentis. <https://doi.org/10.30965/9783957439710_012>.

TROVATO, Paolo. 1979. *Dante in Petrarca: per un inventario dei dantismi nei Rerum vulgarium fragmenta*. Firenze: Olschki.

ULMER, Birgit. 2010. *Die Entdeckung der Landschaft in der italienischen Literatur an der Schwelle zur Moderne*. Dialoghi = Dialogues 15. Frankfurt a. M.: Lang.

WARNING, Rainer. 1987. „Petrarkistische Dialogizität am Beispiel Ronsards." In *Die Pluralität der Welten: Aspekte der Renaissance in der Romania*, ed. Stempel, Wolf-Dieter & Karlheinz Stierle, 327–358. Romanistisches Kolloquium 4. München: Fink.

ZACCAGNINI, Guido & Amos Parducci (ed.). 1915. *Rimatori siculo-toscani del dugento.* Bari: Laterza. <https://www.liberliber.it/online/autori/autori-r/rimatori-siculo-toscani-del-dugento/rimatori-siculo-toscani-del-dugento>.

Zusammenfassung

Wenige Autoren haben die europäische Lyrik so geprägt wie Francesco Petrarca (1304–1374). Das liegt vor allem an dem poetischen Stil von Petrarcas wichtigstem Text in italienischer Sprache, einer Sammlung von Liebesgedichten mit dem Titel *Canzoniere*, die über Jahrhunderte zu einem bedeutenden poetischen Modell für die europäische Lyrik wurde. Petrarcas Einfluss auf die europäische Literatur wird in der Forschung oft mit dem Ausdruck „Petrarkismus" bezeichnet, allerdings wird die genaue Definition des Begriffs nach wie vor kontrovers diskutiert. Ein Grund dafür mag sein, dass nur wenige Untersuchungen zum Petrarkismus ein größeres Textkorpus in den Blick nehmen. Die Analyse eines umfassenden Korpus petrarkistischer Dichtung könnte allerdings nicht nur existierende Definitionen auf breiterer Basis erproben, sondern ferner einen quantitativ fundierten Beitrag zum besseren Verständnis des Petrarkismus leisten.

Der vorliegende Beitrag untersucht ausgehend von bestehenden Erklärungs-ansätzen des Petrarkismus ein Korpus aus 55 italienischen Gedichtsammlungen mittels digitaler stilometrischer, Kookkurrenz- und Netzwerkanalysen.

Abstract

Few authors have shaped the history of European poetry as much as Petrarch (1304–1374). This is largely due to the poetic style of his most important Italian text, a collection of love poems entitled *Canzoniere*, which became an important poetic model for centuries. Scholars usually use the word "Petrarchism" to refer to Petrarch's influence on the literary landscape, although the exact definition of the term continues to be under discussion. One reason for this may be the fact that few studies of Petrarchism focus on a larger corpus of texts. The analysis of a comprehensive corpus of Petrarchan poetry, however, could not only offer the possibility of testing existing definitions on a larger basis, but furthermore make a quantitatively sound contribution to a better understanding of Petrarchism.

Drawing on existing definitions of Petrarchism, this paper examines a corpus of 55 Italian poetry collections using digital stylometric, co-occurrence, and network analyses.

Dossier

Digital, global, transdisziplinär: Impulse für die Romanistik

Teil 2

Metadaten Bibliotheken Infrastrukturen

Bild: Computergeneriertes Bild (DreamStudio) nach Fernando Botero; Prompt "books and computers in a library" (CC0 1.0)

apropos
[Perspektiven auf die Romania]
Winter 2022
9
www.apropos-romania.de
ISSN: 2627-3446

2022, n°9
pp. 119-152
doi: https://doi.org/10.15460/apropos.9.1894

José Calvo Tello

Where are Romance Studies Heading?

A Bibliographic Data Science Analysis Using Regression

José Calvo Tello
is subject librarian and researcher at Göttingen State and University Library.
calvotello@sub.uni-goettingen.de

Keywords

Romance Studies – Bibliographic Data Science Analysis – Library Records – Linear Regression

1. Introduction

Researchers of any discipline share certain opinions and intuitions about how their discipline has developed in recent years. This is normally influenced both by their own experience and by other senior researchers' opinions about the previous decades. Of course, shared intuitions do not necessarily need to be based on actual facts, they are often misled by particular experiences, the specifics of a research institution or the development of a specific sub-discipline.

In this study, I tackle the question about how the Romance Studies have developed in the past decades. For this purpose, I use library records as research objects and apply statistical methods. In addition to the description of the past decades, I use statistical models to make predictions of the impact of current trends in future years.

Romance Studies are a challenging discipline. In contrast to other philologies such as German or English Studies, a plurality of languages are at the core of this discipline (Kramer 2002, 13). Besides, in the countries where these languages are spoken, the research mainly focuses on one language (French Studies, Spanish Studies, etc.). A further challenge for the Romance Studies is the multilingualism of their research production. Any study trying to be representative for this discipline needs to cover at least publications written in French, Spanish, Italian, Portuguese, German and English, with a long tail of further important languages such as Romanian, Catalan, Occitan, Sardinian, etc.

Previous research about the history of the Romance Studies has mainly focused on the periods before 1950 (Richert 1913; Kalkhoff 2010; Wolf 2012; Lieb and Strosetzki 2013; Kremnitz 2016; Kramer 2020). As Kremnitz points out, it is more difficult to assess current developments than to describe the historical processes (Kremnitz 2016, 287). Many of these historical studies worked on a narrow

selection of scholars with great impact in the fields of Romance Linguistics or Literary Studies. Such an approach is possible when the historical distance to the research object is enough to identify the most influential researchers. Since this is not possible for the last decades, I decided to take a quantitative perspective by using data curated by professionals, i.e. library records. In the past years, Digital Humanities have shown a new interest in working with data from bibliographies and library records (Henny-Krahmer 2017; Jannidis, Konle, and Leinen 2019; González 2021; Ehrlicher and Lehmann 2021; Herrmann et al. 2021; Gittel 2021). This work can be framed in the new paradigm of the *bibliographic data science*, an emerging sub-discipline closely related to the Digital Humanities that analyzes library records and bibliographies to study the historical development of several types of publications, including literature, research, journals, etc. (Tolonen et al. 2020; 2019; Vaara et al. 2019; Maryl and Wciślik 2016).

In this study, I focus on the development of the past 40 years, i.e. between 1980 and 2019. The interest for these years is based on several historical changes, such as the inclusion of many European Romance countries into a shared political structure such as the European Union, the Bologna process, the development of new technologies such as the Internet or e-books, or the rapid development of the scholarly publishing sector (Kramer 2002; Becker et al. 2020; Krefeld 2020; Monjour 2020). These four decades lend itself to analysis because of the availability and quality of the data in the catalogs, which is higher for publications from the last decades than for previous ones.

This analysis focuses on the Romance Studies from the perspective of the German-speaking area. For this, I use as data the library records from German university libraries, as I will explain in detail in the next section.

One could ask whether it is acceptable to use data from German libraries to analyze a discipline about foreign languages. Although a study based on using the data from several linguistic regions and nations could be a valuable option, it could be asked whether such an approach would cover the Romance Studies and the Romance languages accurately. As many researchers acknowledge, in the Romance-speaking countries the disciplines tend to focus only on the national language (Kramer 2002, 17; Holtus and Sánchez Miret 2008; Kremnitz 2016). The German-speaking area is seen as the place where the Romance Studies started and where a comparative approach finds more support (Wandruszka 1988; Holtus and Sánchez Miret 2008; Kremnitz 2016), or as Gumbrecht states “Romance philology arose in Prussia (not in France, Spain, or Italy)” (Gumbrecht 2002, 2). For this reason, I consider the German libraries as one representative source of data for this analysis, although not the only one.

2. Dataset

2.1 Library Records from the Hebis and GVK - GBV Union Catalog

For this analysis, I use records from two large library catalogs. Research and university libraries in Germany are organized in consortia or networks within which infrastructure and data are shared (for example for the catalogs). The consortia tend to cover the research libraries of one or more federal states and there are currently six consortia (Gantert 2016, 44):

1. *Gemeinsamer Bibliotheksverbund* (GBV, translated into English as the GBV Common Library Network): responsible for the libraries of seven federal states plus the Prussian Cultural Heritage Foundation;
2. *Südwestdeutscher Bibliotheksverbund* (SWB): responsible for the libraries of three federal states;
3. *Hessisches BibliotheksInformationsSystem* (hebis): responsible for the libraries of Hesse and a region of Rhineland-Palatinate;
4. *Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen* (hbz): responsible for libraries in the federal states North Rhine-Westphalia and Rhineland-Palatinate;
5. *Kooperativer Bibliotheksverbund Berlin-Brandenburg* (KOBV): responsible for the libraries of two federal states;
6. *Bibliotheksverbund Bayern* (BVB): responsible for the libraries in Bavaria.

Since 2019, the GBV and the SWB have joined their efforts and have created the K10plus network. Within this network of consortia, the libraries share their library records in a database called the K10plus, which can be consulted as a catalog.[1]

Three of these consortia use the same format and software based on the PICA format, which stands for Project for Integrated Catalogue Automation (Voß 2020). The rest of the consortia work with different formats and software solutions such as Aleph and SISIS. Although it would have been of great interest to have included all catalogs in this study, there is no integrated pool of bibliographic data from these resources. This translates to several challenges when such an analysis is performed. First, different frameworks and technologies need to be applied. Second and most importantly, the record fields from the several formats need to be mapped into a single model in order to compare the categories across databases. This task cannot be done by a single researcher and the consortia currently do not offer solutions for this problem.

How representative are the records of these three consortia for the library landscape for Romance Studies? This dataset does cover the records from university libraries of the majority of the German states, since records from 11 out of 16 German states are being covered, with a diverse geographical distribution.

[1] <https://opac.k10plus.de/DB=2.299/START_WELCOME>.

Besides, the records cover at least a part of two *Specialised Information Services* (in German *Fachinformationsdienste für die Wissenschaft* or simply FIDs), which constitute projects carried out by libraries focused on a specific discipline (Gantert 2016, 155–58). For the Romance Studies, two FIDs are especially relevant:

1. The Specialised Information Service (FID) Romance Studies, at the University Library of Bonn and at the State and University Library of Hamburg.
2. The Specialised Information Service (FID) Latin America, Caribbean and Latino Studies, at the Library of the Ibero-American Institute (part of the Prussian Cultural Heritage Foundation) in Berlin.

Both the Prussian Cultural Heritage Foundation and the Hamburg State and University Library are part of the GBV and therefore their records can be found in the K10plus database. However, the records of the University Library of Bonn are integrated in the hbz consortium and therefore are not part of this analysis. In addition, the records of the FID for Russian, East and Southeast European Studies at the Bavarian State Library also had to be excluded from this analysis. This will be mentioned again when looking at the results and trends for Romanian.

To summarize the answer to the question about the representativeness of the dataset, it represents the greatest dataset of library records about Romance Studies and it does represent the majority of the libraries in Germany. However, important sections of the German librarian landscape are not being considered, in particular the resources from the hbz and the BVB. This opens the possibility that similar analyses could be performed in the future using datasets created in a distributed manner, with each consortium contributing its respective records.

The data from both networks can be accessed via Application Programming Interfaces (APIs). From these sources, the data can be downloaded in Pica+ format, expressed in an XML serialization. Although unknown in other areas, Pica+ is the standard format for cataloging in libraries from several countries, among others in Germany.

The retrieved databases are also the sources for the standard Online Public Access Catalog (OPACs) of the libraries. These catalogs contain data of independent works (in German *selbständige Werke*) such as monographs, collective works and journals (Gantert 2016, 228–29). This means that chapters of collective works or articles in journals are not part of the analyzed dataset.

2.2 Classification Systems for the Identification of Romance Studies Publications

After downloading a dataset for the last decades from the APIs of both consortia (K10plus[2] and hebis),[3] the next step is to extract from the original dataset only those publications related to Romance Studies. For that, I use several library classification systems. These classification systems are hierarchical structures of classes that represent subjects, such as Chemistry, Theology or Romance Studies (Gantert 2016, 203). In theory, any class can be divided into more specific classes, creating a tree-like structure or taxonomy. One or more classes of these classification systems are then assigned to any publication. This annotation can be used by users to retrieve from the catalog the publications of any specific area. In my case, these classification systems allow me to identify the publications relating to Romance Studies and thus define the dataset for the analysis.

For filtering of the original data, I apply three classification systems:

- Dewey Decimal Classification and Subject Categories:[4] The German National Library's catalog is using a set of around 100 Subject Categories (in German *Sachgruppen*). These groups are a simplification of one of the most widely accepted classification systems, especially in English-speaking countries: the Dewey Decimal Classification (DDC; see further details in Chowdhury et al. 2008, 96–99).[5] Specifically, for this analysis I consider publications assigned with the DDC classes starting with 44, 45, 46, 84, 85, or 86.
- Regensburger Verbundklassifikation (RVK):[6] This classification system is probably the most widespread in the German-speaking area. In its current version, it contains more than 800.000 classes and these do not only represent subjects, but also publications about or by specific people.[7] For this analysis, I consider publications assigned with the RVK classes starting with the letter I.
- Basisklassifikation (BK):[8] This classification system was originally developed in the Netherlands, and it has a wide acceptance in the libraries of the GBV consortium. It contains around 2.000 classes and, as its name describes, it is seen as a basic classification system in contrast to other more complex systems such as the complete DDC or the RVK. I consider publications assigned with BK classes starting with 18.2 or 18.3.

[2] <https://wiki.k10plus.de/display/K10PLUS/SRU>.
[3] <http://sru.hebis.de/sru/DB=2.1>.
[4] <https://www.dnb.de/DE/Professionell/DDC-Deutsch/DDCinDNB/ddcindnb_node.html
[5] The German National Library decided to add three classes which were not present in the original Dewey Decimal Classification: Literary fiction (class B, in German *Belletristik*), Children and Youth literature (class K, in German *Kinder- und Jugendliteratur*) and Textbooks (class S, in German *Schulbücher*).
[6] <https://rvk.uni-regensburg.de/>.
[7] For example, the RVK class IR 8005 represents secondary literature about Fernando Pessoa.
[8] <https://wiki.k10plus.de/pages/viewpage.action?pageId=437452809>.

After a first analysis, it became clear that primary literature (i.e. published novels, poetry, theater plays, etc.) and secondary literature (research publications) differ in many of their characteristics. For example, in many cases publishers do not offer any e-book license for primary literature to university libraries. Besides, the prices of primary literature are much lower than those of secondary literature, since the first ones are meant for a general public, while the latter are meant for a specialized readership and have therefore lower print runs. Because of these and other differences in many of the analyzed categories, I decided to exclude all primary literature from the dataset. For this step, I use again the three classification systems, since all of them have one or more classes that specify that the publication is primary literature:

- *Sachgruppen*: class B;[9]
- BK: classes 17.97 and 17.98;
- RVK: classes containing in their labels the phrases *Gesammelte Werke* ('Collected works') or *Einzelwerke* ('individual works'). This is the case for almost 8.000 RVK classes. For example, the class IE 5101 represents the individual works of primary literature by the author Adem de la Halle, while for the next author in the RVK, Audefroi (le Bastard), the class IE 5107 represents both his collected and individual works. These cases exemplify the challenges of working with the RVK: what could be done for the *Sachgruppen* and BK with one or two steps, the researcher is forced to repeat them 8,000 times for the RVK, which is only feasible when scripts can be programmed.

For the exclusion of the primary literature, I also use the field content type. This field[10] "contains factual terms to describe the content of this publication".[11] This is a mandatory field for many types of publications, such as comics, biographies or exhibition catalogs, but not for primary literature. Even when it is not mandatory, many records are marked as fictional representation (*Fiktionale Darstellung*), anthology (*Anthologie*) and collection of letters (*Briefsammlung*). The records from both cataloges (hebis and K10plus) with these values were also excluded from the analysis.

The specific implementation of the steps can be followed in the companion Jupyter Notebooks of this publication, which will be described in detail in Section 2.5.

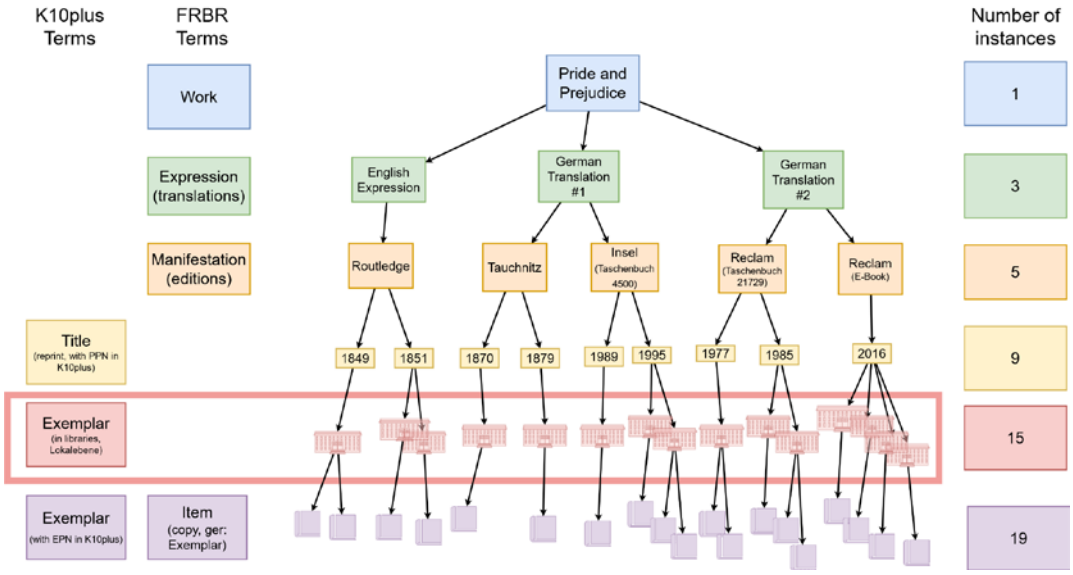
As mentioned before, I only analyze records published between 1980 and 2019 (including both years). These four decades represent a compromise between a historical overview of the field and the quality and homogeneity of the data. Cataloging rules and workflows in libraries evolve constantly. The data of the catalog for publications in 1900 is very different to the data for publications of the

[9] Only present in the German *Sachgruppen* and not in the original DDC.
[10] With the Pica3 code 1131, Pica+ code 013D.
[11] Following the documentation of the K10plus, my translation: <https://swbtools.bsz-bw.de/cgi-bin/k10plushelp.pl?cmd=kat&val=1131&katalog=Standard&adm=1>.

year 2000. For instance, the classification systems applied to identify Romance Studies publications (*Sachgruppen*, BK and RVK) were developed only in the second half of the 20th century. In any case, exploring ways of expanding the analyzed period while ensuring data quality is worth further studying.

2.3 Analyzed Instance: FRBR and Pica Model

A further filter needs to be applied to the dataset: what kind of instance is analyzed exactly. I have already stated that the hebis and K10plus databases contain monographs, journals and series and tend to exclude chapters and journal articles (among other scholarly publications). However, the databases consider different levels of abstractions relating to the publications. In this section, I explain these instances using the example of *Pride and Prejudice* visualized in Figure 1, even though the analysis does not cover primary literature as explained before. The novel *Pride and Prejudice* was written originally in English by Jane Austen (used as example in Wiesenmüller and Horny 2017). The model of the Functional Requirements for Bibliographic Records (FRBR, Wiesenmüller and Horny 2017) uses the term *work* to relate to the abstract unit. This unit is then expressed in several linguistic *expressions*. These expressions include the original text (in this case, the English text written by Austen) and the different translations. If the text is translated more than once, every translation is counted as a further expression of the work. When a publisher publishes one of these expressions, then it creates a *manifestation*. If several publishers publish the same expression (the original text or any translation), they are considered several manifestations (Wiesenmüller and Horny 2017, 18). However, if the same publisher launches several reprints of the text, they are still considered part of the same manifestation (Wiesenmüller and Horny 2017, 19).



1 | Comparison of FRBR and Pica+ models with the example of *Pride and Prejudice*

Although the fields from both databases are organized following this FRBR model, the databases in these consortia reflect an alternative model. The reason for that

is that, following the FRBR model, reprints of several years would be part of the same manifestation. However, the users of the library could be interested in knowing whether the copy of *Pride and Prejudice* in the library was published in 1849 or in 1851. For this reason, the presented model in the analyzed databases considers more specific instances. On the top of this model, we find the concept of title (*Titel* in German). This could be seen as any manifestation but distinguishing the reprints of different years. Each different title has a unique identifier (called *Pica-Produktionsnummer* or PPN) in the database.

When a library acquires a publication (in the case of printed publications and e-books licenses) or decides to consider a publication in their catalog (in the case of Open Access publications), it creates an *exemplar* in the library's catalog. From each exemplar, the library can purchase one or more copies (*items* in FRBR terms). Each item receives a different call number that enables library users to find it in the library. In the database, each item is identified with a unique identifier (called *Exemplar-Produktionsnummer* or EPN).

Theoretically, I could have analyzed any instance of the FRBR model or the databases, from the most abstract one (work) to the most specific (item). This decision has strong implications on at least two aspects. First, working with FRBR instances would have forced us to reconcile many of the data of the databases since the FRBR instances are actually not explicitly identified for the majority of the cases. That would have meant to modify and edit many of the analyzed data, resulting in perhaps errors and noise. Second, working on more abstract instances (such as work or expression) would have made the analysis blind to many aspects.

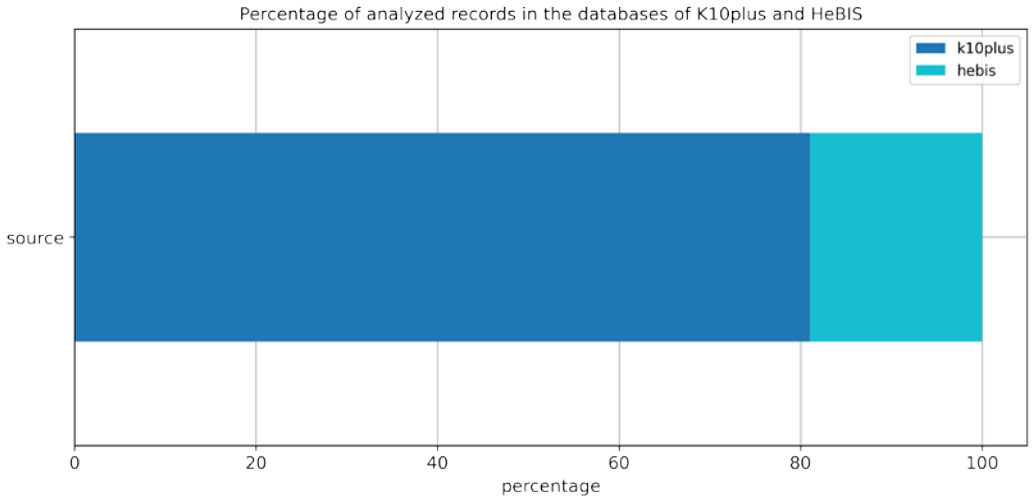
It is valuable information about a work knowing whether several translations can be found in the catalog, whether several publishers were interested in the text, whether several reprints were made. Besides, it is informative whether a publication is present in only one German library (perhaps purchased by the FIDs) or if many German libraries have it in their catalog. Since part of the goals of some FIDs is to purchase relevant publications of a discipline, working at the title level would have disproportionately magnified the impact of the FIDs which would have skewed the results of the analysis. In order to avoid this, I decide to work at the exemplar level. In this way, the important library stocks of the FIDs are a part of the analysis, but all libraries in both consortia are considered equally.

I reject the options of working at the item level, and therefore ignoring the number of copies of each publication in each library. Although this could have some interest, the number of copies is stronger related to very specific aspects of each department, such as the number of students, the budget funds, or whether the publication was used in class.

2.4 Description of the Data Set

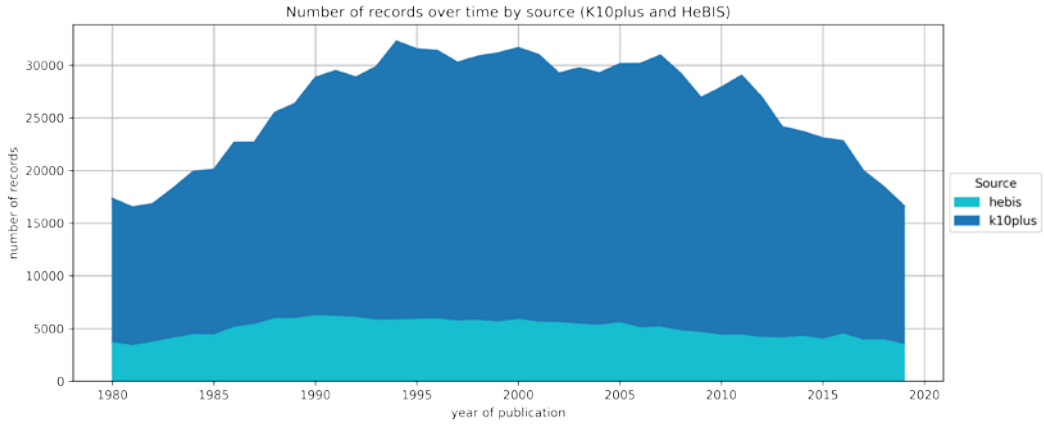
After these steps, the available dataset contains 1,041,157 exemplars of publications in the libraries of both consortia. These exemplars are based on 334,221 different titles (identified with different PPNs in both consortia). That

means that on average, each title tends to be present in three libraries of both consortia.



2 | Percentage of analyzed records in the databases of K10plus and hebis

Figure 2 shows the distribution of records in both consortia. The dataset contains 843,558 records from the K10plus database, which constitutes 81% of the analyzed exemplars. Correspondingly, hebis contributes with 197,599 records, around 19%. The difference of the number of records of both databases is not surprising since the K10plus covers a much wider area and therefore more libraries than hebis. Figure 3 shows the chronological distribution of the four decades with the number of records from both consortia.



3 | Number of records over time by source (K10plus and hebis)

The chronological distribution of Figure 3 shows an increase of records in both databases until the 1990s. After that, the number of annual records in the hebis decreases, while it remains stable in the K10plus, at least until the end of the 2000s. The decrease of annual publications is especially noticeable in the last 10 years in both databases. While there is a certain variation, the distribution between both databases remains rather stable over the entire period. The number of records

obtained from the hebis database oscillates between 15% (the lowest point in 2011) and 24% (the highest point in 1987).

Has the number of publications in the field of Romance Studies been decreasing since 2010 but especially since 2015, as Figure 3 shows? Before jumping to conclusions, the characteristics of the analyzed dataset need to be remembered: records from libraries filtered based on classification. Libraries do not only purchase or obtain publications of the current or last year, but also from former years and decades. These publications then need to be cataloged and assigned to classification systems in order to be identified as publications for Romance Studies. In other words, it might take some years for libraries to complete the process of identifying, obtaining, cataloging, and classifying publications properly. For this reason, I consider all results from the last five years as preliminary.

Finally, I would like to add a remark about joining the records from both consortia. Although I have argued for the combination of the records from the K10plus and the hebis databases, this step also means an increase of the heterogeneity of the dataset. Although both consortia share many characteristics, some specifics about cataloging rules, classification systems or formats differ. For example, some Pica+ codes are different in both databases, and therefore the extraction of the data needs to be done separately. Another case relates to the classification systems: while BK is one of the most widely used classification systems in the K10plus, it is not used in hebis, where the RVK is more widespread. Although these subtle differences have little impact in this analysis, it needs to be considered the trade-off between the greater interest of combining data from different sources and the resulting decline of the quality of the data.

2.5 Extraction of Information and Normalization of the Data

The data of the catalog was saved as the Pica+-XML files. In order to analyze the data, a selection of the sub-fields of the catalog were extracted from the Pica+-XML fields and saved as columns in a tabular format. These tables then were saved as parquets files, a format for large or sparse tables. However, the content present in each field needs certain amount of pre-processing in order to be analyzed. For example, the Pica+ field 34D (sub-field a) contains the number of pages of each publication. Table 1 shows a few examples of the content that can be found there.

PPN	Title	Pages
1132286107	Antología poética	128 S
1604360445	Everest diccionario práctico de americanismos	238 S.
1604428902	Crônica da casa assassinada	XXXVII, 810 S.
1612842208	La @mort d'Agrippine	XXIV, 90 S.
228112982	Gewohnheit - heilen	476 S.
278864376	Cultures of the Aztecs, Mayas and Incas	216 S.
34184957X	Regards sur la littérature québécoise	312 p
489256120	French studies	Online-Ressource
745009786	Revista de Cancioneros Impresos y Manuscritos	Online-Ressource
798179643	Der @französische Wortschatz der Vorklassik	377 Seiten

Table 1 | Random sample of titles, with the original information of pages

Except for the online resources, the rest contain information about the number of pages. However, they are encoded slightly differently: with the German word *Seiten*, its first letter, or the letter "p". Some cases have a full stop at the end, others do not. Besides, two examples mark the length of an introductory section with Roman numerals.

If a researcher wants to calculate the mean of pages in publications, it is needed to extract the numerical values from the column "Pages" in Table 1. For this field, a regular expression was enough to extract the numerical values of these strings and assign them to the new column "pages extracted" in Table 2.

PPN	Title	Pages	Pages extracted
1132286107	Antología poética	128 S	128.0
1604360445	Everest diccionario práctico de americanismos	238 S.	238.0
1604428902	Crônica da casa assassinada	XXXVII, 810 S.	810.0
1612842208	La @mort d'Agrippine	XXIV, 90 S.	90.0
228112982	Gewohnheit - heilen	476 S.	476.0
278864376	Cultures of the Aztecs, Mayas and Incas	216 S.	216.0
34184957X	Regards sur la littérature québécoise	312 p	312.0
489256120	French studies	Online-Ressource	NaN
745009786	Revista de Cancioneros Impresos y Manuscritos	Online-Ressource	NaN
798179643	Der @französische Wortschatz der Vorklassik	377 Seiten	377.0

Table 2 | Random sample of titles, with the original and extracted information of pages

As can be observed, the simpler cases are correctly extracted. The pages in Roman numerals are ignored and therefore the information relating to the number of pages of these publications is simplified. Besides, the electronic resources are set as *NaN* ('Not a number') and these cases can easily be ignored in later calculations.

In other categories with the option of multiple values (for example, publications in several languages), I extract the data using ad-hoc tokenizers for the encoded information in Pica+ which normalize the data to a certain point. More details can be found in the companion Jupyter Notebooks, described in the following section. However, I decide to not adding many normalizing steps to the data and trust the work of the librarians who edit the catalog. Quantitative normalization of the data for the analysis can introduce new types of errors and noise, for example, joining together entities that should be treated separately.

When dealing with catalog data it is important to remember that only a few fields contain information relevant to all records. While the catalog contains information about the length in pages of 95% of the records, information about the price can be only found for 27% of the cases. This missing information is due to many causes: it was not feasible to obtain the information (for example, many publications do not contain any references to the year or publisher), some information may not be applicable to all records (for example, name of publisher in manuscripts) or the cataloging practice considers the information optional (such as price, see Section 4.5). This means that the analysis of some categories with a coverage close to 100%

of the records (such as the medium or the language of the publication) is much more representative than others (such as the price).

2.6 Publication of Code and Data

The data and code used for this publication are available online for anyone interested. Both components have been saved in two repositories: DARIAH Repository[12] and Zenodo.[13] The folder "data" contains the tables with the bibliographic records of the Romance Studies publications. The folder "code" contains scripts written in the programming language Python. These are in two formats: functions written in a Python script (with the ending ".py") and Jupyter Notebooks (with the ending ".ipynb"). Jupyter Notebooks are documents that can combine documentation, programming code and its output into a single file (VanderPlas 2016; Dombrowski, Gniady, and Kloster 2019). This way, any reader can have access to all steps, parameters and outputs that I considered during the analysis, many of which cannot be addressed in this article.

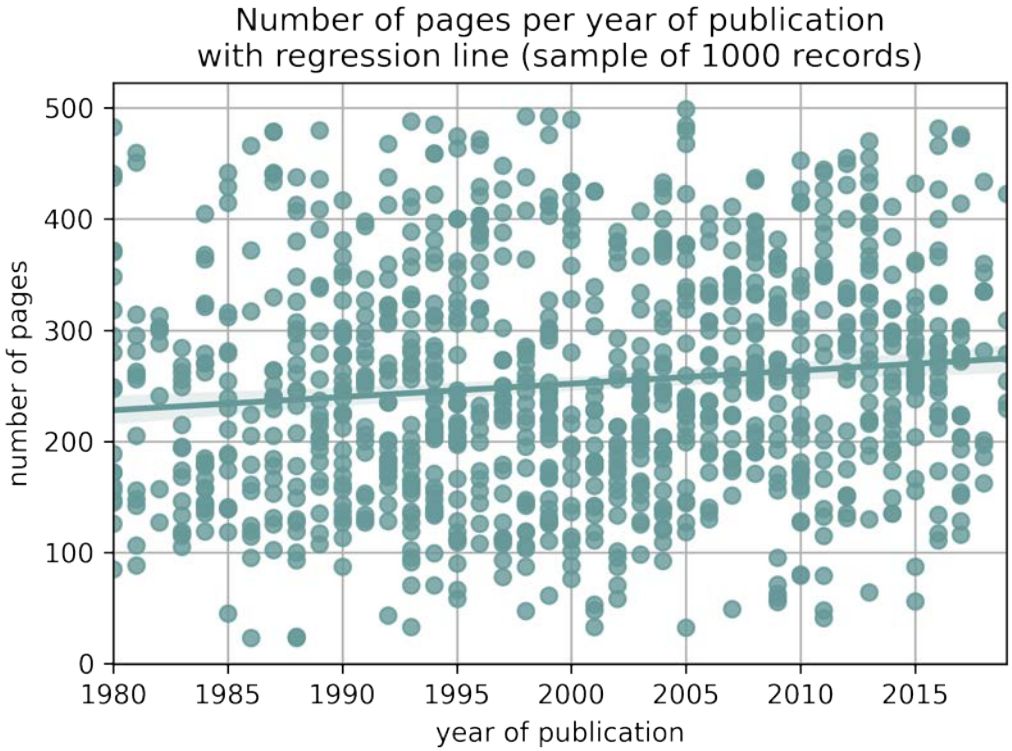
3. Linear Regression as Method and Example of Pages

As mentioned before, in this article I mainly use linear regression for the analysis of the categories. Linear regression is part of the family of techniques from statistics known as regression, whose goal is to create a model from some observed data in order to predict numerical values for new cases (Evans 2014, 160–69). Therefore, regression can be seen as a method of Machine Learning (Müller and Guido 2016). Like classification, it is also a supervised method since the input data contains cases with the correct output-labels (in contrast to unsupervised methods such as clustering or dimensionality reduction). However, while the the task of classification in Machine Learning predicts categorical values (such as the genre of publication or the name of the author), regression predicts a numerical value, such as a price or a probability that something will happen. In this article, I use linear regression to predict what can be expected in future years for Romance Studies publications.

For an intuitive idea of this technique, in this section I take the information of the number of pages of each publication. Before looking at the real data, let us imagine an unreal scenario in which all publications in the field of the Romance Studies published in 1980 would be exactly 80 pages long. In 1981, all authors and publishers went a step further and every single publication would be 81 pages long. This would have repeated in 1982 (82 pages), 1983 (83), and kept on going during the entire period, so that all publications from 2019 would be 119 pages long. It is easy to predict in this unlikely scenario that for the year 2030, publications should be 130 pages long. Of course, this prediction is based on the premise that authors and publishers would follow the previous trend. In other words, this technique assumes a conservative perspective that the development will remain similar and therefore ignores the possibility of sudden changes.

[12] <https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000F-73EA-8>.
[13] <https://zenodo.org/record/7098830#.YyqblbTP1aS>.

Let us now move to the real data of the length of publications.[14] In Figure 4, each publication is a data point and these are sorted following two axes: the horizontal axis sets the year of the publication and the vertical axis shows the number of pages of each publication, with values between ten and 500 pages. As can be expected, each year shows a variety of publication length. However, looking only at the lowest and highest values on the vertical axis, it can be seen that publications become slightly lengthier over time: While in the first years there is no publication close to 500 pages, this milestone is surpassed by several publications after the year 1990.



4 | Number of pages per year of publication with regression line

This tendency is not only observable in these long outliers, but in the central tendencies by year. While the median number of pages for publications in 1980 is 219 pages, it is 256 for publications published in the year 2019. The linear regression model formalizes this tendency as a slope, a function that is visualized as a line in Figure 4. This slope is positive, meaning that an increase in the values of the horizontal axis increases the predicted values in the vertical axis. In other words, we can expect lengthier publications in the future. The exact slope of this category in this data is 0.9, which is the increase of number of pages that can be expected for each year. Using different components of the models, it can be predicted that publications will be 277 pages long on average in the year 2030, and 286 in 2040. As mentioned before, this will only happen in the (unlikely) premise that things will develop exactly as they have done until now. Although I do not expect these exact

[14] In order to obtain a clearer visualization, Figure 4 does not contain the entire dataset, but a randomly selected data sample of 1000 publications from Romance Studies. This selection was used with the function sample of the Python library Pandas, which allows the user to get random samples of a given size. The reported slope and p-value are based on the entire dataset and very similar to the results based on the sample.

values to be perfect predictions for the future, they show the expected tendency for future years.

This uncertainty can be seen as a weakness of this study, but actually this is a problem inherent to any Machine Learning application. Regardless of the complexity of the data or the algorithms, current Machine Learning is based on the premise that new or future cases will follow the same principles as the ones that were observed in the past. This weakness of Machine Learning is perhaps clearer in this study because of the simplicity of the algorithm used and because the predicted values lay in the future, which of course cannot be predicted with certainty.

The exact function in Python that I use for the regression model can be seen in the Jupyter Notebooks. This function (from the library Scipy) gives also other values, such as an intercept, an r-value, a p-value and the standard deviation of the r-value. In this article, I report the p-value to observe whether the calculated tendency of the linear regression is statistically significant. Since the p-values correlate negatively with the size of the data and the dataset is relatively large, I assume a p-value lower than 0.001 for statistical significance. This is the case for the analyzed data of the pages in this section.

4. Analysis

In the following sections, I describe and analyze several categories extracted from the fields of the catalog, mentioning the Pica+ codes that were extracted from the original sources, the percent of records with this information in the catalog (coverage), the historical development of the past decades and then the prediction for the next decades.

4.1 Language of Publication

The language of publication is a frequent topic of reflection and discussion in Romance Studies. In general, it is accepted that three groups of languages are the main options (Lieber and Wentzlaff-Eggebert 2002):

1. The Romance languages, with a traditional predominance of French, followed by Spanish and Italian (Schrott 2003)
2. German, as the native language of many scholars in Romance Studies and the language of scientific communication in the Humanities in the German-speaking countries
3. English, as the international language of communication that the different communities can at least read

Each of these three options bring advantages and disadvantages, and these have been discussed in previous publications. In general, many researchers in Romance Studies argue that Romance Studies need to be a multilingual field, with many of them supporting German as language of publication; although it is accepted that the influence of English is increasing, it is seen as an undesired process which can bring harm both to German-speaking researchers and to the field of Romance

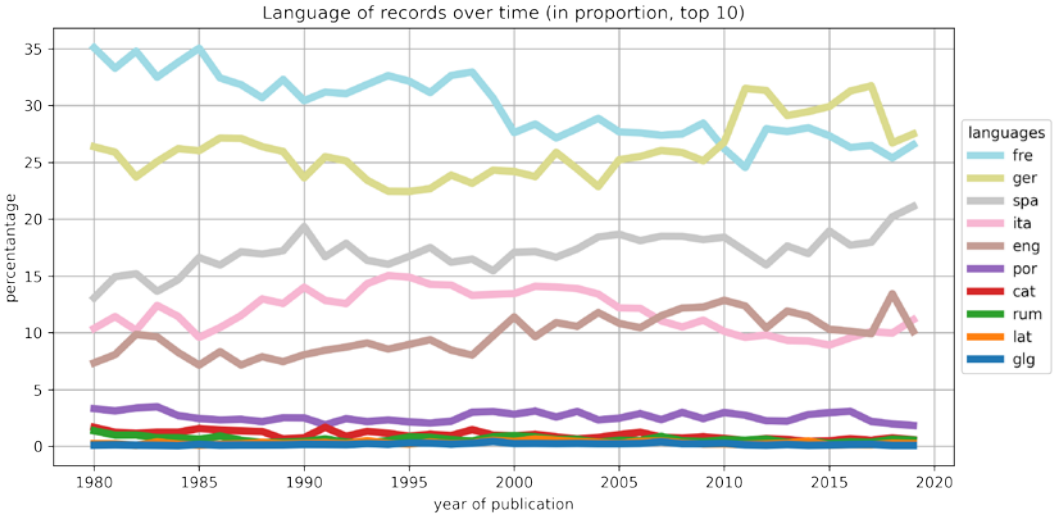
Studies (Lieber and Wentzlaff-Eggebert 2002; Kramer 2002; Constantinescu 2002; Nitschack 2002). Previous research has been centered on the role of the Romance Studies as publication language in different academic areas (Burr 2008; Kramer 2008; Haarmann 2008).

My hypotheses (based on articles or opinions spread throughout the community) for this period are the following:

1. The number of publications in French is being reduced over time (cfr. Wandruszka 1988; Haarmann 2008; Kramer 2008)
2. The number of publications in Spanish is increasing
3. The number of publications in English is also increasing

For the rest of the languages, including German, I do not have specific expectations but they are also considered for an exploratory analysis. It is especially interesting to observe the development of Romance languages such as Romanian, Catalan or Galician.

The data for the language of publication is extracted from the Pica+ field 010@ (sub-field a). The coverage of this category is surprisingly high, with values in 98% of the records. Multiple values are possible, the combination of German and French being the most frequent (29,302 publications). In total, the analyzed dataset contains publications in more than 170 languages (such as Russian, Dutch, Polish, etc.).

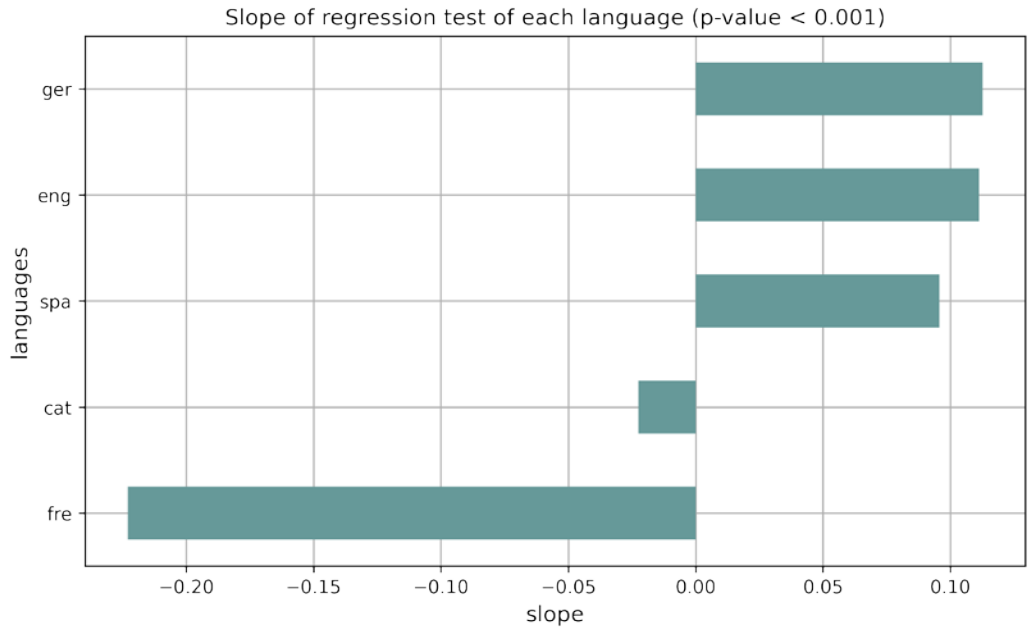


5 | Development of percentage of records by language (top ten)

Figure 5 shows the development of percentages of the records in the ten most frequent languages in the dataset. As can be seen, French is the language with the largest number of publications, with a total of 33.48%. A decreasing tendency of publications in this language can be observed for the entire period. French is closely followed by German with 29.17%, with a clear increase after 1995 and surpassing French in 2010 and therefore becoming the predominant language. Spanish is the third language with 19.44% of the records and a solid increasing tendency of publications for the entire period. 13.66% of the publications are in Italian, with an

increase until the 2000s and since then first certain decrease and finally stability around 10% for the last years. English is the fifth language with 11.21% of the records, with an increasing tendency during the entire period and surpassing Italian after 2007. The difference between these five languages and the next ones is notable: only 2.92% of the records are in Portuguese, followed by Catalan (1.08%), Romanian (0.67%), Latin (0.35%), and Galician (0.22%). As mentioned before, if the records from the BVB library consortium with the FID for Eastern Europe were part of the dataset, the number of publication in Romanian would be likely higher.

The next step is to apply linear regression in order to quantify and evaluate the observed tendencies. For this, I run separate regression models for each language. Figure 6 shows the positive and negative slopes for those languages whose results have p-values smaller than 0.001. Only five languages show trends with statistical significance: German, English, Spanish, Catalan, and French. That means that the models are not able to make statistically significant predictions for languages such as Italian, Portuguese, Latin, or Galician, for example, because their situation is rather stable during the analyzed period.



6 | Positive and negative statistically significant slopes of linear regression models analyzing language of publication

As can be seen in Figure 6, German, English and Spanish have positive slopes, all three with values close to 0.1. This means that an increase of 0.1% of publications in these languages can be expected every year. For the year 2030, 29.46% of the publications can be expected in German, 20.04% in Spanish and 13.22% in English. On the lower part of Figure 6 the languages with negative slopes can be observed. While Catalan has a very small slope of only -0.02, French has the strongest absolute value with -0.22. With this tendency, the model predicts 23.04% of publication in French in 2030. By the year 2040, French is expected to be surpassed by Spanish and become the third most widely used language of communication in the Romance Studies, with still a large distance to English.

4.2 Place of publication

A further category strongly associated with the language of publication is the place of publication, which is the next analyzed category. Since the European countries have moved in the last decades towards a political and economic integration, it could be expected to see an increase of publications from the Romance-speaking countries. Besides, I want to explore the tendency of the number of publications from Romance-speaking countries from other continents.

This information can be found in the Pica+ field 33A, sub-field p. As a typical bibliographic information, its coverage in the catalogs is very high, with values in this field in 99% of the analyzed records. For those publications with several places of publications (for example, Berlin-Boston), the cataloging rules foresee several separated fields.[15] The field then is tokenized following this and trying to correctly extract places with names composed of several tokens (such as Frankfurt am Main, Buenos Aires, New York, etc.). Figure 7 shows an overview based on a random sample of 5,000 records,[16] visualized through the DARIAH Geo-Browser. The data loaded in the Geo-Browser is available online for further exploration.[17] The maps in Figure 7 show the dominance of places of publication in the German-speaking area, France, Spain and Italy. Besides, several publications come from the south-east of England (London, Oxford) and the East Coast of the United States and Canada. In Latin America, only exclusively the capitals of some countries are covered in this sample of 5,000 publications, which does not contain any publication from Africa, Asia or Oceania. Of course, this does not mean that the sample does not contain research from authors coming from or based in these areas, since many of them publish through printing houses based in other countries.

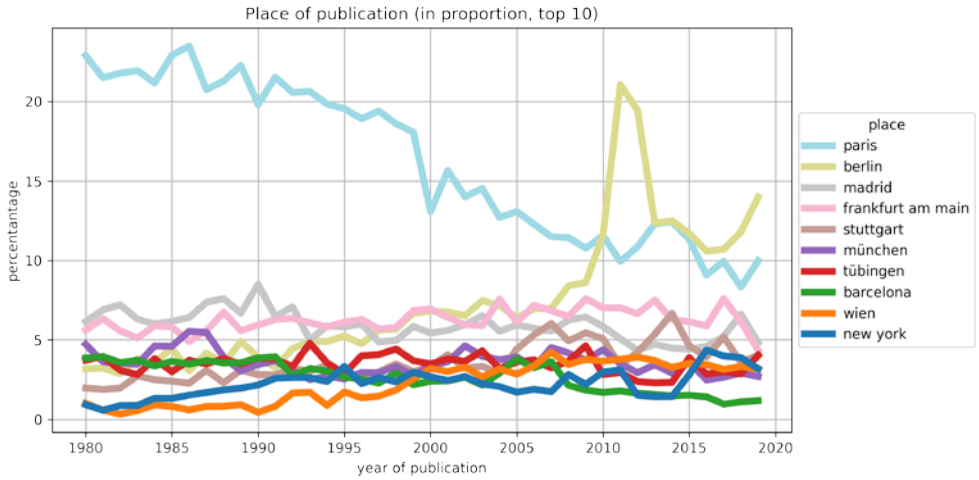
[15] However, the cataloging practice might have developed over the years, introducing in the catalog only the first place of publication for a period of time. This might influence the results, for example if a city tends to appear only as secondary place of publication.

[16] As for the sample of Figure 4, this random sample is created with the function *sample* of the Python library Pandas.

[17] <https://geobrowser.de.dariah.eu/?csv1=https://cdstar.de.dariah.eu/dariah/EAEA0-6966-72CA-0E10-0>.

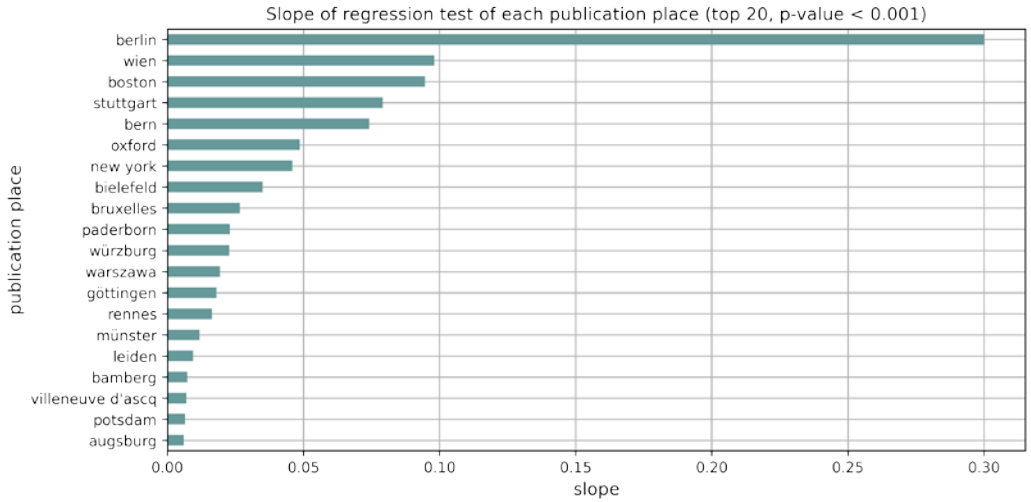
7 | Maps with the place of publications of 5,000 publications, centered in Europe, North and South America

Figure 8 shows the development of the top ten publication places in the analyzed period. The decrease of Paris as a place of publication for the Romance Studies is especially remarkable, with more than 20% of the records in the 1980s, but under 10% in some of the most recent years. The other remarkable development concerns Berlin, with only 3% in 1880 and peaking in 2011 of 21%. This peak will be further explained in the following sections about publishers and medium. Beyond these two places, the rest of the top ten places of publications vary between 8% and 1%. Five of them are in Germany, one is in Austria (Vienna) and only three are in Romance-speaking countries: Paris, Madrid and Barcelona.



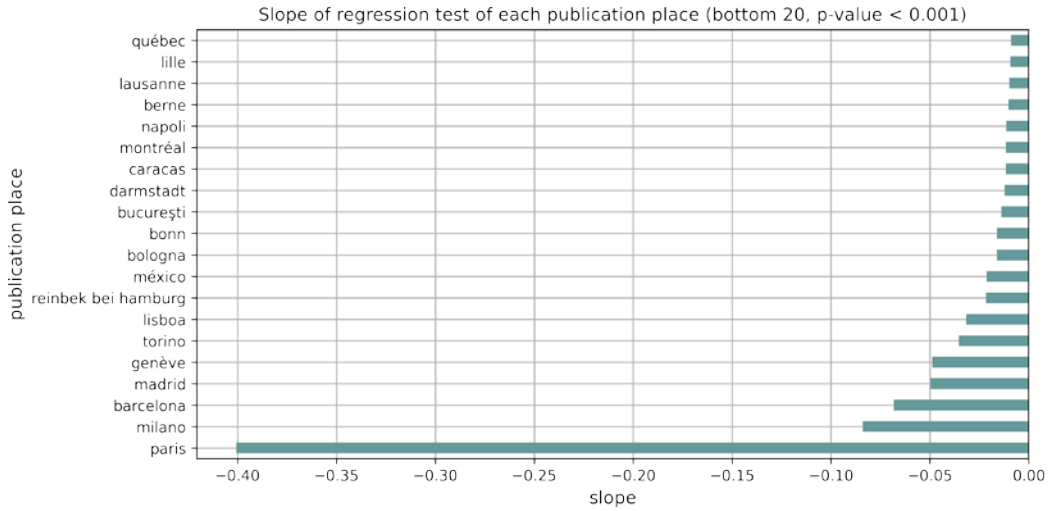
8 | Development of percentage of records of place of publication (top ten)

To observe the development of these and other places, I calculate the linear regression models for each place. Figure 9 shows the places with positive slope and a p-value under 0.001. Berlin shows the highest slope, with a predicted increase of 3% in the records of the catalog within 10 years. This is followed by several places with slopes over 0.03, all of them in the German or in the English-speaking area: Vienna, Boston, Stuttgart, Bern, Oxford, New York and Bielefeld. Besides Bruxelles, Rennes and Villeneuve d'Ascq, the rest of the places in this visualization are in Germany (with some exceptions for Poland and the Netherlands).



9 | Positive statistically significant slopes of linear regression models analyzing place of publication (top 20)

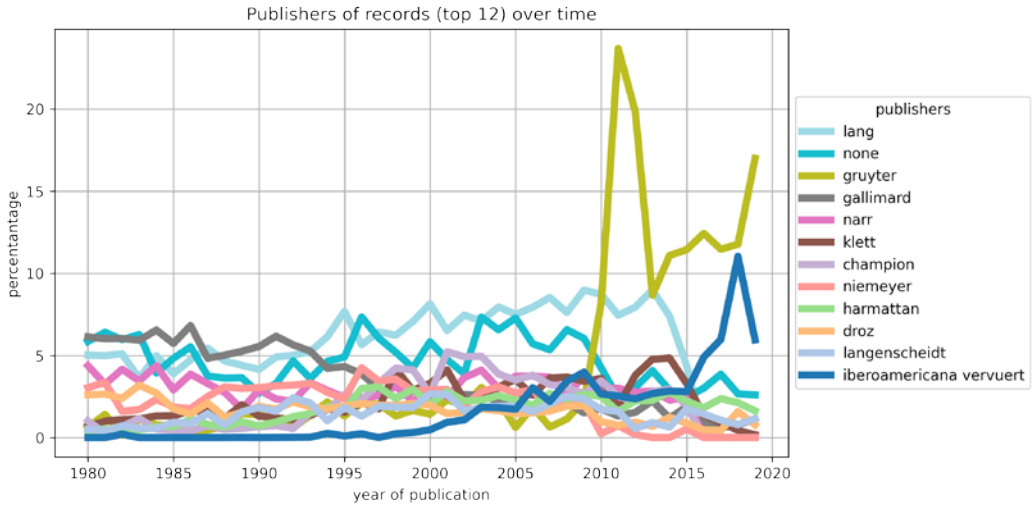
On the opposite side, Figure 10 shows the places with negative slopes. By far, the lowest slope is for Paris with almost a -0.4. The models predict 4.15% of the records being published in Paris for the year 2030. In general, the great majority of the places with negative slopes are from the different Romance-speaking countries: France, Italy, Spain, Portugal, the French-speaking areas of Switzerland and Canada, Mexico, Romania, Colombia. Besides, some places in Germany are in this figure, such as Bonn or Darmstadt.



10 | Negative statistically significant slopes of linear regression models analyzing place of publication (top 20)

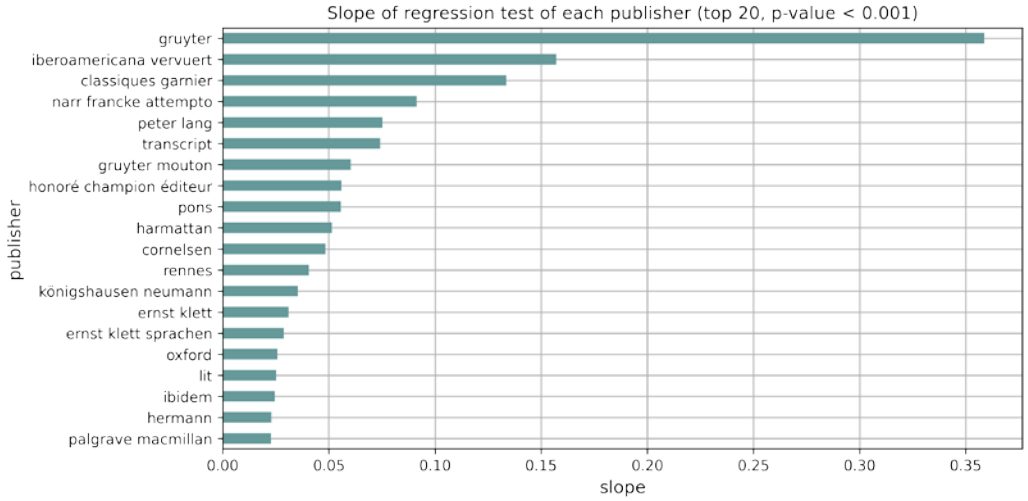
4.3 Publishers

Also strongly related to the language and place of publication is the development in the number of publications by publisher. As in the previous categories, the catalog has information for this field for a large majority of the records (98%) which is contained in the Pica+ field 033A (sub-field n). This field contains many abbreviations, such as *Verl., Univ., Ed., Éd., Pub., GmbH*, etc. I decided to delete all the references to the form of the institutions and a series of stop words in different languages (*&, of, de, von*, etc.). Besides, it needs to be considered that in many cases, the publishers change their name, sometimes because they are integrated into other publishing houses, sometimes because they create a specific imprint for some niches. For example, historical changes of the name can be observed for publishers such as Peter Lang, Narr, or Iberoamericana / Vervuert.



11 | Development of percentage of records of publisher (top twelve)

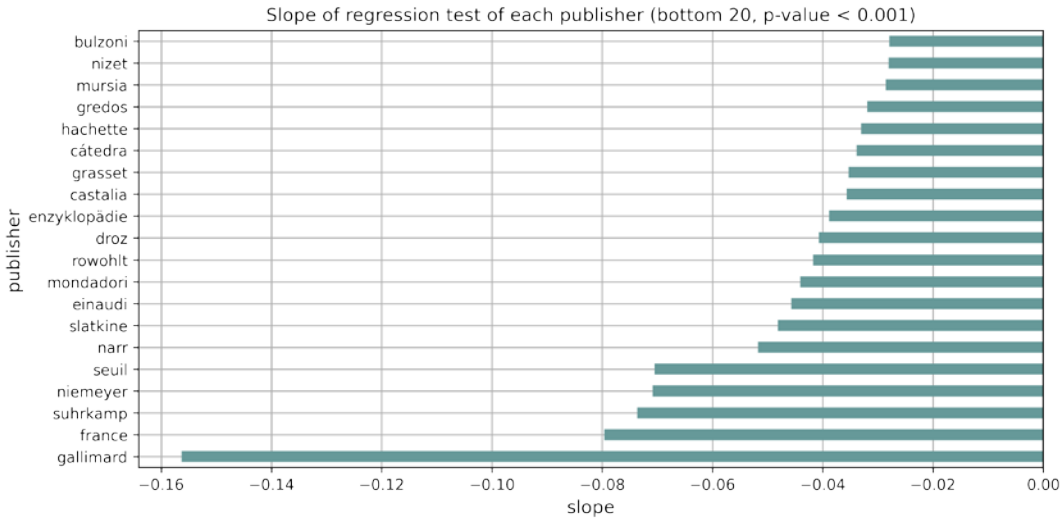
The historical development of the proportion of publications of the top twelve publishers is shown in Figure 11.[18] While in the 1980s the top publishers contributed between 6% and 1% to the catalog, after the 2000s publishers tended to increase or decrease their proportion of the publications notably. In general, a concentration of the publishing market can be observed in the sense that a few publishers (De Gruyter, Peter Lang, Iberoamericana / Vervuert) have been notably increasing their contribution to the catalog, while in general many smaller publishers reduce theirs. This will be tackled again in the next section about the development of the printed and digital publications.



12 | Positive statistically significant slopes of linear regression models analyzing publishers (top 20)

In Figure 12, the publishers with the highest slopes can be observed. De Gruyter obtains an exceptionally high slope over 0.35, followed by other publishers with slopes of 0.05 or greater, such as Iberoamericana / Vervuert, Classiques Garnier, Narr Francke Attempto, Peter Lang, transcript, Honoré Champion éditeur and Pons. Many of these publishers have their main location in the German-speaking area and therefore these results cuold be seen as an explanation of the results of Sections 4.2.

[18] The reasons for showing twelve cases and not ten is the development of Iberoamericana-Vervuert in the last five years. In the range between the top ten and 20 publishers, no other case has increased its presence similarly to this publisher.



13 | Negative statistically significant slopes of linear regression models analyzing publishers (top 20)

The opposite is shown in Figure 13, which lists the publishers with the lowest slopes. The publishers with the lowest slope are Gallimard (-0.15), followed by Presses Univ. de France (appearing only as *france* in the figure), Suhrkamp, Niemeyer, Seuil and Narr with slopes under -0.6. Some traditional publishers from France, Italy and Spain also appear in this figure, such as Droz, Mondadori, Castalia, Cátedra and Gredos.

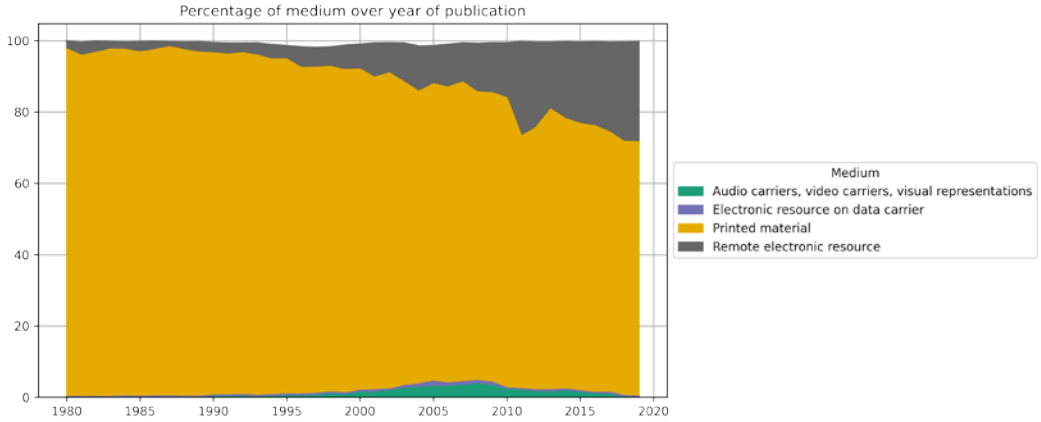
4.4 Medium

What is the current development of e-books in Romance Studies? This question is observed in this section, analyzing the data encoded in the Pica+ field 002@, subfield 0. Since this field is mandatory, it is included in 100% of the records of the catalog. Although the documentation of this field foresees many possible values, the majority of them are very rare in the catalog. In this analysis, I only focus on four categories:

- Printed material
- Remote electronic resource (e-books)
- Audio carriers, video carriers, visual representations (such as DVDs or Blu-rays)
- Electronic resource on data carrier (such as databases in CDs)

Of course, an increase in the number of e-books can be expected, and the main interest is how strong this is happening and if there are historical milestones in the past decades.

The chronological distribution of the medium can be observed in Figure 14. In general, printed material constitutes 88.16% of the total of records, while e-books represent 10.17%. The other two categories only represent around or less than 1% of the records. The contribution of the two other media is observable between the years 2000 and 2010, with a clear decline since then. Figure 14 shows that a section of the e-books are marked as published in the 1980s, when e-books were not purchased by libraries. That means that previous publications are being published in this format and therefore entering the catalog. However, after the year 2000 the share of e-books surpasses 10%. After 2010, e-books tend to represent over 20% of the publications.

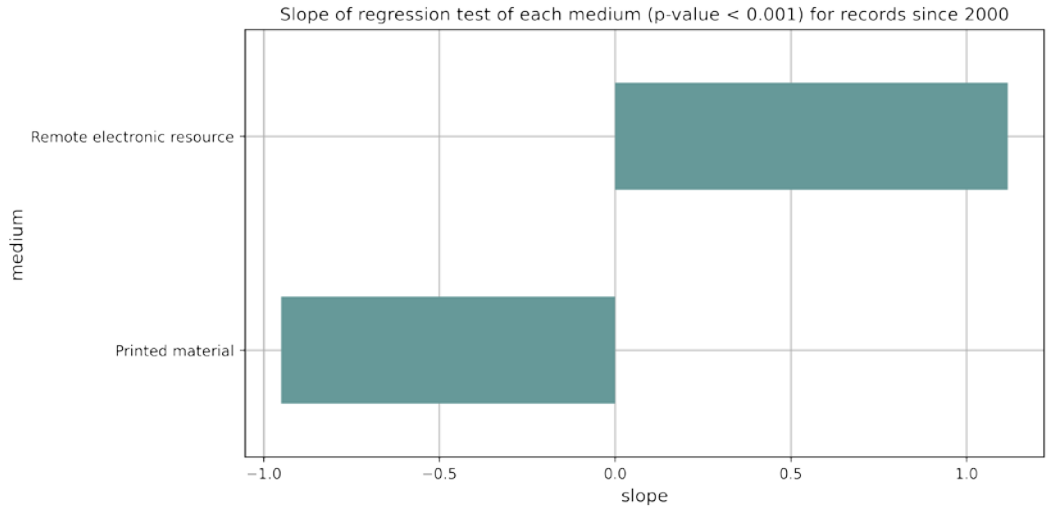


14 | Development of percentage of media

In 2011 a peak of e-books can be observed. In previous categories, parallel peaks were visible also in this year, such as an increase of publications from Berlin and from the publisher De Gruyter. This publisher based in Berlin made many previous publications available as e-books and launching them as published in 2011. The treatment of the e-books by De Gruyter could be also one of the factors that explains the observed success of the publisher in Figure 11. In contrast to many other publishers from the Humanities, De Gruyter tends to offer e-book licenses to its publications. Besides, the prices of these are similar to the printed version and their e-books can be downloaded in a single file by any user of the library. In other publishing houses, the user of the library can only download sections of the publication (for example a certain number of chapters or pages) and the prices tend to double or triple the price of the printed version. In the following section, I will give more details about the prices, also distinguishing between printed material and e-books. In any case, these results arise the question whether the digital paradigm is actually reinforcing the national publishing markets.

For the prediction of this field in future years, I consider two models: one for the entire period and another one only with the data for the last 20 years. Since e-books were rather insignificant for libraries before the year 2000, it is questionable to take this data for predicting the future. Figure 15 shows that the expected tendency is an annual increase of e-books of 1%, with the consequent similar decrease of the printed material. The linear model of four decades gives similar but

more conservative results, with negative and positive slopes for both media around 0.7% (further details in the Jupyter Notebooks).



15 | Positive and negative statistically significant slopes of linear regression models analyzing media

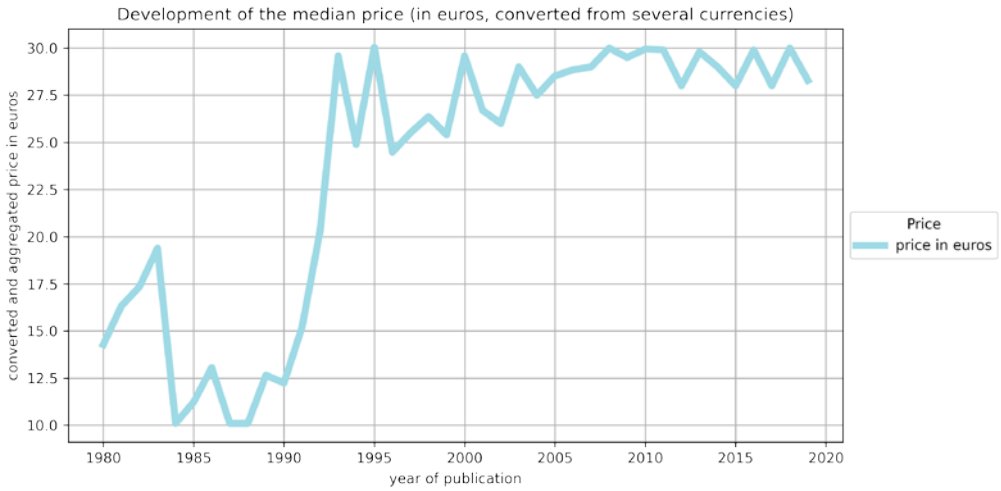
Although this shows a clear tendency, it does not describe a revolutionary change in the medium of the publications. Although it is unlikely that printed material will disappear completely, one could ask, with this current tendency, when a fully and entirely digital catalog for Romance Studies can be expected. The model predicts this taking place in the year 2080.

4.5 Price

In the previous section I have commented on the differences between the prices of printed material and e-books. In this section, I explore the development of the prices in the last decades and compare the price of both media. It is important to admit that the catalog is not the ideal source of data for prices since they are only kept as a comment to the ISBN in the Pica+ field 005A (sub-field f). In contrast to the previous categories, the coverage of this field in the catalog is much lower, with some information only for 27% of the records. The scarcity of the data is stronger during the 1980s with less than 10% of the records, while it increases up to 57% for the publications of the last years. Besides, this field brings further challenges:

- There is no homogeneous way to encode the price. Here are just some possibilities for the same price: 18.95 €; 18,95 €; 18.95 EUR; EUR 18.95 (DE); 18.95.
- The euro was introduced in several European countries, among others in Germany, where the analyzed libraries are located
- Some publications contain their prices in foreign currencies
- Some publications have information about the price in several currencies

All these problems required a special treatment of this field. The specific functions and regular expressions for each currency can be found in the Jupyter Notebooks and the Python code. For this analysis, I consider the prices assigned in euros, German marks, Swiss francs, British pounds, and US dollars. To compare the prices, I convert them into euros, following the average rate of the last years.[19] After these steps, I was able to obtain a price in euros for 19.86% of the records. I am aware that comparing absolute values from several currencies, countries and decades can be problematic. The goal is not to present an exhaustive analysis of this factor, but have a first glimpse of the development.



16 | Development of the median price

Figure 16 shows the historical development of the median price. As mentioned before, the dataset contains little data about the price of publications during the 1980s, therefore I argue that the apparent large increase of the price at the beginning of the 1990s is just an artifact of the scarcity of the data. Figure 16 shows a slow increase of the price since the 1990s, with a current median price being slightly under 30 euros.

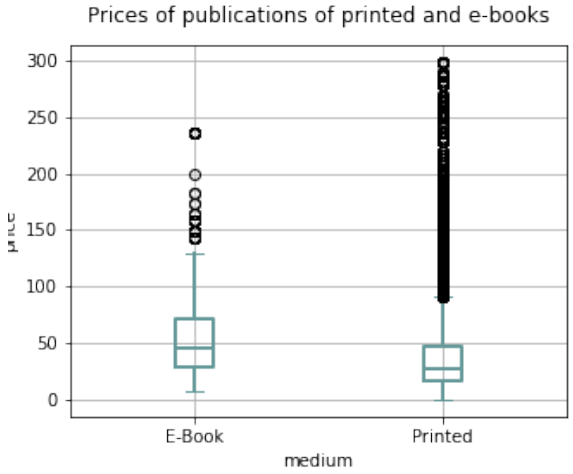
The linear regression model outputs a positive slope of 0.15 (p-value < 0.001), which corresponds to an annual increase of 15 cents of euro and a median price of 31.70 € for the year 2030.

When the current currencies are analyzed separately, it can be observed that while euros, Swiss franks and US dollars are stable, the prices in British pounds are increasing notably. While the median price in British pounds in 1995 was 39.60, it increased up to 76.56 for the year 2015.[20] A linear regression model gives a slope of 1.51 (p-value < 0.001), predicting for the year 2030 a price of 93.79 € if publications are originally priced in pounds.

[19] Which are 0.85 for US dollars, 1.32 for British pounds, 0.76 for Swiss franks and 0.51 for German marks. Further details in Jupyter Notebook and <https://www.ofx.com/en-au/forex-news/historical-exchange-rates/yearly-average-rates/>.

[20] Although this data represents the prices in pounds, the actual values here are expressed in euros.

Finally, I compare the prices of printed material and e-books. Figure 17 shows the distribution of the prices through box plots for both media in the dataset. While the median price for printed publications is 28 €, it is 46.32 € for e-books. A Welch's t-test throws a p-value lower than 0.001, meaning that, at least for this dataset, it can be said that e-books are statistically more expensive than printed versions.



17 | Comparison of prices printed and e-books

Of course, it needs to be considered that both media are highly imbalanced in the dataset. While I was able to obtain the price of 22.29% of the printed versions, this was only the case for 0.72% of the e-books. Although the results confirm the experience in the daily work of the library, further research and discussion about the price of e-books is needed.

Even though the development of the prices reflects only a subtle increase (except for publications from the United Kingdom), the higher prices of e-books need to be properly addressed. The budgets for the purchase of literature in German libraries have been frozen in the last decade. Digitization in the library and remote access to research publications have become central in the last decade and reinforced by the COVID-pandemic since 2019 (Bieselin 2005; M. Ernst 2021). An increase in the number of available e-books can only be possible with an increase in libraries' budgets, either for purchasing e-books or for the support of Open Access publications.

4.6 Keywords

The last category I explore in this analysis is related to the keywords (*Schlagwörter* in German). Keywords are controlled vocabularies composed by terms from one natural language, used to describe the content of a document in the catalog as accurately as possible (Gantert 2016, 198–99). Table 3 contains ten randomly selected examples from publications with their title and the keywords that can be found in the catalog.

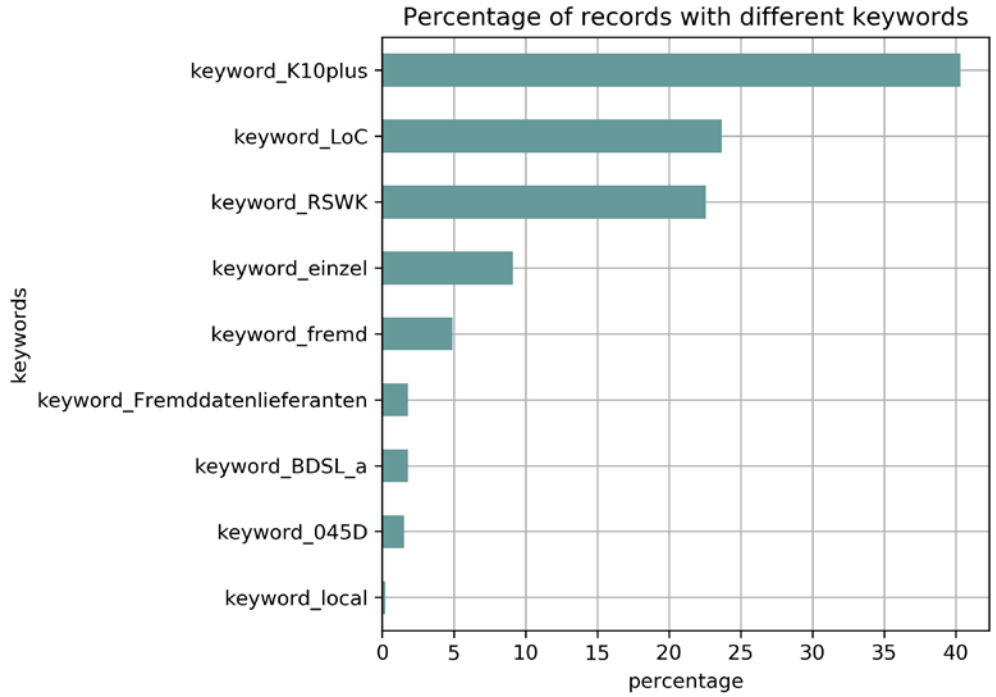
Title	Keywords
Edizione nazionale delle opere	Alberti
Méthodes de français à l'école secondaire en Suisse alémanique	Deutsche Schweiz|Personalpronomen|Fehler-analyse|Französischunterricht|Gymnasium|Pronominaladverb
Alexis de Tocqueville	Tocqueville
Lesage, écrivain	Le Sage
Von der "novela social" zur "nueva novela española"	Goytisolo|Roman|Goytisolo, Luis|1935-|Criticism and interpretation|Social problems in literature|Spanish fiction|20th century|History and criticism
Zwischen weißer und schwarzer Schrift	Jabès|Literaturtheorie|Schreiben
Diccionario fonético descriptivo de la lengua española	Spanish language|Spanisch|Aussprache|Deskriptive Phonetik|OBV
Elio Vittorini und die moderne europäische Erzählkunst (1926 - 1939)	Vittorini|Vittorini, Elio|1908-1966|Criticism and interpretation
Michel Tournier et le détournement de l'autobiographie. Suivi d'un entretien avec Michel Tournier	Tournier, Michel|Criticism and interpretation|Self (Philosophy) in literature|Tournier|Auto-biografische Literatur
Der König im Kontext	Calderón de la Barca|Comedia|König|Herrschaft|Kontingenz|Geschichtsbild|Calderón de la Barca, Pedro|Kings and rulers in literature

Table 3 | Examples of ten publications and their assigned keywords in the catalog

In contrast to the previous categories, these keywords allow the researcher to gain insight into the actual content of the publication. For example, until now I have only explored which language was used for the text of the publication. Perhaps the decline in French can only be traced as publication language and researchers are still equally interested in French language, literature or culture, while publishing their results in other languages such as German or English.

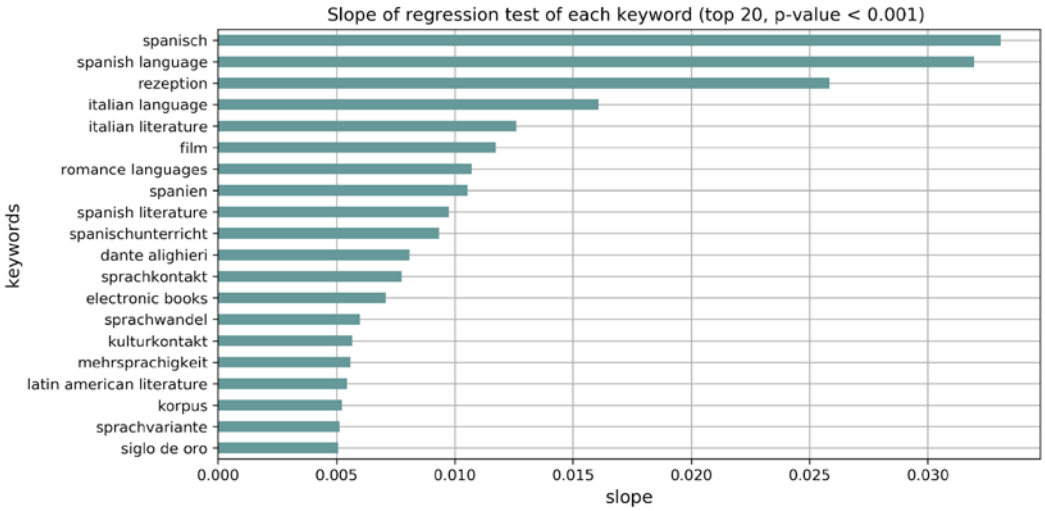
Although a similar analysis would have been possible by using the classes of the classification systems, I decided to reject this for two reasons. First, because the classes of classification systems were already applied to create the dataset. It could be argued that an analysis of the classes would be circular to a certain degree. Second and most importantly, keywords express more specific concepts than the classes of classification systems. Some examples of this can be seen in Table 3, such as “Kings and rulers in literature”, “Deskriptive Phonetik”, and “Fehleranalyse”.

The catalog contains different formalizations of the keywords depending on the source, the language used to express the keywords (English or German) or the controlled vocabulary applied. Figure 18 shows the coverage of several fields from the catalog containing this information. The fields with the highest coverage are the Pica+ fields 044K (in the Figure as keyword_k10plus), 44A (in the Figure as keyword_LoC) and 41A (keyword_RSWK). Each of these three fields contains data for more than 20%, while the rest covers fewer than 10% of the cases.



18 |Percentage of records with different keywords

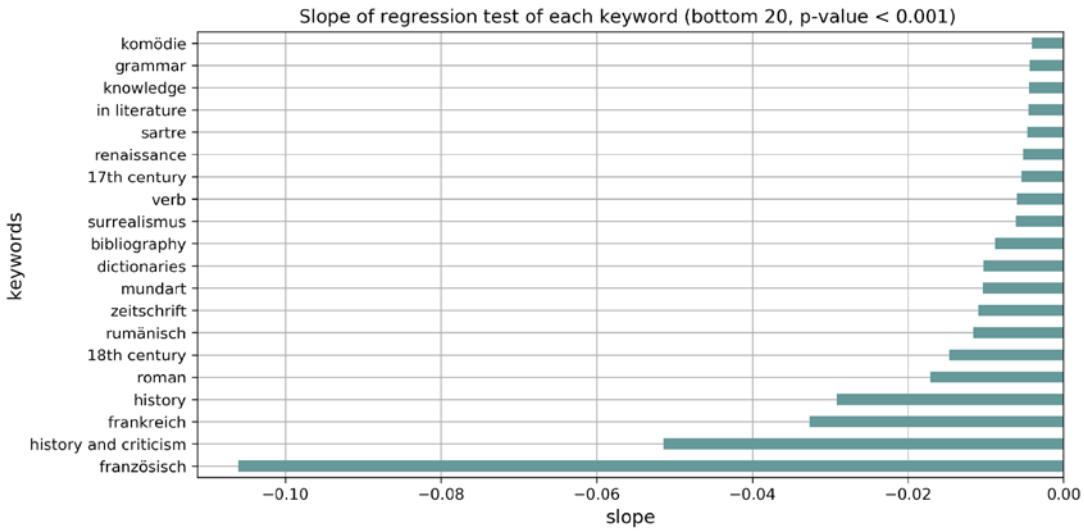
For this reason, I decide to work only with these three categories. While the LoC keywords are expressed in English, the other two are in German. Together, the three fields cover 56.29% of the dataset for Romance Studies publications.



19 | Positive statistically significant slopes of linear regression models analyzing keywords (top 20)

The twenty keywords with the highest slopes with statistical significance can be observed in Figure 19. Many of these keywords relate to the Spanish language, literature and culture, with two keywords with the highest slopes followed by others (*Spanien, Spanish literature, Spanischunterricht, Latin American literature, siglo de oro*). In any case, the slope seen for Spanish in 4.1 was much higher than the slopes in Figure 19. This could be explained with the fact that the language is

present in several keywords, keeping lower values. However, it could be pointed out that Spanish is increasing its role of publication language beyond the general interest of Spanish language and literature. In the case of Italian, in contrast to the stable situation seen in Section 4.1, the keywords show an increasing interest in Italian language and literature. This is reinforced by the presence of the keyword Dante Alighieri in these results. However, both keywords are expressed with the LoC vocabulary, which can point to an increasing interest in the English-speaking countries, while this could not be not the case in the German-speaking area. Numerous keywords associated with linguistics relate to contact and changes across and within languages (*Romance languages, Sprachkontakt, Sprachwandel, Kulturkontakt, Mehrsprachigkeit, Sprachvariante*). On its side, literary studies show positive trends relating reception and film. Finally, two further keywords can be associated with the new digital paradigm: *electronic books* and *korpus*.



20 | Negative statistically significant slopes of linear regression models analyzing keywords (top 20)

In contrast, Figure 20 shows the keywords with the lowest slopes. In coherence with the results seen in the previous sections, French appears in the keywords with the two lowest slopes, reinforced by the presence of a keyword for Sartre. In contrast to the case of Italian in Figure 19, the keywords relating to French are only expressed in German, which could reflect that this trend can be only observed in the German-speaking area. Romanian is the only other language on these bottom 20 keywords, which is the observed tendency in the analyzed catalogs, but it could have been notably different if the records from the BVB had been part of the dataset. Several keywords from Figure 20 are related to publications types, such as journals, dictionaries and bibliographies. Besides, two classical literary genres show strong decreasing tendencies: novel and comedy.[21] Literary Studies seem less

[21] However, this trend could be traced rather to a change of the librarian practice, assigning in the past years more specific keywords.

interested in previous periods (Renaissance, 17th and 18th century) and surrealism, while this is the case for verb and grammar in Linguistics.

5. Conclusions

In this article, I have used library records of publications in order to describe the development of the Romance Studies in the past decades and make predictions about the next decades. Although the predictions for the next years are interesting, all of them are based on the simplification that the tendencies of the past will remain, which might no be the case. This simplification is a general limitation of current Machine Learning approaches, which are based on the premise that new cases can be predicted following what was observed in the past. The analyzed dataset coming from the hebis and K10plus library consortia show different trends relating to the different analyzed categories.

The presence of German language publications and of publishers based in the German-speaking area is increasing in the Romance Studies. There is a decline in the importance of French which can be observed in the language of publication, place of publication, publishers and research topic. Spanish shows a clear increase both as publication language and topic of research, however this does not translate into an increase of publications coming from Spain. Italian shows a rather stable situation, with certain positive trends as a research subject.

E-books have become an important part of the publications of the Romance Studies, the change to the new digital paradigm will take decades in the current development. The results suggest that the new digital paradigm could be reinforcing the national publishing market, since German libraries could prefer acquiring e-books from German publishers. Besides, this study shows that e-books are notably more expensive, which needs to be addressed by an increase in the libraries' budgets.

In Linguistics, grammar and syntax seem to be in decline, while topics relating to language contact, variation and multilingualism are increasing. For Literature Studies, previous periods (1500-1800) and classical genres and publications have made room for new research subjects such as films and cultural contact.

Although these results are representative for a large section of the German territory, this analysis does not exhaust further possibilities. In future studies, the dataset could be expanded to more sources, countries, periods and other publication types (specially chapters and journal articles). However, the decline of data quality needs to be considered when combining several sources. Moreover, further possibilities of analysis are worth exploring, such as the generation of keywords (both from supervised or unsupervised tasks) or the annotation of the titles through lexical resources in several languages such as WordNet.

The results of this study can be used by researchers, Romance Studies departments and libraries to reflect their own decision. When researchers decide the language of the publication, the publisher of their next anthology for a section of a conference, or the topic of a new professorship, it can be decided to reinforce or

not the current trend. The same is true for libraries, which influence the reception of a publication by their purchase decisions.

In this article, I have shown the potential of the bibliographic data science analysis applied to the last decades of a specific discipline. These research studies are only possible thanks to the valuable work and experience of the professionals in the libraries, who daily curate the data in the catalog. If we believe that data is especially valuable in the new digital paradigm, we also need to acknowledge and promote the role of the professionals curating the data on a day-to-day basis. In any case, this kind of analysis requires the combination of knowledge relating to libraries, the specifics of the discipline, and the use of digital and statistical methods. For this reason, I argue for closer collaboration between libraries and researchers.

References

BECKER, Lidia, Julia Kuhn, Christina Ossenkop, Anja Overbeck, Claudia Polzin-Haumann, and Elton Prifti, eds. 2020. *Fachbewusstsein der Romanistik: Romanistisches Kolloquium XXXII*. Tübinger Beiträge zur Linguistik 578. Tübingen: Narr Francke Attempto.
<http://elibrary.narr.digital/book/99.125005/9783823394181>.

BIESELIN, Tanja-Barbara. 2005. 'Im Kampf gegen Etat-Kürzungen, Schließungen und morsches Image. Guerilla-Marketing für Bibliotheken' 29 (3): 361–75.
<https://doi.org/10.1515/BFUP.2005.361>.

BURR, Isolde. 2008. 'Romanische Sprachen in internationalen Organisationen'. In *Romanische Sprachgeschichte: ein internationales Handbuch zur Geschichte der romanischen Sprachen; Histoire linguistique de la Romania*, edited by Gerhard Ernst, Gerold Ungeheuer, and Armin Burkhardt, 3:3339–54. Handbücher zur Sprach- und Kommunikationswissenschaft; Handbooks of linguistics and communication science 23. Berlin: De Gruyter.
<http://www.degruyter.com/doi/book/10.1515/9783110211412.3>.

CHOWDHURY, G. G., Paul F. Burton, David McMenemy, and Alan Poulter. 2008. *Librarianship: An Introduction*. London: Facet Publishing.

CONSTANTINESCU, Ioan. 2002. 'Deutschprachige Romanistik – Eine Wissenschaft Mit Zukunft'. In *Deutschsprachige Romanistik - Für Wen?*, edited by Maria Lieber and Harald Wentzlaff-Eggebert, 41–46. Heidelberg: Synchron, Wiss.-Verl. der Autoren.

DOMBROWSKI, Quinn, Tassie Gniady, and David Kloster. 2019. 'Introduction to Jupyter Notebooks'. *The Programming Historian* 8.
<https://programminghistorian.org/en/lessons/jupyter-notebooks>.

EHRLICHER, Hanno, and Jörg Lehmann. 2021. 'La recolección de datos como laboratorio epistemológico. Algunas reflexiones acerca del entorno virtual de investigación Revistas Culturales 2.0'. *Signa: Revista de la Asociación Española de Semiótica* 30 (0): 59–81.
<https://doi.org/10.5944/signa.vol30.2021.29298>.

ERNST, Michael. 2021. 'Ein Trend und seine Folgen'. *Verfassungsblog* (blog). 17 June 2021.
<https://verfassungsblog.de/ein-trend-und-seine-folgen/>.

EVANS, Michael S. 2014. 'A Computational Approach to Qualitative Analysis in Large Textual Datasets'. *PLoS ONE* 9 (2).
<https://doi.org/10.1371/journal.pone.0087908>.

GANTERT, Klaus. 2016. *Bibliothekarisches Grundwissen*. *Bibliothekarisches*

Grundwissen. Berlin, Boston: De Gruyter Saur.
<https://www.degruyter.com/view/title/302969>.

GITTEL, Benjamin. 2021. 'An Institutional Perspective on Genres: Generic Subtitles in German Literature from 1500-2020'. *Journal of Cultural Analytics*, April.
<https://doi.org/10.22148/001c.22086>.

GONZÁLEZ, Juana María. 2021. 'Análisis cuantitativo de la revista Índice Literario (1932-1936)'. *Artnodes* 27.
<https://doi.org/10.7238/a.v0i27.374373>.

GUMBRECHT, Hans Ulrich. 2002. *Vom Leben Und Sterben Der Großen Romanisten: Karl Vossler, Ernst Robert Curtius, Leo Spitzer, Erich Auerbach, Werner Krauss*. Edition Akzente. München: Carl Hanser Verlag.

HAARMANN, Harald. 2008. 'Romanische Sprachen als Publikationssprachen der Wissenschaft: 19. und 20. Jahrhundert'. In *Romanische Sprachgeschichte: ein internationales Handbuch zur Geschichte der romanischen Sprachen; Histoire linguistique de la Romania*, edited by Gerhard Ernst, Gerold Ungeheuer, and Armin Burkhardt, 3:3359–70. Handbücher zur Sprach- und Kommunikationswissenschaft; Handbooks of linguistics and communication science 23. Berlin: De Gruyter.
<http://www.degruyter.com/doi/book/10.1515/9783110211412.3>.

HENNY-KRAHMER, Ulrike. 2017. 'Bib-ACMé: Bibliografía digital de novelas argentinas, cubanas y mexicanas (1810-1930).' In *Sociedades, políticas, saberes.*, edited by Nuria Rodríguez Ortega, 99–104. Málaga: Universidad de Málaga.
<http://hdh2017.es/wp-content/uploads/2017/10/Actas-HDH2017.pdf>.

HERRMANN, J. Berenike, Giulia Grisot, Susanne Gubser, and Elias Kreyenbühl. 2021. 'Ein Großer Berg Daten? Zur Bibliothekswissenschaftlichen Dimension Des Korpusliteraturwissenschaftlichen Digital Humanities-Projekts „High Mountains – Deutschschweizer Erzählliteratur 1880–1930"'. *027.7 Zeitschrift Für Bibliothekskultur / Journal for Library Culture* 8 (1).
<https://doi.org/10.21428/1bfadeb6.6e2feff6>.

HOLTUS, Günter, and Fernando Sánchez Miret. 2008. *'Romanitas', Filología Románica, Romanística*. Beihefte Zur Zeitschrift Für Romanische Philologie. - Berlin: De Gruyter, 1905- ; ZDB-ID: 200077-5 347. Tübingen: Niemeyer.

JANNIDIS, Fotis, Leonard Konle, & Peter Leinen. 2019. 'Makroanalytische Untersuchung von Heftromanen'. In *Digital Humanities: Multimedial & Multimodal*, 167–73. Mainz-Frankfurt: Dhd.
<https://zenodo.org/record/2600812#.XLg1bUNS9hE>.

KALKHOFF, Alexander M. 2010. *Romanische Philologie Im 19. und Frühen 20. Jahrhundert: Institutionsgeschichtliche Perspektiven*. Romanica Monacensia. - Tübingen : Narr Francke Attempto, 2014- ; ZDB-ID: 3020712-5 78. Tübingen: Narr.
<http://elibrary.narr.digital/book/99.125005/9783823375043>.

KRAMER, Johannes. 2002. 'Deutsch Als Publikationssprache Und Vielsprachige Romanistik — Ein Ärgernis in Der Internationalen Wissenschafts-landschaft?' In *Deutschsprachige Romanistik - Für Wen?*, edited by Maria Lieber and Harald Wentzlaff-Eggebert, 13–25. Heidelberg: Synchron, Wiss.-Verl. der Autoren.

KRAMER, Johannes. 2008. 'Romanische Sprachen als Publikationssprachen der Wissenschaft bis zum 18. Jahrhundert'. In *Romanische Sprachgeschichte: ein internationales Handbuch zur Geschichte der romanischen Sprachen; Histoire linguistique de la Romania*, edited by Gerhard Ernst, Gerold Ungeheuer, and Armin Burkhardt, 3:3354–59. Handbücher zur Sprach- und Kommunikationswissenschaft; Handbooks of linguistics and

communication science 23. Berlin: De Gruyter.
<http://www.degruyter.com/doi/book/10.1515/9783110211412.3>.

KRAMER, Johannes. 200. 'Selbstdarstellungen der Romanistik während der Gründungsphase, um 1900 und nach 1988'. In *Fachbewusstsein der Romanistik: Romanistisches Kolloquium XXXII*, edited by Lidia Becker, Julia Kuhn, Christina Ossenkop, Anja Overbeck, Claudia Polzin-Haumann, and Elton Prifti, 1. Tübinger Beiträge zur Linguistik 578. Tübingen: Narr Francke Attempto.
<http://elibrary.narr.digital/book/99.125005/9783823394181>.

KREFELD, Thomas. 2020. 'FAIRness weist den Weg – von der Romanischen Philologie in die Digital Romance Humanities'. In *Fachbewusstsein der Romanistik: Romanistisches Kolloquium XXXII*, edited by Lidia Becker, Julia Kuhn, Christina Ossenkop, Anja Overbeck, Claudia Polzin-Haumann, and Elton Prifti, 291–310. Tübinger Beiträge zur Linguistik 578. Tübingen: Narr Francke Attempto.
<http://elibrary.narr.digital/book/99.125005/9783823394181>.

KREMNITZ, Georg. 2016. *Geschichte Der Romanischen Sprachwissenschaft: Unter Besonderer Berücksichtigung Der Entwicklung Der Zahl Der Romanischen Sprachen*. Bachelor Master Studies. - Wien : Praesens, 2014- ; ZDB-ID: 2806920-1 8. Wien: Praesens Verlag.

LIEB, Claudia, and Christoph Strosetzki, eds. 2013. *Philologie Als Literatur- Und Rechtswissenschaft: Germanistik Und Romanistik 1730 - 1870*. Euphorion. Beihefte Zum Euphorion. - Heidelberg : Winter, 1964- ; ZDB-ID: 503579-X 67. Heidelberg: Winter.

LIEBER, Maria, and Harald Wentzlaff-Eggebert, eds. 2002. *Deutschsprachige Romanistik - Für Wen?* Heidelberg: Synchron, Wiss.-Verl. der Autoren.

MARYL, Maciej, and Piotr Wciślik. 2016. 'Remediations of Polish Literary Bibliography: Towards a Lossless and Sustainable Retro-Conversion Model for Bibliographical Data'. In *Digital Identities: the Past and the Future*.
<https://dh-abstracts.library.cmu.edu/works/2767>.

MONJOUR, Alf. 2020. 'Romanistik nach Bologna? Zum Nachdenken über zukünftige Positionen der romanistischen Sprach- und Kulturwissenschaften'. In *Fachbewusstsein der Romanistik: Romanistisches Kolloquium XXXII*, edited by Lidia Becker, Julia Kuhn, Christina Ossenkop, Anja Overbeck, Claudia Polzin-Haumann, and Elton Prifti, 195–203. Tübinger Beiträge zur Linguistik 578. Tübingen: Narr Francke Attempto.
<http://elibrary.narr.digital/book/99.125005/9783823394181>.

MÜLLER, Andreas C., and Sarah Guido. 2016. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. Beijing, Boston: O'Reilly.

NITSCHACK, Horst. 2002. 'Deutschsprachige Romanistik — Für Wen? Dreifache Dialog'. In *Deutschsprachige Romanistik - Für Wen?*, edited by Maria Lieber and Harald Wentzlaff-Eggebert, 7–11. Heidelberg: Synchron, Wiss.-Verl. der Autoren.

RICHERT, Gertrud. 1913. *Die Anfänge Der Romanischen Philologie Und Die Deutsche Romantik*.

SCHROTT, Angela. 2003. 'Romanistische Sprachgeschichtsforschung: Zeitschriften'. In *Romanische Sprachgeschichte: ein internationales Handbuch zur Geschichte der romanischen Sprachen; Histoire linguistique de la Romania*, edited by Gerhard Ernst, Gerold Ungeheuer, and Armin Burkhardt, 1:421–27. Handbücher zur Sprach- und Kommunikationswissenschaft; Handbooks of linguistics and communication science 23. Berlin: De Gruyter.
<http://www.degruyter.com/doi/book/10.1515/9783110146943.1>.

TOLONEN, Mikko, Mark J. Hill, Ali Ijaz, Ville Vaara, and Leo Lahti. 2020. 'Data-Driven Analysis of Canonical Works in Early Modern Britain.' In *15th*

Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Ottawa, Canada, July 20-25, 2020, Conference Abstracts.
<https://dh2020.adho.org/wp-content/uploads/2020/07/555_DatadrivenanalysisofcanonicalworksinearlymodernBritain.html>.

TOLONEN, Mikko, Jani Marjanen, Hege Roivainen, and Leo Lahti. 2019. 'Scaling Up Bibliographic Data Science.' In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, Copenhagen, Denmark, March 5-8, 2019.*, 450–56.
<http://ceur-ws.org/Vol-2364/41_paper.pdf>.

VAARA, Ville, Ali Ijaz, Iiro Tiihonen, Antti Kanner, Tanja Säily, and Leo Lahti. 2019. 'The Emerging Paradigm of Bibliographic Data Science'. In T*he Index of Digital Humanities Conferecnces*.
<https://dh-abstracts.library.cmu.edu/works/9931>.

VANDERPLAS, Jake. 2016. *Python Data Science Handbook: Essential Tools for Working with Data*. First edition. Beijing, Boston: O'Reilly.

VOß, Jakob. 2020. *Einführung in die Verarbeitung von PICA-Daten*. Göttingen: GBV (VZG).
<https://pro4bib.github.io/pica/>.

WANDRUSZKA, Mario. 1988. 'Deutsche Romanistik: Kritische Bilanz Und Perspektive'. In *Ein 'Unmögliches Fach': Bilanz Und Perspektiven Der Romanistik*, edited by Fritz Nies and Reinhold R. Grimm, 213. Tübingen: Narr.

WIESENMÜLLER, Heidrun, and Silke Horny. 2017. *Basiswissen RDA: Eine Einführung für deutschsprachige Anwender*. *Basiswissen RDA*. De Gruyter Saur.

Wolf, Johanna. 2012. *Kontinuität Und Wandel Der Philologien: Text-archäologische Studien Zur Entstehung Der Romanischen Philologie Im 19. Jahrhundert*. Romanica Monacensia. - Tübingen : Narr, 1968- ; ZDB-ID: 404830-1 80. Tübingen: Narr Francke Attempto.

Abstract

What have been the main trends in Romance Studies in the last decades? What can be expected for the next decade? These are the two main research questions of this article. To answer them, a large dataset of over one million publications of research in Romance Studies has been extracted from German library catalogs. This dataset is analyzed through descriptive statistics and linear regression in order to predict the development in future years. Several fields of the respective catalogs are analyzed, such as the language and place of publication, publishers, e-book vs. printed versions, price and subjects.

Zusammenfassung

Was waren die wichtigsten Trends in der Romanistik in den letzten Jahrzehnten? Was ist für das nächste Jahrzehnt zu erwarten? Dies sind die beiden Hauptforschungsfragen des vorliegenden Artikels. Zur Beantwortung dieser Fragen wurde ein großer Datensatz von über einer Million romanistischer Forschungspublikationen aus deutschen Bibliothekskatalogen extrahiert. Dieser Datensatz wird mittels deskriptiver Statistik und linearer Regression analysiert, um die Entwicklung in den kommenden Jahren vorherzusagen. Dabei werden verschiedene Felder der jeweiligen Kataloge analysiert, wie z.B. Sprache und Ort der Veröffentlichung, Verlage, E-Book und gedruckte Version, Preis und Themen.

2022, n°9
pp. 153-162
doi: https://doi.org/10.15460/apropos.9.1903

Christoph Müller

La transformación digital en la investigación y en las bibliotecas especializadas en América Latina y el Caribe

Retrodigitalización, objetos de origen digital, datos de investigación

Christoph Müller
es vicedirector de la biblioteca del Instituto Ibero-Americano de Berlín, director del departamento Biblioteca Digital.
mueller@iai.spk-berlin.de

Palabras clave

Información digital – bibliotecas – digitalización – datos de investigación – acceso abierto

Retos de la transformación digital

La transformación digital de los métodos de investigación, de sus resultados, así como de la comunicación científica, plantea múltiples retos a los actores del ámbito científico y de las bibliotecas. Constantemente surgen formas de trabajo nuevas, que aportan también constantemente nuevos tipos de datos y producen requisitos cada vez más específicos sobre la información, las fuentes y los formatos de datos. Por un lado, estos deben ser generados y utilizados por los investigadores y, por otro lado, los bibliotecarios[1] deben ser capaces de comprenderlos de tal manera que las bibliotecas puedan seguir cumpliendo su papel de proveedores centrales de información de manera precisa y sostenible.

Mientras que la transformación digital en las bibliotecas al principio se ha reflejado principalmente en la conversión de la forma análoga a la forma electrónica de los catálogos y de las fuentes de información, sucesivamente comenzaron a situarse en el centro del trabajo de las bibliotecas la recopilación de publicaciones electrónicas y el apoyo a la implementación de infraestructuras de investigación y publicación digitales. Recientemente, las bibliotecas se han dedicado cada vez más a crear servicios electrónicos para recopilar, asegurar y hacer accesibles los datos digitales de investigación.

En este proceso, ha quedado claro que la transformación digital tanto en la ciencia y la investigación como en las bibliotecas no solo significa la creación y el suministro

[1] Para facilitar la lectura en este artículo se utiliza siempre la forma masculina refiriéndose a todos los géneros.

de información en formato electrónico o el desarrollo de métodos de trabajo digitales. Cada vez es más evidente la importancia que tiene en este proceso el intercambio entre investigadores y bibliotecarios, que juntos coordinan los distintos requisitos científicos y los posibles servicios de información (Stille et al. 2021).

En el contexto de la investigación especializada en América Latina, se plantea ahora la cuestión sobre qué servicios de información digital están ya disponibles tanto en América Latina como en otras regiones del mundo, qué servicios quedan por implementar y cómo puede organizarse el intercambio sobre las necesidades de la ciencia entre investigadores y bibliotecarios. Los conceptos de acceso y ciencia abiertos son especialmente importantes.

Servicios de información digital relacionados con América Latina

Para la investigación relacionada con América Latina, existen numerosos servicios, especialmente en la región, que proporcionan información y literatura científica en su mayoría en acceso abierto. La plataforma de publicación de revistas científicas *SciELO* desarrollada y mantenida en Brasil por una cooperación de la *Fundação de Amparo à Pesquisa do Estado de São Paulo* (FAPESP) con el *Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde* (BIREME/OPS/OMS) – pero que también es importante para toda América Latina – ofrece a los editores latinoamericanos la oportunidad de hacer que sus revistas sean accesibles al público en texto completo sin tener una infraestructura técnica propia. *SciELO* no solo proporciona la tecnología necesaria, sino que también evalúa la calidad editorial de las revistas sobre la base de criterios definidos.[2]

Redalyc sigue un principio similar. Se trata de un sistema de información de revistas operado por la *Universidad Autónoma del Estado de México*, que también comprueba la calidad editorial de las revistas científicas de y sobre América Latina sobre la base de criterios fijos y, tras haber satisfecho estos, las hace accesibles en línea en texto completo.[3]

Tanto *SciELO* como *Redalyc* ofrecen también soportes específicos para monografías. En el caso de *SciELO* existe *SciELO libros*. *Redalyc* colabora con la plataforma de libros electrónicos *AmeliCA*, operada por un grupo internacional de universidades en su mayoría latinoamericanas.[4]

Paralelamente, existen también las plataformas de publicación de *CLACSO* y *CEPAL*. El *Consejo Latinoamericano de Ciencias Sociales* (CLACSO) ofrece un repositorio digital en el que se publican a texto completo libros, artículos, documentos de trabajo y presentaciones de las instituciones asociadas participantes. La *Comisión Económica para América Latina y el Caribe* (CEPAL) también pone a disposición fuentes de información de diversa índole en texto completo. Aunque las

[2] <https://www.scielo.org/>, 29.01.2022
[3] <https://www.redalyc.org/>, 29.01.2022
[4] <http://books.scielo.org/>, 29.01.2022, <http://amelica.org/>, 29.01.2022.

publicaciones de estas dos instituciones son principalmente de ciencias sociales, también existen numerosas fuentes de información relevantes para la investigación de los estudios románicos con respecto a América Latina y el Caribe.[5]

LA Referencia (*La Red Federada de Repositorios Institucionales de Publicaciones Científicas*) – una cooperación de repositorios de acceso abierto latinoamericanos apoyada administrativamente por la *Cooperación Latino Americana de Redes Avanzadas* (RedCLARA) – también ofrece un acceso transnacional a texto completo a más de 3 millones de documentos, alrededor de 1,8 millones de artículos y más de 300.000 publicaciones universitarias, que fueron previamente recolectadas de los respectivos repositorios nacionales que cooperan en el proyecto.[6]

Otra importante contribución a la oferta de información de y sobre América Latina y el Caribe la proporciona el servicio de información de revistas *Latindex*. Esta iniciativa, que existe desde los años 90, es un proyecto de cooperación entre la *Universidad Nacional Autónoma de México* (UNAM) e instituciones asociadas de América Latina, el Caribe y Europa. El objetivo de esta iniciativa es registrar de forma centralizada todas las revistas académicas de y sobre las regiones anteriormente mencionadas y comprobar su calidad editorial. Esto ofrece la posibilidad de buscar la amplia gama de estas revistas según temas, países e instituciones, filtrar las fuentes de información requeridas y tener en cuenta la calidad editorial de las revistas y los criterios básicos para las buenas prácticas científicas y editoriales. Aunque no se ofrecen textos completos, se proporciona toda la información necesaria para acceder a las revistas.[7]

Todas estas iniciativas, proyectos y plataformas suelen estar dirigidas por universidades e instituciones de investigación. A menudo, las respectivas bibliotecas y las editoriales universitarias colaboran estrechamente. A través de un intercambio regular entre los autores científicos, los editores y los bibliotecarios implicados, se han desarrollado y se siguen perfeccionando los servicios adaptados a las necesidades de información de la ciencia y el mundo académico. El trabajo está conformado centralmente por los principios de acceso abierto y ciencia abierta, que en América Latina y el Caribe ya han encontrado expresión en leyes nacionales e iniciativas de política científica (Müller 2020). De este modo, los contenidos y la información producidos en la región son visibles y accesibles tanto allí como en el resto del mundo.[8]

Pero no solo estas amplias y establecidas plataformas de información de publicaciones científicas producidas de manera digital ofrecen información de y sobre la región. También existen numerosas colecciones digitales en las que se ponen a disposición fuentes primarias retrodigitalizadas, es decir, transformadas

[5] <http://biblioteca.clacso.edu.ar/>, 29.01.2022, <https://www.cepal.org/es/publications 29.01.2022.
[6] <https://www.lareferencia.info/es/>, 29.01.2022.
[7] <https://www.latindex.org/>, 29.01.2022.
[8] Desde 2016, el *Ibero-Amerikanisches Institut Stiftung Preußischer Kulturbesitz* (IAI) participa activamente como Centro de Acopio de *Latindex* en la registración y calificación de las revistas sobre América Latina, el Caribe, España y Portugal publicadas en Europa, pero fuera de España y Portugal. Con ello contribuye también a mejorar la calidad editorial de estas revistas mediante el intercambio y el asesoramiento a los autores y editores de las mismas.

de su forma análoga a una forma digital, tradicionalmente realizadas por grandes bibliotecas.[9]

Por ejemplo, las bibliotecas nacionales latinoamericanas están digitalizando sucesivamente sus fondos y presentándolos, por un lado, en sus colecciones digitales institucionales individuales. Por otro lado, tienen la posibilidad de agregarlos a una presentación digital conjunta gestionada por la *Biblioteca Nacional de España* impulsada por la *Asociación de Bibliotecas Nacionales de Iberoamérica* (ABINIA). Esta *Biblioteca digital del Patrimonio Hispanoamericano* contiene principalmente fuentes históricas.[10] Esto se debe a que, por regla general en la mayoría de los países del mundo, las publicaciones solo pueden ser accesibles digitalmente al público 70 años después de la muerte de sus autores, debido a la ley de derechos de autor.

La *Digital Library of the Caribbean* (dLOC), una cooperativa de digitalización de más de 70 socios de la región del Caribe, de Canadá y de Europa, tiene un enfoque similar.[11] Se trata de una biblioteca digital en la que se está reuniendo el patrimonio cultural escrito de los estados insulares del Caribe en particular para hacerlo accesible y asegurarlo de forma permanente. Además de las respectivas bibliotecas nacionales, otras bibliotecas académicas del Caribe y de otras partes del mundo también colaboran para preservar el patrimonio cultural caribeño en formato digital, que está en peligro por posibles desastres naturales como terremotos o huracanes. También en este caso están digitalizadas y presentadas fuentes de información principalmente históricas.

Estos dos proyectos muestran las posibilidades y limitaciones de la digitalización. Es cierto que, con las soluciones técnicas actuales, todo tipo de soportes escritos analógicos pueden grabarse y visualizarse digitalmente. Pero las leyes de derechos de autor vigentes en los distintos países no suelen permitir que se pongan a disposición en acceso abierto obras publicadas por lo menos en los últimos 70 años.

Además de estas plataformas de publicaciones de libre acceso, existe también una serie de repositorios con pagos especializados en América Latina y el Caribe. Se trata de ofertas comerciales de empresas que hacen disponibles, a cambio de una cuota, las publicaciones de su propio programa editorial o las fuentes de información digitalizadas por ellas. De este modo tanto los resultados de la investigación científica como las fuentes primarias de bibliotecas y archivos están disponibles en formato electrónico. El espectro de posibles acuerdos de uso va desde las licencias de usuario único para una sola publicación hasta las licencias nacionales para paquetes completos de publicaciones y bases de datos.

Por regla general, son las bibliotecas las que adquieren estas licencias y luego ponen las publicaciones a disposición de sus usuarios a través de los catálogos de las bibliotecas. Con unos presupuestos estancados (en algunos casos en descenso), los costes (a menudo bastante elevados) plantean a las bibliotecas difíciles retos

[9] En Göbel/Chicote 2017 se puede encontrar un debate detallado sobre las posibilidades y los desafíos de la digitalización en las bibliotecas y los archivos latinoamericanos.
[10] <http://www.iberoamericadigital.net/>, 29.01.2022.
[11] <https://www.dloc.com/>, 29.01.2022.

financieros. En algunos casos, no se pueden adquirir todas las licencias de forma generalizada, sino que hay que hacer una selección de contenidos orientada a las necesidades actuales de la comunidad científica. Para ello, los bibliotecarios responsables deben estar familiarizados con el estado actual de la investigación en las respectivas disciplinas, así como consultar de manera frecuente y amplia las necesidades de las comunidades investigadoras.

Las fuentes de información que figuran en las plataformas de publicación de acceso abierto, en las bibliotecas y colecciones digitales, así como en las ofertas comerciales de libros y revistas electrónicas, no solo están disponibles en los respectivos sitios web de los proveedores. Varias bibliotecas las incluyen como publicaciones electrónicas en sus catálogos electrónicos y las hacen disponibles desde los catálogos enlazándolas con los textos completos de libre acceso o con licencia de pago utilizando identificadores persistentes. Esto permite una búsqueda integrada en el contenido de todas estas plataformas, facilitando que los investigadores no tengan que buscar en cada una por separado.

Este es el principio que persigue la biblioteca del *Ibero-Amerikanisches Institut Stiftung Preußischer Kulturbesitz* (IAI, Instituto Ibero-Americano Fundación del Patrimonio Cultural Prusiano) en general y con el *Fachinformationsdienst Lateinamerika, Karibik und Latino*-Studies (Servicio de Información Especializado América Latina, Caribe y Latino Studies) – este último un servicio financiado por la *Deutsche Forschungsgemeinschaft* (Fundación Alemana para la Investigación Científica) – como principal biblioteca en Alemania para el suministro de recursos de información de y sobre América Latina y el Caribe.[12] En su buscador *IberoSearch* se da la posibilidad de pesquisar de forma centralizada tanto los fondos análogos del IAI y de sus propias colecciones digitales como las publicaciones de los proveedores comerciales y no comerciales mencionados aquí.[13]

También se pueden encontrar recursos de información sobre América Latina en la *Linga-Bibliothek* de la *Staats- und Universitätsbibliothek Hamburg* (Biblioteca Estatal y Universitaria de Hamburgo) y en la biblioteca del *German Institute for Global and Area Studies* (GIGA).[14] El *Fachinformationsdienst Romanistik* (Servicio de información especializada de la Romanística), dirigido por la ya mencionada biblioteca de Hamburgo y la *Universitäts- und Landesbibliothek Bonn* (Biblioteca Universitaria y Estadual de Bonn), ofrece publicaciones específicas sobre lingüística, literatura y estudios culturales románicos, incluidos los que tienen relación con América Latina.[15]

Además de estas bibliotecas alemanas, existen otras bibliotecas en otros países europeos especializadas en América Latina o que cuentan con colecciones latinoamericanas. Destaca la *Biblioteca Hispánica* de Madrid que, junto con la del

[12] <https://www.iai.spk-berlin.de/>, 29.01.2022, <https://fid-lateinamerika.de/>, 29.01.2022.
[13] <http://iberosearch.de/>, 29.01.2022.
[14] <https://www.sub.uni-hamburg.de/sammlungen/linga-bibliothek.html>, 29.01.2022, <https://www.giga-hamburg.de/de/das-giga/regionalinstitute/giga-institut-fuer-lateinamerika-studien>, 29.01.2022.
[15] <https://fid-romanistik.de/>, 29.01.2022.

Instituto Ibero-Americano, son unas de las mayores bibliotecas de Europa especializadas en la región.[16]

Con el fin de proporcionar información sobre sus colecciones de y sobre América Latina y el Caribe, pero también sobre sus actividades culturales relacionadas con la región, coordinarlas cuando sea posible y asesorar a los investigadores en sus búsquedas bibliográficas, las bibliotecas y los centros de documentación europeos especializados en América Latina y el Caribe colaboran en la *Red Europea de Información y Documentación sobre América Latina* (REDIAL). Su web recoge toda la información sobre las bibliotecas, sus fondos y sus eventos relacionados con América Latina y el Caribe. La particularidad de esta página es que está gestionada conjuntamente por REDIAL y el *Consejo Europeo de Investigaciones Sociales de América Latina* (CEISAL), que reúne a todos los investigadores de Europa sobre América Latina y el Caribe. El portal es un resultado central de la larga cooperación y del intenso intercambio entre investigadores y bibliotecarios sobre las necesidades de información y cómo satisfacerlas.[17]

Con todas estas iniciativas las bibliotecas ponen a disposición del público la mayor cantidad posible de información y publicaciones en formato digital en acceso abierto y adquieren licencias para hacer disponibles las publicaciones protegidas por derechos de autor de la forma más amplia posible. Sin embargo, por razones financieras y legales todavía está lejos de ser posible satisfacer todas las necesidades de la comunidad científica en cuanto a la disponibilidad digital de las fuentes de información.

Datos de investigación

Paralelamente a estos servicios bibliotecarios, numerosos proyectos académicos de investigación están creando de forma independiente colecciones de datos y contenidos digitales. Por ejemplo, algunos investigadores escanean ellos mismos las respectivas fuentes de información durante sus visitas de investigación a archivos, bibliotecas y museos. Utilizan cámaras digitales, teléfonos móviles o escáneres sencillos para disponer de la información en formato electrónico de la manera más completa posible. Aunque este modo de transformar un objeto análogo a un objeto digital está permitido por las leyes de derechos de autor, a menudo plantea a los investigadores el problema de cómo almacenar y reutilizar estos datos. Los recursos financieros de los proyectos muchas veces no permiten crear y poner en funcionamiento permanente la infraestructura técnica necesaria para procesar, almacenar y hacer accesibles los escaneos. Además, los escaneos se proporcionan con metadatos, pero muchas veces ya no con metadatos según los estándares más frecuentemente utilizados a nivel internacional, lo que dificulta o incluso imposibilita el uso posterior y la interoperabilidad de los datos. Por lo tanto, una fusión posterior de estos con los datos normalizados creados y conservados en las bibliotecas y los archivos no suele ser posible o solo lo es con dificultad debido a la incompatibilidad o falta de metadatos.

[16] <https://www.aecid.es/ES/biblioteca>, 29.01.2022.
[17] <https://rediceisal.hypotheses.org/>, 29.01.2022.

Si consideramos la importancia de que las fuentes de información escritas o impresas cuenten con datos primarios y metadatos lo más completos posible y estandarizados a la hora de utilizar las herramientas de las humanidades digitales, se hace evidente que durante la creación de los datos digitales debería producirse un intenso intercambio entre los investigadores, por un lado, y los bibliotecarios y archivistas, por otro. De este modo, una planificación y coordinación temprana evitaría procesos de reelaboración que resultan un esfuerzo extra y apoyarían la usabilidad e interoperabilidad de los datos a largo plazo.

No obstante, los resultados de la digitalización de fuentes manuscritas o impresas no son los únicos datos que se pueden crear durante la investigación científica. Las grabaciones y registros de entrevistas o rituales, los geodatos o los resultados de encuestas son generados hoy en día en su mayoría en formato digital. Esta información también debe almacenarse y estar disponible de la forma más sostenible posible. Así se facilita su uso dentro de los respectivos proyectos de investigación, se hace posible la replicación del análisis original y se pone a disposición de otros investigadores para su posterior uso de acuerdo con los principios de la ciencia abierta.

Además de los aspectos técnicos, diversos aspectos legales suponen un reto para los investigadores como productores y usuarios de datos, así como para los operadores de las respectivas infraestructuras. Por ejemplo, hay que respetar las normas de protección de datos y proteger los derechos personales de los implicados. Sin embargo, las normas de las universidades e instituciones de investigación en las que se llevan a cabo los proyectos, así como las de los respectivos financiadores terceros, también deben ser observadas en lo que respecta al almacenamiento y la accesibilidad de estos datos de investigación. Por regla general, los datos generados en un proyecto de investigación son propiedad de la institución o del tercero que los financia. En consecuencia, la información debe almacenarse y estar disponibles a través de los correspondientes repositorios institucionales. A diferencia de Europa, donde existen estructuras transnacionales además de las infraestructuras nacionales para el almacenamiento, la accesibilidad y la reutilización de los datos de investigación,[18] en América Latina suelen ser infraestructuras institucionales específicas y solo parcialmente nacionales las que deben conservar esos datos de investigación. El suministro a través de plataformas internacionales a menudo no es posible o incluso no está permitido.

Intercambio y cooperación

En todos estos contextos, se pone de manifiesto la importancia cada vez mayor del intercambio entre investigadores y bibliotecarios o archivistas. Por un lado, es tarea de los investigadores describir de forma transparente y especificar con el mayor

[18] Algunos ejemplos son la *Nationale Forschungsdateninfrastruktur* (NFDI, Infraestructura nacional para Datos de Investigación) alemana, actualmente en construcción, la infraestructura de investigación alemana para las humanidades *CLARIAH-DE*, el consorcio europeo de infraestructuras de investigación *DARIAH-EU* o el servicio de almacenamiento en línea de datos científicos *zenodo.org*, financiado por la Comisión Europea (<https://www.nfdi.de/>, <https://www.clariah.de/>, <https://www.dariah.eu/>, <https://zenodo.org>, todas 29.01.2022).

detalle posible las nuevas herramientas y métodos de procesamiento científico de los datos digitales y las necesidades resultantes en cuanto a formatos de datos, interfaces e infraestructuras. Por otro lado, los bibliotecarios y archivistas necesitan conocer el estado actual de la investigación y los futuros desarrollos en sus respectivas comunidades. También deben estar atentos a las posibilidades y limitaciones financieras, legales y técnicas en la provisión de fuentes de información digital y ser capaces de transferirlas y aplicarlas de forma orientada a las distintas necesidades de los investigadores (Müller 2021).

Para ello, no solo hay que seguir desarrollando o incluso crear las competencias específicas en las bibliotecas, sino que también hay que examinar críticamente, en un contexto de recursos limitados, si es necesario abandonar los servicios tradicionales en favor de los servicios modernos de apoyo a la investigación (Stille et al. 2021, Bonte 2014, Ceynowa 2014, Mittler 2014). Esto tampoco debería decidirse sin tener en cuenta la perspectiva de los investigadores que pueden verse afectados.

El intercambio entre los diversos actores, que es necesario para el desarrollo de las infraestructuras de información, requiere no solo conocimientos específicos de la materia, sino también pensamiento estratégico, capacidad de compromiso, pragmatismo y, especialmente en materias como los estudios románicos y latinoamericanos, multilingüismo y competencia intercultural (Tappenbeck 2015).[19]

Solo en un intercambio al mismo nivel, caracterizado por el respeto mutuo de los respectivos conocimientos y habilidades, las herramientas e infraestructuras existentes para la provisión y reutilización de fuentes de información digital pueden alinearse con precisión y seguir desarrollándose para satisfacer las futuras necesidades del mundo académico.

Literatura

BONTE, Achim. 2014. „Wissenschaftliche Bibliotheken der nächsten Generation. Sind die Institutionen und ihre Mitarbeiter für die Zukunft gerüstet?“ *Zeitschrift für Bibliothekswesen und Bibliografie* 61 (4-5), 239-242. <http://dx.doi.org/10.3196/18642950146145114>.

CEYNOWA, Klaus. 2014. „Digitale Wissenswelten – Herausforderungen für die Bibliothek der Zukunft“ *Zeitschrift für Bibliothekswesen und Bibliografie* 61 (4-5), 235-238. <http://dx.doi.org/10.3196/18642950146145109>.

Göbel, Barbara & Gloria Chicote (ed.). 2017. *Transiciones inciertas. Archivos, conocimientos y transformación digital en América Latina*. La Plata: FAHCE, Universidad Nacional de la Plata, Ibero-Amerikanisches Institut. <https://www.libros.fahce.unlp.edu.ar/index.php/libros/catalog/book/99>.

MARTÍN, Eloisa & Barbara Göbel (ed.). 2018. *Desigualdades interdependentes e geopolítica do conhecimento. Negociações, fluxos, assimetrias*. Rio de Janeiro: 7 Letras.

[19] Estos conocimientos y habilidades son de especial importancia en el contexto de la transmisión de información y conocimiento de y sobre América Latina en Alemania y Europa, para superar las asimetrías existentes en la circulación y geopolítica del conocimiento (Martin/Göbel 2018).

MITTLER, Elmar. 2014. „Nachhaltige Infrastruktur für die Literatur- und Informationsversorgung: im digitalen Zeitalter ein überholtes Paradigma – oder so wichtig wie noch nie?“ *Bibliothek, Forschung und Praxis* 38 (3), 344-364.
<https://doi.org/10.1515/bfp-2014-0059>.

MÜLLER, Christoph. 2020. „Elektronisches Publizieren und Open Access: Die Perspektive Lateinamerikas“ *b.i.t. online* 23 (4), 374-380.
<https://www.b-i-t-online.de/heft/2020-04-fachbeitrag-mueller.pdf>.

MÜLLER, Christoph. 2021. „Between Digital Transformation in Libraries and the Digital Humanities. New Perspectives on Librarianship“ En *World Editors. Dynamics of Global Publishing and the Latin American case between the Archive and the Digital Age* ed Guerrero, Gustavo, Benjamin Loy & Gesine Müller, 379-384, Berlin: De Gruyter [Latin American Literatures in the World, v. 8].
<https://doi.org/10.1515/9783110713015-023>.

STILLE, Wolfgang et al. 2021. „Forschungsunterstützung an Bibliotheken. Positionspapier der Kommission für forschungsnahe Dienste des VDB“ *o-bib* 2021 (2), 1-19.
<https://doi.org/10.5282/o-bib/5718>.

TAPPENBECK, Ina. 2015. „Fachreferat 2020: from collections to connections“ *Bibliotheksdienst* 49 (1), 37-43.
<https://doi.org/10.1515/bd-2015-0006>.

Resumen

La transformación digital de los contenidos, los métodos de trabajo y las herramientas de la investigación plantea a las bibliotecas retos cada vez más importantes. Tienen que satisfacer las diferentes necesidades de los investigadores con información electrónica específica y nuevos servicios digitales.

En el contexto de la investigación centrada en América Latina y el Caribe, la oferta de recursos de información digital puede realizarse gracias a numerosas plataformas de y sobre la región que proporcionan información en acceso abierto o de pago. Por otro lado, la gestión y la seguridad de los datos digitales de investigación es aún más complicada, ya que los productores de datos de investigación tienen que elegir entre los repositorios de las respectivas instituciones o países y los repositorios disciplinarios en el ámbito internacional. Por lo tanto, hasta ahora no existe una organización y provisión central de datos de investigación sobre América Latina y el Caribe.

Por eso, es tarea de investigadores y bibliotecarios en un intercambio conjunto coordinar la oferta de información y la gestión sostenible de los datos de investigación con las necesidades específicas de los científicos y desarrollar nuevos servicios digitales.

Abstract

The digital transformation of content, working methods and tools in research is presenting libraries with ever new challenges. They have to meet the different needs of researchers with specific electronic information and new digital services.

In the context of research focused on Latin America and the Caribbean, the supply of digital information resources can be realized thanks to numerous platforms from and about the region that provide information either in open access or for a fee. The management and securing of digital research data, on the other hand, is still more complicated, since the producers of research data have to choose between the repositories of the respective institutions or countries and the disciplinary repositories in the international field. Therefore, a central organization and provision of research data on Latin America and the Caribbean until now does not exist.

Against this background, it is the task of researchers and librarians in a joint exchange to coordinate the supply of information and sustainable research data management with the specific needs of the scientists and to develop new digital services.

2022, n°9
pp. 163-185
doi: https://doi.org/ 10.15460/apropos.9.1899

Markus Trapp & Johannes von Vacano

(FAIRe) Forschungsdaten, Open Access und neue Formen der Kommunikation in der Romanistik

Beiträge des FID zur Gestaltung des digitalen Wandels

Markus Trapp
ist Fachreferent für Hispanistik und Lusitanistik und Leiter der Arbeitsstelle FID Romanistik an der Staats- und Universitätsbibliothek Hamburg.
markus.trapp@sub.uni-hamburg.de

Johannes von Vacano
ist wissenschaftlicher Mitarbeiter an der Universitäts- und Landesbibliothek Bonn.
johannes.von.vacano@ulb.uni-bonn.de

Keywords

Romanistik – Fachinformationsdienst – Open Access – Forschungsdaten – Wissenschafts-kommunikation

1. Einführung

Der seit 2016 von der ULB Bonn und der SUB Hamburg betriebene, DFG-geförderte Fachinformationsdienst Romanistik (FID) unterstützt die Etablierung einer transdisziplinären wissenschaftlichen Forschung romanistischer Prägung im digitalen Raum. Zu den in Abstimmung mit den romanistischen Fachverbänden und ihrer *AG Digitale Romanistik* sowie mit *romanistik.de* aufgebauten und fortlaufend weiterentwickelten Services zählen digitale Angebote für die Literaturrecherche, die Bereitstellung ausgewählter elektronischer Medien sowie ein online verfügbares fachspezifisches Informationsangebot, in dem traditionelle und digitale Ressourcen gebündelt präsentiert werden. Darüber hinaus widmet sich der FID insbesondere den genuin digitalen Bereichen Forschungsdatenmanagement (FDM) und Open-Access-Publizieren sowie den Möglichkeiten digitaler Kommunikation und Vernetzung über Social Media.

Dieser Beitrag[1] präsentiert drei Handlungsfelder des FID, in denen die eingangs skizzierten Ziele konkret umgesetzt werden: Die Teilprojekte Forschungsdatenmanagement (2.) und Open-Access-Publizieren (3.) sowie die Initiativen des FID im Bereich der neuen Kommunikationsformen (4.).

2. Forschungsdaten

Forschungsdatenmanagement hat das Potenzial, die Transdisziplinarität zu unterstützen, indem es Daten bzw. deren Nachnutzbarkeit über ihren unmittelbaren Entstehungskontext hinaus optimiert, angefangen bei der Auffindbarkeit bis hin zur technischen Verwendbarkeit und umfassenden Beschreibung, um das Nachnutzungspotenzial zuverlässig ablesen zu können. Methodisch und disziplinär ganz anders gelagerte Nutzungsszenarien sind dabei idealerweise stets mitzudenken.

Im Folgenden wird kurz auf den Begriff des Forschungsdatenmanagements eingegangen und auf den Zusammenhang mit dem Akronym FAIR, bevor angerissen wird, mit welchen Angeboten der FID Forschende in der Romanistik in dieser Hinsicht unterstützt. Anschließend wird die Bedeutung von Metadaten für das Forschungsdatenmanagement und in diesem Kontext insbesondere für die Auffindbarkeit von Forschungsdaten skizziert, um davon ausgehend einen Impuls für eine stärker standardisierte Metadatenvergabe in der Romanistik zu formulieren.

FAIRe Forschungsdaten

Der Umgang mit Forschungsdaten[2] wird konzeptionell häufig anhand des Datenlebenszyklus[3] erläutert (vgl. Abb. 1), der seinerseits ein traditionelleres Verständnis eines linearen Forschungsablaufs – von der Datenerhebung über deren Analyse bis hin zur wissenschaftlichen (Text-)Publikation – umbiegt zu einer zirkulär verlaufenden Struktur, bei der, wie etwa bei Text-Publikationen längst etabliert, auch die zugehörigen Daten erneut in den Forschungsprozess eingespeist werden.[4]

[1] Der Artikel basiert auf einem gleichnamigen Vortrag, der in der Sektion „Digital, global, transdisziplinär: Impulse für eine transdisziplinäre Digitale Romanistik“ auf dem XXXVII. Romanistentag gehalten wurde. Die Folien sind auf *Zenodo* veröffentlicht (Trapp & von Vacano 2021) <https://doi.org/10.5281/zenodo.5548234>.

[2] Für eine mögliche Definition des Begriffs im vorliegenden Kontext siehe : <https://fid-romanistik.de/forschungsdaten/was-sind-romanistische-forschungsdaten>.

[3] Diese schematische Darstellung – teilweise auch als Forschungsdatenzyklus bezeichnet – ist eine vereinfachte Version des deutlich komplexeren Curation Lifecycle Model des Digital Curation Centre (DCC) (vgl. Higgins 2008 und Rümpel 2011). Sie beinhaltet in der Regel Elemente wie die Erhebung, Verwendung, Dokumentation, Speicherung und Publikation von Daten sowie je nach fachlichem Zuschnitt weitere Aspekte des Umgangs damit, die kreisförmig angeordnet sind und häufig mit der Erhebung bzw. Erstellung der Daten beginnen. Das Beispiel in Abb. 1 ist den Webseiten des FID entnommen, vgl.: <https://fid-romanistik.de/forschungsdaten/arbeit-mit-forschungsdaten/grundlegendes-zum-forschungsdatenmanagement-in-der-romanistik#c2602>, vgl. auch den entsprechenden Artikel auf forschungsdaten.info: <https://www.forschungsdaten.info/themen/informieren-und-planen/datenlebenszyklus/>.

[4] Wie bei den meisten Modellen handelt es sich auch hierbei nicht um eine Anleitung, der Schritt für Schritt zu folgen ist. Vielmehr „ist der Zugang über den Datenlebenszyklus sinnvoll, weil damit wesentliche Phasen des FDM und damit verbunden Aufgaben, Rollen und Verantwortlichkeiten adressiert werden“ (Dierkes 2021, 305).

1| Eine Variante des Datenlebenszyklus

Ist Forschungsdatenmanagement eigentlich gedacht als Begleitung durch den gesamten Forschungskreislauf, wird dabei doch immer in einer umgekehrten Teleologie angestrebt, alle Entscheidungen von Anfang an so zu treffen und zu dokumentieren, dass am Ende eines solchen Kreislaufs ein qualitativ hochwertiger Datensatz steht, der sich daraufhin neu ins wissenschaftliche Getümmel stürzen kann. Reflektiert werden muss dabei immer auch, welche weiteren Nutzungsszenarien für die erhobenen Daten infrage kommen; gerade in einem vielfältigen Fach wie der Romanistik betrifft das, je nach Art der Daten, die Analyse unter anderen disziplinären Vorzeichen. Dabei lassen sich mindestens vier stereotype Probleme imaginieren, auf die Forschungsdatenmanagement antizipierend eingeht und die im Folgenden als Fragen[5] aus Nutzendenperspektive illustriert werden sollen:

1. „Ich weiß nicht, was es gibt und wo ich es finde“.

Sich einen Überblick verschaffen zu können, welche Forschungsdaten zu einem bestimmten Ansatz bereits vorliegen, ist ein zentrales Anliegen zu Beginn eines wissenschaftlichen Vorhabens und erfordert, genauso wie im Falle der Sichtung verfügbarer Literatur zu einem Thema, Kenntnisse über die entsprechenden Infrastrukturen für die Suche. Ähnlich einer Bibliothek bzw. deren Katalog genügt es jedoch nicht zu wissen, dass sich die gesuchten Inhalte an einem bestimmten Ort befinden. Sie müssen nach bestimmten Konventionen auffindbar gemacht werden,

[5] Angelehnt an die in Mathiak & Kronenwett 2017 beschriebene Umfrage an der Kölner Universität und die Fragen aus Wilkinson et. al. 2016, 2-3.

indem, kurz gesagt, passende Etiketten daran angebracht werden. Gutes[6] Forschungsdatenmanagement hilft also dabei, die Daten dort zu veröffentlichen, wo auch danach gesucht wird, und in einer Art und Weise und mit einer solchen Beschreibung, dass sie mit einer relevant formulierten Suchanfrage gefunden werden.

2. „Ich habe (Hinweise auf) etwas Interessantes gefunden, aber ich komme nicht dran".

Herauszufinden, dass es möglicherweise passende Forschungsdaten gibt und wo sie sich befinden (könnten), ist nur der erste Schritt und oftmals nicht ausreichend, sofern man nicht weiß, wie man sich Zugang dazu verschaffen kann. Bei der Suche nach einem Buch, dessen Signatur man recherchiert hat, kommt man nicht weiter, wenn es in einen Tresor eingeschlossen ist. Es sei denn, das freundliche Bibliothekspersonal erläutert die notwendigen Schritte, um sich den Tresor öffnen zu lassen. FDM sorgt also dafür, dass bei jedem Datensatz deutlich erkennbar ist, welche administrativen und technischen Schritte notwendig sind, um auf die Daten zugreifen zu können – im Idealfall genügt ein Klick auf die Schaltfläche „Herunterladen".

3. „Ich habe einen Datensatz heruntergeladen, aber keines meiner Programme kann damit etwas anfangen".

Besonders frustrierend ist es, ein Buch gefunden zu haben, das seinem Titel und seiner Beschreibung nach möglicherweise brauchbar ist, nur um dann festzustellen, dass es sich gar nicht erst aufklappen lässt, es als Mikrofiche vorliegt und ein bestimmtes technisches Gerät benötigt wird, um den Text visualisieren zu können, oder es schlichtweg in einer Sprache verfasst ist, die man nicht lesen kann. Dass die im Forschungsprozess erzeugten Daten anschließend auch auf anderen Rechnern von anderen (gängigen) Programmen geöffnet, ausgelesen und mit weiteren Daten verknüpft werden können, kann auch durch entsprechende Weichenstellungen des Forschungsdatenmanagements befördert werden, die die technische Nachnutzbarkeit sicherstellen.

4. „Ich habe im Grunde passende Dateien gefunden, aber ich weiß nicht, wie die Daten zustande gekommen sind oder ob ich sie weiterverwenden darf".

Sollten schließlich alle anderen Hürden überwunden sein, kann es dennoch geschehen, dass die Inhalte des Buchs nicht nachvollziehbar hergeleitet sind oder unklar bleibt, wie zitierfähig dieser Inhalt ist. Ein bisschen stärker an den Haaren

[6] Was genau mit „gutem FDM" gemeint ist, muss freilich irgendwo definiert werden. Vgl. Wilkinson et.al. 2016, 1: „Good data management is not a goal in itself […] What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility."

herbeigezogen ist hier die Nutzungslizenz-Metapher, dass womöglich die Inhalte nicht verwendet werden können, weil ein großer roter Sticker im Einband besagt, dass das nicht erlaubt ist. Forschungsdatenmanagement hilft dabei, die Entstehung und Bearbeitung der Daten zu dokumentieren, damit Forschende evaluieren können, ob die für den eigenen Ansatz benötigten Information(skett)en vorhanden sind. Außerdem wird durch die Vergabe einer zu den Umständen der Daten passenden Lizenz gewährleistet, dass für alle Nachnutzenden unzweideutig feststeht, wie die Daten rechtlich verwendet werden dürfen.

Diese vier etwas konstruierten, aber doch an den Bedarfen der Fachwissenschaft orientierten[7] Probleme – zusammenfassbar als Auffindbarkeit, Zugangsmöglichkeiten, Verwendbarkeit und Nachnutzbarkeit – beschreiben wesentliche Anliegen, die in den vielzitierten FAIR-Prinzipien (vgl. Wilkinson et. al. 2016, insbes. S. 2–3) in Angriff genommen werden. Das auch von der olympisch angehauchten Konnotation lebende Akronym plädiert dafür, Forschungsdaten sollten „findable", „accessible", „interoperable" und „reusable" gemacht werden und regt zu diesem Zweck eine Reihe relativ eng gefasster Zielformulierungen an, die oben unter 1–4 in groben Zügen skizziert wurden. An anderer Stelle[8] hat die *AG Digitale Romanistik* die einzelnen Buchstaben F, A, I und R und ihre dahinterliegenden Prinzipiencluster fachspezifisch und gekonnt illustriert, sodass hier beruhigt davon abgesehen werden kann, näher darauf einzugehen.

Erlaubt sei jedoch der ebenso obligatorische wie kurze Hinweis auf den Nexus von Open Science als Paradigma auch für das Forschungsdatenmanagement und den ubiquitären FAIR-Prinzipien. Denn obwohl diese beiden Leitlinien und Ideale in komplementärer Eintracht Hand in Hand gehen können und sollen, lohnt es sich hervorzuheben, dass sie nicht synonym sind. Grundlegender Unterschied ist, dass Daten auch dann FAIR sein können, wenn sie nicht ohne jede Einschränkung zugänglich sind (was eine sehr kurze, aber in diesem konkreten Falle zutreffende Auslegung von „Open" darstellt). Vielmehr operiert FAIR unter der Prämisse des „So offen wie möglich, aber so geschlossen wie nötig"[9]. Damit wird unterstrichen, dass es durchaus Fälle gibt, in denen beispielsweise rechtliche oder ethische Gründe dagegensprechen, Daten uneingeschränkt verfügbar zu machen, weshalb

[7] Vgl. beispielsweise die Schwierigkeiten im Umgang mit geisteswissenschaftlichen Forschungsdaten, die Forschende der Universität Köln in einer dortigen Umfrage genannt haben: „Die Nachhaltigkeit an sich wird schon als Problem gesehen. 66% der Befragten geben an, dass sie befürchten die Daten zu verlieren, wenn sich nach Projektende niemand mehr für die dazugehörigen Webseiten zuständig fühlt [*Accessible*]. 60% fürchten Datenkonversionsprobleme [*Interoperable*]. Aber auch für Probleme mit der Auffindbarkeit (45%) [*Findable*] und der Dokumentation (41%) [*Reusable*] besteht eine prinzipielle Sensibilität." (Mathiak & Kronenwett 2017, Ergänzungen in eckigen Klammern von JvV).

[8] Die Artikelserie der AG auf dem Romanistik-Blog des FID beginnt mit dem einführenden Beitrag von Thomas Krefeld und Stephan Lücke: „FAIRness: ein contrat social für die Wissenschaftskommunikation im Internet" (Krefeld & Lücke 2020). Siehe zum Romanistik-Blog und der Reihe auch die Ausführungen unter „4. Neue Formen der Kommunikation in der Romanistik" weiter unten.

[9] Die Formulierung ist den Guidelines des Förderprogramms Horizont2020 der europäischen Kommission entnommen und wird seitdem in einem Atemzug mit den FAIR-Prinzipien genannt: „The Commission's approach can therefore be described as ‚*as open as possible, as closed as necessary*'." (Europäische Kommission 2020).

die unter „Accessible“ („zugänglich“) aufgelisteten Punkte an keiner Stelle den bedingungslosen Zugang einfordern.[10]

Dienstleistungsangebot des FID Romanistik rund um Forschungsdaten

Die oben erwähnten Probleme und Lösungsansätze werden an vielen Stellen bereits durch das Informationsangebot des FID Romanistik zum Forschungsdatenmanagement adressiert, das sich in seiner inhaltlichen Ausgestaltung eng an Bedarfen orientiert, die aus der Fachcommunity an ihn herangetragen wurden.[11] Diese betreffen die Bereitstellung

- allgemeiner Informationen zu Formaten, Standards, Tools und zur Digitalisierung im Allgemeinen,

- von Hinweisen zum wissenschaftsadäquaten Umgang mit Forschungsdaten,

- eines fachwissenschaftlichen Nachweissystems für Forschungsdaten sowie

- eines zentralen Suchinstruments für romanistische Daten.

Zur Deckung des Bedarfs an allgemeinen Informationen wurde im Portal des FID die Sektion „Forschungsdaten“[12] eingerichtet, die grundlegende und fachspezifisch aufbereitete Hinweise zum Forschungsdatenmanagement bereithält. Die Struktur der Unterpunkte orientiert sich an den typischen Aufgaben im Forschungsdatenmanagement, wie sie im Datenlebenszyklus veranschaulicht werden. Der Fundus an Informationen und Verweisen zu externen Ressourcen wird beständig aktualisiert und erweitert.

Speziellere Hinweise zum Umgang mit Forschungsdaten bietet der Unterbereich „Arbeit mit Forschungsdaten“. Unter der Überschrift „Erstellen, Nutzen und Analysieren von Forschungsdaten“ finden sich insbesondere Informationen für die eine Hälfte des Datenlebenszyklus, etwa zu digitalen Tools und Methoden. Die andere Hälfte des Lebenszyklus von Forschungsdaten deckt der Punkt „Sichern und Publizieren von Forschungsdaten“ ab, der beispielsweise Schritt-für-Schritt-Anleitungen für die Speicherung von Forschungsdaten in den frei zugänglichen Repositorien *Zenodo* und *DARIAH-DE Repository* enthält.[13]

Um dem Bedarf eines Nachweissystems für Forschungsdaten nachzukommen, wurde zum einen in Kollaboration mit den Projektpartnern, *romanistik.de* und der

[10] Vgl. zum Forschungsdatenmanagement zwischen FAIR und „Open“ auch Higman et. al 2019.

[11] 2017 fanden hierzu zwei Workshops statt, die der Bedarfserhebung bzw., in Reaktion darauf, der Eruierung von Maßnahmen zur Deckung dieser Bedarfe gewidmet waren, vgl. <https://fid-romanistik.de/forschungsdaten/workshops/.> Eine erste Aufarbeitung der konzeptionellen Umsetzung der Bedarfsermittlung im Informationsangebot des FID bieten Christoph Hornung und Jan Rohden in ihrem Artikel „Der Beitrag des Fachinformationsdienstes Romanistik zur romanistischen Digitalkultur“ (Hornung & Rohden 2019). Diese und weitere Betrachtungen sind später eingeflossen in das Übersichtspapier *Forschungsdatenmanagement in der Romanistik. Aktuelle Situation und zukünftige Perspektiven* von Maria Erben, Doris Grüter und Jan Rohden (Erben et al. 2018).

[12] <https://fid-romanistik.de//forschungsdaten/>.

[13] <https://fid-romanistik.de/forschungsdaten/arbeit-mit-forschungsdaten/sichern-und-publizieren-von-forschungsdaten#c2586>. Als weiteres Angebot zur Bündelung romanistischer Forschungsdaten wurde gemeinsam mit der *AG Digitale Romanistik* und *romanistik.de* eine sogenannte Community in *Zenodo* angelegt; weitere Informationen hierzu im *Zenodo*-Leitfaden oder im Blog-Eintrag „Wohin mit romanistischen Forschungsdaten? Teil 1: Zenodo“ (von Vacano 2020).

AG Digitale Romanistik, ein Meldeformular entwickelt, das mithilfe der in der Fachgemeinschaft wohlbekannten Benutzungsoberfläche der Kommunikationsplattform *romanistik.de* eine niedrigschwellige Meldung von Forschungsdaten durch die Forschenden selbst ermöglicht.[14] Zusätzlich verwendet der FID Romanistik seine Datenbank für die Erfassung von Internetquellen, um dort romanistisch relevante Forschungsdaten nach bibliothekarischen Standards zu verzeichnen. Dies geht Hand in Hand mit der Bereitstellung eines zentralen Suchinstruments für romanistische Daten, da die relevanten Einträge der Datenbank über die Webseiten des FID durchstöbert werden können.[15] Darüber hinaus sind diese Daten auch im Discovery-Portal des FID Romanistik nachgewiesen, wo gezielter danach recherchiert werden kann.[16] Die facettierte Suche macht sich die Anreicherung der Datensätze mit standardisierten Metadaten zunutze, die bei der Aufnahme in die Datenbank vorgenommen wird. Dabei werden grundlegende Informationen zu den Datensätzen – wie URL, Titel, Urheber*innen, Format oder Publikationsdatum – übernommen und um eine inhaltliche Erschließung und Klassifikation ergänzt. Zu diesem Zweck kommen Richtlinien zur Anwendung, die sicherstellen, dass eine Standardisierung erreicht wird, welche wiederum die systematische Recherche ermöglicht. Diese Richtlinien speisen sich aus der bibliothekarischen Praxis und setzen neben einer freien Beschreibung und einigen Klassifikationen insbesondere auf die Verwendung der Gemeinsamen Normdatei, kurz GND, für die Zuweisung einheitlicher Schlagwörter. Die GND wird an der Deutschen Nationalbibliothek gehalten und die darin kuratierten Begriffe „repräsentieren und beschreiben Entitäten, also Personen, Körperschaften, Konferenzen, Geografika, Sachbegriffe und Werke, die in Bezug zu kulturellen und wissenschaftlichen Sammlungen stehen." Solche „Normdaten erleichtern die Erschließung, bieten eindeutige Sucheinstiege und vernetzen unterschiedliche Informationsressourcen."[17] Indem alle Datensätze, die sich beispielsweise mit dem Französischen, Spanischen oder Italienischen auseinandersetzen das GND-Schlagwort für die jeweilige Sprache erhalten, können diese Ressourcen anschließend über die jeweiligen Sucheinstiege gebündelt werden. Ebenso verhält es sich mit Schlagwörtern, die den Typ der Ressource kennzeichnen, etwa „Korpus"[18] für ein linguistisch ausgezeichnetes Textkorpus oder „Programm"[19] für eine spezielle Software, beispielsweise das Tool, mit dem besagtes Textkorpus erstellt wurde.

Metadaten

Der vorangegangene Exkurs illustriert zum einen die Bedeutung normierter Metadaten, zum anderen die Anstrengungen des FID, romanistische Forschungsdaten recherchierbar zu machen. Denn die selbstständige Suche an sich gestaltet

[14] Das Formular steht zur Verfügung unter <https://www.romanistik.de/res>, eine Anleitung zu seiner Verwendung auf den Seiten des FID unter: <https://fid-romanistik.de/fileadmin/user_upload/dokumente/Texte/Anleitung_Meldesystem_Forschungsdaten_Tools.pdf>.

[15] <https://fid-romanistik.de/forschungsdaten/suche-nach-forschungsdaten#c2520>.

[16] Ein Tutorial hierzu steht auf den Webseiten bereit: <https://fid-romanistik.de/recherchewerkzeuge/online-tutorials>.

[17] <https://www.dnb.de/gnd>.

[18] <https://d-nb.info/gnd/4165338-5>.

[19] <https://d-nb.info/gnd/4047394-6>.

sich schwierig, da Forschungsdaten nach wie vor verstreut abgelegt werden: Es gibt kein zentrales Repositorium für die Romanistik. In übergreifenden Repositorien wiederum sind infrage kommende Forschungsdaten so gut wie nie gebündelt und selten für die Suche innerhalb des Repositoriums optimiert,[20] geschweige denn für eine externe Suchanfrage, beispielsweise über Meta-Suchmaschinen.[21] Teilweise werden Datensätze eher zufällig über punktuelle Meldungen von Forschenden in Social-Media-Kanälen gefunden[22] oder über Erwähnungen in Forschungsliteratur.[23] Nach wie vor sind diese Datensätze nicht immer mit einem Persistenten Identifikator[24] versehen, der ihre langfristige Auffindbarkeit gewährleistet. Vor allem sind jedoch Vollständigkeit und Qualität von Metadaten und Dokumentation ein entscheidender Faktor für die Auffindbarkeit und die Nachnutzbarkeit von Forschungsdaten.
Hier ist nicht der richtige Ort für die wahrscheinlich ohnehin müßige Frage, weshalb den Metadaten häufig nicht genügend Aufmerksamkeit zuteilwird – produktiver erscheint die Frage, ob das auch weiterhin so sein muss. Im Folgenden wird versucht, das „Nein", mit dem auf letztere Frage geantwortet wird, noch ein wenig aufzublähen, indem ein Vorschlag für das „Wie" gewagt wird.

Unbestreitbar erschweren diverse Faktoren die Auszeichnung von Forschungsdaten mit Metadaten. Zum einen ist bereits der Begriff „Metadaten"[25] – außerhalb der Expert*innenzirkel – potenziell eher negativ oder zumindest dröge belegt und wird als wenig einladend wahrgenommen. Zum anderen ist die Vergabe umfangreicher Metadaten auch mit einem nicht unerheblichen Aufwand verbunden, der sich in vielen Repositorien bereits an der Länge der Eingabeformulare und der schieren Menge an ausfüllbaren Feldern auf einen Blick ablesen lässt – und auf diesen ersten Blick auch direkt sein ganzes Abschreckungspotenzial entfaltet. Häufig bleiben entsprechende Formulare – die aus der durchschnittlichen Nutzendensicht dennoch attraktiver erscheinen müssen als lange Textdokumente voller

[20] Vergleiche diesbezüglich Burger et al. 2021, wo die Metadatenqualität der Datensätze untersucht wird, die ihren DOI über die TIB registriert haben. Unter anderem wird festgestellt, dass es einen „starken Fokus auf die Metadatenpflichtfelder" gibt, was nicht ausreicht, um einen hohen Grad an FAIRness zu erreichen, aber nach Ansicht der Autor*innen immerhin „die grundsätzliche Auffindbarkeit der in den Repositorien befindlichen wissenschaftlichen Ressourcen sicherstellt." (S. 8). Hierbei muss jedoch berücksichtigt werden, dass nicht ausschließlich DOIs für Forschungsdaten untersucht wurden und auch keine Unterscheidung nach Fächern vorgenommen wurde, weshalb die Vermutung gestattet sei, dass bei einer Einschränkung auf Forschungsdaten im Allgemeinen und geisteswissenschaftliche im Speziellen diese Zahlen niedriger ausfallen dürften.

[21] Zu nennen sind hier OpenAire (<https://www.openaire.eu/>), BASE, die Bielefeld Academic Search Engine (<https://www.base-search.net/>) oder DataCite Search (<https://search.datacite.org/>).

[22] Der oft erwähnte direkte Austausch zwischen Forschenden, die in demselben engen disziplinären Bereich arbeiten, ist im Hinblick auf die FAIR-Prinzipien ebenfalls ausbaufähig.

[23] Schwer abzuschätzen ist, aus welchen Gründen das Meldeformular für Forschungsdaten auf *romanistik.de* nicht häufiger genutzt wird. Womöglich ist es noch zu wenig bekannt oder es werden weniger relevante Forschungsdaten produziert, als man annehmen möchte.

[24] Bspw. DOI (Digital Object Identifier) oder URN (Uniform Ressource Name), vgl. <https://www.forschungsdaten.info/themen/veroeffentlichen-und-archivieren/persistente-identifikatoren/>.

[25] „Erhalten Metadaten immer die Aufmerksamkeit, die sie verdient haben,", fragt auch Ursula Winter in ihrem sehr empfehlenswerten Blog-Beitrag zu Metadaten in der Romanistik, und bringt das Problem schnell auf den Punkt: „oder wirken sie im Schatten der Forschungsleistung und der dafür erforderlichen zeitintensiven Datenerhebung, -aufbereitung und -auswertung eher wie ein notwendiges Übel, eine lästige Pflicht, die es auf dem Weg zur Datenpublikation eben schnell zu erfüllen gilt, um Forschungsförderer und Infrastruktureinrichtungen zufriedenzustellen?" (Winter 2022).

XML-Code – zudem eher vage bezüglich der Angaben, die man darin machen soll und darf. Selbst die FAIR-Prinzipien, die Metadaten essenzielle Bedeutung beimessen, beschränken sich im Grunde darauf, „rich metadata" (F2)[26] einzufordern, ohne zu spezifizieren, was Forschende sich darunter konkret vorzustellen hätten.[27] Es verwundert also wenig, dass häufig nur die absoluten Pflichtfelder bestückt werden.[28] Hier kann der FID unterstützen und tut es bereits, indem er die angesprochenen Probleme aufgreift und entsprechende Informationen für die Community aufbereitet. Sein Angebot lässt sich im Dialog mit der Fachcommunity jedoch noch erweitern.

Zahlreiche Institutionen und Initiativen bauen nun Beratungsstellen auf, um Forschende in allen Belangen rund um das Forschungsdatenmanagement zu unterstützen, darunter auch die Auszeichnung mit Metadaten für die anschließende Speicherung und Publikation. In der Regel sind diese institutionellen und/oder regionalen FDM-Stellen jedoch für eine Bandbreite an Disziplinen zuständig, die sich mit der Vielfalt der Fächer an den jeweiligen Institutionen deckt. Punktuelle, fachspezifische Expertise kann zufällig in den jeweiligen Teams vorhanden sein; mehrheitlich dürfte jedoch über die individuelle Vernetzung bzw. über die Weitervermittlung an Stellen mit dem benötigten Fachwissen nach Lösungen gesucht werden. Diese Beschreibung trifft auch auf die Romanistik zu. Treten Forschende mit fachspezifisch romanistischen Bedarfen an eine solche FDM-Stelle heran, muss diese sich womöglich selbst auf die Suche nach Informationen oder Kontakten begeben. Für die Romanistik ist der FID der richtige Ansprechpartner, der durch seine Ausrichtung die fachspezifische Perspektive einnimmt und sowohl inhaltlich zugeschnittene Informationen bereitstellt, als auch an konkrete Ansprechpersonen weitervermitteln kann. Das ist der Hintergrund, vor dem an dieser Stelle angeregt wird, eine Handreichung[29] für eine standardisierte und konsistente Metadatenvergabe in der Romanistik zu erarbeiten, die aktuelle Bedarfe des Fachs aufgreift und mit den Ansprüchen an ein nachhaltiges Forschungsdatenmanagement kombiniert.

[26] Vgl. <https://www.go-fair.org/fair-principles/f2-data-described-rich-metadata/>, wo immerhin ergänzend ausgeführt wird, Metadaten „can (and should be) generous and extensive, including descriptive information about the context, quality and condition, or characteristics of the data."

[27] Zur Ehrenrettung der FAIR-Prinzipien sei darauf verwiesen, dass das eine gewollte Unschärfe darstellt, um ihre Anwendung in Übereinstimmung mit den geltenden Standards und Traditionen der jeweiligen Fachcommunity zu ermöglichen.

[28] Zu diskutieren wäre freilich, ob die Einteilung in Pflicht- und optionale Felder nicht bereits das falsche Signal sendet, nämlich, dass erstere genügten. Vergleiche dazu den Best-Practice-Guide der IT-Gruppe Geisteswissenschaften der LMU München und des Leibniz-Rechenzentrums der Bayerischen Akademie der Wissenschaften, der für den internen Gebrauch die Standards des DataCite-Metadatenschemas stärker standardisiert durch weitere Pflichtfelder und Einschränkungen im erlaubten Vokabular (Schulz et al. 2020).

[29] Diese Handreichung kann perspektivisch ausgebaut werden zu einer Metadaten-Policy für die Romanistik. Den Wert einer klaren Datenpolicy auf den Weg zu FAIRen Daten stellt auch die FAIRsFAIR-Initiative fest: „FAIRsFAIR's landscape assessment found that data policies that are clear and easy to understand can positively influence researchers", vgl. <https://www.fairsfair.eu/fairsfair-deliverables-community-review> und im Detail Davidson et al. 2019.

Metadaten in der Romanistik

Einen wichtigen Grundstein hat die *AG Digitale Romanistik* 2017 vorgelegt mit dem Positionspapier zu „Open Access und Forschungsdaten", das im Mitteilungsheft des Deutschen Romanistenverbands veröffentlicht wurde.[30] Als spezifisch romanistisch wird darin vor allem die inhärent internationale Ausrichtung des Fachs genannt, die auch die Nachnutzung der Daten betreffe. Dies mache die Verwendung ebenso international verbreiteter Standards[31] „noch wichtiger als ohnehin schon" (Schöch et al. 2017, 56). Hinzu kommt die Mehrsprachigkeit des Fachs und seiner Forschungsansätze, woraus sich die zentrale Forderung nach mehrsprachigen Metadaten ergibt: „Metadaten, die die Forschungsdaten beschreiben und kontextualisieren, müssen mindestens auf Englisch, daneben auf Deutsch, in der Dokumentsprache und/oder anderen (romanischen) Sprachen vorliegen" (Schöch et al. 2017, 56).[32]

Keine Hinweise enthält das Papier bislang dazu, wie die Forderungen tatsächlich umgesetzt werden können oder wo diese mehrsprachigen Metadaten herkommen sollen. Auf diese offenen Fragen will auch dieser Beitrag keine abschließende Antwort geben, sondern vielmehr einen Diskussionsprozess innerhalb der Fachgemeinschaft anregen, um aus ihr heraus angemessene, bedarfsgelenkte und wissenschaftsgeleitete Lösungen zu generieren. Am Anfang dieses Prozesses sollte eine breite Erhebung[33] in der Fachcommunity stehen, die zunächst zweierlei erfasst.

1. Welche Metadaten werden bei der Veröffentlichung von Forschungsdaten in der Praxis verwendet bzw. für eine angemessene Beschreibung zusätzlich benötigt?[34]

[30] Die darin formulierten Empfehlungen wurden anschließend vom DRV aufgegriffen, vgl. die „Stellungnahme des DRV-Vorstands zum Positionspapier der AG Digitale Romanistik ‚Open Access und Forschungsdaten in der Romanistik'" (Deutscher Romanistenverband 2018).

[31] Als Beispiele werden die *Text Encoding Initiative* (<https://tei-c.org>), die *Dublin Core Metadata Initative* (<https://dublincore.org>) sowie die internationale Normdatei für Personen VIAF (Virtual International Authority File – <https://viaf.org>) genannt (vgl. Schöch et.al. 2017, 56).

[32] Der ebenso wichtigen Forderung nach einer Erhöhung der Auffindbarkeit durch Verzeichnung in internationalen Nachweissystemen kann in der Regel durch Auswahl eines geeigneten Forschungsdaten-Repositoriums und, wie ebenfalls im Papier angemerkt, durch die Vergabe eines Persistenten Identifikators wie DOI, der auch zentrale Metadaten enthält, entsprochen werden (Schöch et.al. 2017, 56-57).

[33] Anstelle der klassischen Umfrage, die bereits ein gewisses Fachwissen im Hinblick auf Schemata und Normdaten voraussetzt, böte sich hier vielleicht die Spielart der User Story bzw. des Use Cases an, bei der, kurz gesagt, die Adressat*innen einer Maßnahme ihre Anforderungen in möglichst einfacher Form formulieren. Erfolgreich zum Einsatz kam diese Methode beispielsweise im NFDI-Konsortium *Text+* oder bei der Entwicklung des Metadatenschemas für das Repositorium von *OstData*, vgl. Rißler-Pipka et al. 2021 bzw. Stanzel 2020. Dieses Vorgehen wird auch im Hinblick auf die Standardisierungsbestrebungen der *Nationalen Forschungsdateninfrastruktur* (NFDI) als möglicher Weg betrachtet: „Einige Konsortien nutzen Use Cases für den Umgang mit Metadaten und den geplanten Entwicklungen. Sie eignen sich als konkrete und anschauliche Beispiele in der Kommunikation zwischen den Forschenden; ein Einbeziehen von Gruppen mit unterschiedlich weit entwickelten Kenntnisständen und Fähigkeiten im Bereich Metadaten wird erleichtert." (Iglezakis et al. 2021, 132).

[34] Die bewusst offene Fragestellung soll widerspiegeln, dass in der umfassenderen Definition von Forschungsdaten auch Produkte des Forschungsprozesses wie Datenmanagementpläne, Software bzw. Code und Quellen enthalten sind.

2. Welche Arten von Forschungsdaten werden gesucht (und anhand welcher Metadaten würde danach recherchiert)?

Anhand der gewonnenen Einblicke ließe sich im Dialog mit den Romanist*innen ein Prozess anstoßen, um innerhalb der Fachcommunity eine Verständigung auf angemessene Standards und Praktiken zu befördern.[35] Darauf aufbauend wiederum sollte gemeinsam eine entsprechende Handreichung[36] konzipiert werden, die einen durchschnittlichen Ist-Zustand als Ausgangspunkt nimmt, um einen Soll-Zustand zu erreichen, welcher wiederum nicht allein von einer losgelösten Idealvorstellung und partikularen Perspektive diktiert wird.

Zu diskutieren wären etwa praktische Empfehlungen zum Ausfüllen von Metadatenfeldern, die üblicherweise nicht zu den Pflichtangaben[37] gehören, wie Schlagwörter oder die Verknüpfungsmöglichkeiten zu weiteren Ressourcen, etwa verwandten Datensätzen oder Text-Publikationen sowie verwendeter Software. Der große Wert von Best-Practice-Beispielen für eine aussagekräftige und die Auffindbarkeit unterstützende Gestaltung von Titel und Beschreibung oder *Abstract* ist ebenso unbestritten wie die Notwendigkeit, Hinweise zu geeigneten kontrollierten Vokabularen und deren konkreter Verwendung für die Verschlagwortung bereitzustellen.

Dennoch sollten mehrere weitere Ebenen reflektiert werden, die gerade im Hinblick auf Nachhaltigkeit und Anschlussfähigkeit der Romanistik wichtig sind. Das betrifft zum einen die Erhebung und Dokumentation der in den jeweiligen Teildisziplinen tatsächlich verbreiteten und genutzten Konventionen, Ressourcen und Infrastrukturangebote. Ein weiterer Ansatzpunkt besteht darin, eine solche Empfehlung für die Vergabe von Metadaten organisch mit den wichtigsten bestehenden Anforderungen und Empfehlungen von Förderinstitutionen wie der DFG[38] und einschlägigen Initiativen im Umfeld des Forschungsdatenmanagements, wie FAIR[39], zu bündeln und zu verweben. Nicht zuletzt sind die zahlreichen

[35] In einem sich daran anschließenden Schritt wäre dann zu diskutieren, wie und in welchem Maße das Thema Metadaten – als wesentlicher Aspekt im Umgang mit Forschungsdaten – in den Curricula zu verankern ist.

[36] Diese Handreichung soll sich im Wesentlichen auf jene Metadaten konzentrieren, die primär für die Auffindbarkeit von Forschungsdaten eine Rolle spielen. Dabei sind die Grenzen zu stärker inhaltlichen und disziplinspezifischen Metadaten selbstverständlich fließend – je spezifischer die Recherche, desto granularer müssen die dabei zu berücksichtigenden Metadaten sein und desto spezialisierter auch die speichernde Infrastruktur. Statt des Versuchs einer problematischen scharfen Kategorisierung von Metadaten sei hier in enger Anlehnung an das erste Metadata-Prinzip der Research Data Alliance behauptet: „The only difference between metadata […] is mode of use“, vgl. <https://rd-alliance.org/metadata-principles-and-their-use.html>.

[37] Inwiefern es solche gibt und um welche es sich handelt hängt natürlich stark vom jeweiligen Kontext und dem zugrundeliegenden Metadatenschema ab. Diese Informationen müssten daher ebenfalls Teil der Erhebung sein, auch indirekt über die Benennung verwendeter Repositorien, um in der Handreichung so generisch wie möglich, so spezifisch wie nötig vorgehen zu können, wenn dieses Anlehnen gestattet ist.

[38] Hier sind zweifelsohne die Leitlinien zur Sicherung guter wissenschaftlicher Praxis zu nennen, wie sie im gleichnamigen Kodex der DFG festgehalten und an zahlreichen Institutionen zur Grundlage für eigene Datenpolicies gemacht wurden (vgl. Deutsche Forschungsgemeinschaft 2019 sowie die „dritte Ebene“ des Kodex unter <https://wissenschaftliche-integritaet.de/>). Aber auch die Vorgaben weiterer nationaler und internationaler Förderinstitutionen sollten berücksichtigt werden.

[39] Schon allein die große inhaltliche Kontiguität sowie die weite Verbreitung gebieten es, die FAIR zugrundeliegenden Prinzipien mit zu berücksichtigen. Während einige der eher technisch-administrativen Aspekte, etwa die Vergabe eines Persistenten Identifikators oder die Zuweisung einer klaren Nutzungslizenz, häufig

Öffnungsimpulse zu erwägen, die unter „Open Science“ gefasst werden, während selbstverständlich auch das romanistische Ausland und etwaige dortige Initiativen[40] nicht übersehen werden dürfen.

Wichtigstes Ziel der Handreichung bleibt die Hilfestellung[41] für Forschende in der Romanistik, um die eigenen Daten bei der Publikation mithilfe von Metadaten möglichst FAIR und somit auffindbar und nachvollziehbar zu machen. Dabei bleibt der Dialog mit den lokalen FDM-Stellen[42] und den Infrastrukturanbietern[43] stets im Blick.

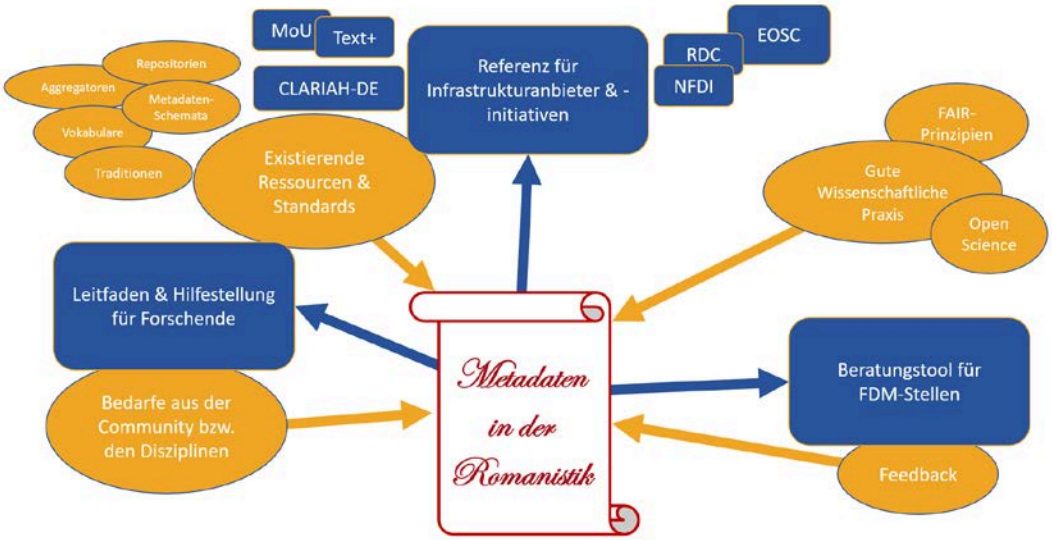
über die verpflichtenden Angaben in einem Repositorium sozusagen automatisch abgedeckt sind, gibt es noch ausreichende Gelegenheit für die Datengebenden, durch simple Angaben in den repositoriumsseitig optionalen Metadaten – etwa den Persistenten Identifikator einer mit dem Datensatz zusammenhängenden Publikation, Software, oder weiterer damit in Beziehung stehender Datensätze sowie die Verwendung normierter Schlagwörter – die Qualität und FAIRness der eigenen Daten zu erhöhen. Und mit erhöhter FAIRness einher gehen wohlgemerkt die Auffindbarkeit und Sichtbarkeit der eigenen Daten. Die fachliche Ausrichtung einer romanistischen Metadatenpolicy ermöglichte – im konkreten Fall der FAIR-Prinzipien – durch spezifische Anwendungshandreichungen den aufgrund ihres übergreifenden und disziplinagnostischen Anspruchs teilweise bewusst allgemein gehaltenen Charakter der FAIR-Prinzipien abzustreifen und sie dadurch zugänglicher zu machen. Zu prüfen wäre darüber hinaus auch eine Einbindung der CARE-Prinzipien, vgl. <https://www.gida-global.org/care bzw. Carrol et. al. 2020>.

[40] Exemplarisch seien etwa Huma-Num genannt (<https://www.huma-num.fr/>), die zentrale französische Infrastruktur für Forschungsdaten, sowie, auf einer höheren Ebene, die europäischen Infrastruktur-Konsortien CLARIN bzw. DARIAH, vgl. <https://www.clarin.eu/ bzw. https://www.dariah.eu/>.

[41] Auch könnten grundlegende Anregungen für den Einsatz eindeutiger Identifikatoren wie ORCID (<https://orcid.org/>) oder ROR (<https://ror.org/>) gegeben sowie eine Orientierung für eine einheitliche Zitierweise von Forschungsdaten erarbeitet werden.

[42] So empfiehlt beispielsweise die FAIRsFAIR-Initiative ausdrücklich, dass neben Forschenden auch die „Data Stewards“ von entsprechende Empfehlungen profitieren könnten: „Provide practical guidance to researchers and data stewards on how to implement FAIR within different domains – specifically on how to describe data using appropriate metadata standards, data tags and ontologies.“ Davidson et al. 2020, 7.

[43] Auch hier besteht vermutlich keine geringfügige Erfahrung, welche Aspekte im Umgang mit Metadaten noch besonderer Aufmerksamkeit bedürfen. Andererseits stellen Infrastrukturen wie Repositorien ihrerseits die entsprechenden Metadatenschemata und Eingabeformulare zur Verfügung, was aus einer pragmatischen Sicht ebenfalls in der Handreichung berücksichtigt werden sollte – ohne gleichzeitig die Wechselwirkung unterbinden zu wollen, dass die Anforderungen aus der Romanistik auch als Impuls genutzt werden dürfen, die erwähnten Schemata, Formulare und Softwarelösungen zu aktualisieren. Zu denken ist hier an *CLARIAH-DE*, aber auch deutlich in Richtung des NFDI-Konsortiums *Text+*, das quasi das Tor in die *Nationale Forschungsdateninfrastruktur* und, darüber hinaus, die *European Open Science Cloud* (EOSC) darstellt. Die Informationen aus der Handreichung würden helfen, die Belange der Romanistik unmittelbarer in die Bemühungen um Standardisierung und Interoperabilität einfließen zu lassen.



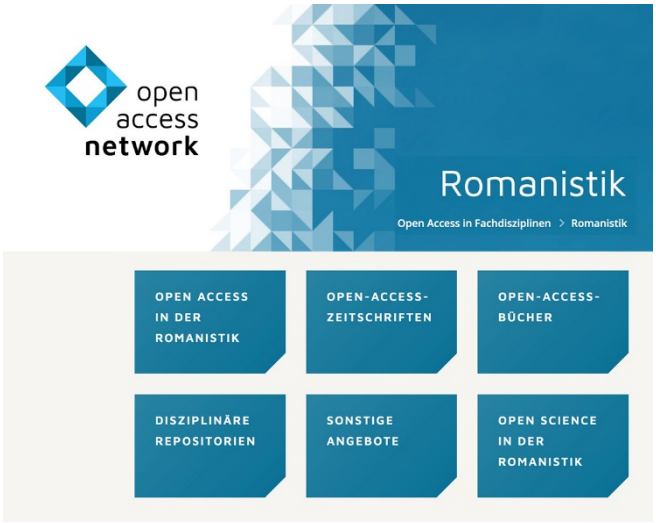
2| Sammlung möglicher Aspekte einer Metadaten-Policy in der Romanistik

Bei seinem kontinuierlichen Ausbau des Informationsangebots im Bereich Forschungsdatenmanagement berücksichtigt der FID Romanistik diese vielfältigen Perspektiven und sieht es auch weiterhin als seine Aufgabe an, bei Bedarf zwischen der romanistischen Fachcommunity und den Akteuren in der Infrastruktur zu vermitteln und so dazu beitragen, Hürden bei der Nutzung generischer Angebote abzubauen. In Kooperation mit der *AG Digitale Romanistik* bietet der FID daher an, eine Diskussion rund um eine konsistente Metadatenvergabe in der Romanistik konstruktiv zu begleiten und die Ergebnisse des Abstimmungsprozesses in einer Handreichung zu erfassen.

3. Die Bedeutung von Open Access in der Romanistik

Die Romanistik ist geprägt von einer heterogenen Publikationskultur, in der die Veröffentlichung von Forschungsergebnissen in gedruckten Monographien, Sammelbänden und Zeitschriften einen hohen Stellenwert einnimmt. In zunehmenden Maße wächst aber auch in der Romanistik das Interesse und die Notwendigkeit, wissenschaftliche Veröffentlichungen in elektronischer Form vorzunehmen, sei es als Erst- oder Zweitveröffentlichung von Monographien im Open Access (OA) oder in OA-Zeitschriften. In Abstimmung mit den romanistischen Fachverbänden und dem wissenschaftlichen Beirat des FID[44] hat sich der Fachinformationsdienst der Aufgabe angenommen, sowohl eine Beratung und Information zum Thema OA anzubieten als auch mit der Einrichtung des Romanistik-Repositoriums einen zentralen Veröffentlichungs- und Verzeichnisort für romanistische Publikationen im Open Access zu schaffen. Die bestehenden und neu eingerichteten Angebote hierzu werden im Folgenden ausgeführt.

[44] <https://fid-romanistik.de/ueber-uns/wissenschaftlicher-beirat>.



3 | Dossier Open Access in der Romanistik

Wer einen kompakten Überblick zur Bedeutung von Open Access in der Romanistik bekommen möchte, sei auf das umfangreiche Online-Dossier Romanistik[45] verwiesen, das der FID auf Einladung des BMBF geförderten Projektes *open-access.network* für die Sektion „Open Access in Fachdisziplinen" verfasst hat. Dort wird nicht nur die aktuelle Situation von Open Access in der Romanistik beschrieben, die der FID im Rahmen von Workshops und Umfragen untersucht hat, sondern es wird eine Übersicht zu OA-Zeitschriften, OA-Büchern, disziplinären Repositorien und zu Open Science in der Romanistik geboten. Um eine Dopplung der Information zu vermeiden, wird an dieser Stelle auf das vom FID erstellte Dossier verwiesen.

Information und Beratung zu Open Access im FID

Aus der Wissenschaft wird immer wieder an uns herangetragen, wie wichtig es für die Forschenden in der Romanistik ist, zu den forschungsnahen Dienstleistungen und Infrastrukturen Informationen vom FID zu bekommen. Genauso, wie wir das zum Thema Forschungsdatenmanagement tun, nehmen wir uns natürlich auch der Aufgabe an, die Fachcommunity zum Thema Open Access zu informieren. Dazu zählt nicht nur das gebündelte Informations- und persönliche Beratungsangebot, dargestellt auf der Website[46], sondern auch die kontinuierliche Information im Romanistik-Blog des FID. Open Access ist ein sehr volatiles Thema. Deshalb werden in einer eigens eingerichteten Kategorie „Open Access"[47] konkrete Angebote aus dem Open-Access-Bereich in Einzeldarstellungen vorgestellt.

[45] <https://open-access.network/informieren/open-access-in-fachdisziplinen/romanistik> Christoph Hornung, FID Romanistik (Stand: Dezember 2021)
[46] <https://fid-romanistik.de/open-access>.
[47] <https://blog.fid-romanistik.de/category/openaccess/>.

Romanistik-Repositorium als zentrale Anlaufstelle bei der Suche nach wissenschaftlicher Literatur im Open Access

Zu einem der Punkte des eingangs genannten OA-Dossiers, den disziplinären Repositorien, hat der FID auf ein Desiderat der Wissenschaft reagiert. Bis dato gab es noch kein spezifisches romanistisches Fachrepositorium. Deshalb hat der FID in Abstimmung mit den Fachverbänden und dem wissenschaftlichen Beirat ein solches Repositorium eingerichtet. Das Romanistik-Repositorium bietet nun die Möglichkeit für Romanist*innen, Forschung als Erst- oder Zweitveröffentlichung an zentraler Stelle zu publizieren. Die Zweitveröffentlichung einer Publikation ist möglich, wenn bei der Erstveröffentlichung dem Verlag kein ausschließliches, sondern ein einfaches Nutzungsrecht eingeräumt worden ist oder wenn die Autor*innen sich das Recht auf parallele Onlineveröffentlichung ausdrücklich vorbehalten haben. Das Repositorium wurde basierend auf der Software *DSpace* eingerichtet und der FID wird die romanistische Wissenschaft informieren, sobald Texte zur Publikation eingereicht werden können. Interessent*innen können sich selbstverständlich schon jetzt beim FID melden.

Verzeichnung bestehender Angebote von Fachverlagen mit OA-Publikationen

Aus den Erfahrungen des OA-Workshops hat der FID den Wunsch der Forschenden zur Kenntnis genommen, dass viele zwar gerne im Open Access publizieren möchten, dies aber bei den Verlagen tun wollen, mit denen sie bisher vertrauensvoll zusammengearbeitet haben und bei denen sie schon in der Vergangenheit publiziert haben. Zahlreiche Verlage haben reagiert und entsprechende Angebote zur OA-Publikation eingerichtet. Logischerweise sollen und werden auch diese romanistischen Veröffentlichungen im Romanistik-Repositorium aufgenommen. Der FID steht dazu seit Jahren mit mehreren Verlagen im Kontakt. Da das Romanistik-Repositorium nun eingerichtet ist, wird die Integration von OA-Publikationen deutscher Verlage mit romanistischem Programm in das Fachrepositorium gerade erarbeitet.

Integration von OA-Zeitschriften

Auch mit romanistischen OA-Zeitschriften steht der FID im Kontakt, deren Angebote genauso wie die Verlagspublikationen in das Fachrepositorium integriert werden sollen. Der FID berät und unterstützt auch OA-Zeitschriften-Neugründungen und Transformationen von Zeitschriften in den Open Access.

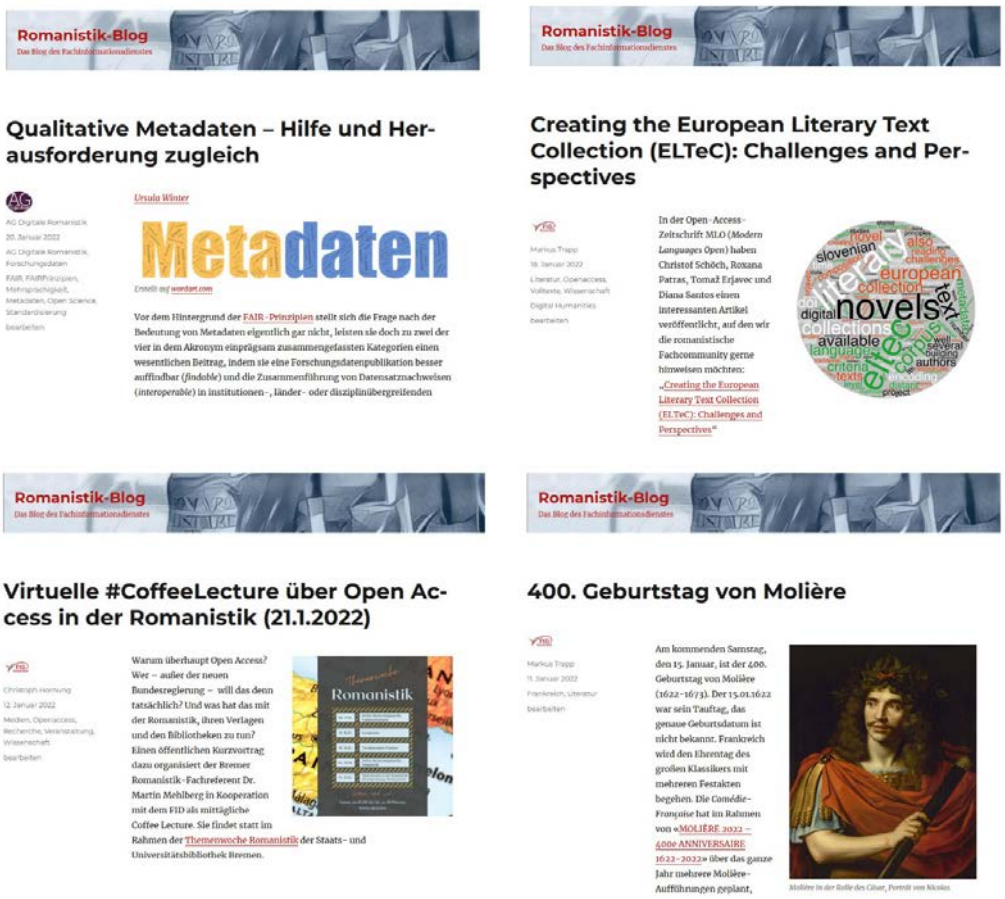
Zusammenfassend lässt sich sagen, das Romanistik-Repositorium soll zentrale Anlaufstelle bei der Suche nach wissenschaftlicher Literatur im Open Access werden und wird damit zur weiteren Vernetzung digitaler Ressourcen beitragen. Interessenten können sich schon jetzt beim FID melden. Sobald das Romanistik-Repositorium mit ersten Veröffentlichungen befüllt ist, wird es über die Website des FID erreichbar sein, in das Suchportal des FID integriert werden und auf den üblichen Kommunikationswegen der Fachcommunity vorgestellt werden.

4. Neue Formen der Kommunikation in der Romanistik

In seinem Romanistik-Blog informiert der FID fortlaufend über eigene Angebote und bringt die Community bei den volatilen Themen Forschungsdaten, Open Access und Digital Humanities zeitnah und für die Romanistik speziell aufbereitet mit Fachartikeln und Tutorials auf den neusten Stand der Entwicklungen im digitalen Raum. Die *AG Digitale Romanistik* hat im Blog eine viel beachtete Reihe zu den FAIR-Prinzipien veröffentlicht. Im Folgenden wird aufgezeigt, wie das Blog und der dazugehörige Twitter-Kanal des FID dazu genutzt werden, sowohl die intra-disziplinäre als auch die interdisziplinäre Vernetzung und Sichtbarmachung romanistischer Forschung über das eigene Fach hinaus voranzutreiben.

Romanistik-Blog, das Blog des FID

Im Austausch mit der Wissenschaft wird dem FID immer wieder bescheinigt, dass er für die romanistische Community eine wichtige Quelle der Information ist. Dies gilt nicht nur für die Recherche nach Fachmedien, sondern auch für Informationen zu forschungsrelevanten Themen wie Forschungsdatenmanagement und Open Access. Zusätzlich zu den gebündelten Informationen zu diesen Themengebieten auf der Website liefert der FID Neuigkeiten von romanistischer Relevanz im Romanistik-Blog.

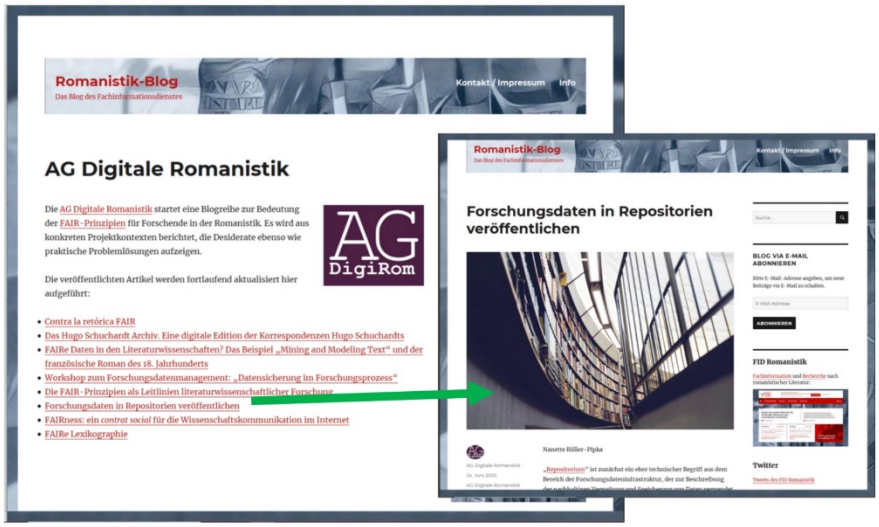


4 | Screenshots Blogartikel Romanistik-Blog

Die zentralen Themen werden in eigenen Kategorien angeboten, in denen sich die Wissenschaft punktgenau auf dem Laufenden halten kann: Forschungsdaten[48] und Open Access[49]. In diesen beiden Gebieten, die ja auch eng verbunden sind mit den Digital Humanities, gibt es kontinuierlich neue Informationen für die Fachwissenschaft, die der FID für die Romanistik entsprechend aufbereitet. Darüber hinaus berichtet das FID-Team im Romanistik-Blog von neuen FID-Angeboten und neu eingespielten Datenquellen in das Suchportal des FID. Passende Tutorials, geeignet zur Nachnutzung in der eigenen Forschung und Lehre, werden auf diesem Weg präsentiert. Zudem besteht das gut genutzte Angebot, dass Wissenschaftler*innen und Institutionen im Rahmen von Gastartikeln Themen mit romanistischem Bezug vorstellen.

Bei der Frage, wie man über neue Artikel im Romanistik-Blog informiert werden kann, gibt es mehre Optionen. Zum einen werden die neuen Artikel, die in einem Rhythmus von ca. einem Artikel pro Woche veröffentlicht werden, auf der Startseite des FID verlinkt. Das in der Open-Source-Software *WordPress* realisierte und bei der SUB Hamburg gehostete Blog des FID kann selbstverständlich per RSS-Feed abonniert werden. Zudem gibt es die von vielen romanistisch Forschenden genutzte Funktion, das Blog auch via E-Mail zu abonnieren. Die entsprechenden Links dazu befinden sich im Blog. Bei Fragen hierzu hilft das Team des FID.

Blogreihe AG Digitale Romanistik



5 | Blogreihe AG Digitale Romanistik

Die *AG Digitale Romanistik* publiziert im Romanistik-Blog im Rahmen einer Blogreihe zur Bedeutung der FAIR-Prinzipien für Forschende in der Romanistik. Hier wird aus konkreten Projektkontexten berichtet, die Desiderate ebenso wie praktische Problemlösungen aufzeigen. Die veröffentlichten Artikel werden fortlaufend aktualisiert auf einer eigens eingerichteten Unterseite[50] im Überblick

[48] <https://blog.fid-romanistik.de/category/forschungsdaten/>.
[49] <https://blog.fid-romanistik.de/category/openaccess/>.
[50] <https://blog.fid-romanistik.de/ag-digitale-romanistik/>.

aufgeführt. Dabei ist bereits ein beeindruckendes Kompendium zu den FAIR-Prinzipien (*Findable – Accessible – Interoperable – Re-usable*) zustande gekommen, das gerade durch die Bezugnahme auf die romanistische Forschung eine wertvolle Informationsquelle für die Fachcommunity darstellt.

6 | Twitter-Profil FID Romanistik

Der Twitter-Kanal des FID

Vernetzung innerhalb der romanistischen Fachgemeinschaft, aber auch mit Blick auf die interdisziplinäre Zusammenarbeit innerhalb der Geisteswissenschaften und darüber hinaus, ist von großer Bedeutung. Auf digitalen Wegen geschieht dies schon gut über das Romanistik-Blog, aber noch stärker wird es über den Twitter-Account des FID betrieben. Unter @FIDRomanistik[51] twittert das Team des FID fortlaufend und tagesaktuell zu romanistischen Themen. Seit Anfang 2022 folgen dem Account über 1.000 Personen und Einrichtungen. Der Twitter-Kanal des FID kann auch gelesen werden, wenn man nicht bei dem Dienst Twitter angemeldet ist. Die Tweets sind zudem auf der Kontaktseite des FID[52] eingebunden. Auf Twitter werden nicht nur die neuen Artikel im Romanistik-Blog vorgestellt, sondern auch Entwicklungen und Angebote einzelner romanistischer Institute und der Romanistik nahestehender Einrichtungen kommuniziert. Nicht erst seit den Zeiten der Pandemie weist der FID hier auf Online-Angebote von Universitäten, Verlagen und Kultureinrichtungen hin. Die Vernetzung über das enger gefasste Feld der Romanistik und die wichtigen intradisziplinären Kontakte hinaus bietet eine Chance, die der Fachinformationsdienst gerne und erfolgreich nutzt: die Sichtbarmachung romanistischer Forschung über das eigene Fach hinaus. Dies kommt dem Bestreben der Romanistik, die internationale Zusammenarbeit zu stärken, entgegen. Und es dient dazu, gemeinsam mit der Fachwissenschaft die Rolle der FIDs im Kontext der *Nationalen Forschungsdateninfrastruktur* (NFDI) herauszustellen.

[51] <https://twitter.com/FIDRomanistik>.
[52] <https://fid-romanistik.de/kontakt>.

5. Zusammenfassung

Wie dargestellt, unterstützt das vorhandene Angebot des FID eine transdisziplinär ausgerichtete Romanistik und es soll weiterhin in diese Richtung ausgebaut werden. Die enge Anbindung an die Fachcommunity ist hierfür wegweisend. Aufgezeigt wurden auch die zahlreichen Möglichkeiten, wie Forschende und Interessierte sich mit dem FID vernetzen bzw. das FID-Team kontaktieren können – und dies auch gerne wahrnehmen sollen. Der FID versteht sich als Anbieter und Vermittler von zeitgemäßer Wissenschaftskommunikation, sei es über die Website, das Blog oder die Social-Media-Kanäle. Neben den Grundaufgaben der Literaturversorgung und dem Angebot entsprechender Nachweis- und Recherchesysteme, macht der FID der romanistischen Fachcommunity Informationsangebote zu forschungsnahen Dienstleistungen.

Der FID Romanistik lädt alle Lesenden dieses Artikels ein, sich mit dem FID zu vernetzen. Fragen und Anregungen zu den neuen Formen der Kommunikation in der Romanistik sowie zu den Themen Open Access und Forschungsdatenmanagement sind jederzeit willkommen.

Bibliografie

BURGER, Marleen, Anette Cordts & Ted Habermann. 2021. „Wie FAIR sind unsere Metadaten? Eine Analyse der Metadaten in den Repositorien des TIB-DOI-Services“. *Bausteine Forschungsdatenmanagement* 3 (Dezember). 1–13. <https://doi.org/10.17192/bfdm.2021.3.8351>.

CARROLL, Stephanie et. al. 2020. „The CARE principles for Indigenous data governance“. *Data Science Journal 19*. <https://doi.org/10.5334/dsj-2020-043>.

DAVIDSON, Joy et al. 2019. „D3.1 FAIR Policy Landscape Analysis (v1.0)“. Zenodo- <https://doi.org/10.5281/zenodo.5537032>.

DAVIDSON, Joy et. al. 2020. „D3.3 Policy Enhancement Recommendations“. Zenodo. <https://doi.org/10.5281/zenodo.5362183>.

DEUTSCHE FORSCHUNGSGEMEINSCHAFT. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis. Kodex.* Bonn. <https://doi.org/10.5281/zenodo.3923602>.

DEUTSCHER ROMANISTENVERBAND. 2018. „Stellungnahme des DRV-Vorstands zum Positionspapier der AG Digitale Romanistik ‚Open Access und Forschungsdaten in der Romanistik‘“, *Mitteilungsheft des Deutschen Romanistenverbands*, Herbst 2018, 25–30. <https://www.deutscher-romanistenverband.de/wp-content/uploads/2019/01/DRV-Mitteilungsheft-Herbst-2018.pdf#page=25> 31.1.2022.

DIERKES, Jens. 2021. „Planung, Beschreibung und Dokumentation von Forschungsdaten“. *Praxishandbuch Forschungsdatenmanagement*, ed. Putnings, Markus, Heike Neuroth & Janna Neumann, 303–326, Berlin, Boston: De Gruyter Saur. <https://doi.org/10.1515/9783110657807-018>.

ERBEN, Maria, Doris Grüter & Jan Rohden. 2018: *Forschungsdatenmanagement in der Romanistik. Aktuelle Situation und zukünftige Perspektiven*. Bonn. Fachinformationsdienst Romanistik. <http://hdl.handle.net/20.500.11811/1178>.

EUROPÄISCHE KOMMISSION. „Guidelines: Open access to publications and

research data in Horizon 2020“.
<https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm>, 31.1.2022.

HIGGINS, Sarah. 2008. „The DCC Curation Lifecycle Model“. *International Journal of Digital Curation* 3 (1), 134–140.
<https://doi.org/10.2218/ijdc.v3i1.48>.

HIGMAN, Rosie, Daniel Bangert & Sarah Jones. 2019: „Three Camps, One Destination: The Intersections of Research Data Management, FAIR and Open“. *Insights* 32: 18, 1–9.
<https://doi.org/10.1629/uksg.468>.

HORNUNG, Christoph & Jan Rohden. 2019. „Der Beitrag des Fachinformationsdienstes Romanistik zur romanistischen Digitalkultur.“ In: *Digitalkulturen/ Cultures numériques. Herausforderungen und interdisziplinäre Forschungsperspektiven/Enjeux et perspectives interdisciplinaires*, ed. Montemayor, Julia, Vera Neusius & Claudia Polzin-Haumann. 51-76, Bielefeld: transcript.
<https://doi.org/10.14361/9783839442159-003>; zweitveröffentlicht unter:
<https://blog.fid-romanistik.de/2020/02/10/publikation-des-fid-romanistik-zur-romanistischen-digitalkultur-in-freiem-zugang/>, 31.1.2022.

IGLEZAKIS, Dorothea et al. 2021. „Interoperabilität von Metadaten innerhalb der NFDI: Konsortienübergreifender Metadaten-Workshop am 2./3. Juli 2020“. *Bausteine Forschungsdatenmanagement* 2 (Juli). 124–35.
<https://doi.org/10.17192/bfdm.2021.2.8313>.

KREFELD, Thomas & Stephan Lücke. 2020. „FAIRness: ein *contrat social* für die Wissenschaftskommunikation im Internet“. *Romanistik-Blog*, 16.5.2022.
<https://blog.fid-romanistik.de/2020/05/16/fairness-ein-contrat-social-fuer-die-wissenschaftskommunikation-im-internet/>, 31.1.2022.

MATHIAK, Brigitte, & Simone Kronenwett. 2017. „Umfrage zu Forschungsdaten an der Philosophischen Fakultät der Universität zu Köln“. *DHd 2017 – Digitale Nachhaltigkeit. 4. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“*. Bern.
<https://doi.org/10.5281/zenodo.4622695>.

RIßLER-PIPKA, Nanette et al. 2021. „Community Involvement in Research Infrastructures: The User Story Call for Text+“. Zenodo.
<https://doi.org/10.5281/zenodo.5384085>.

SCHÖCH, Christof et al. 2017. „Open Access und Forschungsdaten. Ein Positionspapier der AG Digitale Romanistik“. *Mitteilungsheft des Deutschen Romanistenverbands*, Frühjahr 2017. 50–59.
<https://doi.org/10.5281/zenodo.4516104>.

RÜMPEL, Stefanie. 2011. „Der Lebenszyklus von Forschungsdaten“. *Handbuch Forschungsdatenmanagement*, ed. Büttner, Stephan, Hans-Christoph Hobohm & Lars Müller, 25–34. BOCK + HERCHEN Verlag, Bad Honnef.
<https://doi.org/10.34678/opus4-208>.

SCHULZ, Julian et al. 2020. „Standardisierung eines Standards: Warum und wie ein Best-Practice-Guide für das Metadatenschema DataCite entstand“. *Korpus im Text*, 20.1.2020.
<http://www.kit.gwi.uni-muenchen.de/?p=42800&v=1>, 31.1.2022.

STANZEL, Arnost. 2020. „OstData Erfahrungsberichte: User Stories als Methode zur Entwicklung eines Metadatenschemas für Forschungsdaten“, *OstBib*, 12.03.2020.
<https://ostbib.hypotheses.org/3839>, 31.1.2022.

TRAPP, Markus & Johannes von Vacano. 2021. „Präsentationsfolien: (FAIRe) Forschungsdaten, Open Access und neue Formen der Kommunikation in der Romanistik – Beiträge des FID zur Gestaltung des digitalen Wandels (XXXVII Romanistentag, 5.10.2021)“. Zenodo.
<https://doi.org/10.5281/zenodo.5548234>.

von Vacano, Johannes. 2020. „Wohin mit romanistischen Forschungsdaten? Teil 1: Zenodo“, *Romanistik-Blog*, 15.5.2020. <https://blog.fid-romanistik.de/2020/05/15/wohin-mit-romanistischen-forschungs-daten-teil-1-zenodo/>, 31.1.2022.

WILKINSON, Mark D. et al. 2016. „The FAIR Guiding Principles for scientific data management and stewardship“. *Scientific Data* 3. <https://doi.org/10.1038/sdata.2016.18>.

WINTER, Ursula. 2022. „Qualitative Metadaten – Hilfe und Herausforderung zugleich“. *Romanistik-Blog*, 20.01.2022. <https://blog.fid-romanistik.de/2022/01/20/qualitative-metadaten-hilfe-und-herausforderung-zugleich/>, 31.1.2022.

Genannte Websites

(letzter Abruf: 31.1.2022)

<https://blog.fid-romanistik.de/ag-digitale-romanistik/>.
<https://blog.fid-romanistik.de/category/forschungsdaten/>.
<https://blog.fid-romanistik.de/category/openaccess/>.
<https://d-nb.info/gnd/4047394-6>.
<https://d-nb.info/gnd/4165338-5>.
<https://dublincore.org>.
<https://fid-romanistik.de/fileadmin/user_upload/dokumente/Texte/Anleitung_Meldesystem_Forschungsdaten_Tools.pdf>.
<https://fid-romanistik.de/forschungsdaten/>.
<https://fid-romanistik.de/forschungsdaten/arbeit-mit-forschungsdaten/grundlegendes-zum-forschungsdatenmanagement-in-der-romanistik#c2602>.
<https://fid-romanistik.de/forschungsdaten/arbeit-mit-forschungsdaten/sichern-und-publizieren-von-forschungsdaten#c2586>.
<https://fid-romanistik.de/forschungsdaten/was-sind-romanistische-forschungsdaten>.
<https://fid-romanistik.de/forschungsdaten/workshops/>.
<https://fid-romanistik.de/kontakt>.
<https://fid-romanistik.de/open-access>.
<https://fid-romanistik.de/recherchewerkzeuge/online-tutorials>.
<https://fid-romanistik.de/ueber-uns/wissenschaftlicher-beirat>.
<https://open-access.network/informieren/open-access-in-fachdisziplinen/romanistik>.
<https://orcid.org/>.
<https://rd-alliance.org/metadata-principles-and-their-use.html>.
<https://ror.org/>.
<https://search.datacite.org/>.
<https://tei-c.org>.
<https://twitter.com/FIDRomanistik>.
<https://viaf.org>.
<https://wissenschaftliche-integritaet.de/>.
<https://www.base-search.net/>.
<https://www.clarin.eu/>.
<https://www.dariah.eu/>.
<https://www.dnb.de/gnd>.
<https://www.fairsfair.eu/fairsfair-deliverables-community-review>.
<https://www.forschungsdaten.info/themen/informieren-und-planen/datenlebenszyklus/>.
<https://www.forschungsdaten.info/themen/veroeffentlichen-und-archivieren/persistente-identifikatoren/>.
<https://www.gida-global.org/care>.
<https://www.go-fair.org/fair-principles/f2-data-described-rich-metadata/>.

<https://www.huma-num.fr/>.
<https://www.openaire.eu/>.
<https://www.romanistik.de/res>.

Zusammenfassung

Der DFG-geförderte Fachinformationsdienst Romanistik (FID) wird von der SUB Hamburg und der ULB Bonn betrieben und bietet laufend weiterentwickelte Services für die Bereitstellung und Recherche von Literatur für die Forschung sowie ein reichhaltiges fachspezifisches Online-Informationsangebot. Der Beitrag illustriert, wie der FID mit seinen Teilprojekten Forschungsdatenmanagement (FDM) und Open-Access-Publizieren (OA) sowie mit den neuen Kommunikationskanälen eine transdisziplinäre digitale Romanistik unterstützt.

Zunächst wird argumentiert, dass FDM nach den FAIR-Prinzipien Transdisziplinarität immer schon mitdenkt. Nach einer Vorstellung der entsprechenden Unterstützungsangebote des FID wird auf die Bedeutung einer konsistenten Metadatenvergabe eingegangen und eine breitere fachinterne Diskussion zu diesem Thema angeregt.

Als nächstes wird das Informationsangebot des FID zum Thema OA vorgestellt sowie ein Ausblick auf das gerade eingerichtete Repositorium für romanistische OA-Publikationen geboten, das als zentrale Anlaufstelle für frei zugängliche fachspezifische Erst- und Zweitveröffentlichungen konzipiert ist.

Abschließend werden mit dem Romanistik-Blog und dem Twitter-Kanal des FID zwei Angebote präsentiert, die neben der Information über aktuelle Entwicklungen in relevanten Bereichen auch die Vernetzung innerhalb und außerhalb der Fachcommunity unterstützen.

Abstract

The Specialised Information Service for Romance Philology (FID), run by Hamburg State and University Library and Bonn University and State Library with funding from the German Research Foundation, offers a suite of constantly improved services for finding and accessing literature for research along with rich subject-specific information. The article highlights how the FID-projects revolving around research data management (RDM) and open access (OA) publication practices as well as its use of modern communication technology support transdiciplinary digital Romance studies.

Firstly, it will be argued that RDM following the FAIR Principles always includes a transdisciplinary perspective, followed by a swift introduction to FID services in support of FAIR RDM. Outlining the importance of consistently attributed metadata, it is suggested a broader discussion of this topic within the Romance studies community be initiated.

This is followed by a brief introduction to the extensive information regarding OA the FID provides as well as a preview of the FID's recently established repository for

OA publications from the romance community which is designed to function as central access point for freely accessible academic publications (previously published or unpublished).

The article concludes with a presentation of the FID's own blog (Romanistik-Blog) and Twitter channel which not only supply information regarding current developments in relevant areas but also supports networking within the romance studies community and beyond.

Dossier

Digital, global, transdisziplinär: Impulse für die Romanistik

Bild: Computergeneriertes Bild (DreamStudio) nach Eugène Delacroix: Prompt "a man repairing a computer" (CC0 1.0)

Teil 3

Computerlinguistik und Sprachdaten

apropos
[Perspektiven auf die Romania]

Winter 2022

9

www.apropos-romania.de

ISSN: 2627-3446

2022, n°9
pp. 187-212
doi: https://doi.org/10.15460/apropos.9.1877

Beatrice Colcuc & Anna Rodella

Con parole tue

Dai parlanti a VerbaAlpina attraverso il crowdsourcing

Beatrice Colcuc
è ricercatrice presso VerbaAlpina, Ludwig-Maximilians-Universität München.
beatrice.colcuc@lmu.de

Anna Rodella
è stata collaboratrice presso VerbaAlpina, Ludwig-Maximilians-Universität München.
arodella@live.it

Parole chiave

VerbaAlpina – Crowdsourcing – Digital Humanities – Geolinguistica – Digitalizzazione

1. Introduzione

Con i suoi 1300 km di distensione territoriale, la catena delle Alpi rappresenta il sistema montuoso più importante d'Europa. Dal punto di vista topografico ed etnografico, le Alpi appaiono sostanzialmente compatte: cultura, tradizioni, flora, fauna e formazioni paesaggistiche si ritrovano lungo tutto il crinale alpino dalla Francia alla Slovenia passando per la Svizzera, il Liechtenstein, l'Austria, la Germania e l'Italia. Decisamente più eterogenea si mostra invece la situazione linguistica, legata non solo alle lingue nazionali (francese, tedesco, italiano, retoromanzo e sloveno), ma soprattutto relativa all'assetto dialettale dell'area. La varietà di idiomi parlati nel territorio alpino è straordinariamente grande e non solo ogni regione o ogni provincia, ma spesso ogni piccolo villaggio detiene una parlata locale dalle caratteristiche peculiari. Il pluralismo linguistico a livello sincronico, è frutto però di una storica unità linguistico-culturale che il progetto *VerbaAlpin*a dell'Università di Monaco di Baviera tenta di analizzare e di mettere in rilievo.

Il presente contributo rappresenta la versione approfondita e aggiornata della relazione tenuta dalle stesse autrici in occasione del trentasettesimo *Deutscher Romanistentag*, organizzato dall'università di Augusta e tenutosi online dal 4 al 7 ottobre 2021. L'occasione del Romanistentag ha fornito un'importante opportunità

di presentare dati relativi alle attività e alle modalità di raccolta dei dati linguistici da parte di *VerbaAlpina* in una prospettiva *work in progress*. Come si evincerà dal presente scritto, *VerbaAlpina* opera una raccolta dati mediante una piattaforma di *crowdsourcing* interna e, al contempo, rivolgendosi a un portale esterno. Il presente contributo concentrerà la propria attenzione sulla piattaforma di *crowdsourcing* interna, portando a conoscenza del lettore il contesto, gli obiettivi e alcuni risultati scaturiti da una breve indagine sull'utilizzo della piattaforma da parte dei parlanti dei comuni dell'arco alpino. Dopo la presentazione generale del progetto e delle sue finalità, il contributo si concentra sul *crowdsourcing* quale modalità di raccolta dati, approfondendo alcuni aspetti specialmente legati a *VerbaAlpina*.

2. VerbaAlpina - Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit

VerbaAlpina è un progetto di ricerca della *Ludwig-Maximilians-Universität* (LMU) di Monaco di Baviera finanziato dalla *Deutsche Forschungsgemeinschaft* (fondazione tedesca per la ricerca) a partire da ottobre 2014 con una prospettiva di durata fino al 2026. Il progetto è stato concepito e viene portato avanti in collaborazione dall'*Institut für Romanische Philologie* e dall'*IT-Gruppe Geisteswissenschaften* (Centro di Tecnologia dell'Informazione per le Scienze Umane) della LMU, e si iscrive all'interno dell'area di studi delle *digital humanities* (umanistica digitale) (cf. Krefeld & Lücke 2014a; Krefeld 2019a).

VerbaAlpina è un atlante linguistico digitale dell'area alpina le cui attività sono state suddivise in tre fasi temporali: la prima fase (10/2014 - 10/2017) è stata dedicata al lessico relativo alla gestione dei pascoli alpini, concedendo particolare attenzione all'ambito della lavorazione del latte. La successiva seconda fase (11/2017-10/2020) ha visto la consacrazione agli ambiti lessicali della flora, della fauna, delle formazioni paesaggistiche e della meteorologia alpine. La terza fase di elaborazione, iniziata a novembre 2020, si concluderà a ottobre 2023 e ha come oggetto di indagine il lessico dell'ambiente di vita moderno, con un occhio di riguardo all'ecologia e al turismo nelle Alpi. Questa terza fase, seppur affine alle prime due in termini di concezione ed elaborazione, ha visto l'introduzione di alcune novità relative soprattutto alla raccolta dei dati e alle modalità di collocazione del progetto all'interno del territorio alpino di cui sarà fatta relazione in questo scritto.

2.1 Concezione e presentazione generale del progetto

La concezione e la delineazione metodologica di *VerbaAlpina* sono state mosse dal desiderio di superare alcune limitazioni imposte sia dalla tradizione di ricerca in geolinguistica come anche dagli strumenti e dalle modalità di ricerca tipiche del periodo pre-digitale. Solitamente, gli atlanti linguistici si limitano a un'area linguistica o politica nazionale in particolare, oppure prendono in considerazione solo una varietà linguistica esistente all'interno di una determinata area. La conseguenza di tali limitazioni si profila nell'idea illusoria di tratti linguistici che si arrestano improvvisamente entro i confini dell'area presa in considerazione,

oppure di un monolinguismo che non rispecchia l'effettiva situazione linguistica della zona.

Per *VerbaAlpina*, il superamento delle limitazioni della geolinguistica tradizionale significa anzitutto non limitarsi a un'area linguistica o politica nazionale (intesa come Stato-Nazione) in particolare, ma si è voluto estendere l'area di ricerca a tutta la regione montuosa dell'arco alpino. Ciononostante, agli inizi del progetto si rendeva necessaria una delimitazione della zona sottoposta a ricerca: ci si è così affidati al perimetro della *Convenzione delle Alpi*[1] al fine di analizzare l'unità storico-culturale che caratterizza questa regione montuosa e poter estrapolare le similitudini e i punti di contatto che intercorrono tra le diverse varietà linguistiche e, di conseguenza – ma soprattutto –, tra le tre diverse famiglie linguistiche delle Alpi: la romanza, la germanica e la slava.

Proprio nell'ottica di osservare il grado di vicinanza delle varietà linguistiche delle tre famiglie linguistiche alpine, è risultato essenziale innanzitutto riunire il materiale necessario all'analisi. I dati lessicali sono forniti da un lato da atlanti linguistici, cartacei e digitali, pubblicati nel corso degli anni, ai quali si sono aggiunti anche dizionari dialettali locali, e dall'altro lato da dati provenienti dalla piattaforma di *crowdsourcing* di cui verrà fatta relazione di seguito. I dati contenuti negli atlanti cartacei hanno necessitato di un processo di digitalizzazione attraverso un sistema di trascrizione basato esclusivamente su caratteri ASCII (cf. Krefeld 2019b). Una volta inserite nella banca dati attraverso la trascrizione (oppure attraverso la piattaforma di *crowdsourcing*), le espressioni subiscono un processo di tokenizzazione, mediante il quale esse vengono suddivise in singole parole per poi essere successivamente tipizzate, ovvero raggruppate sotto un'unica forma (un tipo morfo-lessicale) rappresentativa di tutte le sottovarianti secondo le loro caratteristiche grammaticali. La tipizzazione del materiale lessicale rappresenta la *conditio sine qua non* per strutturare i dati e per facilitare la loro lettura e la loro analisi in termini linguistici. Un tipo morfo-lessicale viene definito dalla concordanza delle seguenti proprietà: famiglia linguistica, categoria lessicale, parola semplice vs. parola affissa, genere, tipo di base lessicale. Lo schema seguente riassume il funzionamento della tipizzazione:

[1] La *Convenzione delle Alpi* è un trattato internazionale sottoscritto dai Paesi alpini di Austria, Francia, Germania, Italia, Liechtenstein, Principato di Monaco, Slovenia e Svizzera e dalla Comunità economica europea, con l'obiettivo di garantire una politica comune per l'arco alpino. Dal punto di vista metodologico, essa possiede un importante significato per *VerbaAlpina*, in quanto il progetto ha adottato, fissandola come area di ricerca, la definizione di 'regione alpina' basata sulle frontiere geografiche e amministrative delimitate proprio dalla *Convenzione delle Alpi*.

Varianti fonetiche	*barga*	*bark*	*bɑrgúŋ*	*margún*
Famiglia linguistica	roa	roa	roa	roa
Categoria lessicale	sub	sub	sub	sub
Affisso	-	-	+	+
Genere	f	m	m	m
Tipo morfo-lessicale	1	2	3	3

Tab. 1 | Sistema di tipizzazione del materiale lessicale di *VerbaAlpina*

Tuttavia, il raggruppamento del materiale linguistico in tipi morfo-lessicali non è sufficiente all'adempimento completo degli obiettivi accennati *supra*. Per mostrare i punti di coesione tra i diversi idiomi alpini si rende necessario e interessante un altro tipo di tipizzazione, ovvero un raggruppamento del materiale linguistico in tipi di base. Un tipo di base è rappresentato dalla radice lessicale, comune alle diverse attestazioni, espressa attraverso la prima forma storicamente attestata. Ad esempio, il tipo di base latino *butyru(m)* 'burro' rappresenta la prima forma storicamente attestata di diverse parole dialettali (ma anche appartenenti alle lingue nazionali) di origine germanica, romanza e slava quali *bˈyʀi* (La Javie, Francia), *butˈeːr* (Sonico, Italia), *Buttr* (Füssen, Germania), *putər* (Jesenice, Slovenia)[2], tutte relative al concetto BURRO. Importante, ai fini della ricostruzione dei percorsi etimologici, è la precisazione della tipologia di relazione storica tra la forma di base e la parola dialettale (e/o il tipo morfo-lessicale) in questione: mentre per le forme romanze, in questo caso, la relazione con la forma base lat. *butyru(m)* è diretta, le forme germaniche e slave sono identificabili come prestiti.

I dati linguistici di *VerbaAlpina* appaiono su una mappa interattiva (cf. mappa interattiva 2022b)) che dispone di diverse modalità di filtrare i dati in prospettiva onomasiologica e semasiologica. Sulla mappa sono visualizzabili anche dati toponomastici ed extra-linguistici. Accanto a questa forma di visualizzazione dei dati, *VerbaAlpina* propone anche il *Lexicon Alpinum* (cf. Krefeld & Lücke 2014b) come strumento di lettura dei dati: in esso sono contenuti tutti i tipi morfo-lessicali, tipi di base e i concetti di *VerbaAlpina*, solitamente provvisti di un commento di tipo storico-etimologico oppure etnolinguistico.

2.2 Sfide generali

VerbaAlpina è un atlante linguistico pensato per e basato sul web. Ciò significa che, da un punto di vista tecnico-informatico, esso ha la necessità di confrontarsi con specifiche questioni che riguardano la sua integrità e la sua durabilità nel tempo. La pagina web, sviluppata sulla base del *content management system Wordpress*, viene «congelata» ogni sei mesi. Si vengono quindi a creare delle versioni stabili non più modificabili. L'attuale versione è la numero 21/2 (ossia la seconda versione

[2] Per la consultazione della cartina interattiva che raccoglie tutte le forme dialettali legate al tipo di base lat. *butyru(m)* cf. mappa interattiva 2022a.

creata nel 2021). Le modifiche vengono apportate sempre alla versione di lavoro che prende il nome «xxx». In termini cartacei, ogni versione può essere paragonata a una nuova edizione pubblicata di *VerbaAlpina*.[3] Fondamentale per *VerbaAlpina* è, al contempo, lo sviluppo di strategie per garantire la durabilità dei dati così come il loro accesso anche dopo la fine del progetto. Ancor prima che i princìpi FAIR fossero formulati (cf. Wilkinson & Dumontier et al. 2016), *VerbaAlpina* ha lavorato, anche in collaborazione con enti esterni, affinché il progetto fosse reso *findable* (progetti *Generic Research Data Infrastructure - GeRDI, eHumanities - interdisziplinär e Discover*[4]), *accessible* (rinunciando ai diritti d'autore, applicando una licenza CC-BY SA 4.0. e mettendo a disposizione un'API per l'accesso computerizzato ai dati; cf. Krefeld & Lücke 2014-d), *interoperable* (strutturando minuziosamente i dati, applicando metadati, identificatori e controllo di autorità) e di conseguenza *reusable*.[5]

3. Il *crowdsourcing*: dare spazio alla collettività nella ricerca (geolinguistica)

L'immagine tradizionale che si ha del ciclo della scienza (nel corso di questo scritto si farà riferimento specialmente alla geolinguistica) segue normalmente lo schema seguente (estremamente semplificato): un ricercatore (o un gruppo di ricerca) elabora un progetto, si reca presso gli informanti (o i probandi), raccoglie i dati di cui necessita, ritorna presso il suo laboratorio, prosegue con l'analisi di quanto raccolto e dà comunicazione dei risultati emersi alla comunità scientifica. In questa immagine, il ricercatore viene visto come la guida, la persona che decide cosa estrapolare, ma soprattutto, quando estrapolare dati dall'informante. Sempre estremizzando, si tratta dunque di un sistema in cui l'idea che si veicola è di un ricercatore detentore del sapere da un lato e di un informante fonte di informazione, ma privo di pensiero critico, dall'altro lato. Questo tipo di approccio, pur essendo largamente utilizzato oltre che molto tradizionale, non sempre è adatto a ricerche nelle quali vi sia l'esigenza di disporre di un cospicuo *dataset*. Nel momento in cui si rende necessaria la raccolta di un grande numero di dati, oppure in casi in cui i progetti di ricerca siano molto corposi, può essere proficuo un altro tipo di approccio, basato sull'apertura al grande pubblico. Non si tratta più solamente di «sfruttare» le competenze dei parlanti, ma di farli partecipare attivamente alla ricerca, di coinvolgerli attraverso le loro competenze, fornendo loro la possibilità di prestarsi a un qualcosa di socialmente utile. Si tratta, in un certo senso, della democratizzazione dell'attività di fare scienza. Accanto al ricercatore, cittadini non necessariamente formati in uno specifico ambito (ad esempio non formati in linguistica) diventano protagonisti partecipando al successo del progetto di ricerca.

[3] Per una panoramica delle maggiori novità introdotte per ogni versione cf. Krefeld & Lücke 2014-c.

[4] Cf. Tochtermann 2018-, Universitätsbibliothek Erlangen-Nürnberg 2018-, University Library LMU Munich 2021-2022.

[5] Per un approfondimento sull'applicazione dei princìpi FAIR a *VerbaAlpina* si rimanda a Lücke 2020, Lücke 2021, Colcuc & Mutter 2020.

Le tecnologie di cui disponiamo rendono possibile e permettono un rapido sviluppo di questa maniera di fare scienza che, oggigiorno, è conosciuto sotto il nome di *crowdsourcing*. Martin, Lessman & Voß (2008, 1256) propongono la seguente definizione:

> Crowdsourcing ist eine interaktive Form der Leistungserbringung, die kollaborativ oder wettbewerbsorientiert organisiert ist und eine große Anzahl extrinsisch oder intrinsisch motivierter Akteure unterschiedlichen Wissensstands unter Verwendung moderner IuK-Systeme auf Basis des Web 2.0 einbezieht.[6]

La parola *crowdsourcing* è formata dalla sincresi delle due parole inglesi *crowd* 'folla' e *outsourcing* 'esternalizzazione', ovvero la pratica utilizzata da sempre più aziende di ricorrere a imprese o enti esterni per lo svolgimento di alcune parti dei propri processi di produzione. Dall'economia, dunque, il pensiero di esternalizzazione di alcune pratiche, si è esteso anche alla scienza, anche se, nel contesto accademico, sarebbe improprio pensare al *crowdsourcing* come a un fenomeno recente. Ovviamente le tecnologie attuali permettono, come già accennato, di praticare *crowdsourcing* in maniera più immediata, ma esempi di tale approccio sono documentati già nella storia. Si pensi ad esempio all'*Oxford English Dictionary*, la cui realizzazione poté contare sulla partecipazione attiva di volontari chiamati a catalogare il materiale e a fornire esempi concreti per l'uso delle parole (cf. Gilliver 2020; Ogilvie 2018) oppure all'enciclopedia libera *Wikipedia* (cf. Aa.Vv. 2001-).

Le inchieste linguistiche sono normalmente caratterizzate da una raccolta dei dati di tipo diretto oppure indiretto. Le prime si caratterizzano per il fatto che il ricercatore, solitamente munito di taccuino e/o di un registratore, si reca egli stesso nei punti di rilevamento da lui scelti al fine di raccogliere i dati ai quali è interessato. Questa modalità di raccolta dei dati è tipica della tradizione di ricerca romanza: si pensi, ad esempio allo *Sprach- und Sachatlas Italiens und der Südschweiz* (AIS) di Karl Jaberg e Jakob Jud (1928-1940). La tradizione di ricerca germanistica ha invece prediletto le inchieste di tipo indiretto mediante questionari solitamente inviati ai parlanti per posta.[7] Tutto sommato dunque, adoperare il *crowdsourcing* come modalità di raccolta dei dati in geolinguistica significa mettere in campo una ricerca di tipo indiretto, ma sfruttando le potenzialità di internet.

I metodi di *crowdsourcing* si inseriscono nel più ampio contesto della cosiddetta *citizen science*, ossia la scienza fatta dai cittadini, nell'idea di facilitare e incoraggiare la partecipazione del grande pubblico ai processi di ricerca e di mettere a disposizione della comunità i risultati emersi (cf. ECSA 2013-).

I vantaggi dell'utilizzo di piattaforme e metodi di *crowdsourcing* sono numerosi: potendo contare teoricamente su più persone che si impegnano in un determinato progetto, è più verosimile che gli obiettivi siano raggiunti più rapidamente o che i

[6] Traduzione italiana: Il *crowdsourcing* rappresenta una forma interattiva di fornitura di servizi, organizzata in modo collaborativo o competitivo, coinvolgente un gran numero di attori motivati estrinsecamente o intrinsecamente, che dispongono di conoscenze diverse. Essa è resa possibile tramite l'utilizzo dei moderni sistemi d'informazione e di comunicazione, sulla base del web 2.0 (traduzione Beatrice Colcuc).

[7] Per un approfondimento sulle differenze tra raccolta dei dati diretta e indiretta si rimanda a Kunzmann & Mutter 2021.

problemi siano risolti con maggiore velocità e che, data la grande partecipazione e potendo contare su feedback diversificati, si arrivi in minor tempo a identificare eventuali criticità o incompletezze. Per quanto riguarda specificatamente i progetti di ricerca, l'utilizzo del *crowdsourcing* attraverso internet rende possibile una velocizzazione significativa dei tempi così come un risparmio di risorse economiche. Dal punto di vista dell'immaginario sociale, il *crowdsourcing* veicola un'idea più forte di collettività, facendo notare come sia solamente mediante la partecipazione e l'apporto di tutti che progetti, idee e ricerche possono essere completati. Un altro vantaggio è, in base al tipo di studio che si desidera condurre, l'anonimato: molto spesso le piattaforme di *crowdsourcing* funzionano attraverso la partecipazione anonima di persone e questo può andare a beneficio sia della qualità del dato, poiché gli informati potrebbero tendere a rivelare di più di quanto non avrebbero fatto in una situazione di faccia a faccia con il ricercatore, sia della quantità, poiché ci si potrebbe sentire più a proprio agio nel non fornire informazioni personali (si osservi la differenza quantitativa tra *crowder* registrati e non registrati su *VerbaAlpina* in 4.3.).

Senza dubbio, i metodi di *crowdsourcing* portano con sé anche alcuni svantaggi: la partecipazione di un pubblico (più o meno) indefinito di persone significa maggiore probabilità che sorgano domande, dubbi o difficoltà. È necessario dunque un accompagnamento costante da parte degli iniziatori al fine di evitare errori o incomprensioni. Oltre a ciò, il grande numero di persone fa sì che talvolta non si riesca ad avere il controllo totale delle procedure.

4. Crowdsourcing e VerbaAlpina

4.1 Contesto e piattaforme

Il panorama delle opere lessicografiche in area alpina è complesso e abbastanza dettagliato. Come accennato *supra*, il *dataset VerbaAlpina* è rappresentato da dati estrapolati da diversi atlanti linguistici, coprenti parti diverse del territorio alpino, pubblicati (per la maggior parte) su carta e riconducibili a diverse tradizioni di ricerca (germanistica, romanistica e slavistica). Di conseguenza, il materiale lessicale di cui si dispone è consistente. In un primo momento, tuttavia, questo insieme di dati si presenta abbastanza eterogeneo: gli atlanti linguistici della tradizione germanistica seguono infatti modelli e strutture diversi da quelli relativi a varietà linguistiche romanze o slave, non sempre contengono forme dialettali riguardanti gli stessi concetti oppure, ancora, sono stati pubblicati in momenti storici diversi. Oltre a ciò, gli atlanti linguistici e vocabolari dialettali tendono a racchiudere in essi forme che provengono da punti di rilevamento solitamente ben identificati dagli esploratori. Questo tipo di approccio, ormai tradizionale, comporta però un percettibile svantaggio: mentre le varietà linguistiche di alcune località, data la loro particolarità e/o la loro rilevanza scientifica, sono fatte spesso oggetto di studi, altre rischiano di essere tagliate fuori e di non essere né documentate, né considerate in sede scientifica. Fermo restando che con i metodi tradizionali di ricerca sul campo sarebbe impossibile rilevare la completezza dei dati per tutte le micro-località dei territori di ricerca e che quindi una selezione ponderata dei punti

è d'uopo, è altrettanto noto che le nuove tecnologie possono offrire la soluzione a (tanti) limiti della ricerca tradizionale. Per superare i limiti qui illustrati, a partire dal 2017 (versione 16/2), *VerbaAlpina* ha lanciato una piattaforma di *crowdsourcing*, sviluppata all'interno dello stesso, al fine di raccogliere parole provenienti da parlanti di tutti i comuni della realtà alpina. La piattaforma (cf. crowdsourcing 2022) presenta ai parlanti una sola domanda: si desidera sapere qual è la forma dialettale utilizzata in loco per riferirsi a determinate entità della vita reale. L'immagine seguente mostra come si presenta la piattaforma di *crowdsourcing* di *VerbaAlpina*:

1 | Finestra principale della piattaforma di *crowdsourcing* di *VerbaAlpina* senza risposte

Al centro della finestra si può identificare l'arco alpino e, al suo interno, i dati *crowd* registrati per ogni località. Si nota immediatamente come vi siano delle regioni ad alta intensità di dati (Alpi centrali) e altre a bassa intensità (Alpi occidentali e nord-orientali). Nella parte bassa della schermata si può leggere la domanda posta al *crowder*: si chiede di cliccare sul comune per il quale si desidera fornire delle parole, successivamente si chiede di scegliere un concetto dalla lista proposta. Nella prospettiva di un *crowder*, la piattaforma presenta uno stimolo testuale in lingua standard (per i linguisti il concetto) che deve essere «tradotto» nel dialetto del comune prescelto. Una volta inviata una parola, si potrà proseguire con altre, sempre seguendo lo stesso procedimento.

2 | Finestra principale della piattaforma di crowdsourcing di VerbaAlpina con risposta

Prima di arrivare alla finestra principale, vi sono due passi precedenti da eseguire: la scelta della lingua di navigazione (francese, italiano, tedesco, sloveno) e la scelta del proprio dialetto tra quelli proposti, oppure, in alternativa, il suo inserimento nella lista. La piattaforma offre inoltre la possibilità di registrarsi oppure di inviare i dati in forma anonima (in alto a destra), di indicare la propria età e dispone di una sezione di domande frequenti così come di una classifica dei *crowder* più attivi, dei comuni più rappresentati e dei concetti preferiti (nella parte sinistra della schermata). L'interesse primario di VerbaAlpina si rivolge alla distribuzione delle varianti nello spazio geografico senza però focalizzarsi sull'opposizione tra "lingua" e "dialetto". Il materiale raccolto attraverso il *crowdsourcing* non esclude dunque varianti regionali delle lingue nazionali che si collocano in posizione intermedia tra i dialetti e queste ultime. In quest'ottica, categorizzare le parole raccolte come "dialettali" oppure come scaturite dalle lingue nazionali risulta superfluo. Nondimeno, si è rilevata una prossimità lievemente più accentuata alla lingua nazionale per quanto riguarda le parole relative ai concetti della vita moderna rispetto a quelle relative ai domini concettuali dell'alpicoltura e della natura alpine.

4.2. Finalità del crowdsourcing

Gli obiettivi perseguiti da *VerbaAlpina* sono molteplici: si desidera innanzitutto raccogliere dati attuali per poter operare un confronto con i dati presenti negli atlanti linguistici e nei dizionari cartacei. La raccolta di nuovo materiale è utile anche per colmare eventuali lacune presenti nelle opere lessicografiche (nel caso in cui, ad esempio, un atlante non abbia rilevato la forma lessicale relativa a un concetto preciso per un determinato comune). Essendo il comune l'unità territoriale di rilevamento più piccola in *VerbaAlpina*, ognuno di essi rappresenta un punto di rilevamento aumentando dunque la copertura territoriale.

Da ultimo, non meno importante, anzi, fondamentale è il quarto obiettivo: *VerbaAlpina* utilizza la piattaforma di *crowdsourcing* al fine di raccogliere i dati lessicali per la terza fase del progetto. Mentre per le prime due fasi (alpeggio e

natura alpina) la base di dati più consistente era rappresentata da opere lessicografiche già pubblicate, per l'attuale terza fase, incentrata sulla vita moderna nelle Alpi, in particolare sui domini concettuali dell'ecologia e del turismo, non sono ancora state documentate forme lessicali dialettali in maniera massiccia in alcuna opera lessicografica. Per questo motivo, la partecipazione dei parlanti dei comuni alpini è fondamentale.

La piattaforma di *crowdsourcing* non è soltanto utile alla raccolta di dati meramente lessicali, ma anche sociolinguistici e relativi alle rappresentazioni dei parlanti. In particolare, l'analisi delle indicazioni del dialetto (nelle immagini soprastanti si nota la dicitura dialettale *veneto* in basso a sinistra) può dare un contributo allo studio delle rappresentazioni relative al nome della propria parlata.

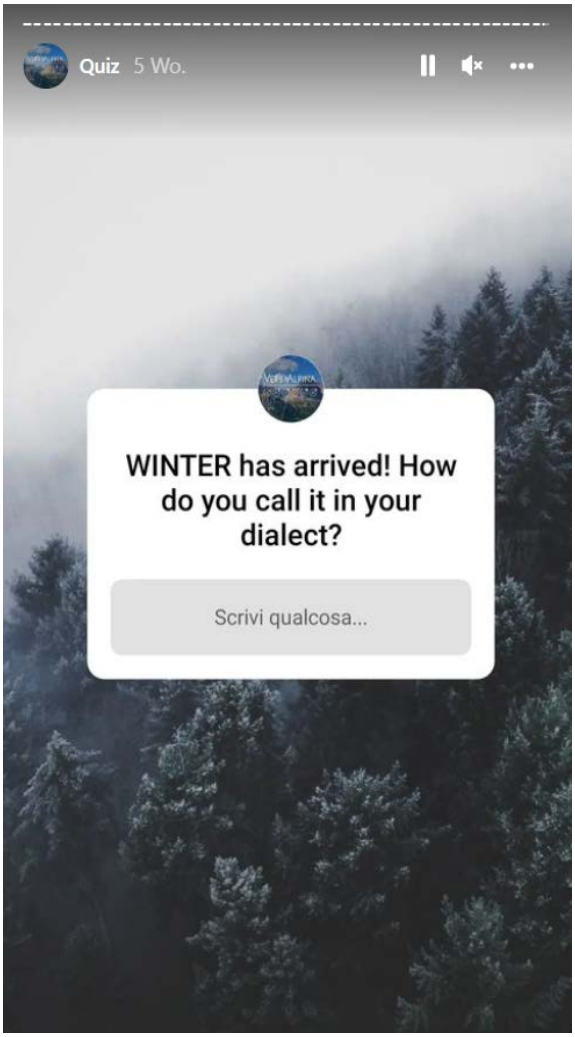
VerbaAlpina non pratica il *crowdsourcing* solo attraverso una piattaforma sviluppata in seno al progetto, ma ha trovato spazio anche sul portale di *citizen scienze Zooniverse* (cf. Zooniverse 2009-). Sono state proposte due attività: il *main task* chiedeva di corredare un insieme di 1212 stimoli di cartine dell'AIS con il proprio numero, mentre l'*optional task* prevedeva che i volontari della rete si impegnassero nella trascrizione del materiale degli atlanti utilizzando le regole di trascrizione di *VerbaAlpina*. Il progetto di *crowdsourcing* di *VerbaAlpina* su *Zooniverse* ha visto la partecipazione di 753 volontari in totale. Nonostante il successo generale dell'operazione, *VerbaAlpina* ha riscontrato alcune difficoltà nell'inserimento e nel lancio del progetto su *Zooniverse*. La creazione di un apposito profilo è un'operazione che richiede molta cura e molto tempo, il procedimento di revisione da parte di *Zooniverse* si è anch'esso prolungato nel tempo. Dalla prima richiesta di partecipazione al lancio effettivo di *VerbaAlpina* sono passati circa due anni e, una volta lanciato il progetto sulla piattaforma, sono occorse diverse difficoltà da parte dei *crowder* che hanno necessitato di un accompagnamento da parte di *VerbaAlpina*. Si è trattato soprattutto di domande relative alla trascrizione di specifiche attestazioni. L'AIS presenta infatti talune imprecisioni relative all'applicazione delle convenzioni di trascrizione stabilite dagli autori stessi. Altri problemi si sono presentati laddove la qualità dell'immagine delle cartine proposte non permettesse una chiara lettura. Oltre a ciò, VerbaAlpina ha ricevuto domande scaturite dal semplice fatto che alcuni utenti non avevano letto completamente e correttamente le regole da noi formulate.

4.3 Strategie e risonanza generale

A distanza di quattro anni dal lancio della piattaforma interna di *crowdsourcing*, sono possibili e necessarie alcune considerazioni sui successi ottenuti e sulle difficoltà riscontrate. Innanzitutto, è necessario sottolineare come la piattaforma, per portare risultati, abbia bisogno di un'assistenza continua. Naturalmente, affinché la partecipazione da parte dei parlanti sia costante nel tempo, vi è la necessità di far conoscere la piattaforma e questo implica il bisogno di pensare continuamente a modalità e strategie nuove e diversificate per pubblicizzare il progetto e la sua raccolta dati. L'attività di partecipazione da parte dei *crowder* è monitorata attraverso una statistica live (cf. statistiche live 2017-). Per ogni giorno di calendario, vengono segnalate quante parole entrano a far parte del *dataset* di

VerbaAlpina, da quali Paesi e da quanti *crowder*. L'altezza delle colonne è proporzionale al numero di parole inserite al giorno. Osservando i dati in un periodo di tempo più ampio, si nota come, ad ogni specifica attività da parte del progetto, corrisponda una partecipazione maggiore al *crowdsourcing*.

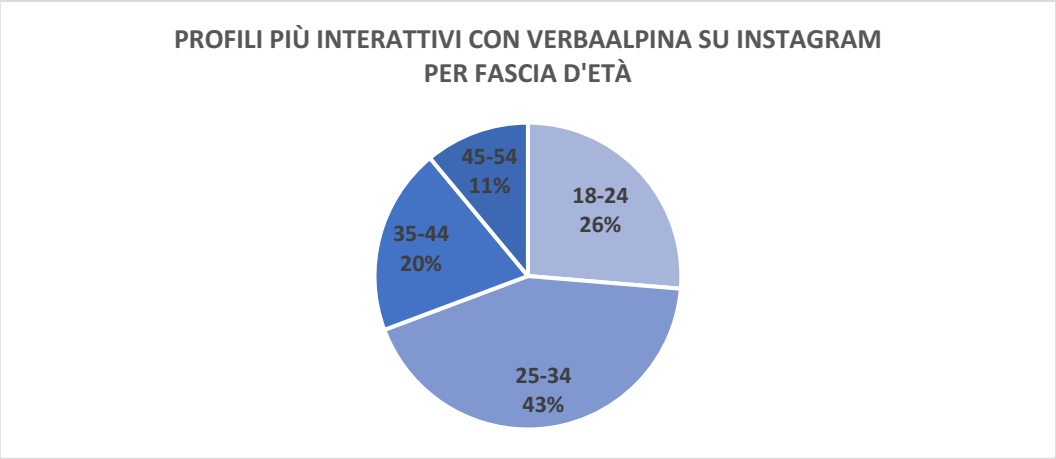
L'attività di *VerbaAlpina* per pubblicizzare il progetto e la sua piattaforma è cambiata nel corso degli anni, anche alla luce delle esperienze via via raccolte. Per la prima e la seconda fase ci si è affidati soprattutto a modalità di pubblicità stampate, specialmente articoli di giornale o di riviste e, parallelamente, sono stati creati dei flyer pubblicitari da distribuire all'interno dell'area di ricerca. *VerbaAlpina* ha inoltre preso parte anche a convegni popolari e ha rilasciato interviste radiofoniche (cf. eco mediatica). Contemporaneamente, si è proceduto a contattare, soprattutto via e-mail, enti e associazioni dell'arco alpino che si occupassero di lavorazione del latte e/o di natura con la richiesta di partecipare al *crowdsourcin*g e/o di far conoscere il progetto nella propria cerchia di contatti. La terza fase del progetto ha visto un cambio di rotta per quanto riguarda le modalità di promozione del progetto: ai profili social di *VerbaAlpina* su *Facebook* e *Twitter*, aperti nel 2016, si è aggiunto nel 2020 un profilo *Instagram* (cf. VerbaAlpina su *Facebook* 2016-, VerbaAlpina su Twitter 2016-, VerbaAlpina su Instagram 2020-). La strategia per la diffusione delle informazioni riguardo a *VerbaAlpina* in questa fase ha visto (e vede tutt'oggi) protagonista internet, in particolare i *social network*. Le attività sui social sono di natura diversa: da un lato si ricercano attivamente contatti, specialmente pagine che condividono, almeno in parte, gli stessi interessi di *VerbaAlpina*; dall'altro lato si propone un ventaglio di contenuti divulgativi inerenti a *VerbaAlpina*, al mondo delle Alpi e ai temi cari al progetto. Per esempio, a luglio 2021 è stata ideata la rubrica "False myths to fight" 'falsi miti da sfatare' che si occupa di smontare idee false e luoghi comuni sui dialetti nell'ottica di fare chiarezza e di avvicinare i parlanti al mondo della dialettologia attraverso la sensibilizzazione. Su *Instagram* vengono proposte le cosiddette stories contenenti domande rivolte al pubblico. Una buona quantità dei dati raccolti a partire da maggio 2021 proviene proprio dalle risposte dei parlanti alle stories in cui viene chiesto loro, ad esempio, qual è la parola per INVERNO nel loro dialetto.



3 | Una story di *VerbaAlpina* su *Instagram*

Al fine di costruire una rete di contatti sempre più stabile e duratura, *VerbaAlpina* ha anche organizzato e partecipato a dirette su *Facebook* assieme ad altri progetti che si occupano di sensibilizzazione linguistica nelle Alpi.

Il cambio di strategia applicato con l'inizio della terza fase, ossia la (più) costante attività sui *social network*, ha evidenziato un maggiore interesse da parte di parlanti che hanno deciso di partecipare attivamente alla raccolta dati attraverso la piattaforma di *crowdsourcin*g. A questo proposito è interessante soffermarsi brevemente sulla tipologia di pubblico che viene raggiunta grazie alle attività social di *VerbaAlpina*. Il grafico seguente mostra i profili che interagiscono maggiormente con la pagina di *VerbaAlpina* su *Instagram*, divisi per fasce d'età:

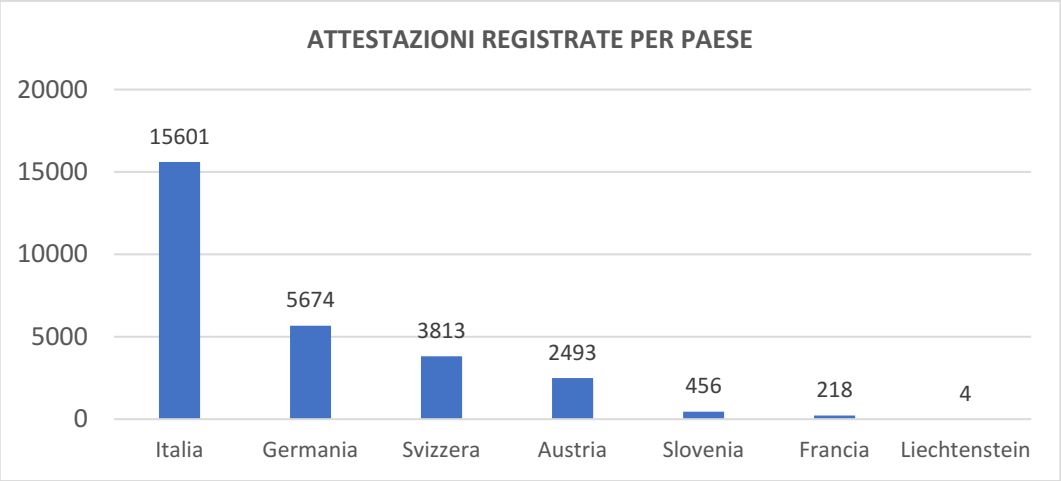


4 | Profili che interagiscono maggiormente con VerbaAlpina su Instagam per fasce d'età

Come accennato *supra*, la piattaforma di *crowdsourcing* concede ai *crowder* la libertà di creare un profilo e di dichiarare la propria età. Non essendo questo un dato che i *crowder* devono per forza fornire, *VerbaAlpina* non dispone di un numero netto relativo all'età media delle persone che partecipano alla piattaforma. Inoltre, i dati scaturiti dalle statistiche di *Instagram* sull'interattività con il profilo di *VerbaAlpina* non garantiscono obbligatoriamente una maggiore attività sulla piattaforma di *crowdsourcing*, ma suggeriscono prudentemente che gli informanti di *VerbaAlpina* potrebbero essere probabilmente più giovani rispetto agli informanti dei tradizionali atlanti linguistici.

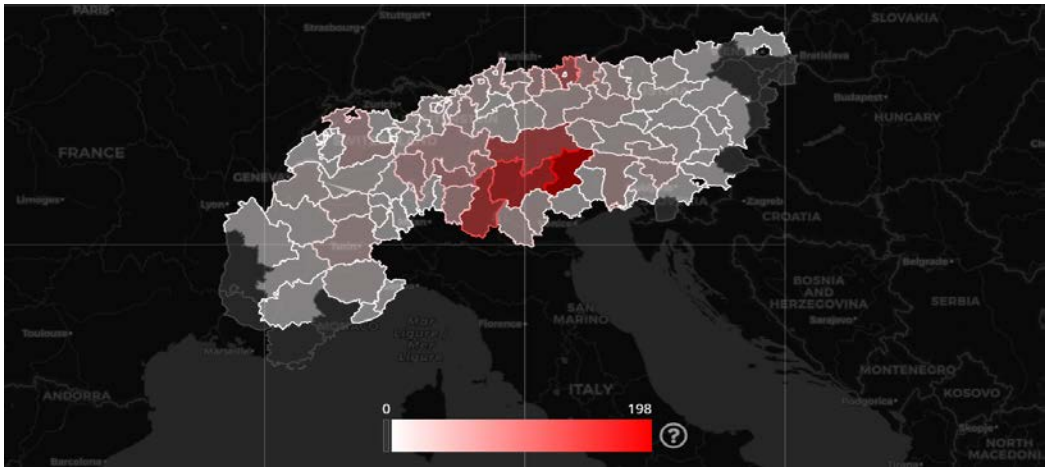
Dall'inizio dell'attività di *crowdsourcing* (dati aggiornati al 26 gennaio 2022), *VerbaAlpina* conta un totale di 1916 informanti, di cui 1668 non registrati e 248 registrati e ha raccolto 28.259 parole. Di queste, 15.601 provengono dall'Italia, 5674 dalla Germania, 3813 dalla Svizzera, 2493 dall'Austria, 456 dalla Slovenia, 218 dalla Francia e 4 dal Liechtenstein.[8] Le parole raccolte attraverso il *crowdsourcing* hanno portato alla creazione di un totale di 4241 tipi morfo-lessicali nuovi (non presenti in atlanti linguistici), di cui 1794 per la prima fase (340 romanzi, 1384 germanici e 106 slavi), 1833 per la seconda fase (815 romanzi, 1001 germanici e 17 slavi) e 704 per la terza fase (611 romanzi, 93 germanici). Globalmente, i dati raccolti sono inoltre altamente qualitativi: su circa 40.000 tokens totali di cui si dispone (dati aggiornati a ottobre 2022), il numero di dati non validi è decisamente irrilevante (si tratta ad esempio di espressioni come "non so" oppure "non ricordo" inviate al posto della parola dialettale per un dato concetto).

[8] Per consultare il numero di attestazioni registrate in costante aggiornamento cf. statistiche live.



5 | Numero di attestazioni registrate per Paese (dati aggiornati al 26 gennaio 2022)

La cartina seguente mostra invece le regioni NUTS[9] dell'arco alpino che concentrano il maggior numero di *crowder* di *VerbaAlpina*: si tratta delle province italiane di Belluno, Trento, Brescia e Bolzano. A settentrione si distingue inoltre il Landkreis Rosenheim (per la versione interattiva della stessa cartina cf. mappa interattiva 2022c).



6 | Provenienza dei crowder di VerbaAlpina, secondo regioni NUTS
(dati aggiornati al 26 gennaio 2022)

Per motivi di assetto interno del progetto, l'impegno per la promozione attraverso i *social network* si è inizialmente concentrato sull'Italia e, di contro, tale attività rafforzata ha incontrato fin da subito un forte interesse da parte del pubblico italiano.

[9] Nomenclatura delle unità territoriali statistiche, si tratta di un sistema di suddivisione geografica del territorio dell'Unione europea a fini statistici.

4.4 Crowder

Si è accennato alla particolarità della terza fase di *VerbaAlpina*, caratterizzata dalla mancanza di documentazione lessicale in atlanti linguistici o dizionari dialettali per i domini concettuali della vita moderna alpina (ecologia e turismo), e dal fatto che, conseguentemente, *VerbaAlpina* ha deciso di impiegare la piattaforma di *crowdsourcing* come strumento principale della raccolta dati. Dopo tre anni di utilizzo della piattaforma, si è resa necessaria e significativa una breve indagine riguardante la percezione della piattaforma da parte dei *crowder*, il suo utilizzo e il rapporto tra *VerbaAlpina* e i parlanti al fine di poter considerare successi, criticità ed eventualmente apportare delle migliorie.

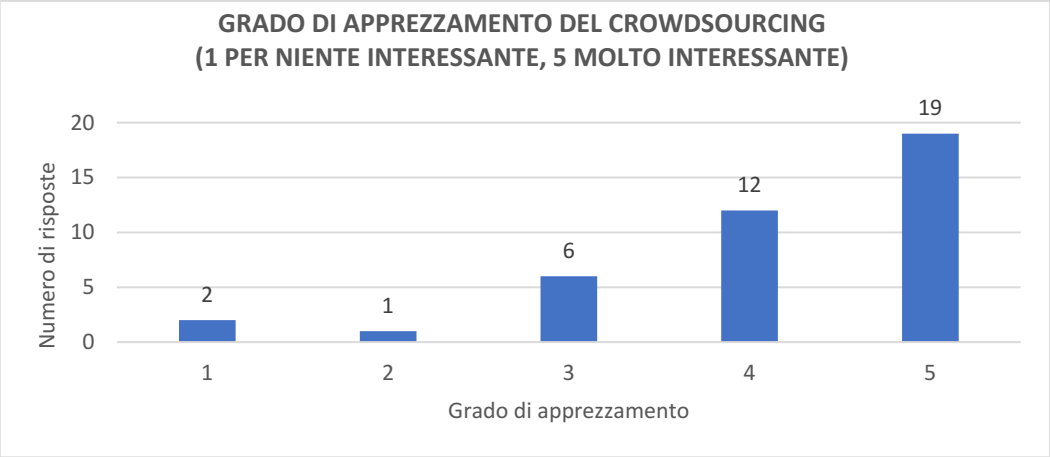
A questo scopo è stato creato un questionario online con diverse domande, suddivise in tre sezioni: la prima parte indaga il rapporto del *crowder* con *VerbaAlpina*, la seconda parte si concentra sulla valutazione dell'importanza attribuita dal *crowder* alla partecipazione ad attività collettive anche presso la propria comunità e la terza parte concerne il rapporto del *crowder* con il proprio dialetto. Il questionario è stato redatto nelle quattro lingue di *VerbaAlpina* (italiano, francese, tedesco e sloveno) e inviato a tutti i *crowder* registrati (di cui quindi *VerbaAlpina* possiede un indirizzo e-mail).[10] L'osservazione delle risposte è servita per avere una panoramica generale di ciò che è apprezzato e ciò che invece andrebbe migliorato sulla piattaforma nell'ottica dei suoi utilizzatori.

In sintesi, il partecipante medio (registrato) dell'attività di *crowdsourcing* si profila come segue: ha tra i 26 e i 50 anni, dà un valore molto alto alla propria varietà dialettale e si interessa attivamente a questioni culturali della sua zona. È dell'idea che alle varietà dialettali si debba concedere più spazio a livello sociale, ad esempio attraverso l'insegnamento a scuola e considera importante che le piccole realtà linguistiche vengano documentate attraverso i progetti di ricerca.

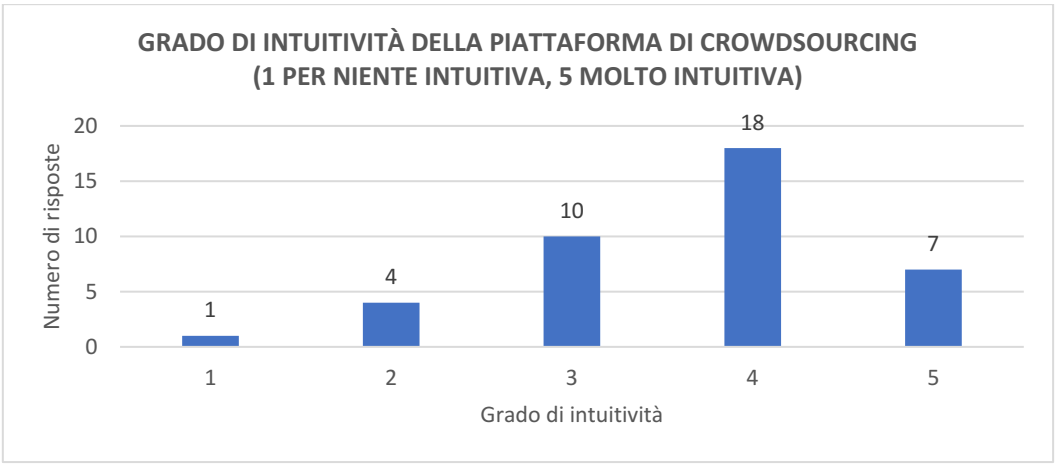
Ai fini della ricerca, per *VerbaAlpina* sono fondamentali tre aspetti dell'attività di *crowdsourcing*: in primo luogo, l'attività deve destare l'interesse dei parlanti e invogliarli a prendere parte al progetto; in secondo luogo, per stimolare le persone a partecipare o a continuare a partecipare, è necessario che la piattaforma sia rapida e di facile utilizzo. Inoltre, dal momento che l'idea generale del *crowdsourcing* è quella di creare una *community* che partecipi alle inchieste, è importante che i *crowder* che hanno apportato il loro contributo si sentano invogliati a invitare familiari, amici e conoscenti a prendere parte.

I dati cumulativi hanno mostrato che i riscontri relativi a tali aspetti sono, nel complesso, positivi. In linea generale, i *crowder* hanno trovato interessante la loro partecipazione e la piattaforma è risultata abbastanza intuitiva come è possibile notare dai due diagrammi sottostanti:

[10] Per la versione italiana del questionario cf. Colcuc & Rodella 2021.

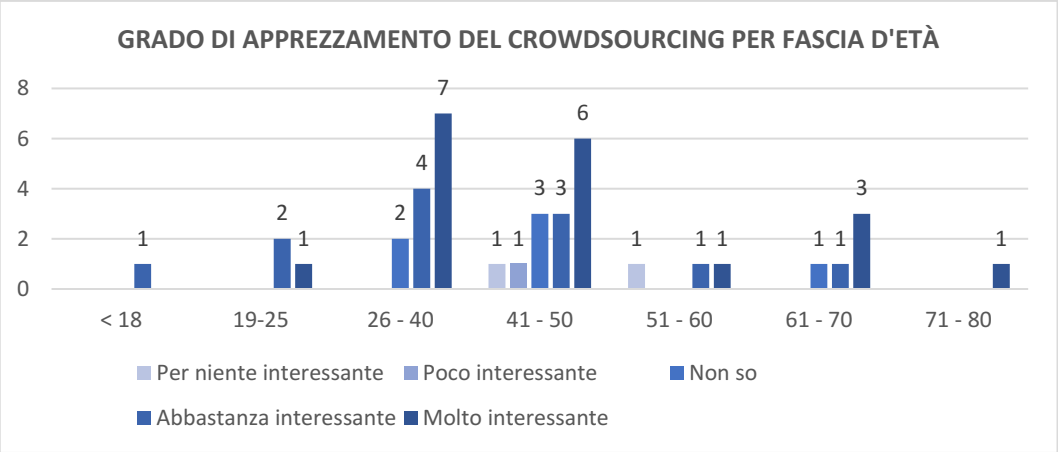


7 | Grado di apprezzamento della partecipazione al *crowdsourcing* di *VerbaAlpina*



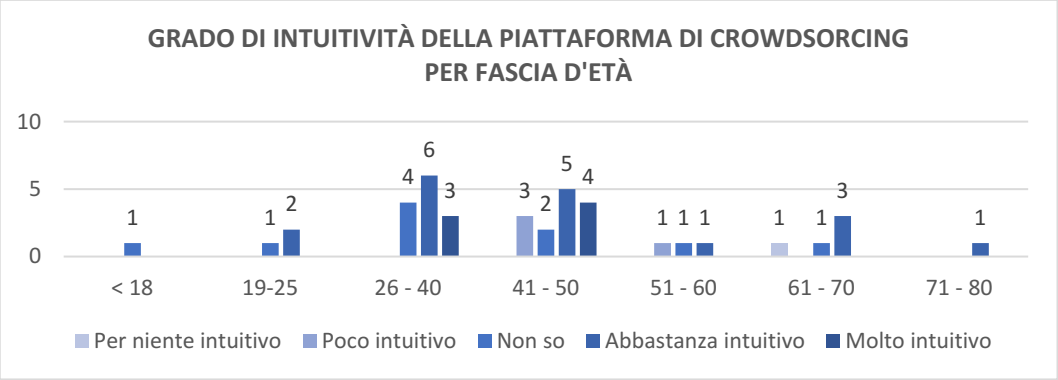
8 | Grado di intuitività della piattaforma di *VerbaAlpina*

I dati divengono più interessanti e incisivi nel momento in cui si vanno a incrociare con l'età e il titolo di studio dei partecipanti. In 4.3. si accennava alle fasce d'età che interagiscono maggiormente con la pagina *Instagram* di *VerbaAlpina* (fascia d'età tra i 25 e i 43 anni). Se, presi i dati del questionario, si nota come lo stesso gruppo dimostri di aver provato interesse nei confronti della partecipazione alla piattaforma. Su un totale di 13 partecipanti al questionario per la fascia 26-40, 11 hanno giudicato la loro partecipazione come abbastanza o molto interessante. Il dato è molto significativo, oltre che utile, per *VerbaAlpina*, poiché conferma che il progetto è riuscito a fare breccia in un gruppo di popolazione considerato generalmente distante dalle questioni legate ai dialetti e dal mondo della ricerca scientifica in dialettologia.



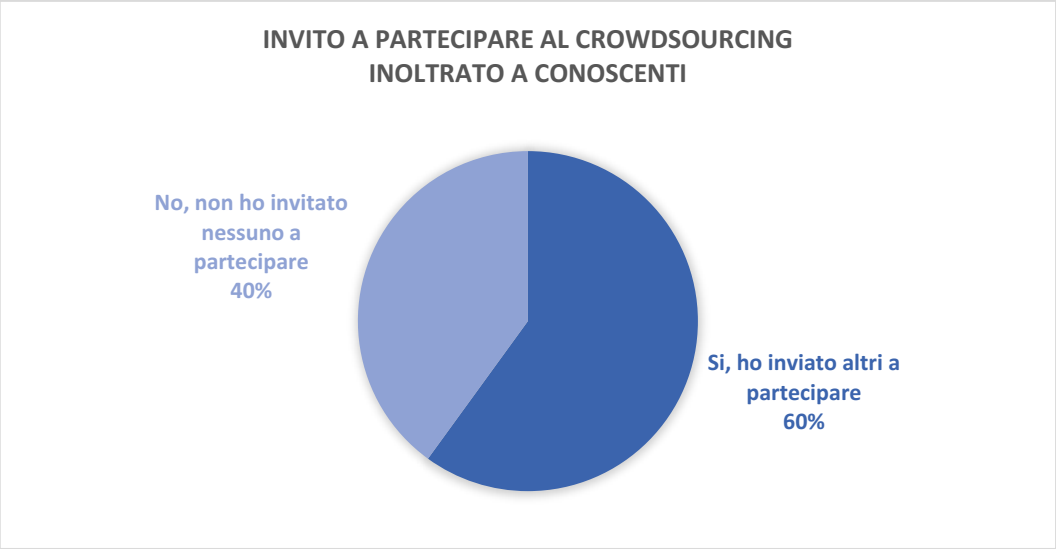
9 | Grado di apprezzamento della partecipazione al crowdsourcing per fascia d'età

Osservando i dati sul grado di intuitività della piattaforma in base all'età, si nota come tutte le fasce diano un giudizio sommariamente positivo. Ciononostante, idealmente, il livello di intuitività di una piattaforma che si pone come obiettivo la raccolta di dati lessicali, dovrebbe essere eccellente. Da considerare parallelamente a questo dato sono i *feedback* ricevuti da *VerbaAlpina* da alcuni utenti di *Facebook* a seguito di un'attività intensificata di ricerca di contatti: dopo aver pubblicato un breve testo di presentazione di *VerbaAlpina* insieme a un appello a partecipare, una decina di utenti, avendo riscontrato difficoltà di carattere tecnico, hanno contattato *VerbaAlpina* per ricevere delucidazioni.



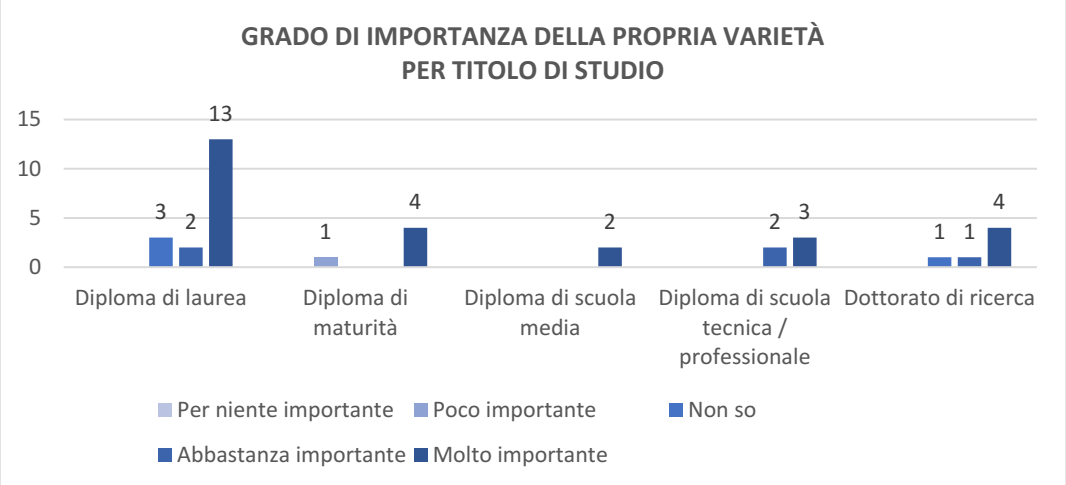
10 | Grado di intuitività della piattaforma di crowdsourcing di VerbaAlpina per fascia d'età

Per ciò che concerne la creazione di una community, il dato che traspare dalla breve inchiesta è che non tutti i *crowder*, dopo aver contribuito con le loro conoscenze, hanno invitato altre persone a partecipare.



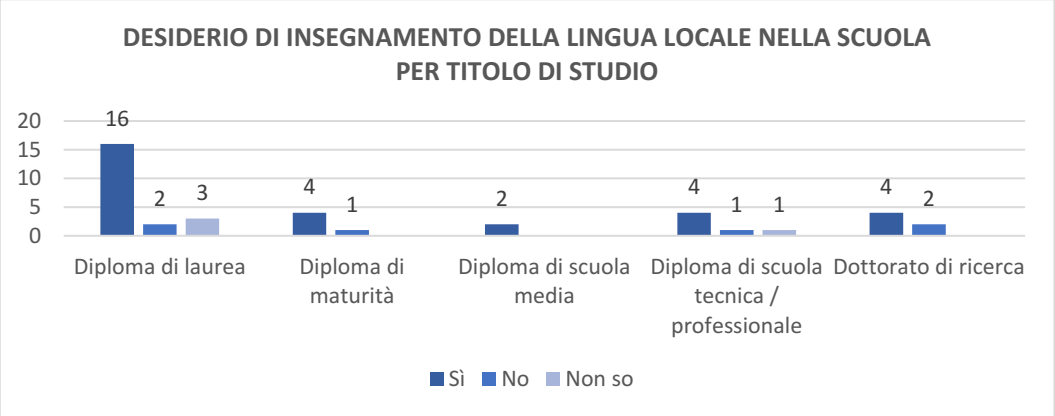
11 | Invito a partecipare al *crowdsourcing* inoltrato a conoscenti

Dopo aver mostrato e discusso i dati soprastanti, rimane un ultimo aspetto interessante da mettere in rilievo: il grado di istruzione. Si potrebbe (erroneamente) pensare che, in linea di massima, coloro che si sentono più legati e si interessano alle questioni a esso legate siano persone mediamente meno istruite. Le risposte al questionario hanno fornito una prospettiva diversa e interessante sulla questione. Su un totale di 40 partecipanti, 21 persone si sono dichiarate in possesso di un diploma di laurea. I dati relativi all'importanza concessa alla propria varietà, al desiderio di insegnamento della lingua locale nelle scuole e, più specificatamente, a *VerbaAlpina*, il grado di apprezzamento della partecipazione al *crowdsourcing*, hanno mostrato come un percorso di istruzione completo non sia correlato a un disinteresse verso la propria lingua locale, anzi vi può essere un interesse e un attaccamento alla realtà dialettale molto forte. I diagrammi seguenti rendono i dati in maniera grafica:



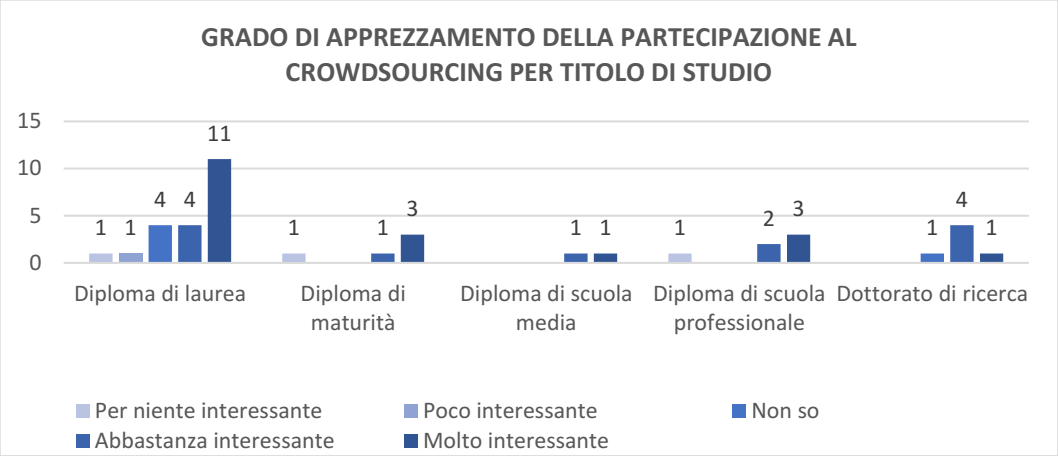
12 | Grado di importanza concesso alla propria varietà per titolo di studio

Si notino soprattutto le colonne relative alle persone in possesso di diploma di laurea e di dottorato di ricerca. Oltre a concedere importanza alla propria varietà, i parlanti sono dell'idea che si dovrebbe introdurre l'insegnamento della lingua locale nel percorso scolastico dei bambini.



13 | Desiderio di insegnamento della lingua locale nella scuola per titolo di studio

Per concludere, anche la stessa partecipazione alle inchieste di *VerbaAlpina* ha riscosso successo presso i più istruiti. I giudizi dei possessori di laurea e di dottorato sono, come si vede, positivi.



14 | Grado di apprezzamento della partecipazione al crowdsourcing per titolo di studio

5. Considerazioni globali

Alla luce dei dati e delle informazioni apportate nel corso del presente scritto, alcune considerazioni riassuntive sono doverose. Innanzitutto, la decisione di sottoporre i *crowder* registrati a una breve inchiesta si è rivelata decisamente efficace, poiché è proprio sulla base dei dati raccolti che *VerbaAlpina* ha potuto (e voluto) adoperarsi per apportare un miglioramento.

Si accennava all'inizio al fatto che la piattaforma di *crowdsourcing* non riesce a funzionare in maniera del tutto indipendente e che è necessario un continuo supporto e accompagnamento con attività di promozione da parte dei membri del

team di *VerbaAlpina*. Ciò che si può dire con certezza in questo frangente è che l'intuizione di concentrarsi sull'utilizzo dei *social media* come strumento di promozione è stata fondamentale. Oltre alla concreta partecipazione da parte di parlanti alpini attraverso l'invio di parole dialettali, *VerbaAlpina* ha riscontrato anche un crescente interesse da parte di testate giornalistiche e/o radiofoniche nei confronti dell'attività, segno che il progetto si sta interfacciando con sempre più persone in tutto il territorio alpino. Un altro vantaggio dell'utilizzo intensivo dei *social network* è, ai fini del progetto, il fatto che il pubblico giovane (e si è visto come vi siano anche molti giovani tra i *crowder* registrati), partecipando, apporta materiale attuale, aspetto centrale se l'obiettivo di *VerbaAlpina* è quello di fornire una prospettiva diacronica sui dati lessicali della regione alpina.

VerbaAlpina ha seriamente considerato i dati raccolti mediante il breve questionario e, dopo averli contestualizzati, si è proceduto ad elaborare delle strategie di comunicazione nuove che hanno riguardato soprattutto i *social network* e apportato delle migliorie di carattere tecnico che vanno nella direzione della *user-friendliness*. Per ciò che concerne le prime, mentre in un primo momento ci si era concentrati sul contattare pagine che avessero un legame con le Alpi con la preghiera di condividere un post o di chiamare i propri utenti a partecipare al *crowdsourcing*, ci si è attivati anche nei gruppi *Facebook* che raccolgono persone di determinate città o villaggi delle Alpi. Questo tipo di attività ha immediatamente e visibilmente aumentato la partecipazione e, al contempo, pubblicizzato il progetto direttamente presso i parlanti. Tuttavia, le novità principali lanciate a gennaio 2022 hanno riguardato l'integrazione di alcune nuove funzionalità della piattaforma di *crowdsourcing*. Tenuto conto dei *feedback* anonimi registrati relativamente al grado di intuitività mediante il questionario parallelamente ai commenti diretti ricevuti dagli utenti della rete (soprattutto su *Facebook*), è stato creato un breve video-tutorial che mostra il procedimento da seguire per inviare correttamente le parole, sono stati aggiunti alcuni effetti per rendere più visibili i passi da compiere (per esempio, prima si clicca su «comune» e poi su «concetto») e si è resa più visibile la scelta dei domini concettuali attraverso una breve spiegazione riguardante la possibilità di scelta.

5.1 Sfide

In primo luogo, *VerbaAlpina* deve cercare di attirare un maggior numero di parlanti, soprattutto provenienti dalle aree meno coinvolte, e convincerli a condividere il proprio sapere linguistico attraverso la piattaforma di *crowdsourcing* di *VerbaAlpina*. Per fare ciò, il progetto è costantemente impegnato a diffondere la propria presenza sul web attraverso l'uso attivo dei canali social e cercando di utilizzare tutte le lingue diffuse nell'arco alpino (con l'aggiunta dell'inglese).

Un fattore importante da tenere in considerazione è la «pigrizia» del parlante: si è notato che molti *crowder* sono disponibili a fornire parole, ma non sono sempre disposti a farlo da soli, non raramente per una semplice mancanza di dimestichezza con gli strumenti tecnologici e internet online, nonostante la piattaforma sia stata volutamente concepita in maniera semplice e risulti abbastanza immediata alla maggior parte degli utilizzatori. In questi casi è fondamentale operare una

mediazione, andando incontro alle esigenze di questi parlanti e colmando le loro lacune, per esempio rimanendo a disposizione tramite le e-mail o la messaggistica istantanea oppure inviando loro una lista di parole da tradurre nel proprio dialetto via e-mail che una persona del progetto andrà poi a inserire al loro posto.

Un'altra criticità riguarda la densità di parole registrate. Al momento, per *VerbaAlpina* si verifica la compresenza di poche località ad alta densità di dati (province di Belluno, Brescia, province autonome di Trento e Bolzano e il *Landkreis Rosenheim*) e di moltissime aree a bassa densità. La situazione ideale, caratterizzata da un numero omogeneo di entrate in tutto l'arco alpino, è dunque lontana dall'essere raggiunta e l'aspetto sul quale è necessario incentrare gli sforzi è l'attività proprio nelle zone meno rappresentate (si veda il grafico in Fig. 5) cercando ad esempio di allacciare dei contatti in loco.

5.2 Opportunità

L'esperienza di *VerbaAlpina* con il *crowdsourcing* ha portato alla luce le numerose opportunità che questo strumento è in grado di offrire.

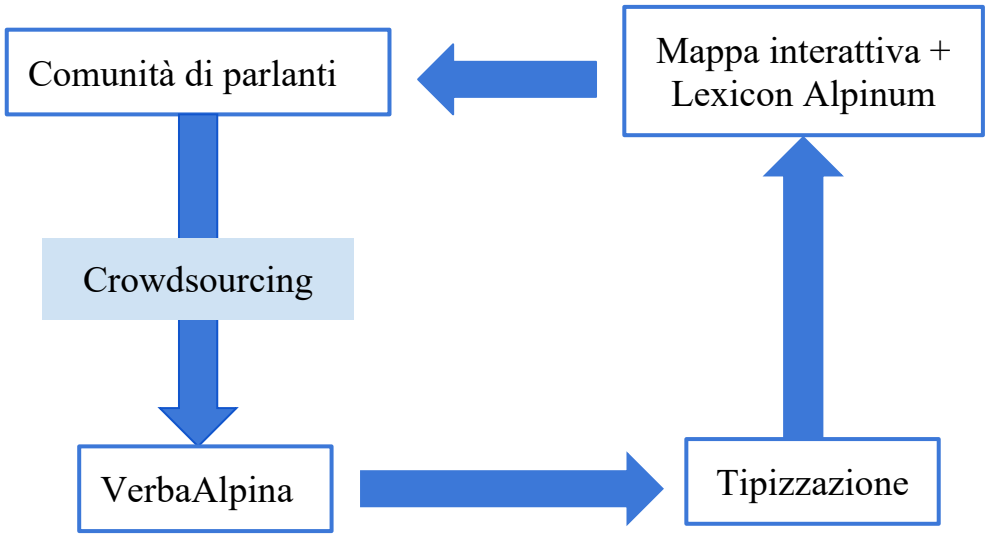
In primo luogo, l'apertura al grande pubblico per la raccolta di materiale lessicale, grazie a una piattaforma online, permette il superamento del limite di un informante per punto di rilevamento tipico della geolinguistica tradizionale. Attraverso il *crowdsourcing*, ogni punto di rilevamento, ossia ogni comune per *VerbaAlpina*, può essere infatti rappresentato da più informanti.[11] Si tratta di un aspetto da non sottovalutare se si tiene conto dell'importanza della rappresentatività. In aggiunta, contrariamente a quanto accade(va) con gli atlanti linguistici tradizionali, dal momento che l'invio di dati lessicali è legato in primo luogo alla categoria territoriale del comune e non a una lingua o a una famiglia linguistica (almeno nella prospettiva del parlante), il *crowdsourcing* dà modo di registrare la realtà linguistica di un territorio dando conto della sua eventuale situazione di plurilinguismo. Riportiamo di seguito due esempi: per il comune di Livinallongo del Col di Lana (provincia di Belluno) si evidenziano due attestazioni, inviate da due utenti distinti, afferenti a due tipi morfo-lessicali diversi per il concetto BAITA (FABBRICATO, SEMPLICE, GESTITO SULL'ALPEGGIO), ossa *tablé* (tipo morfo-lessicale romanzo *tablé* di base lat. tabulātum) e *ciasota* (tipo morfo-lessicale romanzo *casotta* che si ricollega al tipo di base lat. casa(m)). Nel comune di Longarone (sempre in provincia di Belluno), si registrano due attestazioni per il concetto PANNA (STRATO DI GRASSO SOPRA IL LATTE CRUDO) che sono da un lato *cao* (tipo morfo-lessicale romanzo *capo*, con tipo di base lat. caput) e *brama* (tipo morfo-lessicale *brama*, con tipi di base prelat. crama e lat. bruma).[12]

Il *crowdsourcing* propone un metodo di ricerca creativo, che consente di indagare le potenzialità di mezzi non canonici, ma estremamente attuali come i *social media*,

[11] In termini teorici, non vi erano impedimenti che non permettessero la documentazione di forme raccolte da informanti diversi per punto di rilevamento, ma nella pratica, la piattaforma di *crowdsourcing* permette di eseguire questa operazione in maniera più semplice e rapida.

[12] Per visualizzare gli esempi sulla sulla mappa interattiva si consulti il seguente link: <https://www.verba-alpina.gwi.uni-muenchen.de/it?page_id=27&noredirect=it_IT&db=221&tk=4533>.

per la raccolta di dati linguistici. Storie con quiz, rubriche, contenuti divulgativi, collaborazioni con altre pagine con obiettivi simili attraverso post o eventi sono tutti elementi potenzialmente fertili per la ricerca linguistica. Come accennato più volte, il *crowdsourcing* pone il parlante al centro, rendendolo (s)oggetto di studio privilegiato; inoltre, permette la raccolta di dati ecologici (ovvero non viziati dai vincoli e dalle sovrastrutture di un ambiente di laboratorio), reali e attuali, provenienti da parlanti spesso inconsapevoli della propria competenza dialettale e del proprio patrimonio linguistico locale. Attraverso il *crowdsourcing* l'accademia si avvicina alla comunità dei parlanti e si pone al suo servizio, creando uno scambio proficuo tra la conoscenza linguistica del parlante e il rigore del metodo di studio.



15 | Schematizzazione del rapporto tra *VerbaAlpina* e la comunità di parlanti[13]

5.3 Limitazioni

L'idea originale del *crowdsourcing* è l'integrazione di una comunità di interessati nei processi di elaborazione di un dato progetto. Nella teoria non si tratta quindi del mero invio di dati (nel caso di *VerbaAlpina*, dati linguistici), bensì la partecipazione potrebbe (o dovrebbe) estendersi anche ad altri ambiti come per esempio l'utilizzo della cartina interattiva per scopi personali, il salvataggio di cartine, la partecipazione attiva attraverso la promozione del progetto nella propria cerchia di conoscenze. Come è già stato evidenziato *supra*, la piattaforma di *crowdsourcing* di *VerbaAlpina*, pur funzionando egregiamente dal punto di vista tecnico, non è in grado di sostentarsi da sola e di garantire un flusso di dati costante se non vi è attività di accompagnamento continuo da parte del progetto stesso. Una difficoltà per la creazione di tale community è rappresentata dal fatto che *VerbaAlpina* si basa sulla partecipazione completamente volontaria di persone che decidono di impiegare il proprio tempo per aiutare la scienza. Nell'assetto sociale attuale, il tempo è divenuto una risorsa preziosa che tende a mancare e la maggior parte delle persone operano scelte di pragmaticità scegliendo di impiegare il loro

[13] Oltre ad essere strumenti al servizio delle comunità di parlanti, la mappa interattiva e il *Lexicon Alpinum* sono a disposizione di scienziati e accademici per la propria ricerca.

tempo libero in attività di cui, nella loro prospettiva, riscontrano immediatamente un vantaggio a breve o a lungo termine. Per questo, nella fase iniziale della promozione della piattaforma di *crowdsourcing*, *VerbaAlpina* ha profuso un grande impegno nel contattare diverse istituzioni dell'arco alpino con la preghiera di fare da tramite presso i parlanti. Si tratta perlopiù di istituzioni ed enti legati territorialmente alle Alpi oppure che si occupano di salvaguardia del patrimonio culturale, linguistico o storico locale. Si dava per scontato che la comunicazione avrebbe potuto dare maggiori frutti se mediata da enti e istituzioni locali, nelle quali (si pensava) i parlanti si identificano, rispetto alla comunicazione diretta tra *VerbaAlpina* (all'epoca sconosciuto ai più) e i parlanti poiché essi avrebbero capito che la documentazione della propria varietà avrebbe giovato sia a livello locale (per il proprio villaggio), sia a livello alpino. Tuttavia si è compreso ben presto che questa idea sarebbe stata limitata da diversi ostacoli. Molto spesso non è stata recepita alcuna risposta oppure è stata riscontrata poca predisposizione alla collaborazione. Le ragioni di tale atteggiamento possono essere diverse. Probabilmente il primo motivo è da ricollegare, ancora una volta, alla mancanza di tempo e di risorse da parte degli impiegati delle istituzioni che sono state contattate. Ciò che si riscontra però, considerato il contesto generale e ampliando la riflessione, è la mancanza di visione comune nella creazione di una community dalla quale gruppi di ricerca, istituti di cultura ed enti del territorio alpino potrebbero approfittare. Mentre il rapporto tra *VerbaAlpina* e il parlante è abbastanza diretto, ciò che blocca la richiesta di *VerbaAlpina* di fungere da intercessore presso i parlanti potrebbe essere una sorta di esitazione da parte degli enti esterni nell'affidarsi a una piattaforma online per la raccolta dei dati. *VerbaAlpina* ha dunque deciso di operare un cambio di direzione e di concentrare il proprio impegno sulla comunicazione diretta tra il progetto e i parlanti.

5.4 Prospettive

Il più importante (e più complesso) obiettivo che *VerbaAlpina* si propone è quello di creare una comunità internazionale di parlanti che partecipi al *crowdsourcing* in maniera costante nel corso del tempo. Si tratta di una prospettiva ambiziosa che vuole essere raggiunta progressivamente portando avanti e consolidando l'attività sul web e sui *social network*. Al fine di raggiungere questo obiettivo, si ritiene fondamentale ricorrere alla pratica piuttosto tradizionale del classico passaparola tra i parlanti con il conseguente sviluppo di una community organica e, contemporaneamente, conferire una struttura a tale insieme. Per fare ciò, è necessario continuare a sviluppare idee su come i *crowder* più coinvolti e interessati al progetto possano diventare uno strumento utile per diffondere il progetto, diventando a tutti gli effetti degli ambasciatori di *VerbaAlpina*. Ad esempio, si potrebbe conferire loro uno status speciale di *ambassador* o *top crowder*, dando la possibilità di gestire una piccola sub-community territoriale. Un ulteriore mezzo di comunicazione con i *crowder* potrebbe essere un canale *Telegram* di *VerbaAlpina* (da pubblicizzare adeguatamente sui *social network*), in cui tutte le persone interessate al progetto possono sentirsi libere di condividere i propri pensieri riguardo ai temi cari a *VerbaAlpina*. Un canale *Telegram* potrebbe essere anche un modo per interagire in modo diretto con più utenti in contemporanea, magari

proponendo una parola del giorno e chiedendo a tutti gli iscritti al canale di tradurla nel proprio dialetto.

6. Porsi all'ascolto

Nel corso di questo scritto si è presentata l'esperienza di *VerbaAlpina* ponendo l'accento su criticità, opportunità e sfide della piattaforma di *crowdsourcing*. Il fulcro della questione gira attorno all'idea di un'apertura all'inclusione dei parlanti da parte del progetto di ricerca. Una riflessione finale che desideriamo portare riguarda il rapporto tra *VerbaAlpina* e la comunità dei suoi *crowder* attuali e di quelli che lo diverranno in futuro. Lasciare spazio a persone non-scienziate non può e non deve limitarsi a chiedere loro qualcosa. È fondamentale assecondare richieste e necessità, ma soprattutto rendere queste persone partecipi attraverso attività di divulgazione e di pubblicazione dei risultati, attraverso momenti di dialogo e con opere di sensibilizzazione rispetto ai temi della dialettologia e della cultura locale. Inoltre, è importante prestare ascolto ai parlanti, farli sentire protagonisti di un'attività transnazionale e parte di una comunità panalpina. Le realtà accademiche spesso faticano a comunicare con persone esterne all'ambito poiché interessi, modalità e termini non sempre coincidono, ma lo sforzo va compiuto al fine di propagare l'idea di una scienza più inclusiva e meno esclusiva, abbracciando proprio *in toto* l'idea di *citizen science*.

Bibliografia

AA. VV. 2001-. «Wikipedia, l'enciclopedia libera e collaborativa.», versione italiana.
<https://it.wikipedia.org/> 21.1.2022.

COLCUC, Beatrice & Anna Rodella. 2021. *Ci racconti chi sei? VerbaAlpina, il Crowdsourcing e la vita nelle Alpi*. Online: VerbaAlpina.
<https://www.verba-alpina.gwi.uni-muenchen.de/?fragebogen=questionario-verbaalpina_it.> 15.2.2022.

COLCUC, Beatrice & Christina Mutter. 2020. «Interopérabilité des données géolinguistiques à l'exemple du projet VerbaAlpina.» In *Bien Dire et Bien Aprandre – Revue de Médiévistique*, 35, 131–146.
<https://doi.org/10.5282/verba-alpina?urlappend=%2Fwp-content%2Fuploads%2Fbdba-35-colcuc-mutter.pdf%3Fdb%3D202>.

CROWDSOURCING. 2022. «Piattaforma di crowdsourcing di VerbaAlpina» *VerbaAlpina* 21/2.
<https://www.verba-alpina.gwi.uni-muenchen.de/crowdsourcing> 15.2.2022.

ECO MEDIATICA. 2022. «Eco mediatica di VerbaAlpina» *VerbaAlpina* 21/2.
<https://www.verba-alpina.gwi.uni-muenchen.de/it/?page_id=175&db=212> 15.2.2022.

ECSA = European Citizen Science Association (ed.). 2013-.
<https://ecsa.citizen-science.net/> 31.01.2022.

GILLIVER, Peter. 2020. *The OED's latest exercise in crowdsourcing*. Oxford: Oxford University Press.
<https://public.oed.com/blog/the-oeds-latest-exercise-in-crowdsourcing/>.

JABERG, Karl & Jakob, Jud. 1928-1940. *Sprach- und Sachatlas Italiens und der Südschweiz (AIS)*, 8 vol., Zofingen: Ringier.

KREFELD, Thomas & Stephan Lücke (ed.). 2014a. *VerbaAlpina. Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit*. München.
<https://dx.doi.org/10.5282/verba-alpina>.

KREFELD, Thomas & Stephan Lücke (ed.). 2014b. «Lexicon Alpinum.» *VerbaAlpina* 21/2.
<https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=12180&db=212> 15.2.2022.
KREFELD, Thomas & Stephan Lücke (ed.). 2014c. «Timeline.» *VerbaAlpina* 21/2.
<https://www.verba-alpina.gwi.uni-muenchen.de/it/?page_id=547&db=212> 15.2.2022.
KREFELD, Thomas & Stephan Lücke (ed.). 2014d. «API Dokumentation.» *VerbaAlpina* 21/2.
<https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=8844&db=212> 26.1.2022.
KREFELD, Thomas. 2019a. «Umanistica digitale.» *VerbaAlpina* 21/2.
<https://doi.org/10.5282/verba-alpina?urlappend=%2Fit%3Fpage_id%3D21%26db%3D212%26letter%3DU%2314>.
KREFELD, Thomas. 2019b. «Trascrizione.» *VerbaAlpina* 21/2.
<https://doi.org/10.5282/verba-alpina?urlappend=%2Fit%3Fpage_id%3D21%26db%3D212%26letter%3DT%2357>.
KUNZMANN, Markus & Christina Mutter. 2021. «Zur Verwertbarkeit indirekt erhobener Sprachdaten: Erfahrungen aus dem Forschungsprojekt VerbaAlpina.» In *Linguistik grenzenlos: Berge, Meer, Käse und Salamander 2.0*, Lücke, Stephan et al. (ed.), online, versione 1.
<http://www.kit.gwi.uni-muenchen.de/?p=74940&v=1>.
LÜCKE, Stephan. 2020. «Lizensierung.» *VerbaAlpina* 21/2.
<https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D212%26letter%3DL%2341>.
LÜCKE, Stephan. 2021. «FAIR-Prinzipien.» *VerbaAlpina* 21/2.
<https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D212%26letter%3DF%23128>.
Mappa interattiva. 2022a. «Mappa interattiva del tipo di base lat. *butyru(m)*.» *VerbaAlpina* 21/2.
<https://www.verba-alpina.gwi.uni-muenchen.de?page_id=133&db=212&tk=4041> 14.2.2022.
Mappa interattiva. 2022b. «Mappa interattiva di VerbaAlpina.» *VerbaAlpina* 21/2.
<https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=133&db=212> 15.2.2022.
Mappa interattiva. 2022c. «Mappa interattiva di VerbaAlpina sulla provenienza dei crowder.» *VerbaAlpina* 21/2.
<https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=133&db=xxx&tk=4047&layer=5> 15.2.2022.
MARTIN, Nicole, Stefan Lessmann & Stefan Voß. 2008. «Crowdsourcing: Systematisierung praktischer Ausprägungen und verwandter Konzepte.» In *Multikonferenz Wirtschaftsinformatik*, Bichler, Martin (ed.), 1254 – 1263, Berlin: GITO-Verlag.
OGILVIE, Sarah (ed.). 2018. «Crowdsourcing the OED.» *Dictionary Lab.*
<https://dictionarylab.web.ox.ac.uk/crowdsourcing-oed> 31.01.2022.
Statistiche live 2017-. «Statistiche live dell'attività di crowdsourcing.» *VerbaAlpina* 21/2.
<https://www.verba-alpina.gwi.uni-muenchen.de/it/?page_id=523&noredirect=it_IT&db=212> 15.2.2022.
TOCHTERMANN, Klaus (ed.). 2018-. *Generic Research Data Infrastructure - GeRDI*. Kiel, Hamburg.
< https://www.gerdi-project.eu/> 26.1.2022.
Universitätsbibliothek Erlangen-Nürnberg (ed.). 2018-. *eHumanities – interdisziplinär*.
<https://www.fdm-bayern.org/ehumanities-interdisziplinaer/> 26.1.2022.
University Library LMU Munich (ed.). 2021-2022. *Discover*.

<https://discover.ub.uni-muenchen.de/> 26.1.2022.

VerbaAlpina su Facebook 2016-. *Profilo di VerbaAlpina su Facebook*. <https://www.facebook.com/verbaalpina/> 15.2.2022.

VerbaAlpina su Twitter 2016-. *Profilo di VerbaAlpina su Twitter*. <https://twitter.com/verbaalpina/> 15.2.2022.

VerbaAlpina su Instagram 2016-. *Profilo di VerbaAlpina su Instagram*. <https://www.instagram.com/verba.alpina/ > 15.2.2022.

WILKINSON, Mark et al. 2016. «The FAIR Guiding Principles for scientific data management and stewardship.» *Sci Data* 3 (160018). <https://doi.org/10.1038/sdata.2016.18>.

ZOONIVERSE. 2009-. *Progetto VerbaAlpina su Zooniverse*. <https://www.zooniverse.org/projects/filip-hr/verbaalpina/> 15.2.2022.

Riassunto

Il progetto *VerbaAlpina* della Ludwig-Maximilians-Universität di Monaco di Baviera sta elaborando la sua terza fase, incentrata sul lessico alpino relativo all'ecologia e al turismo in area alpina. Dal momento che gli atlanti linguistici dell'arco alpino non contengono tale lessico, si è proceduto alla raccolta del materiale lessicale autonomamente utilizzando un'apposita piattaforma. Il presente contributo desidera dare comunicazione del contesto di utilizzo, dei successi e delle limitazioni di tale piattaforma e, nell'ottica della condivisione delle esperienze, si rivolge in prima linea a ricercatori che intendono intraprendere lo stesso percorso.

Abstract

In its third stage, the project *VerbaAlpina* of Munich University investigates the vocabulary of ecology and tourism in the Alpine area. Since the existing linguistic atlases do not include this vocabulary, *VerbaAlpina* collects the lexical material using a special platform. This paper aims primarily at informing about the context of use and discussing achievements and limitations of this platform. It is addressed to researchers who wish to adopt the same data collection methods.

2022, n°9
pp. 213-239
doi: https://doi.org/10.15460/apropos.9.1902

Annette Gerstenberg & Cord Pagenstecher

‚Mi ricordo', ‚je me souviens': ich erinnere mich.

Sammlungsübergreifende Interviewanalysen in Oral History und Korpuslinguistik

Annette Gerstenberg
ist Professorin für romanische Sprachwissenschaft (Italienisch und Französisch) an der Universität Potsdam.
annette.gerstenberg@uni-potsdam.de

Cord Pagenstecher
ist Historiker am Center für Digitale Systeme an der Universitätsbibliothek der Freien Universität Berlin, Bereich Digitale Interview-Sammlungen.
cord.pagenstecher@cedis.fu-berlin.de

Keywords

Oral History – Korpuslinguistik – Erinnerung – Zeitzeugen – Interviews

1. Einleitung

Geschichtswissenschaft und Linguistik führen und analysieren Audio- und Video-Interviews, gehen dabei aber sehr unterschiedlich vor und arbeiten selten zusammen. Die soziolinguistischen Interviews der Sprachwissenschaft und die lebensgeschichtlichen Interviews der Oral History entstehen in unterschiedlichen, im Folgenden erläuterten Fragekontexten. Beide sind biographische Zeugnisse und historische Quellen, zugleich aber auch sprachlich gestaltete mündliche Erzählungen. Durch das Entstehen großer digitaler Archive sind die Interview-sammlungen zu nachnutzbaren Forschungsdaten geworden, bei deren Analyse die qualitativ-hermeneutische Oral History und die quantitativ-thesenorientierte Korpuslinguistik fruchtbar zusammenarbeiten können. Dabei stellen unter-schiedliche Erkenntnisinteressen, Transkriptionsweisen, Analysemethoden und Terminologien freilich eine Herausforderung dar. Besonders gilt dies bei einer über-greifenden Analyse verschiedener Sammlungen, die bislang in keiner der beiden Disziplinen gängig ist, durch die neue Forschungsumgebung *Oral-History.Digital* nun aber ermöglicht wird.

In der Gegenüberstellung einer Oral-History-Sammlung (*Zwangsarbeit 1939-1945*) und eines linguistisch konzipierten Korpus' biographischer Interviews (*LangAge*) skizziert dieser Beitrag Potenziale der interdisziplinären Analyse. In beiden Sammlungen stehen individuelle und kollektive Erinnerungen im Zentrum; in beiden Sammlungen spielt die lebensgeschichtliche Erfahrung des Zweiten Weltkriegs eine wichtige Rolle. Die hier vorgestellte Interviewanalyse beschränkt sich auf französisch- und italienischsprachige Interviews, ist sammlungsübergreifend angelegt und verbindet Fragestellungen und Methoden aus Korpuslinguistik und Oral History. Um die Sprache der Erinnerung an Weltkrieg und Zwangsarbeit zu erforschen, verfolgen wir einerseits einen thesengeleiteten Ansatz und untersuchen sprachliche Merkmale, deren Verwendungshäufigkeit und Verwendungskontexte Aufschluss geben können über individuelle und kollektive Erinnerungs- und Erzählmuster. Wir fokussieren dabei auf Personalpronomina der ersten Person, Singular und Plural, als erste Annäherung an die Beteiligten (cf. Knowles et al. 2021). In einem weiteren, datengetriebenen Ansatz untersuchen wir andererseits Interviewsammlungen nach viel verwendeten Wortkombinationen (N-Gramme) und unterziehen dann auffällig häufige Wendungen einer genaueren Analyse. Als inspirierend erwies sich in den hier untersuchten drei Beständen – italienische Interviews aus *Zwangsarbeit 1939-1945*, französische Interviews aus *Zwangsarbeit 1939-1945*, sowie *LangAge* als rein französisches Korpus – vor allem die wiederkehrende Wendung *ich erinnere mich*. Die Ergebnisse zeigen teils sprach-, teils sammlungsspezifische Besonderheiten, für deren genaueres Verständnis relevante Interviewpassagen hinzugezogen werden.

2. Interviews aus Linguistik und Oral History

2.1 (Sozio-)linguistische Interviews

Soziolinguistische Interviewtechniken orientieren sich an der Zielsetzung, unter kontrollierten Bedingungen ein Spektrum sprachlicher Ausdrucksweisen zu erfassen, um Phänomene von Sprachwandel und -variation zu untersuchen. Dieses reicht von überwachtem, möglichst spontanem bis hin zum unüberwachtem Sprechen, ergänzt durch Techniken wie Vorlesen von Texten und Wortlisten sowie die Antwort auf Fragen nach Spracheinstellungen und sozioökonomischem Hintergrund (cf. Briggs 2005, 1053; Labov & Auger 1993, 116). In vielen Kontexten sind Interviews auf Grund ihrer hohen Vergleichbarkeit nicht zu ersetzen, wie Wagner und Tagliamonte (2017, 213) feststellen: „Interviews remain the gold standard for eliciting a range of vernacular [maximally unmonitored] and more formal speech styles from informants." Parallel zur Weiterentwicklung unterschiedlicher Interviewformate wurde ein breites Spektrum nicht eigens für die Erhebung angebahnter (elizitierter), sondern natürlich vorkommender Interaktion für die Untersuchung der gesprochenen Sprache erschlossen. Daraus ergaben sich wiederum Effekte für die Einordnung von Interviews, die in konstruierten Forschungssettings geführt wurden, insofern diese selbst in ihrer interaktionalen Dimension verstanden wurden (cf. Deppermann 2013) und die kooperative Herstellung in die Auswertung einbezogen wird. So zeigen Wagner und Tagliamonte (2017), wie dieses Bewusstsein in die Interviewführung mit einbezogen werden kann, indem

nicht der Versuch unternommen wird, die Rollen zu neutralisieren, sondern sie bewusst einzusetzen.

2.2 Oral History-Interviews

Aus einer ganz anderen Forschungstradition stammen die lebensgeschichtlichen Interviews in der Geschichtswissenschaft. Seit der „Geburt des Zeitzeugen" (Sabrow & Frei 2012) nach 1945 wurden viele Interviewprojekte nach der Methode der Oral History durchgeführt, besonders mit Holocaust-Überlebenden und Opfern anderer staatlicher Verfolgungen.

Die im angelsächsischen Raum schon länger etablierte Oral History wurde seitens der – lange Zeit auf staatliches Handeln und schriftliche Überlieferungen fixierten – deutschen Geschichtswissenschaft zunächst skeptisch betrachtet. Vor allem außerakademische Initiativen wie Geschichtswerkstätten oder Gedenkstätten begannen, Zeitzeugen-Interviews zu führen. Technisch dominierten zunächst Audioaufnahmen, die seit den 1980er Jahren zunehmend durch verschiedene Videoformate nicht ersetzt, aber ergänzt wurden. Vor allem mit dem Cultural Turn und dem Boom der Memory Studies sind audiovisuell aufgezeichnete Erinnerungen nun aber zu wesentlichen Quellen für die Alltags-, Kultur- und Geschlechtergeschichte geworden (cf. Apel, Leh & Pagenstecher 2022; Eusterschulte, Knopp & Schulze 2016).

Ein Oral History-Interview dauert in der Regel mehrere Stunden. Im Kern steht eine von den Interviewten selbst strukturierte lebensgeschichtliche Erzählung; erst in einer zweiten Phase werden Nachfragen zu dieser Erzählung gestellt, um weitere Narrationen zu evozieren. Anschließend folgen thematische Fragen, schließlich oft eine gemeinsame Betrachtung privater Fotos und Dokumente. Beabsichtigt (wenngleich in der Praxis nicht immer erreicht) sind also im Dialog inspirierte Narrationen der „Erzähler*innen", keine investigativen Frage-Antwort-Interviews mit „Respondent*innen". Interessant sind in der Interpretation weniger die berichteten Fakten als ihre erinnernde Deutung.

Darüber hinaus dienen die lebensgeschichtlichen Erzählungen der Geschichtswissenschaft als Quellen für Ereignisse oder Erfahrungen, die in schriftlichen Archivquellen unzureichend oder einseitig dokumentiert sind, etwa von unterdrückten Minderheiten. Gerade im Hinblick auf Massenverbrechen wie den Holocaust ist das Bezeugen des Geschehenen ein zentrales Motiv für und in den Interviews. Ein Zeugnis abzulegen, stellvertretend für die Ermordeten, ist immer auch ein „Versprechen der Wahrheit" (Eusterschulte, Knopp & Schulze 2016, 21), „eine erlebte, *eingefleischte* Wahrheit" (Klüger 1996, 405, Herv. im Original). Angesichts der vielfach bewusst vernichteten Akten erweisen sich mündliche Zeugnisse häufig als wertvolle und erstaunlich exakte Quellen für die Ereignisgeschichte oder Topographie bestimmter Lager oder Verbrechenskomplexe. Viele Oral History-Interviews entstehen in Gedenkstätten oder im Rahmen von Erinnerungs-Initiativen, sind also nicht primär Forschungsdaten, sondern entstehen als wichtiger Teil des kulturellen Erbes, genau am Übergang vom kommunikativen zum kulturellen Gedächtnis.

Ein Oral History-Interview ist gewissermaßen als mündliche Autobiographie zu interpretieren; die Zeitzeug*innen sind die Autor*innen dieses Werks. Die Gestaltung dieser Autobiographie ist ein höchst gewagter Prozess, denn der audiovisuelle Text eines Interviews entsteht live am Mikrofon. Es gibt kein Manuskript, keine Korrekturmöglichkeit, kein Lektorat; das Video erlaubt es uns, „den Überlebenden beim Erinnern zuzusehen" (Nägel 2016, 356). Die meist betagten Autor*innen müssen spontan, möglicherweise noch vor laufender Kamera, und häufig in einer anderen als ihrer Erstsprache, eine erzählerische und sprachliche Form für ihre Lebensgeschichte finden – und dabei mit einer möglichen Reaktivierung traumatischer Erfahrungen umgehen. Zudem wird dieser autobiographische Text nicht nur verfasst, sondern gleich auch zur Aufführung gebracht, denn die Lebensgeschichte wird im Interview mit entsprechender Betonung, Mimik und Gestik in Szene gesetzt. Jenseits der schwierigen Erinnerungsarbeit ist ein Interview also auch eine performative Gestaltungsleistung.

Ferner ist der audiovisuelle Text eines Oral History-Interviews das Ergebnis eines Gesprächs-Settings, eines medial aufgezeichneten Dialogs, einer „Erzählgemeinschaft" (Nägel 2016, 352). Die Interviewenden sind Co-Autor*innen beim gemeinsamen Verfertigen und Aufführen der lebensgeschichtlichen Erzählung (cf. Pagenstecher & Pfänder 2017). Die schwierige Rolle der Interviewenden ist diesen in der Praxis wohl bewusst, bleibt in der Forschung aber oft unterbelichtet. Als Zuhörende sind sie – und die Kameraleute – die primären Adressat*innen der Erzählung und repräsentieren zugleich doch auch zukünftige Rezipient*innen, vielleicht gar die Nachwelt schlechthin.

Oral History-Interviews sind also historische Quelle, oft auch Zeugnis für anderweitig kaum dokumentierte Verhältnisse und Geschehnisse bis hin zu Verbrechen; zugleich sind sie sprachlich, narrativ und performativ gestaltete Erzählungen und Deutungen der eigenen Lebensgeschichte, die in einem dialogischen Aufnahmesetting entstehen. Und sie sind umfangreiche Daten der gesprochenen Sprache, die auf unterschiedliche Weise analysiert werden können.

3. Digitale Interview-Archive

3.1 Interview-Sammlungen

Neben dem Setting der einzelnen Aufnahme ist auch der breitere Entstehungskontext der Interviews zu berücksichtigen. In der Regel werden Oral History-Interviews nicht isoliert, sondern im Rahmen eines Forschungsprojekts, einer Museumssammlung oder einer erinnerungskulturellen Initiative finanziert und durchgeführt. Am Beispiel des *Visual History Archive* der Shoah Foundation (Michaelis 2013; Shenker 2015; Taubitz 2016; Bothe 2019) wurde untersucht, wie stark sich die Auswahlkriterien, Interviewmethoden oder Kamerarichtlinien des Projekts auf die Gestaltung des Interviewkorpus auswirken.

Als wichtiger Teil unseres kulturellen Erbes sind die in Erinnerungsinstitutionen und Forschungsprojekten entstandenen Oral History-Interviews heute großenteils digitalisiert. Entsprechende Sammlungen sind aber verstreut und schlecht zugänglich.

An Universitäten wurden die Audio- oder Video-Interviews lange Zeit nur von den Interviewenden selbst für ihre Publikationen genutzt. Erst in letzter Zeit versteht man sie zunehmend auch als Forschungsdaten, die für verschiedene Disziplinen nachnutzbar sind – oder sein sollten (cf. Apel, Leh & Pagenstecher 2022). Große Interview-Archive wie das *Archiv Deutsches Gedächtnis* an der FernUniversität in Hagen (über 3.000 Interviews) oder die *Werkstatt der Erinnerung* in Hamburg (über 2.000 Interviews) sind bislang überwiegend nur vor Ort zu konsultieren. Ähnliches gilt für die meisten Museen und Gedenkstätten mit ihren Interviewsammlungen. Es gibt keinen Katalog, der eine sammlungsübergreifende Recherche in den verstreuten Beständen ermöglichen würde.

3.2. Die Forschungsumgebung *Oral-History.Digital*

An der Freien Universität Berlin entsteht im DFG-Projekt *Oral-History.Digital* (cf. Oh.d 2020–2022) nun eine Informationsinfrastruktur für wissenschaftliche Interview-Sammlungen. Die bis 2023 fertiggestellte Arbeitsumgebung unterstützt Sammlungsinhaber*innen bei der Archivierung, Erschließung und Bereitstellung, Forschungsprojekte bei der Recherche, Annotation und Auswertung von Oral History-Sammlungen. Methodisch und technologisch stützen wir uns bei *Oral-History.Digital* auf Erfahrungen aus früheren Projekten wie den Interview-Archiven *Zwangsarbeit 1939–1945* (ZWAR 2009–2022) oder *Colonia Dignidad. Ein chilenisch-deutsches Oral History-Archiv* (seit 2022).

Die Erschließungs- und Forschungsumgebung *Oral-History.Digital* wird diese und weitere Bestände umfassen, u. a. aus dem *Archiv Deutsches Gedächtnis* in Hagen und der *Werkstatt der Erinnerung* in Hamburg, aber auch aus verschiedenen Museen und Stiftungen. Auch linguistische Interviewprojekte wie *LangAge* werden darüber nutzbar sein. Die versammelten Ressourcen sind thematisch breit gefächert und betreffen zum Beispiel die DDR-, Bergbau- oder Gewerkschaftsgeschichte. Weiterhin ergibt sich dadurch eine intra- und interdisziplinäre Vielfalt, wie folgende Gegenüberstellung einer Oral History-Sammlung (*Zwangsarbeit 1939–1945*) und eines linguistisch konzipierten Korpus' biographischer Interviews (*LangAge*) zeigt.

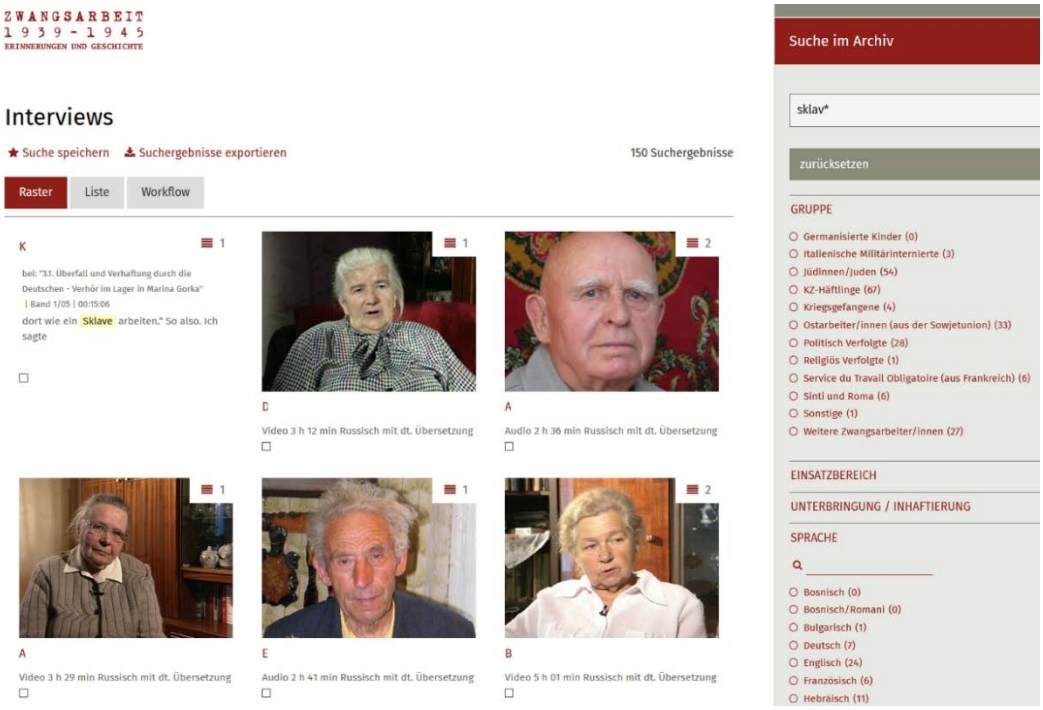
3.3. Das Online-Archiv *Zwangsarbeit 1939–1945*

Das Online-Archiv *Zwangsarbeit 1939–1945. Erinnerungen und Geschichte* ist der Erinnerung an über zwanzig Millionen Menschen gewidmet, die für das nationalsozialistische Deutschland Zwangsarbeit geleistet haben (ZWAR 2009–2022). Konkret erzählen 249 Zwangsarbeiterinnen und 341 Zwangsarbeiter aus 26 Ländern ihre Lebensgeschichte in ausführlichen Audio- und Videointerviews. Dabei kommen neben der Kerngruppe der „zivilen", also dem Arbeitsamt oder privaten Firmen untergeordneten Zwangsarbeiter*innen auch die von der SS beaufsichtigten KZ-Häftlinge und die der Wehrmacht unterstellten Kriegsgefangenen zu Wort.

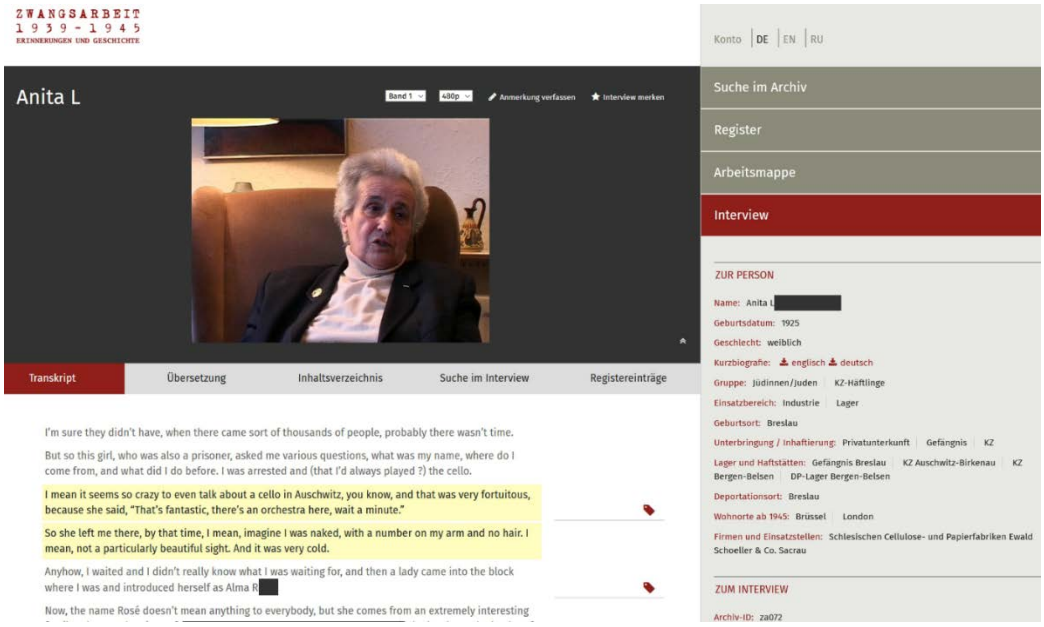
Initiiert von der Stiftung *Erinnerung, Verantwortung und Zukunft*, wurden die Interviews in den Jahren 2005 und 2006 von 32 Initiativen unter der Koordination des Instituts für Geschichte und Biographie der FernUniversität in Hagen in 26

Ländern geführt (cf. Leh, Plato & Thonfeld 2008). Die lebensgeschichtlichen Interviews haben eine durchschnittliche Länge von dreieinhalb Stunden. In einem mehrsprachigen Online-Archiv sind sie mit timecodierten Transkripten bereitgestellt, d. h., mit einer Zeitmarke nach jedem Satz oder nach etwa 120 Zeichen (cf. Apostolopoulos & Pagenstecher 2013, Pagenstecher 2017). Der Zugang erfolgt nach einer Registrierung, die vor der Freischaltung manuell geprüft wird, um die oft sehr persönlichen Lebenserzählungen gegen einen eventuellen Missbrauch zu schützen.

Nach der Anmeldung bietet das Archiv unterschiedliche Rechercheoptionen (Abb. 1a). Über die Kategoriensuche mit Rubriken wie Opfergruppe, Einsatzbereich oder Sprache werden komplette Interviews gefunden. Zudem sind die in 30 Sprachen oder Sprachkombinationen vorliegenden Transkripte im Original oder in ihrer deutschen Übersetzung durchsuchbar und als Untertitel anzeigbar (Abb. 1b). Inhaltsverzeichnisse erleichtern die Orientierung in den lebensgeschichtlichen Erzählungen, editorische Anmerkungen erläutern missverständliche Stellen oder geben weitere Quellenhinweise. Kurzbiographien sowie private Fotos und Dokumente kontextualisieren die Interviews.



1a | Suchmaske in ZWAR



1b | Das Interview mit Anita L.: Transkriptansicht

Das Online-Archiv *Zwangsarbeit 1939-1945* unterstützt eine digitale Recherche, aber nur begrenzt eine digitale Analyse. Bestimmte Begriffe und Wortkombinationen können archivweit gesucht und punktgenau angesteuert, aber nur mühsam quantitativ ausgewertet werden. So wird zum Beispiel für eine Recherche zum Wortfeld *Zwangs- und Sklavenarbeit* bei einer Volltextsuche nach *sklav** die Gesamtzahl von 147 Interviews und für jedes dieser Interviews die Anzahl der Segmente mit diesem Wortteil im übersetzten Transkript angezeigt. Durch zusätzliche Filter können die Suchergebnisse nach Gruppe, Geschlecht, Sprache oder anderen Facetten eingegrenzt werden, wodurch sich ein differenziertes Bild ergibt, welche Interviewten welche ihrer Zwangsarbeits-Erfahrungen mit Sklaverei assoziieren. Für im engeren Sinne korpuslinguistische Analysen müssen die Transkripte aber nach entsprechender Genehmigung exportiert und aufbereitet werden (s. u., Abschnitt 5).

517 der 590 Zeitzeugen-Interviews zur Zwangsarbeit sind online zugänglich; die restlichen Interviews sind aus verschiedenen Gründen zugangsbeschränkt oder noch nicht erschlossen. Die meisten Interviews liegen auf Russisch (120) und Polnisch (74) vor, aber auch die romanischen Sprachen sind vertreten: 19 Interviewte sprachen Französisch, 18 Rumänisch, 9 Italienisch, 7 Katalanisch und 2 Spanisch. Darunter finden sich neben weithin unbekannten Interviewten auch Prominente wie die Auschwitz- und Ravensbrück-Überlebende und spätere italienische Senatorin Liliana Segre, der Buchenwald-Überlebende und spätere Schriftsteller Jorge Semprún oder ein Lagerkamerad und späterer Privatsekretär des Chansonniers George Brassens.

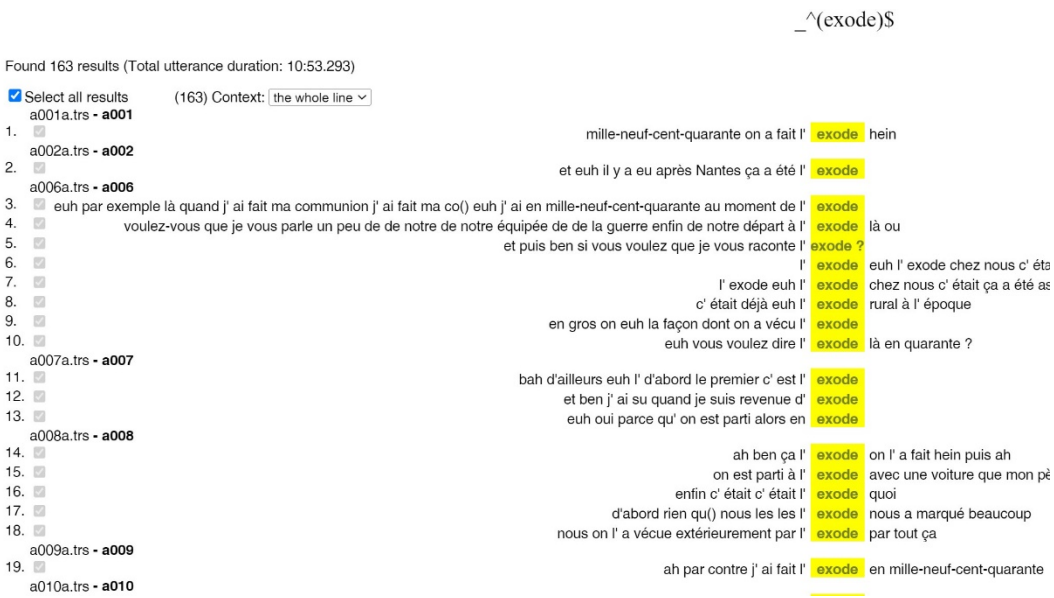
3.4. Das Korpus *LangAge*

Das französischsprachige Korpus *LangAge* entstand in einer 2005, also zeitgleich zu den ZWAR-Interviews, begonnenen Serie biographischer Interviews in Orléans, um

den Sprachgebrauch des höheren Lebensalters zu dokumentieren und der linguistischen Analyse zugänglich zu machen. Die Lebensgeschichten der repräsentierten Generation (Geburtsjahrgänge zwischen 1915 und 1935) sind eng mit dem Zweiten Weltkrieg verbunden, und das Projekt wurde als Initiative der „histoire orale" vorgestellt. Der thematische Leitfaden von *LangAge* folgt den Lebensgeschichten von der Zwischenkriegszeit über Kriegserfahrungen unter der deutschen Besatzung in die Phase des wirtschaftlichen Aufschwungs, der so genannten *Trente Glorieuses*, und bis zu den gesellschaftlichen Ereignissen und Entwicklungen, die mit dem Jahr 1968 verknüpft sind. Im letzten Abschnitt geht es um die Lebensgestaltung seit dem Ende des Erwerbslebens, wobei zahlreiche Aktivitäten, die Einbindung in familiäre und freundschaftliche Netzwerke und Einstellungen zu sozialen und sprachlichen Fragen angesprochen werden.

Die Methode des biographischen Interviews gestaltete die Interviewsituation aus, insofern die Interviewten die Autorität der Zeitzeugenschaft hatten und die Interviewerin durch aktives Zuhören ihr persönliches Interesse zum Ausdruck brachte. Durch die Adressierung unterschiedlicher Milieus wurde im gegebenen räumlichen (Wohnort Orléans und Umgebung) und zeitlichen (Lebensalter) Rahmen ein breiter historischer Erfahrungsraum einbezogen. Unter den Beteiligten ist ein Shoah-Überlebender, ein weiterer Beteiligter wurde zur Zwangsarbeit verpflichtet. Viele erlebten die deutsche Besatzung und den Krieg in Orléans; Väter und Familienangehörige oder auch sie selbst waren an der Front. Ein häufig thematisierter Zusammenhang ist der so genannte *Exode*, die Fluchtbewegung infolge des deutschen Angriffs ab Mitte Mai 1940, der aus unterschiedlichen Perspektiven berichtet wird. Insofern ergibt sich ein Spannungsfeld zwischen verbindenden historischen Erfahrungen der „älteren Generation" (ab 70 Jahren) und der Breite der individuellen Biographien.

Zugleich entstanden durch die biographischen Interviews, die monologische Erzählformen entlang einem thematischen Leitfaden bevorzugten, vergleichbare Sprachdaten von Angehörigen der älteren Generation, auf deren Basis sprachliches Altern und seine Dynamik im Längsschnitt untersucht werden konnte. Die französische Stadt Orléans in der Region Centre bot ideale Voraussetzungen, weil hier seit 1968 die groß angelegte *Étude Sociolinguistique sur Orléans* (ESLO) etabliert ist: Der ersten Erhebungsphase (ESLO1) folgte ab 2008 eine zweite Initiative von Sprachaufnahmen (ESLO2), die den Interviews im Format von ESLO1 weitere Module zur Seite stellte. Vor dem Hintergrund dieser sozial breit angelegten Studie konnte ein generationell fokussiertes Projekt wie *LangAge* an Aussagekraft gewinnen. Mit ESLO verbindet *LangAge*, dass ein sozial breites Spektrum von Personen einbezogen wurde. Die Erhebung wurde in weiteren Phasen 2012 und 2015/2016 mit denselben Personen fortgesetzt. Es handelt sich um Audio-Interviews von durchschnittlich einer Dreiviertelstunde Länge in französischer Sprache (Details s.u.), die mit anonymisierten Transkripten derzeit in einer linguistisch erschlossenen Datenbank zugänglich sind (aktuell 129 Transkripte mit 882.353 Tokens) (Abb. 2). Autorisierte Interviews der ersten Serie werden ab 2023 auch in *Oral-History.Digital* nutzbar sein.



2 | Ergebnisse für die Suche nach *exode* in *LangAge* Corpora mit LaBB-CAT

4. Korpuslinguistische Herangehensweisen an Oral History-Interviews

4.1 Forschungsstand und Forschungsfragen

Die Geschichtswissenschaft untersucht die Interviews der Oral History vor allem mit qualitativen Analysen, die sich einem Interview als Einzelquelle nähern. Sie zielen auf eine biographische und historische Kontextualisierung und eine hermeneutische Interpretation der individuellen Erfahrung und Erinnerung. Diese Ansätze haben sich bewährt und sind dem Charakter eines lebensgeschichtlich-narrativen Interviews angemessen.

Seit etwa zwei Jahrzehnten werden nun aber mehr und größere Sammlungen von Zeitzeugen-Interviews in digitalen Archiven besser erschlossen, so dass Forschende nun statt weniger selbst geführter Interviews eine größere Anzahl vorhandener Interviews für ihre Fragestellungen nutzen können. Das ermöglicht Vergleiche zwischen verschiedenen Gruppen von Interviewten. Fruchtbar waren komparative Gruppenanalysen etwa zu gemeinsamen Erfahrungsmustern verschiedener Interviewter (Browning 2010), zu den Einflüssen gesellschaftlicher Erinnerungskulturen auf die biographischen Erzählungen (Thonfeld 2014), zu Entstehungs- und Rezeptionskontexten verschiedener Interviewsammlungen (Leh, Plato & Thonfeld 2008; Michaelis 2013; Shenker 2015; Taubitz 2016; Bothe 2019) oder zu Genderaspekten (Pagenstecher & Tausendfreund 2015). Als besonders ergiebig erwies sich der Vergleich mehrerer Interviews mit der gleichen Person in verschiedenen Interview-Sammlungen (Laub & Bodenstab 2008; Kangisser Cohen 2014; Pagenstecher & Pfänder 2017; Schuch 2021).

Oral History-Quellen wurden auch diskursanalytisch (Schiffrin 2000) sowie in einigen korpuslinguistischen Studien verwendet (Grant 2010). Eine breitere wechselseitige Rezeption von Oral History und Linguistik wurde durch das 2015 von

Erich Kasten, Katja Roller und Joshua Wilbur veranstaltete Kolloquium *Oral History Meets Linguistics* angeregt. In diesem Rahmen zeigten Pagenstecher und Pfänder (2017), wie Oral History-Interviews aus einer interaktionalen Perspektive neu gelesen und verstanden werden können. Im vergleichenden Zugriff auf ZWAR-Interviews einerseits und *LangAge*-Daten andererseits untersuchte Gerstenberg (2017), wie sich in französischsprachigen Interviews die Verwendung der Begrifflichkeit für *Zwangsarbeit* bei Betroffenen (ZWAR) von gleichaltrigen Nichtbetroffenen (*LangAge*) unterscheiden: Nur die selbst zur Zwangsarbeit verbrachten Betroffenen verwendeten die personalisierte Abkürzung *un STO* für 'jemand, der den Service du Travail Obligatoire leistete'. Ein weiterer Aspekt des Vergleichs brachte hervor, dass die euphemistischen Slogans der NS-Propaganda, in denen die Zwangsarbeit als ertragreicher Arbeitsaufenthalt im Nachbarland schöngefärbt wurde, von den Betroffenen im Hinblick auf ihren fatalen Beitrag zu einem lange Zeit verharmlosenden Bild der Zwangsarbeit diskutiert wurden. Dagegen reproduzierten die nicht betroffenen Personen aus *LangAge* in einigen Fällen die Slogans unkritisch (cf. Gerstenberg 2017).

Die letztgenannten Beispiele zeigen, wie Oral History-Interviews die Grenze vom alltäglichen Sprachgebrauch zur bewussten Verwendung sozialhistorisch bedeutsamer Begriffe überschreiten (zur Unterscheidung cf. Blank 1997, 29). Als Überlagerung individueller und kollektiver Erfahrungen lassen sich in Lebenserzählungen in *LangAge* unterschiedliche Generationsbegriffe herausarbeiten, die sich auf 1968 als Generationsereignis zwischen Jungen und Alten, und zugleich zwischen Familiengenerationen beziehen (Gerstenberg 2009). Auf das Verbindende der eigenen Altersgruppe zielen auch Kommentare zum Sprachgebrauch ab: „Diachrone Markierungen" wie *comme on disait à l'époque* 'so sagte man damals' (Gerstenberg 2011, Kapitel 8.2) heben Wörter hervor, die als veraltet dargestellt werden. Diese Markierungen sind weitgehend unabhängig davon, ob ein Wort tatsächlich heute nicht mehr verständlich oder außer Gebrauch geraten ist, sie können vielmehr als Zeichen einer sprachlichen Generationsidentität verstanden werden.

In den engeren Rahmen der historischen Terminologie fällt der Diskurs um die in vielen Nachkriegsgesellschaften erhobenen Kollaborationsvorwürfe gegen die ehemaligen Zwangsarbeiter*innen (Thonfeld 2014). Diese Diskussion ist in den ZWAR-Interviews sehr präsent, entweder explizit oder im Hintergrund, vor allem in den Berichten von eigenen Sabotageversuchen bei der Zwangsarbeit. Auch die Assoziation zur Sklaverei hat hier ihre Funktion. Manche ehemaligen Zwangsarbeiter*innen beschreiben sich selbst als *Sklaven*, vor allem, wenn sie über besonders demütigende Erfahrungen wie ihre Verteilung auf verschiedene Fabriken und über die ihnen vorenthaltene Entschädigung sprechen; möglicherweise verwendeten gerade politisch interessierte Intellektuelle diese Begrifflichkeit häufiger als andere Interviewte (cf. Film „Sklavenarbeit" 2011, Pagenstecher 2010). In vielen Interviews ordnet sich die Bewertung der eigenen Zwangsarbeit so in ein konfliktbehaftetes Wortfeld zwischen Sklavenarbeit und Kollaboration ein, dessen genauere Analyse vielversprechend erscheint.

4.2 Forschungsmethoden

Um die Interviews nicht nur digital zu durchsuchen, sondern sie auch digital zu analysieren, braucht die Geschichtswissenschaft neue Methoden, die das bewährte Handwerkszeug der Interviewforschung nicht ersetzen, aber ergänzen. In den Digital Humanities werden zahlreiche quantitative Verfahren entwickelt, die auch für die Oral History genutzt werden könnten – von der Netzwerkanalyse über Topic Modeling und Text Embedding bis zu korpuslinguistischen Methoden (cf. Möbus 2020, Knowles et al. 2021). Vielversprechend sind dabei vor allem datengetriebene Ansätze (*corpus driven*), die bemerkenswerte Muster in den vorliegenden Daten zu erkennen suchen, auch jenseits bereits vorab formulierter Hypothesen. In diesem induktiven Vorgehen trifft das Interesse der Digital Humanities auf jenes der Oral History; denn auch diese sucht nicht nach Belegen für vorgefasste Thesen, sondern möchte in den Interviews zunächst immanente, auch unerwartete Deutungsebenen und Sinnstrukturen finden, die dann erst kontextualisierend interpretiert werden. Viele dieser Ansätze verlangen allerdings einen hohen Aufwand in der Datenaufbereitung bei zunächst schwer einschätzbarer Aussagekraft der Ergebnisse, insbesondere wenn einzelne Fundstellen aus den lebensgeschichtlichen Erzählungen herausgelöst und ohne deren Kontext untersucht werden.

Das berührt die zentrale Frage der Herangehensweise an die Interviews als Texte. Verstehen wir die Oral History-Interviews als autobiographische Texte, ist ein literaturwissenschaftlich informiertes Verständnis des autobiographischen Genres erforderlich, gerade auch um die Spezifika eines mündlichen Erinnerungstextes herauszuarbeiten (cf. Michaelis 2013). Blicken wir eher auf das dialogische Aufnahmesetting der Interviews, helfen die Methoden der Gesprächsanalyse und interaktionalen Linguistik, um die zentrale Rolle der Interviewenden oder die performativen Erzählstrategien der Interviewten zu beleuchten (cf. Pagenstecher & Pfänder 2017). Soweit die Interviews filmisch aufgezeichnet wurden, kann die Videoanalyse Erkenntnisse über multimodale Erzählmuster, das Körpergedächtnis oder das spontane Auftauchen von Erinnerungsfragmenten beisteuern (cf. Gülich & Pfänder 2022, Freyburger & Jäger 2021, 189). Diese Ansätze fokussieren auf einzelne Sätze, Wörter, Gesten, die sehr genau transkribiert und analysiert werden, sind also nicht korpusbasiert und nicht biographisch orientiert.

Für diesen Beitrag besonders relevant ist eine gesprächsanalytische Fallstudie zu den italienischen Interviews in *Zwangsarbeit 1939–1945* (Pfänder et al. 2022). Danach dient die Wendung *mi ricordo* ‘ich erinnere mich’ manchmal dazu, eine vorangegangene Passage zu beglaubigen, also epistemische Autorität zu behaupten. Vor allem aber bereitet *mi ricordo* einen Wechsel zwischen Textsorten vor und leitet etwa eine Erzählung eines konkreten, selbst erlebten oder gestalteten Ereignisses ein. Oft bringt *mi ricordo* dann einen Wechsel vom beschreibenden Imperfekt zum erzählenden Perfekt oder Präsens. Dieser Wechsel betrifft auch die Mimik; mit der Äußerung von *mi ricordo* wandert der Blick häufig kurz nach unten, weg von der interviewenden Person, der danach beim Erzählen wieder ins Gesicht geblickt wird.

In korpuspragmatischer Perspektive werden Möglichkeiten gesucht, quantitative Verfahren mit einer kontextsensitiven Herangehensweise zu verbinden, indem Frequenzen über verschiedene Texttypen bzw. Gruppen (Lernersprache vs. Erstsprache) hinweg oder im Übersetzungsvergleich hinzugezogen werden, wie Aijmer (2015) am Beispiel von *I think* zeigt. Grant (2010) zeigt die unterschiedlichen Funktionen von *I don't know – I dunno* in Texttypen gesprochener Sprache – beide Formen nehmen Spitzenplätze der Frequenz von N-Grammen ein („3-word chunk", Grant 2010, 2290). Sie konstatiert in den von ihr manuell annotierten Interviews, darunter auch Oral History-Interviews, zu einem großen Teil die Funktionen des distanzierenden Heckenausdrucks und der Markierung von Unsicherheit.

Dieser kurze Überblick zeigt, dass es zwischen den hermeneutischen Methoden der Oral History und korpuslinguistischen Ansätzen durchaus Verbindungslinien gibt, insofern Funktionen und Verteilungsmuster aus den Daten herausgefiltert werden. Die im nächsten Absatz behandelte Beschaffenheit der Daten spielt dabei eine fundamentale Rolle.

4.3 Unterschiedliche Transkriptionsstandards

Sammlungs- und disziplinübergreifende Analysen stehen vor der Herausforderung unterschiedlicher Traditionen und Regeln bei der Interviewführung, -aufzeichnung, -transkription und -erschließung. Problematisch ist das insbesondere bei verschiedenen Transkriptionsrichtlinien.

Die audiovisuellen Medien bilden den Kern der Interviews; für Suche, Annotation, Analyse und Zitation sind aber manuell erstellte, zunehmend auch automatisch generierte Transkriptionen oder Indexierungen von zentraler Bedeutung. Diese folgen jedoch den unterschiedlichen Transkriptionsgewohnheiten und -regeln der jeweiligen Disziplin und Sammlung.

Oral History-Transkripte sind meist orthographisch geglättete und mit Interpunktion versehene Texte, die oft als unstrukturierte Word-Dateien vorliegen und nur manchmal Timecodes und Auszeichnungselemente enthalten, z. B. für Sprecherwechsel oder Pausen. Wortwiederholungen und nonverbale Äußerungen werden nicht immer transkribiert. Sprachwissenschaftliche Transkriptionssysteme (GAT2 2009) sind unüblich und wären aufgrund der begrenzten Ressourcen bei den mehrstündigen Interviews kaum umsetzbar.

Digitale Interview-Archive wie *Zwangsarbeit 1939–1945* koppeln die oft bis zu 100-seitigen Transkripte mit den mehrstündigen Audio- oder Videoaufnahmen, indem Timecodes automatisch oder manuell in die Texte eingefügt werden (cf. Nägel 2016, 360-363). Diese Segmentierung durch regelmäßig nach bestimmten Zeitabständen (z. B. eine Minute), Zeichenzahlen (100 Zeichen) oder Sinneinheiten (Wort oder Satz) eingefügte Zeitmarken erlaubt eine synchrone Untertiteldarstellung und eine sekundengenaue Volltextsuche oder Annotation. Aus diesem Sprung in die alignierten Videos oder Audioaufnahmen ergeben sich weiterführende Perspektiven für die Kontrolle der Transkription und die Analyse der multimodalen, para- oder nonverbalen Facetten der Interviews, die auch Mimik und Gestik einbezieht. Standardisierte Im- und Exportformate wie das tabellarische

bzw. komma-getrennte Format (CSV) oder der Untertitelstandard Video Text Tracks (VTT) stehen bei entsprechender Berechtigung zur Verfügung; die in Editionen gängigen Kodierungen der Text-Encoding Initiative (TEI) können in *Oral-History.Digital* generiert werden (ISO 2016, cf. TEI Guidelines).

Das Korpus *LangAge* liegt in Tondateien (*.wav) und zeitalignierten Transkripten vor. Diese wurden mit dem Werkzeug Transcriber (Transcriber 1998–2008) erstellt, das Metadaten und Zeitkodierungen in ein XML-Format umsetzt. Die Transkription folgt dem orthographischen Prinzip (Gerstenberg, Hekkel & Kairet 2018) und erfasst Wortabbrüche, Wortwiederholungen, Häsitationen und einige rede-begleitende Ereignisse. Dieser auf die Mündlichkeit ausgerichteten Transkription gemäß wird auf Interpunktion und andere Schriftlichkeitsstandards verzichtet. Das ist allerdings nicht leicht: Wenn nicht explizit dafür geschult wird, verschriften auch gerade geübte Personen „richtige" Formen statt der im Gesprochenen üblichen Varianten.

Eindrücklich zeigt sich dies am Beispiel der französischen Negation. Wie das Beispiel *je veux pas* (gesprochen) vs. *je ne veux pas* (Standardsprache, 'ich will nicht') zeigt, wird das Element *ne* im Gesprochenen sehr oft ausgelassen. Dies ist auch in den ZWAR-Interviews der Fall. Mit der Archivsuche und dem Anmerkungswerkzeug wurden dort 24 Vorkommen des Strings *je ne me* geprüft, der vor verschiedenen Verben vorkam, wie *je ne me souviens pas* 'ich erinnere mich nicht'. In zwei Fällen war nicht zu hören, ob das verneinende *ne* realisiert wurde, in den verbleibenden 22 Fällen konnte die Transkription nicht bestätigt werden, *ne* war nicht zu hören. Diese ungenaue Transkription erschwert den quantitativen Vergleich mit den *LangAge*-Interviews; die bei den ZWAR-Interviews nur scheinbar häufigere Verneinung erweist sich so als ein Artefakt der Transkriptionsweise.

Dieses Bild bestätigte sich in weiteren Kontrollen und weist eine eigene Systematik auf: Die Unterschiedlichkeit von Gesprochenem und Geschriebenem wird so weitgehend vorausgesetzt, dass in der Verschriftung unmittelbar „korrigiert" wird, ohne weitere Markierung. Die Geschichtswissenschaft fragt stärker danach, was erzählt wird, als wie es erzählt wird. Auch spielt die erinnerungskulturelle und würdigende Funktion des Online-Archivs eine Rolle: Durch eine lautgenaue Transkription hätten sich die Interviewten als unbeholfen, ja stotternd bloßgestellt fühlen können. Im Ergebnis geben die Oral History-Transkripte in ZWAR eher die Form wieder, die der sprachlichen Norm der Standardgrammatik entspricht, als das, was tatsächlich gesprochen wurde.

Dagegen folgen die linguistischen Transkriptionen in *LangAge* dem Wortlaut in allen Details, motiviert durch ein deskriptives Selbstverständnis: Alle Varianten werden ohne Bewertung erfasst und eingeordnet. Unterschiedliche Ausdrucksweisen für das Gleiche werden nicht im Sinne der normativen Grammatik als „richtig" und „falsch" beurteilt. Stattdessen ermöglicht die Verschriftung „wie gesprochen", die Formulierungsarbeit nachzuvollziehen und daraus Rückschlüsse auf die Sprecherinnen und Sprecher zu ziehen.

Solche Transkriptionsunterschiede müssen bei sammlungsübergreifenden Analysen also stets beachtet und – durch eine multimediale Forschungsumgebung

wie *Oral-History.Digital* unterstützt – direkt an den Mediendateien überprüft werden. Für korpuslinguistische Analysen schränkt dies die Vergleichbarkeit aber ein: Einfache Häufigkeitsvergleiche führen in die Irre. Basisangaben zu einer Sammlung wie die Gesamtlänge (üblicherweise in Tokens) sind unterschiedlich zu bewerten, je nachdem, ob Wortwiederholungen, Diskursmarker und Wortabbrüche mitgezählt werden. Daher wird die Länge der Interviews hier in Minuten angegeben.

5. Sammlungsübergreifende Analyse

In den folgenden Fallstudien werden die in der Oral History etablierten Methoden der explorativen Herangehensweise in der Anwendung korpuslinguistischer Methoden quasi nahtlos weitergeführt. Zunächst beschreiben wir die Eckdaten der in die Analyse einbezogenen italienischen und französischen Interviews aus *Zwangsarbeit 1939–1945* und aus *LangAge*. Das italienische Teilkorpus aus *Zwangsarbeit 1939–1945* (im Folgenden: ZWAR-it) umfasst neun Interviews mit zwei Frauen (einer als Jüdin und einer politisch Verfolgten) und sieben Männern (vier Militärinternierten, zwei als Juden und einem politisch Verfolgten). Die neun Video-Interviews dauern im Durchschnitt 4 Stunden und 20 Minuten. Eine Interviewerin hat fünf, die andere vier Interviews geführt (cf. Felsen & Frenkel 2008). Zum Zeitpunkt der Interviews 2005/2006 waren die Interviewten zwischen 75 und 85 Jahre alt.

Das französische Teilkorpus aus *Zwangsarbeit 1939–1945* (im Folgenden: ZWAR-fr) umfasst 18 Interviews mit ausschließlich männlichen Interviewten, die alle im Service du Travail Obligatoire (STO) Zwangsarbeit leisten mussten. Die 14 Video- und 5 Audio-Interviews dauern im Durchschnitt 2 Stunden und gut 20 Minuten. Eine Interviewerin hat zehn, die andere fünf Interviews geführt (cf. Granet-Abisset 2008). Zum Zeitpunkt der Interviews 2005/2006 waren die Interviewten zwischen 80 und 85 Jahre alt. Die Interviewten stammen aus verschiedenen Regionen, etwa die Hälfte aus der Umgebung Lyons.

Das hier ausgewertete Teilkorpus aus *LangAge* (im Folgenden: LangAge-oh.d) umfasst 32 Interviews. Die Interviews wurden 2005 sämtlich von einer Person, der Ko-Autorin, in Orléans und Umgebung geführt, die beteiligten 14 Männer und 18 Frauen waren zu diesem Zeitpunkt zwischen 71 und 94 Jahre alt. Die durchschnittliche Länge liegt bei 50 Minuten.

Für die sammlungsübergreifende Suche wurden die Transkripte aller drei Teilkorpora so bereinigt, dass jeweils nur die Äußerungen der Interviewten erhalten blieben. Analyse und Annotation wurden in SketchEngine (Kilgariff et al. 2014) vorgenommen. Diese korpuslinguistische Software ermöglicht es, eigene Daten in eine geschützte Cloud zu laden und programmgestützt zu lemmatisieren und nach Wortarten (Part-of-Speech, POS) zu taggen. Das bedeutet, dass im Hintergrund jedem Wort die Grundform (das Lemma) zugeordnet wird (also *werden* für *wird*), und die Wortart wie Verb oder Substantiv sowie morphologische Merkmale wie Plural oder Person ermittelt werden.

5.1. Schritt 1: Schlüsselwörter

In einer ersten korpuslinguistischen, induktiven Skizze wurden in SketchEngine die beiden französischen Korpora auf ihre Schlüsselwörter (Keywords) hin verglichen, nämlich solche Wörter erfasst, die signifikant häufiger in ZWAR-fr als in LangAge-oh.d vorkommen (in SketchEngine: Keywords mit Referenzkorpus LangAge-oh.d; lemma, ohne Unterscheidung von Groß- und Kleinschreibung). Zu den Schlüsselwörtern in ZWAR-fr gehören auf den ersten 10 Rängen überwiegend Orts- und Personennamen sowie Nationalitäten: *russes*, *breslau*, [Familienname], *bunzlau*, *grenoble*, *juifs*, *romans*, [Vorname]; Ausnahmen betreffen die Fabrikproduktion: *ciment* (Rang 5), *contremaître* (Rang 7). Darin zeichnet sich ab, dass die beiden Korpora viele Gemeinsamkeiten haben, Verben (*fumer* 'rauchen', Rang 17; *affecter* 'zuweisen', Rang 20) oder Pronomina kommen erst auf hinteren Rängen. Kennzeichnend für ZWAR-fr ist etwa das Verb *affecter*: Es beschreibt – stets im Passiv – wie die einzelnen oder Gruppen ihren Einsatzorten 'zugewiesen wurden'. Die 30 Okkurrenzen dieses Verbs kommen in über der Hälfte der ZWAR-fr-Interviews vor, was die Bewertung als Schlüsselwort unterstreicht.

In der Analyse der Schlüsselwörter wurden die Häufigkeiten in ZWAR-fr mit den Häufigkeiten in *LangAge* als so genanntes Referenzkorpus verglichen. An der Tatsache, dass hauptsächlich Eigennamen auf den ersten Plätzen vorkommen, zeigt sich, dass im allgemeinen Wortschatz keine signifikanten Unterschiede zu finden sind. Diese lassen sich an wenigen Verben wie *fumer* und *affecter* festmachen.

5.2. Schritt 2: Personalpronomina

Für den zweiten Schritt war die Frage leitend, ob der Gebrauch der Personalpronomina spezifische Verteilungsmuster aufweist; dabei gingen wir von der Hypothese aus, dass die Singularform der Personalpronomina in ZWAR-fr und ZWAR-it seltener verwendet wird als in LangAge-oh.d, weil die zur Zwangsarbeit Verschleppten ein stark im Kollektiv geprägtes Erleben berichten. Zudem gingen wir geschlechtsspezifischen Unterschieden nach, wie sie Knowles et al. (2021) fanden. Die Personalpronomina sind ausreichend häufig, um aussagekräftige Werte zu erreichen.

Dabei ist zu berücksichtigen, dass das Italienische grundsätzlich keine Personalpronomen in Subjektposition erfordert. Die vergleichende Herangehensweise an die erste Person Singular und Plural ermöglicht es aber dennoch, relative Frequenzen zu betrachten. Dadurch wird auch der unterschiedliche Umgang mit Wortwiederholungen ausgeglichen, die in LangAge-oh.d wie gesprochen transkribiert werden, in ZWAR aber, wie oben dargestellt, nicht vollumfänglich.

Der Vergleich der Häufigkeit pro Teilkorpus (Tab. 1a) zeigt, dass in ZWAR-it und in LangAge-oh.d die Singularform 'ich' öfter vorkommt, während in ZWAR-fr die Pluralform 'wir' dominiert. Diese Angaben sind aber problematisch, denn Wortwiederholungen wurden in ZWAR-fr und LangAge-oh.d unterschiedlich transkribiert, und das Pronomen der ersten Person Singular *je* wird häufiger wiederholt als das *nous* der ersten Person Plural. Außerdem hat die im Gesprochenen übliche Form *on* für 'wir' zugleich die Bedeutung 'man', so dass sie für weiterführende

Untersuchungen zuerst disambiguiert werden müsste. Schließlich wird *noi* / *nous* auch häufig in Präpositionalphasen wie 'für uns', 'von uns' verwendet, was den Vergleich verzerrt.

Daher wurden im zweiten Schritt Kombinationen des 'ich'/'wir'-Pronomens und der Verbform 'habe(n)' ausgewertet; auf diese Weise werden Phrasen mit wiederholtem Pronomen nur einmal erfasst. Es handelt sich bis auf wenige Ausnahmen im einstelligen Bereich (*on a espoir* 'wir haben Hoffnung', *on a juste le titre* 'wir haben nur den Titel') dabei um Formen der zusammengesetzten Vergangenheit. Die Bedeutung von *on* 'man' ist in der zusammengesetzten Vergangenheitsform zwar möglich, aber weniger häufig, weil mit generalisierender Verwendung das eine Dauer beschreibende Imparfait üblich ist. Die Wortverbindung *on a* kann daher mit geringen Unschärfen zur ersten Person Plural gerechnet werden. Die deutlich selteneren Vorkommen von *io sono* bzw. *je suis* sowie *noi siamo* bzw. *nous sommes/on est* wurden nicht einbezogen: Dadurch würde die Vergleichbarkeit weiter eingeschränkt, weil diese Konstruktionen andere Verben bevorzugen und häufig im Passiv sind.

Pronomen	**ZWAR-it absolut**	**ZWAR-it pro Std.**	**ZWAR-fr absolut**	**ZWAR-fr pro Std.**	**LangAge-oh.d absolut**	**LangAge-oh.d pro Std.**
io *je / j'*	2094	54	9119	207,61	8603	325,7
noi *nous*	1094	28	2104	47,9	1307	49,9
on			8525	194	5291	200,3
Quotient 1.Sg./1.Pl.	**1,9**	**1,9**	**0,9**	**0,9**	**1,3**	**1,3**

Tab. 1a | Personalpronomina der ersten Person in den drei Teilkorpora mit Zahl der Vorkommen und Vorkommen pro Stunde (Lemma, Konkordanz-Tool SketchEngine)

Pronomen	**ZWAR-it absolut**	**ZWAR-it pro Std.**	**ZWAR-fr absolut**	**ZWAR-fr pro Std.**	**LangAge-oh.d absolut**	**LangAge-oh.d pro Std.**
io ho *j'ai*	134	3,4	1374	31,3	2416	91,5
noi abbiamo *nous avons*	46	1,2	64	1,5	105	4,0
on a			1121	25,5	909	34,4
Quotient 1.Sg./1.Pl.	**2,9**	**2,9**	**1,2**	**1,2**	**2,4**	**2,4**

Tab. 1b | Personalpronomina der ersten Person mit Auxiliar im Perfekt, in den drei Teilkorpora mit Zahl der Vorkommen und Vorkommen pro Stunde (Phrase, Konkordanz-Tool SketchEngine)

Diese konkreten, auf die Vergangenheit bezogenen Verwendungen (Tab. 1b) bestärken – bereinigt – die beschriebene Tendenz (Tab. 1a). Das Personalpronomen der ersten Person wird in ZWAR-it dreimal so häufig im Singular wie im Plural verwendet, wie der Quotient zeigt.

Die italienischen Interviewten berichten also häufiger in der Ich-Form als in der Wir-Form. Individuelle Unterschiede sind allerdings groß: Die beiden jüdischen Überlebenden Piero T. und Liliana S. sagen deutlich öfter *io* als die politisch Verfolgten oder die Militärinternierten. Wie lässt sich das interpretieren? Spiegelt die Ich-Form die Einsamkeit eines mit 14 Jahren allein in Auschwitz überlebenden Mädchens? Können politisch Verfolgte ihre Lebensgeschichte eher in einem kollektiven Schicksal verorten? Für eine Antwort sind weitere Analysen auf breiterer Basis nötig.

Die französischen Korpora unterscheiden sich deutlich: In LangAge-oh.d wird die Singularform mehr als doppelt so oft verwendet wie die Pluralform; in ZWAR-fr ist das Verhältnis ausgeglichener – freilich bei großen individuellen Unterschieden. So entfallen zwei Drittel aller *nous avons*-Vorkommen auf einen Teilnehmer (Roger C., za076; entspricht 9,9 Okkurrenzen pro Stunde). Mehr als dreimal pro Stunde verwendet Jean A. (za074) *j'ai* 'ich habe' in seinem Interview, damit entfallen knapp 12 Prozent der gesamten Okkurrenzen auf ihn.

Damit bestätigt sich die Hypothese, dass die zur Zwangsarbeit Verschleppten, übrigens alles Männer, mit *on* und *nous* viel mehr über gemeinsam Erlebtes berichten als die während des Krieges im Heimatort Verbliebenen. Interessant ist die Beobachtung, dass die in Anknüpfung an Knowles et al. (2021) formulierte Erwartung geschlechtsspezifischer Unterschiede nicht bestätigt wurde. Liliana S. verwendete die erste Person Singular sehr häufig, die zweite Frau in ZWAR-it weniger als halb so oft. Für eine belastbare quantitative Aussage dazu müssten mehr Interviews ausgewertet werden. Besonders häufig benutzen die ehemaligen Zwangsarbeiter das *wir* in Wendungen wie *on nous a dit* oder *ils nous ont mis*, schildern also die Wir-Gruppe als Objekt fremder Befehle (s.u.).

In der Analyse der Häufigkeiten persönlicher Fürwörter wurden die Schwierigkeiten des direkten Korpusvergleichs deutlich, die sich aus den unterschiedlichen Transkriptionsstandards ergeben, d. h. der Berücksichtigung oder Nicht-Berücksichtigung von Wiederholungen. Diesem Problem wurde dadurch begegnet, dass nicht isolierte Formen ausgewertet wurden, sondern Wortfolgen von Pronomen und Verb. Darin zeigten sich deutliche Unterschiede im Gebrauch der Personalpronomen, mit einer klaren Präferenz für die erste Person Singular in ZWAR-it und in LangAge-oh.de, während in ZWAR-fr die erste Person Plural dominiert.

5.3. Schritt 3: N-Gramme

Der nächste Schritt beruht nicht auf einer Hypothese, sondern folgt einer explorativen Bottom-up-Herangehensweise. Nun verglichen wir die für die drei Teilkorpora – ZWAR-it, ZWAR-fr und LangAge-oh.d – charakteristischen, also häufig wiederholten, Wortkombinationen, zunächst im Rahmen von 3–4 Wörtern. Häufige Wortkombinationen oder N-Gramme bilden typische Muster im Sprachgebrauch ab (cf. Bubenhofer 2009). Dabei wurden die ersten drei Ränge der Frequenzlisten

verglichen und danach N-Gramme mit lexikalischem Element, die einen neuen Aspekt brachten (Tab. 2).

Im Italienischen nimmt die Verbphrase *mi ricordo* 'ich erinnere mich' die erste Position ein und widerspiegelt somit das zentrale Motiv der Interviewten, aber auch des in der Erinnerungskultur positionierten Interviewprojekts. Diese auch gesprächsanalytisch von Pfänder et al. (2022) untersuchte Wendung wird im nächsten Abschnitt genauer betrachtet.

Bezeichnend ist auch die häufige Verbphrase *ci hanno portato* 'sie brachten uns': Das als Kollektiv der Zwangsarbeiter*innen referenzierte 'uns' wird zum Objekt, das von einem ungenannten Subjekt – meistens den Deutschen – von einem Ort zum anderen gebracht wird. In dieser syntaktischen Konstruktion bilden sich die Machtverhältnisse von handelnden Personen (Agens) und den ihnen Ausgelieferten (Patiens) ab. Deutlich wird diese ohnmächtige Passivität auch durch die anderen, zusammen mit *ci hanno* 'sie haben uns' genutzten Begriffe des Verladens (*caricato/i*) und Verlegens (*trasferiti*), Verräumens (*messo*), Verschickens (*mandato*) oder Befehlens (*fatto* + Verb im Infinitiv).

Im Französischen dominieren auf Rang 1 die im Gesprochenen sehr häufigen Präsentativa *il y avait* 'es gab' (ZWAR-fr) und *il y a* 'es gibt' (LangAge-oh.d). Darin zeigen sich die Charakteristika des gesprochenen Französisch; der Unterschied beider Teilkorpora besteht in der Verwendung der Vergangenheitsform 'es gab', die in ZWAR-fr häufiger ist, während die Präsensform 'es gibt' in LangAge-oh.d dominiert.

ZWAR-it (Okkurrenzen)	**ZWAR-fr (Okkurrenzen)**	**LangAge-oh.d (Okkurrenzen)**
mi ricordo che 'ich erinnere mich, dass' (161 Okk., Rang 1)	*il y avait* 'es gab' (1.666 Okk., Rang 1)	*il y a* 'es gibt' (910 Okk., Rang 1)
un po' di 'ein bisschen' (125 Okk., Rang 2)	*il y a* 'es gibt' (608 Okk., Rang 2)	*il y avait* 'es gab' (667 Okk., Rang 2)
un certo momento 'ein bestimmter Moment' (102 Okk., Rang 3)	*il y en* 'es [gibt/gab] davon' (520 Okk., Rang 3)	*c'est c'* 'das ist, das' (255 Okk., Rang 3)
non lo so (100, Rang 4)	*je ne sais* 'ich weiß nicht' (412 Okk., Rang 5)	*je sais pas* 'ich weiß nicht' (213, Rang 9)
ci hanno portato 'sie brachten uns' (42 Okk., Rang 20)	*à l'époque* (203, Rang 24)	*un petit peu* (189, Rang 16)

Tab. 2 | Auswahl häufiger Wortkombinationen (N-Gramme, 3–4 Wörter) in den drei Teilkorpora mit Zahl der Vorkommen und Rang (ohne Unterscheidung von Groß- und Kleinschreibung, SketchEngine)

Die Wendung *il y en* kommt 312 mal im Präsens vor (*il y en a '*es gibt') und knapp 200 mal in der Vergangenheit (*il y en avait '*es gab'). Beide Formen werden aber mit wenigen Ausnahmen im Kontext von Vergangenheitsformen verwendet, wie die Überprüfung ergibt: *Mais il y en a un qui a réussi à se sauver '*aber es gibt einen, dem es gelungen war sich zu retten' (Robert W., za092, Band 2, 0:05:32). Zu dieser

sich nach Häufigkeit als dominant abzeichnenden Vergangenheitsperspektive passen adverbiale Angaben wie 'damals', frz. *à l'époque* und, auf dem dritten Platz in der italienischen Rangfolge, die unbestimmte Zeitangabe [*a(d) un*] *certo momento* '[in einem] gewisse[n] Moment', die unten genauer betrachtet wird.

In allen drei Teilkorpora vertreten sind Einschränkungen in Bezug auf das eigene Wissen wie *je (ne) sais pas* und *non lo so* 'ich weiß (es) nicht' („epistemische markierungen", Gülich & Pfänder 2022, 35). Je nach Kontext ist damit tatsächlich eine Abschwächung der epistemischen Autorität gemeint oder wird damit eine Planungspause überbrückt (cf. die Analyse in Grant 2010). Die Mengenangaben *un petit peu* bzw. *un po' di '*ein bisschen (von)' heben geringe Quantitäten hervor; in ZWAR-fr ist dieses Triple nicht auf den vorderen Plätzen (Rang 71, 108 Okkurrenzen). Damit kann einerseits ganz wörtlich eine geringe Menge angesprochen sein, aber auch eine Abschwächung des Gesagten, wie zum Beispiel in der Kombination mit 'vielleicht': *C'était peut-être un petit peu patriotique peut-être '*das war vielleicht ein bisschen patriotisch vielleicht' (André D., za077, Band 1, 0:11:11). In allen Gruppen sind sich die Interviewten also der Grenzen ihres Erinnerungsvermögens bewusst und formulieren dies auch.

Die programmgestützte Ermittlung besonders häufiger Wortkombinationen führt zu einer weiteren Profilierung der drei Teilkorpora: die Vergangenheitsperspektive tritt deutlich hervor, weiterhin die Thematisierung des Erinnerns – und der Grenzen der Erinnerung. Dieser Punkt soll im nächsten Schritt differenziert und unter Berücksichtigung der jeweiligen Kontexte analysiert werden.

5.4. Schritt 4: 'ich erinnere mich'

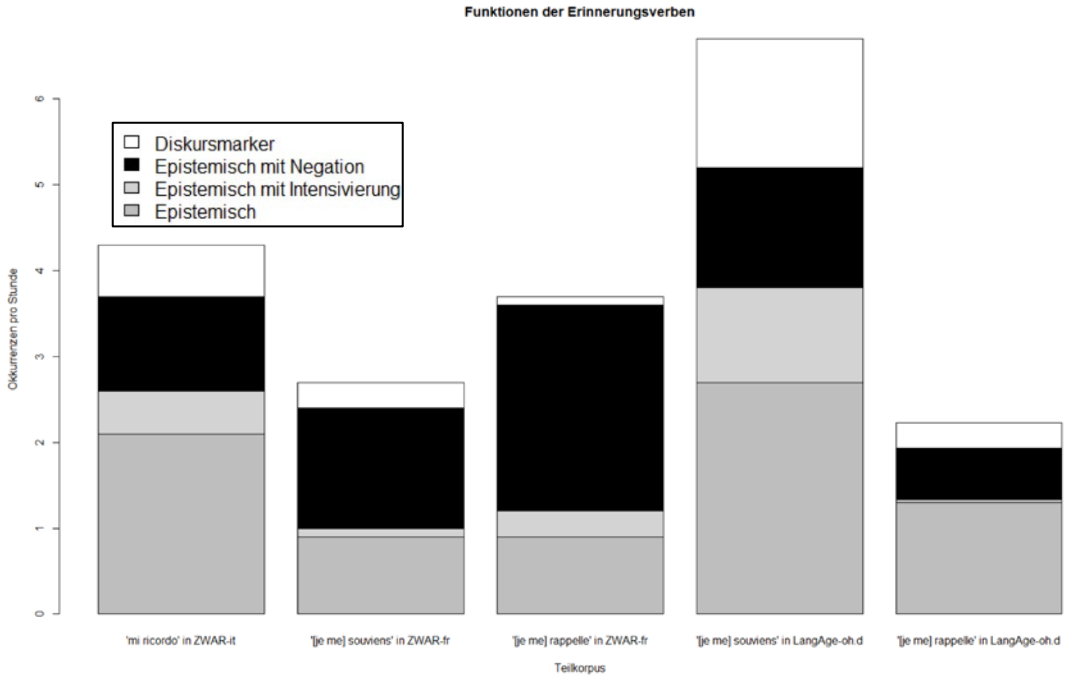
Aufbauend auf dem explorativ und datengetrieben ermittelten Ergebnis der hohen Frequenz von 'ich erinnere mich' wird im Folgenden gefragt, wie sich dies im Vergleich der Teilkorpora darstellt. Wie oben dargestellt, sind die programmgestützt ermittelten Frequenzen der Wortkombination 'ich erinnere mich, dass' zwar kennzeichnend für das Teilkorpus ZWAR-it, aber das Gesamtbild verlangt nach Differenzierung. Dazu gehört, dass die Verteilung über die einzelnen Interviews berücksichtigt wird, und dass unterschiedliche Verwendungsweisen sichtbar gemacht werden. Dafür wurde eine manuelle Annotation in SketchEngine umgesetzt, d. h. die einzelnen Vorkommen wurden manuell im Kontext geprüft und Kategorien zugeordnet.[1]

Um die Verteilung über die einzelnen Interviews im Korpus gleichmäßig zu beobachten, wurden die ersten 20 Okkurrenzen pro Interview annotiert, bzw. alle, die überhaupt vorkamen. Diese Zahl wurde so gewählt, dass in den französischen Teilkorpora die höchsten Werte ausgeschlossen wurden, ohne dass auf eine zu große Anzahl Okkurrenzen verzichtet werden musste. In ZWAR-it wurden die Vorkommen von *ricordo* mit linkem und rechtem Kontext in SketchEngine erfasst (Konkordanz-Tool, nur Verbform), in ZWAR-fr und LangAge-oh.d wurden die französischen

[1] Für die Belegannotation von *ricordo* und *souviens/rappelle* danken wir Elena Bandt. Die Annotation wurde mit Screenshots gesichert; SketchEngine erlaubt keinen Export der annotierten Daten.

Äquivalente [*je me*] *souviens* bzw. [*je me*] *rappelle* (wiederum Verbform, erfasst wurde nur die 1. Person).

Ausgehend von den N-Gramm-Ergebnissen wurden dabei nicht die weiteren textuellen Funktionen analysiert (cf. Pfänder et al. 2022), sondern die Verwendung benachbarter Modifikatoren wie Verneinung oder Verstärkung berücksichtigt. Auf diese Weise wurde die epistemische Funktion von *ich erinnere mich* unterschieden von einer ebenfalls epistemischen, aber durch intensivierende Adverbien wie italienisch *bene* bzw. *benissimo* 'gut', 'sehr gut' und französisch *très bien* '(sehr) gut', *toujours* 'immer', *surtout* 'vor allem', *vraiment* 'wirklich' und *effectivement* 'tatsächlich' verstärkten Form und von der ebenfalls auf das eigene Wissen und Erinnern bezogenen, aber verneinten Form. In Anlehnung an die Studie zu engl. *I don't know* 'ich weiß nicht' (Grant 2010) wurden weiterhin Okkurrenzen unterschieden, in denen *ich erinnere mich* als Diskursmarker fungiert, also nicht inhaltlich bedeutsam ist, sondern in seiner Funktion, eine Planungspause zu füllen bzw. Zögern und Unsicherheit zu markieren. Um ein klares Kriterium zu formulieren, wurden Diskursmarker annotiert, die aus weiteren Hinweisen wie Satzabbrüchen oder Pausen im Kontext als Häsitation eingeschätzt werden konnten.



3 | Funktionen der Erinnerungsverben in der ersten Person Singular: Ergebnisse der manuellen Annotation von bis zu 20 Okkurrenzen pro Interview (n=603)

Durch die Grenze von 20 berücksichtigten Fällen werden sehr hohe Vorkommen bei einzelnen Personen statistisch ausgeglichen, und eine Verzerrung der zusammengefassten Werte wird vermieden. Im Vergleich der Teilsammlungen zeigt sich dann, dass die Relevantsetzung des Erinnerns, also die epistemische Funktion mit und ohne Intensivierung, in LangAge-oh.d am häufigsten ist, es folgt ZWAR-it und dann ZWAR-fr. Umgekehrt ist die Reihenfolge bei den verneinten

Formen: das Nicht-Erinnern wird am häufigsten in ZWAR-fr erwähnt, dann in LangAge-oh.d und am seltensten in ZWAR-it.

Die beiden französischen Teilkorpora unterscheiden sich also deutlich: Die zwar auch biographisch, aber spezifisch zu ihrer lange zurückliegenden Zwangsarbeit Interviewten verwenden häufiger die verneinte Form der Erinnerungsverben. Sie schränken damit häufiger die Aussagekraft der erinnerten Inhalte ein als die allgemein-biographisch (und kürzer) interviewten *LangAge*-Zeitzeug*innen. Die verneinte Form 'ich erinnere mich nicht' ist im italienischen ZWAR-Teilkorpus deutlich seltener als im französischen ZWAR-Teilkorpus, wohingegen die positiven Formen häufiger verwendet werden. ZWAR-it und LangAge-oh.d zeigen in diesem Bereich also Gemeinsamkeiten, während ZWAR-fr sich davon abgrenzt. Dass sich somit Gemeinsamkeiten über die Sprachgrenzen hinweg ausprägen, unterscheidet das Bild der Erinnerungsmarker von der oben ausgewerteten Verteilung der Pronomina.

Wenn dagegen keine Obergrenze von 20 Okkurrenzen eingesetzt wird, treten die Besonderheiten einzelner Interviews hervor. Insbesondere in ZWAR-it, wo die Standardabweichung viel höher ist als in den anderen beiden Teilkorpora, zeigen die neun Interviews sehr unterschiedliche Häufigkeiten von *mi ricordo*. Die Spanne reicht von 8 bei Claudio S. bis zu 153 bei Liliana S. Die Auschwitz-Überlebende Liliana S. sagt auch im Vergleich aller Interviews aller Sprachen in *Zwangsarbeit 1939-1945* – ausweislich der deutschen Übersetzungen – am häufigsten *ich erinnere mich*.

Sie erzählt sehr anschaulich, im häufigen Wechsel zwischen konkret geschilderten Erlebnissen und deren historischer Einordnung, und betont ihre epistemische Autorität als Überlebende der Shoah. Als aktive Zeitzeugin in Schulen und Öffentlichkeit, seit 2018 als Senatorin, ist Liliana S. zudem eine geübte Rednerin. All dies sind individuelle Faktoren, die bei einer korpuslinguistischen Interpretation der Wendung 'ich erinnere mich' berücksichtigt werden müssen.

5.5. Schritt 5: Der Moment

Wie die N-Gramm-Analyse ergab, ist *in un certo momento* 'in einem bestimmten Moment' in den italienischen Interviews eine viel verwendete Wortkombination. In einzelnen Interviews ist dieser ‚Moment' besonders häufig, allein 42 mal etwa bei dem Militärinternierten Claudio S., fast in Art eines Füllwortes, das die Erzählung in Art eines ‚und dann' vorantreibt. Hier ist der ‚bestimmte Moment' meist gerade nicht genau bestimmt, sondern bezeichnet eine nicht präzise datierbare Veränderung nach einer längeren Phase von Wiederholung und Alltäglichkeit, etwa wenn die jugendliche Begeisterung für Mussolinis Faschismus der Ernüchterung weicht (za126, Bd. 1, 00:34:42), die Offizierslaufbahn vom Sturz Mussolinis in Frage gestellt wird (za126, Bd. 2, 00:07:32) oder der ausgehungerte Überlebenskampf im Kriegsgefangenenlager in lebensmüde Gleichgültigkeit umschlägt (za126, Bd. 3, 00:31:41).

Liliana S. dagegen nutzt die Wendung *un certo momento* gar nicht. Ihr ‚Moment' ist oft präziser (*in quel momento*); mit ihm beschreibt sie ihre Gefühle oder bewertet

ihr Verhalten in der zuvor geschilderten Situation, häufig einem emotionalen Erlebnis oder einem zentralen Wendepunkt ihrer Biographie. Solch ein *momento cruciale* war etwa die Entscheidung der 13-jährigen Liliana, auf der Flucht im Moment der Zurückweisung durch die Schweizer Grenzpolizei mit ihrem Vater umzukehren – trotz der nun drohenden Deportation. Im Rückblick bekräftigt sie die Bedeutung dieses Moments durch eine Wiederholung: „non mi sono mai pentita di questo, è stato un momento cruciale della mia vita, ma non mi sono mai pentita di questo" – 'ich habe das niemals bereut, das war ein entscheidender Moment in meinem Leben, ich habe das niemals bereut' (za124, Bd. 2, 00:31:33). Die Wiederholung „un momento ... un momento" verwendet Liliana S. auch im Moment der endgültigen Trennung vom Vater bei der Selektion in Birkenau (za124, Bd. 4, 00:07:55) oder nach dem Krieg bei der Geburt ihres ersten Kindes (za124, Bd. 4, 00:30:48).

In diesen verschiedenen Momenten der Interviewten bündeln sich also unterschiedliche Zeitlichkeiten, manche unerträglich lange, manche plötzlich überraschend. Der ‚Moment' zeigt die individuellen Bedeutungen dieser Zeitlichkeiten – in der erinnerten Erfahrung, der rückblickenden Erinnerung und der aktuell zu strukturierenden Erzählung.

6. Fazit

Diese Fallstudie zur Sprache des Erinnerns in den lebensgeschichtlichen Interviews des Archivs *Zwangsarbeit 1939-1945* und des *LangAge*-Korpus zeigte in ihren fünf einzelnen Arbeitsschritten, wie die in der Oral History-Forschung etablierte Herangehensweise, Hypothesen aus den Daten heraus zu entwickeln, korpuslinguistisch umgesetzt werden kann. Statistische Verfahren wie die Ermittlung von Schlüsselwörtern (Keywords) und häufigen Wortkombinationen (N-Gramme) sowie die Extraktion von Personalpronomina (*ich* und *wir*) führten zu einer ersten Profilierung der drei Teilsammlungen. Diese quantitativen Merkmale bringen Gemeinsamkeiten und Unterschiede hervor, die teils die Teilkorpora der gleichen Sammlung (ZWAR-fr und ZWAR-it vs. LangAge-oh.d), teils die Teilkorpora der gleichen Sprache (ZWAR-fr und LangAge-oh.d vs. ZWAR-it) verbinden.

Auch quer dazu wurden Ähnlichkeiten festgestellt, so in der Analyse von 'ich erinnere mich', wo ZWAR-it und LangAge-oh.d sich näher waren in Bezug auf die positive und verstärkte Wendung. Diese Muster sind linguistisch aussagekräftig für die registerbezogene Einordnung der Interviews (cf. Biber, Egbert & Keller 2020) und die jeweils genutzten sprachlichen Möglichkeiten, um die eigene Position zum Ausdruck zu bringen.

Die zu diesen Befunden oben vorgeschlagenen historischen Interpretationen sind – auch aufgrund der heterogenen und im Umfang begrenzten Datengrundlage – nur als erste Annäherungen zu verstehen, die zu weiteren Analysen anregen wollen.

Zudem ergeben sich aus diesen Beobachtungen verschiedene Anschlussfragen. Sie betreffen zum einen die Rolle der „Zeitzeug*innen" in den unterschiedlichen Interviewsettings, d. h. die Frage, ob die Personen ihr individuelles Zeugnis ablegen

oder eher ihrer Rolle als Vertretung einer Gruppe entsprechen wollen. Zum anderen zeigt sich daran die Notwendigkeit, persönliche, fallbezogene Informationen auszuwerten. Der biographische Zugang der Oral History und das geschichtswissenschaftliche Kontextwissen müssen dazu mit den Ergebnissen der korpuslinguistischen Analyse zusammengeführt werden.

Die methodischen und forschungspraktischen Verschiedenheiten zwischen Oral History und Linguistik erfordern eine hohe Sensibilität für die Standards der Datenaufbereitung. Dies wird am Beispiel unterschiedlicher Transkriptionsstandards deutlich, die an vielen Stellen die Vergleichbarkeit einschränken.

Die Perspektiven der sammlungsübergreifenden Analyse sind vielfältig; sie können zum Beispiel auf weitere Phänomenbereiche wie den Wortschatz bezogen werden, um generationstypische Erfahrungen etwa des Bombenkriegs besser zu verstehen. Der Aufbau der Forschungsumgebung *Oral-History.Digital* wird in Zukunft solche sammlungsübergreifenden Analysen unterstützen. Mit diesem digitalen Blick auf die Interviews ist freilich eine stärkere Distanz zu den lebensgeschichtlichen Erzählungen der Zeitzeug*innen verbunden, die aber stets eine sorgsame und respektvolle Interpretation erfordern.

Bibliografie

Aijmer, Karin. 2015. „Pragmatic Markers." In *Corpus Pragmatics*, ed. Aijmer, Karin & Christoph Rühlemann, 195–218, Cambridge: Cambridge University Press.

Apel, Linde, Almut Leh & Cord Pagenstecher. 2022. „Oral History im digitalen Wandel. Interviews als Forschungsdaten." In: *Erinnern, erzählen, Geschichte schreiben. Oral History im 21. Jahrhundert*, ed. Apel, Linde, 193–222, Berlin: Metropol.

Apostolopoulos, Nicolas & Cord Pagenstecher (ed.). 2013. *Erinnern an Zwangsarbeit. Zeitzeugen-Interviews in der digitalen Welt.* Berlin: Metropol.

Biber, Douglas, Jesse Egbert & Daniel Keller. 2020. „Reconceptualizing register in a continuous situational space." *Corpus Linguistics and Linguistic Theory* 16 (3), 581–616. <https://doi.org/10.1515/cllt-2018-0086>.

Blank, Andreas. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Tübingen: Niemeyer.

Bothe, Alina. 2019. *Die Geschichte der Shoah im virtuellen Raum. Eine Quellenkritik.* Berlin, Boston: De Gruyter Oldenbourg.

Briggs, Charles L. 2005. „Sociolinguistic Interviews/ Soziolinguistisches Interview." In *Sociolinguistics: An International Handbook of the Science of Language and Society*. HSK 3.2, ed. Ammon, Ulrich et al., 1052–1062, Berlin, New York: De Gruyter.

Browning, Christopher R. 2010. *Remembering Survival. Inside a Nazi Slave-Labor Camp*. New York: Norton.

Bubenhofer, Noah. 2009. *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. <https://doi.org/10.5167/uzh-111287>.

Deppermann, Arnulf. 2013. „Interview als Text vs. Interview als Interaktion." *FQS Forum: Qualitative Sozialforschung Social Research* 14(3). <https://doi.org/10.17169/fqs-14.3.2064>.

Eusterschulte, Anne, Sonja Knopp & Sebastian Schulze (ed.). 2016. *Videographierte Zeugenschaft: Ein Interdisziplinärer Dialog.* Weilerswist: Velbrück Wissenschaft.

Felsen, Doris & Viviana Frenkel. 2008. „Die italienischen Deportationen 1943-45." In *Hitlers Sklaven. Lebensgeschichtliche Analysen zur Zwangsarbeit im internationalen Vergleich*, ed. Leh, Almut, Alexander von Plato & Christoph Thonfeld, 285-297, Wien: Böhlau.

Freyburger, Philipp & Frank Jäger. 2021. „Emergentes Erinnern. Sensorische, kognitive und mediale (Spiel-)Räume in Oral-History-Interviews und literarischen Erinnerungstexten." *Romanische Forschungen* 133, 176-205.

GAT 2. 2009. „Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)." *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion*, 10, 353-402.
<http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf>.

Gerstenberg, Annette, Valerie Hekkel, & Julie Kairet. 2018. *Corpus LangAge: Transcription Guide.* University of Potsdam: Romance Linguistics.
<https://doi.org/10.5281/zenodo.6444538>.

Gerstenberg, Annette. 2009. „The Multifaceted Category of 'Generation': Elderly French Men and Women Talking about May '68." *International Journal of the Sociology of Language* 200, 153–170.
<https://doi.org/10.1515/IJSL.2009.049>.

Gerstenberg, Annette. 2011. *Generation und Sprachprofile im höheren Lebensalter*: *Untersuchungen zum Französischen auf der Basis eines Korpus biographischer Interviews*. Frankfurt am Main: Vittorio Klostermann.
<https://doi.org/10.5281/zenodo.7315875>.

Gerstenberg, Annette. 2017. „A Difficult Term in Context: The Case of French STO." In *Oral History Meets Linguistics*, ed. Kasten, Erich, Katja Roller & Joshua Wilbur, 159–184, Fürstenberg: Kulturstiftung Sibirien.
<https://dh-north.org/siberian_studies/publications/orhili_gerstenberg.pdf>.

Granet-Abisset, Anne M. 2008. „Berichte aus ‚Vichy-Frankreich'. Über die Schwierigkeit, Erinnerungen zu sammeln und Geschichte zu schreiben." In *Hitlers Sklaven. Lebensgeschichtliche Analysen zur Zwangsarbeit im internationalen Vergleich*, ed. Leh, Almut, Alexander von Plato & Christoph Thonfeld, 103-112, Wien: Böhlau.

Grant, Lynn E. 2010. „A Corpus Comparison of the Use of 'I don't Know' by British and New Zealand Speakers." *Journal of Pragmatics* 42 (8), 2282–2296.
<https://doi.org/10.1016/j.pragma.2010.01.004>.

Gülich, Elisabeth & Stefan Pfänder. 2022. „Erinnerungsmarkierungen in Zeitzeugenerzählungen. Episodische Rekonstruktion und epistemische Authentifizierung in Gesprächen mit Überlebenden der NS-Zwangsarbeitslager". In *Romanistisches Jahrbuch*. Im Erscheinen.

ISO 24624. 2016. Language resource management – Transcription of spoken language.
<https://www.iso.org/standard/37338.html>.

Kangisser Cohen, Sharon. 2014. *Testimony and Time: Holocaust Survivors Remember*. Jerusalem: Yad Vashem Publications

Kilgarriff, Adam et al. 2014. „The Sketch Engine: ten years on." *Lexicography* 1. 7–36.
<https://doi.org/10.1007/s40607-014-0009-9>.

Klüger, Ruth. 1996. „Zum Wahrheitsbegriff in der Autobiographie." In *Autobiographien von Frauen. Beiträge zu ihrer Geschichte*, ed. Heuser, Magdalene, 405-410, Tübingen: Niemeyer.

Knowles, Anne K. et al. 2021. „Mind the Gap: Reading across the Holocaust

Testimonial Archive." In *The Holocaust in the 21st Century: Relevance and Challenges in the Digital Age*, ed. Cole, Tim & Simone Gigliotti, 216–241. United States: Northwestern University Press.

Labov, William & Julie Auger. 1993. „The Effect of Normal Aging on Discourse: A Sociolinguistic Approach." In: *Narrative Discourse in Neurologically Impaired and Normal Aging Adults.* ed. Brownell, Hiram H. & Yves Joanette, 115–133, San Diego (CA): Singular.

Laub, Dori, & Johanna Bodenstab. 2008. „Wiederbefragt. Erneute Begegnung mit Holocaust-Überlebenden nach 25 Jahren." In *Hitlers Sklaven. Lebensgeschichtliche Analysen zur Zwangsarbeit im internationalen Vergleich*, ed. Leh, Almut, Alexander von Plato & Christoph Thonfeld, 389-401, Wien: Böhlau.

Leh, Almut, Alexander von Plato & Christoph Thonfeld (ed.). 2008. *Hitlers Sklaven. Lebensgeschichtliche Analysen zur Zwangsarbeit im internationalen Vergleich.* Wien: Böhlau.

Michaelis, Andree. 2013. *Erzählräume nach Auschwitz: Literarische und videographierte Zeugnisse von Überlebenden der Shoah.* WeltLiteraturen Bd. 2. Berlin: Akademie Verlag.

Möbus, Dennis. 2020. „Holleriths Vermächtnis – ein Beitrag zur Geschichte von Frauen in der EDV. Topic Modeling als Methode digitaler Sekundäranalyse lebensgeschichtlicher Interviews." *BIOS – Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen* 2, 162–180. <https://doi.org/10.3224/bios.v33i2.01>.

Nägel, Verena Lucia. 2016. „Zeugnis – Artefakt – Digitalisat. Zur Bedeutung der Entstehungs- und Aufbereitungsprozesse von Oral History-Interviews." In *Videographierte Zeugenschaft: Ein interdisziplinärer Dialog*, ed. Eusterschulte, Anne, Sonja Knopp, & Sebastian Schulze, 347–68, Weilerswist: Velbrück Wissenschaft.

oh.d = *Oral-History.Digital, 2020-2022. Informationsinfrastruktur für die Erschließung, Recherche und Annotation von audiovisuellen narrativen Interviews*. Berlin: Freie Universität Berlin, Universitätsbibliothek. <www.oral-history.digital>.

Pagenstecher, Cord & Stefan Pfänder. 2017. „Hidden Dialogues. Towards an Interactional Understanding of Oral History Interviews." In *Oral History Meets Linguistics*, ed. Kasten, Erich, Katja Roller & Joshua Wilbur, 185–207, Fürstenberg: Kulturstiftung Sibirien.

Pagenstecher, Cord & Doris Tausendfreund. 2015. „Interviews als Quellen der Geschlechtergeschichte. Das Online-Archiv ‚Zwangsarbeit 1939-1945' und das ‚Visual History Archive' der USC Shoah Foundation." In *Geschlecht und Erinnern im digitalen Zeitalter. Neue Perspektiven auf ZeitzeugInnenarchive*, ed. Bothe, Alina & Christina Isabel Brüning, 41-67, Berlin: Lit.

Pagenstecher, Cord. 2010. „'We were treated like slaves.' Remembering forced labor for Nazi Germany." In *Human Bondage in the Cultural Contact Zone. Transdisciplinary Perspectives on Slavery and Its Discourses*, ed. Mackenthun, Gesa & Raphael Hörmann, 275–291, Münster: LIT.

Pagenstecher, Cord. 2017. „Oral History und Digital Humanities." *BIOS – Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen* 1-2, 76–91.

Pfänder, Stefan et al. 2022. „Doing remembering as a multimodal accomplishment. On the use of *mi ricordo* ('I remember') in Oral History Interviews." *Interactional Linguistics* 2 (1), 110-136.

Sabrow, Martin, & Norbert Frei. 2012. *Die Geburt Des Zeitzeugen Nach 1945*. Göttingen: Wallstein.

Schiffrin, Deborah. 2000. „Mother/daughter discourse in a Holocaust oral history." *Narrative Inquiry* 10 (1), 1–44. <https://doi.org/10.1075/ni.10.1.01sch>.

Schuch, Daniel. 2021. *Transformationen der Zeugenschaft. Von David P. Boders frühen Audiointerviews zur Wiederbefragung als Holocaust Testimony*. Göttingen: Wallstein.

Shenker, Noah. 2015. *Reframing Holocaust Testimony.* Bloomington: Indiana University Press.

„Sklavenarbeit“: War die NS-Zwangsarbeit Sklaverei? Themenfilm mit Interview-Ausschnitten. 2011. In *Zwangsarbeit 1939-1945. Webseite*, Freie Universität Berlin.
<https://www.zwangsarbeit-archiv.de/zwangsarbeit/ereignisse/sklavenarbeit/index.html>.

Taubitz, Jan. 2016. *Holocaust Oral History und das lange Ende der Zeitzeugenschaft.* Göttingen: Wallstein Verlag.

TEI Guidelines, P5, chapter 8: Transcription of Speech.
<http://www.tei-c.org/release/doc/tei-p5-doc/de/html/TS.html>.

Transcriber = Geoffrois, Edouard, Mark Liberman & Zhibiao Wu (ed.). 1998–2008. *Transcriber 1.5.1. A Tool for Segmenting, Labeling and Transcribing Speech.* DGA: Sourceforge.

Thonfeld, Christoph. 2014. *Rehabilitierte Erinnerungen? Individuelle Erfahrungsverarbeitungen und kollektive Repräsentationen von NS-Zwangsarbeit im internationalen Vergleich.* Essen: Klartext.

Wagner, Suzanne E. & Sali A. Tagliamonte. 2017. „What Makes a Panel Study Work? Researcher and Participant in Real Time.“ In *Panel Studies of Variation and Change*, ed. Wagner, Suzanne & Isabelle Buchstaller, 213–232, New York: Routledge.

ZWAR = *Zwangsarbeit 1939–1945. Erinnerungen und Geschichte*, 2009–2022. Berlin: Freie Universität/Centrum für Digitale Systeme.
<https://www.zwangsarbeit-archiv.de/>.
Zugang zum Archiv: <https://archiv.zwangsarbeit-archiv.de/de>.

Zusammenfassung

Im interdisziplinären Spannungsfeld zwischen Geschichtswissenschaft und Linguistik werden in diesem Beitrag Oral History-Interviews als zentrale Ressource beider Fächer ausgewertet. Die Fallstudie untersucht die Sprache des Erinnerns in italienischen und französischen lebensgeschichtlichen Interviews des Archivs *Zwangsarbeit 1939-1945* und des *LangAge*-Korpus. In fünf Arbeitsschritten setzt sie die in der Oral History-Forschung etablierte Herangehensweise, Hypothesen aus den Daten heraus zu entwickeln, mit korpuslinguistischen Methoden um. So werden die Vorkommen der Personalpronomina *wir* und *ich* verglichen, und Schlüsselwörter (Keywords) und häufige Wortkombinationen (N-Gramme) ermittelt. Diese quantitativen Analysen werden durch die exemplarische Analyse häufiger Wendungen wie ‘ich erinnere mich’ oder ‘in einem bestimmten Moment’ vertieft. Dabei werden der biographische Zugang der Oral History und das geschichtswissenschaftliche Kontextwissen mit den Ergebnissen der korpuslinguistischen Analyse zusammengeführt. Als problematisch erweisen sich dabei unterschiedliche Transkriptionsstandards zwischen Oral History und Linguistik. Die neue Forschungsumgebung Oral-History.Digital unterstützt in Zukunft solche sammlungsübergreifenden und doch quellennahen Analysen lebensgeschichtlicher Interviews.

Abstract

Situated in the interdisciplinary fields of historiography and linguistics, this article evaluates oral history interviews as a resource valuable to both disciplines. The case study analyses the language of remembering in Italian and French life story interviews from the archive *Forced Labour 1939-1945* and the *LangAge* corpus. In five working steps, it implements the oral history approach of developing hypotheses from the data itself with methods from corpus linguistics. Occurrences of personal pronouns *we* and *I* are compared and keywords as well as frequent word combinations (*n-grams*) are identified. These quantitative approaches are complemented by exemplary analyses of frequent phrases such as 'I remember' or 'at a certain moment'. In doing so, the biographical approach of oral history and historical contextual knowledge can be combined with the results of corpus linguistic analysis. Although different transcription standards of oral history and linguistics prove to be problematic, in the future the new research environment Oral-History.Digital will simultaneously support cross-collection and close-to-source analyses of life history interviews.

2022, n°9
pp. 240-253
doi: https://doi.org/10.15460/apropos.9.1900

Anja Weingart & Georg A. Kaiser

Eine FAIRe Anwendungssoftware für textbasierte Forschungsdaten

Das UV2 Annotationstool

Anja Weingart
ist promovierte Romanistin und arbeitet derzeit als Softwareentwicklerin beim Projektträger des Deutschen Zentrums für Luft- und Raumfahrt (DLR-PT).
Anja.Weingart@uni-konstanz.de

Georg A. Kaiser
ist Professor für Romanistische Sprachwissenschaft am Fachbereich Linguistik der Universität Konstanz.
Georg.Kaiser@uni-konstanz.de

Keywords

FAIR-Prinzipien – Forschungssoftware – webbasiertes Annotationstool

Als Richtlinien für den Umgang mit digitalen Forschungsdaten definieren Wilkinson et al. (2016) die FAIR-Prinzipien *findable* (auffindbar), *accessible* (zugänglich), *interoperable* (interoperabel) und *reusable* (wiederverwendbar). Diese vier Prinzipien gelten für alle Arten von digitalen Objekten und damit auch für Softwareanwendungen. Im Unterschied zu Forschungsdaten liegt Software jedoch in zwei Formen vor: als maschinenlesbares, ausführbares Artefakt und als menschenlesbarer Quellcode mit Verknüpfungen zu anderen Softwarebibliotheken. Diese Besonderheiten erfordern eine Anpassung der einzelnen Prinzipien aus Wilkinson et al. (2016), die in Chue Hong et al. (2021) als Ergebnis der Arbeit der *FAIR for Research Software working group* (FAIR4RS) vorliegen. In diesem Beitrag werden zum einen die FAIR Prinzipien nach Chue Hong et al. (2021) für Forschungssoftware vorgestellt. Hierbei handelt es sich um etablierte Richtlinien, die auch von der DFG in ihren Leitfäden zum Umgang mit Forschungsdaten bzw. in Bezug auf Forschungssoftware erwähnt werden. Zum anderen wird die Umsetzung dieser Prinzipien am Beispiel der freien, webbasierten Anwendungssoftware, dem UV2-Annotationstool, illustriert. Dieses Annotationstool wurde zur Untersuchung von Verbzweiteffekten in den romanischen und anderen Sprachen entwickelt und steht grundsätzlich zur Analyse anderer sprachlicher Phänomene zur Verfügung.

In unserem Beitrag möchten wir zeigen, wie diese vier Prinzipien bei der Entwicklung einer Softwareanwendung umgesetzt werden können. Der Schwerpunkt liegt hierbei auf den Prinzipien *reusable* und *interoperable.* Die Konzepte der Wiederverwendbarkeit und Interoperabilität von Software und Softwareelementen beschäftigen die Softwaretechnik seit Jahrzehnten und es existieren hierzu zahlreiche Ansätze und Modelle, die eine Modularisierung von Softwareelementen verschiedener Komplexität und die Architektur des Softwaresystems betreffen (cf. McIlroy 1969, Meyer 1997, Brügge & Dutoit 2013, Richards & Ford 2020). Im vorliegenden Beitrag wird die konkrete Umsetzung der im UV2-Annotationstool implementierten Konzepte dargestellt und im Hinblick auf die FAIR-Prinzipien bewertet. Eine umfassendere Bewertung existierender Konzepte und Verfahren sowie Implementierungsvorschläge ist im von der FAIR4RS Working Group geplanten Leitfaden zur Umsetzung der FAIR-Prinzipien für Forschungssoftware zu erwarten.

Zunächst wird das UV2-Annotationstool hinsichtlich der Entwicklungsstrategie, des Aufbaus des Quellcodes und der Art des ausführbaren Artefakts vorgestellt. Es folgt eine Einführung der FAIR-Prinzipien für Forschungssoftware und eine Einordnung, wie die dort formulierten Anforderungen bei der Implementierung des UV2-Annotationstools umgesetzt werden. Ein Blick hinter die Kulissen der Softwaretechnik ist sicher auch für die traditionelle Romanistik lohnend, weil diese Techniken und Werkzeuge auch für romanistische Forschung genutzt werden können. Uns bekannte Projekte, die diese Techniken bereits verwenden, sind VerbaAlpina (cf. Krefeld & Lücke (2020)) und die historischen Projekte, die Eckhart et al. (2021) vorstellen.

1. Das UV2-Annotationstool

Das UV2-Annotationstool wurde im Rahmen des DFG Projekts '*Uncovering verb-second effects. An interface-based typology* (UV2)' entwickelt (siehe Fußnote 1). Das Projekt untersucht syntaktische und pragmatische Bedingungen für so genannte Verbzweiteffekte in den romanischen Sprachen im Vergleich zu typologisch anderen Sprachen, wie z. B. Baskisch. Grundlage dieser Untersuchung sind umfangreiche Paralleltextkorpora basierend auf Bibelübersetzungen sowie andere in viele Sprachen übersetzte Texte, u. a. *Astérix*, *Le petit Nicolas, Commissario Montalbano*, *Sherlock Holmes*). Zu Beginn des Projekts lagen die Paralleltexte bereits in einem tabellenförmigen Format vor und waren teilweise für Kriterien wie Satztyp, Art der präverbalen Konstituente und Verbposition annotiert. Die Arbeit mit Office-Programmen (Textverarbeitung oder Tabellenkalkulation) verleitet jedoch zu einer „visuell orientierten" Datenstrukturierung und Annotation, also dem Einsatz von Formatierungen wie Farben oder Schriftstilen, weil die Datenspeicherung und die Datendarstellung nicht voneinander getrennt sind (zum häufigen Einsatz dieser Formate und Programme in der romanistischen Forschung siehe die Studie der AG Digitale Romanistik 2015). Die Konsequenz ist, dass so strukturierte Daten nicht direkt maschinenlesbar, also nicht in andere Formate

überführbar und damit nicht mit anderen Werkzeugen bearbeitbar sind.[1] Sie erfüllen nicht die FAIR-Prinzipien der Interoperabilität und der Wiederverwendbarkeit. Für die Forschungspraxis ist es jedoch viel problematischer, dass auf der Grundlage so strukturierter Daten keine komplexen Suchanfragen möglich sind, die Daten also nicht mehr systematisch und mit allen Vorteilen eines digitalen Korpus analysiert werden können.

Mit Hilfe des UV2-Annotationstools werden die Paralleltexte in eine relationale Datenbank übertragen, die online zugänglich sein wird. Die Webanwendung ermöglicht eine kollaborative Bearbeitung und Analyse der Daten, aber auch eine einschränkbare Veröffentlichung. Es werden Funktionalitäten für die Speicherung, Bearbeitung, Annotation und Alignierung von textbasierten Sprachdaten sowie eine Verwaltung von Benutzergruppen und den bibliographischen Angaben der Texte zur Verfügung gestellt. Auf konzeptioneller Ebene besteht die Anwendung aus vier Bereichen:

(1) a. Bereich Datenspeicherung mit Import- und Exportfunktionalitäten,

b. Bereich Sprachdaten mit Bearbeitung, Alignierung und Annotation,

c. Bereich Nutzerverwaltung mit Zugangs- und Rechteverwaltung und

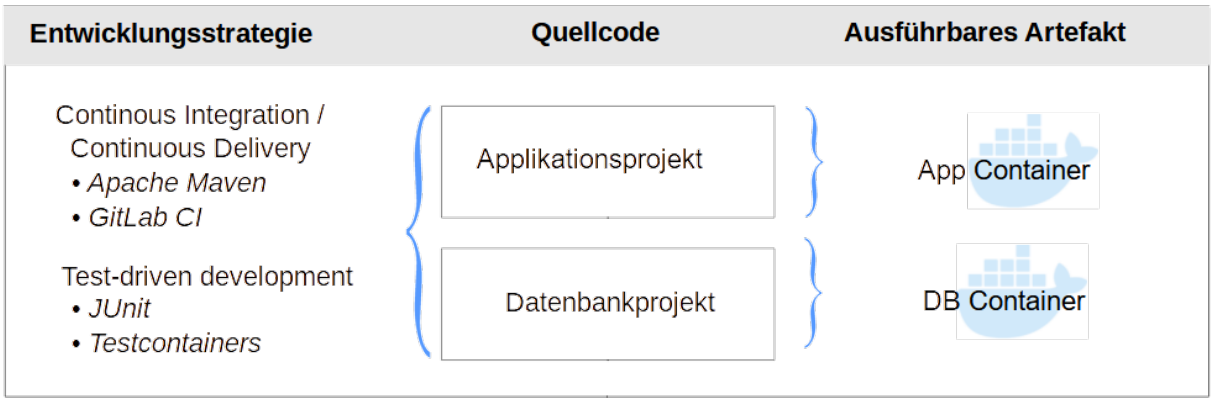
d. Bereich Bibliographie mit der Verwaltung bibliographischer Angaben und Kürzeln.

Damit die Anwendung auch für andere Forschungsprojekte nutzbar und an neue Anforderungen anpassbar ist, sollen die Bereiche möglichst unabhängig voneinander sein und getrennt wiederverwendet werden können. Auf dieses Entwicklungsziel wird bei der Betrachtung der FAIR-Prinzipien noch detailliert eingegangen werden. Im Folgenden wird das UV2-Annotationstool hinsichtlich der Entwicklungsstrategie, des Aufbaus des Quellcodes und der Art des ausführbaren Artefakts vorgestellt. Diese drei Perspektiven sind in Abbildung 1 schematisch aufgeführt. Vorab sei erwähnt, dass bei der Entwicklung nur freie Software, etablierte Verfahren der Softwaretechnik und gängige Entwicklungswerkzeuge eingesetzt werden.

[1] Als Beispiel für „visuelle Annotationen“ sei der Satz in (i) gegeben, der im Asterix Paralleltextkorpus folgende Formatierungen enthält. W-Fragen haben eine rote Textfarbe, das Fragewort ist unterstrichen und das Subjekt ist fett markiert.

(i) "Wo *bleibt* **das Wunder**?" Asterix und die Goten (AST03-08:07a2)

Diese Formatierungen können zwar maschinenlesbar gemacht werden, indem das Tabellendokument als .xml (siehe Bray et. al. 2008) gespeichert oder entpackt wird, aber ihre intendierte Bedeutung ist nicht kodiert. Es bedarf zusätzlicher Transformationsschritte um einen Text, der eine bestimmte Menge an Stilattributen hat, mit linguistischen Annotationen zu versehen. So müsste zum Beispiel der Phrase *das Wunder*, weil sie das Attribut font-weight="bold" hat, eine Annotation für Subjekt zugewiesen werden. In diesem Sinne sind diese Dokumente nicht direkt maschinenlesbar. Es spricht jedoch nichts dagegen, einen Text, der mit bestimmten linguistischen Annotationen ausgezeichnet ist, farbig oder durch Schriftstile darzustellen.



1 | Drei Perspektiven auf eine Softwareanwendung

1.1. Entwicklungsstrategie und ausführbares Artefakt

Die Entwicklungsmethode des Continuous Integration und Continuous Delivery hat zum Ziel, dass zu jeder Zeit, also trotz kontinuierlicher Veränderung des Quellcodes, eine lauffähige Softwareanwendung zur Verfügung steht (cf. Lester 2018). Dies bedeutet, dass mit jeder Änderung der Build- und Testprozess[2] automatisiert ausgeführt werden soll. Für die Automatisierung dieser Prozesse werden für das UV2-Annotationstool das Buildmanagement Apache Maven[3] und die Versionsverwaltung Git[4] sowie GitLab CI wie folgt genutzt. Der Quellcode der Anwendung besteht aus zwei Teilprojekten, dem Applikationsprojekt und dem Datenbankprojekt. Zusätzlich wird eine Entwicklerdokumentation erstellt, auf die hier jedoch nicht näher eingegangen wird. Alle Projekte nutzen Apache Maven um Abhängigkeiten zu verwalten und den Build- und Testprozess zu automatisieren. Alle Projekte werden mit der Versionsverwaltung Git verwaltet. Die Projektdateien werden in einem lokalen Quellcoderepositorium gespeichert, das mit einem externen Quellcoderepositorium synchronisiert werden kann. Für das UV2-Annotationstool wird der Anbieter Gitlab genutzt. Jedes Git-Projekt besitzt eine URL und ist damit eindeutig identifizierbar. Der Zugang auf das Projektrepositorium kann öffentlich gemacht werden oder für ausgewählte Nutzer und Nutzerinnen beschränkt sein. Gitlab ermöglicht es, dass bei jeder Änderung des Quellcodes der Build- und Testprozess automatisiert ausgeführt werden kann. Bei erfolgreichem Abschluss dieser Prozesse wird das ausführbare Artefakt in Form von Docker[5]

[2] Zur Erläuterung dieser Begriffe sei hier kurz angemerkt, dass der Quellcode einer Software nicht nur aus eigenem, sondern größtenteils aus bereits existierendem Programmcode besteht. Letzterer wird importiert und die Verweise auf die entsprechenden Bibliotheken (Abhängigkeiten) müssen verwaltet werden. Während des Buildprozesses werden die Codeteile zusammengeführt, kompiliert (d.h. in ausführbaren Code umgewandelt) und es wird ein ausführbares Objekt erzeugt. Hierzu gehört auch das Ausführen von Programmtests. Diese überprüfen zum Beispiel, ob einzelne Funktionalitäten die gewünschten Ergebnisse liefern.

[3] Apache Maven ist ein freies Buildmanagementtool, das in The Apache Software Foundation (2022) dokumentiert ist.

[4] Git ist ein Open Source Versionsverwaltungssystem (cf. Chacon & Straub 2014). Die bekanntesten Dienstleister, die Git als Clouddienst anbieten, sind gitlab.com oder github.com.

[5] Mit Hilfe von Docker, einer teilweise freien Software, können Anwendungen in sogenannte Container verpackt werden. Auf diese Weise ist die Anwendung vom Betriebssystem isoliert, kann aber trotzdem dessen Ressourcen nutzen.

Containern für das Datenbankprojekt und die Applikation ausgeliefert. Beide Container stehen unter einer eigenen URL zum download zur Verfügung. Auf diese Weise ist das UV2-Annotationstool zu jeder Zeit als lauffähiges Programm erhältlich.

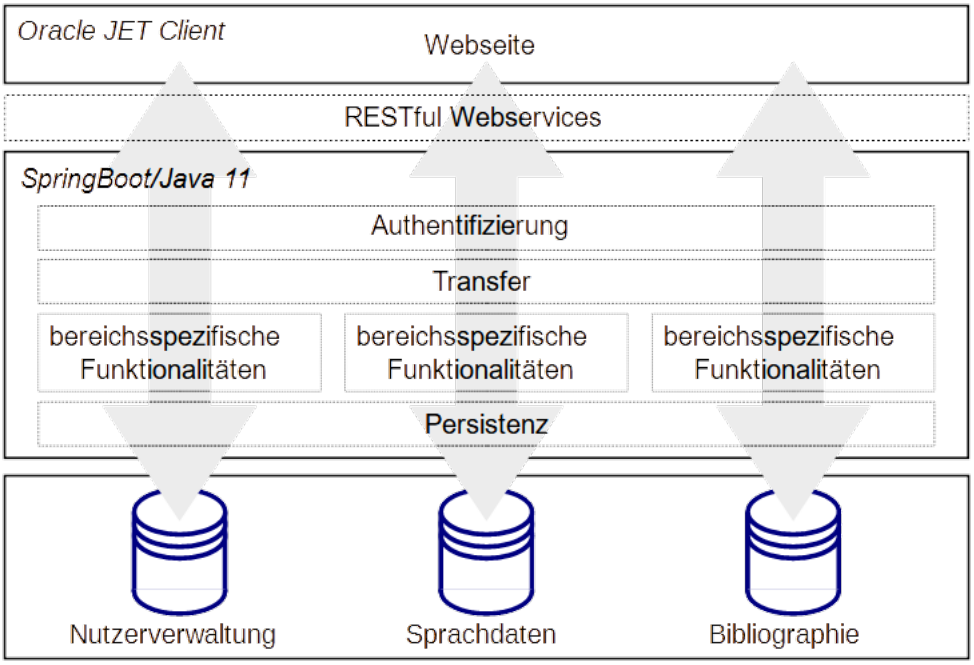
Die Strategie des *test-driven developments* (cf. Beck 2003) kann auf folgende Weise umgesetzt werden: Bevor eine neue Funktionalität implementiert wird, wird für sie ein Test geschrieben. Hierfür wird das Testframework JUnit genutzt (cf. Junit-team 2015-2022). Ein Test überprüft, ob das Programm bei erwarteten Eingaben eine erwartete Ausgabe erzeugt. Mit Hilfe des Tests kann überprüft werden, an welcher Stelle der Programmablauf fehlerhaft oder unvollständig ist. Diese Methode liefert einen Quellcode, dessen Funktionalitäten durch Tests überprüfbar und nachvollziehbar sind. Die Testinfrastruktur des UV2-Annotationstools erlaubt es, die Tests für den Quellcode und für die Container mit Hilfe von Testcontainers (cf. North & Egorov et al. 2015-2022) sowohl lokal, als auch automatisiert auszuführen.

1.2. Struktur des Quellcodes

Im Folgenden wird die Struktur des Quellcodes, insbesondere des Applikationsprojekts detaillierter betrachtet. Das UV2-Annotationstool kombiniert Aspekte des 3-Schichtenmodells[6] und der serviceorientierten Architektur[7] (cf. Richards & Ford 2020). Die Abbildung 2 zeigt eine schematische Darstellung des Aufbaus des Quellcodes.

[6] Im 3-Schichtenmodel wird eine Softwareanwendung in eine Präsentationsschicht, eine Logikschicht und eine Datenhaltungsschicht eingeteilt. In Abbildung 2 sind dies die horizontalen Kästen für den Client, die SpringBoot Applikation und die Datenbanken.

[7] Ein zentraler Gedanke der serviceorientierten Architektur ist die Wiederverwendbarkeit von Softwarekomponenten. In Abbildung 2 sind zum Beispiel die drei Schichten der Bibliographieverwaltung eine wiederverwendbare Komponente. Aber auch die Funktionalität für Authentifizierung und Transfer sind als wiederverwendbare Komponente implementiert.



2 | Architektur des UV2-Annotationstools

Das Applikationsprojekt besteht aus zwei Teilprojekten, dem Client und der SpringBoot/Java11[8] Anwendung, dem sogenannten Backend. Im Clientprojekt werden die Webseite und die Funktionalitäten zur Nutzerinteraktion mit HTML5 (cf. Hickson et al. 2022) und TypeScript (cf. Microsoft 2012-2022) erstellt. In der Backend-Anwendung werden HTTP-Anfragen verarbeitet (Transfer) und Funktionalitäten zu den in (1) genannten Bereichen Nutzerverwaltung, Sprachdaten und Bibliographie ausgeführt sowie die Kommunikation mit der Datenbank geregelt (Persistenz). Der Datenaustausch zwischen Client und Applikation erfolgt über RESTful Webservices (cf. Fielding 2000), indem der Client Anfragen mit dem HTTP Protokoll an Ressourcenadressen, wie zum Beispiel <www.uni-konstanz.de/ling/uv2/alignments>, stellt und diejenigen Daten als Antwort erhält, die unter dieser Adresse zur Verfügung gestellt werden, in unserem Fall die Paralleltexte. Die einzige Verbindung zwischen Client und Backend-Anwendung besteht somit in dem Wissen um die Ressourcenadressen. Auf diese Weise kann bei der Nachnutzung des UV2-Annotationstools auch leichter ein anderer Client verwendet werden.

Im Datenbankprojekt wird für jeden der drei Bereiche – Nutzerverwaltung, Sprachdaten, Bibliographie – eine relationale Datenbank zur Verfügung gestellt. Durch eine entsprechende Abstraktion der Funktionalitäten in der Backend-Anwendung können die drei Bereiche unabhängig voneinander verwendet werden. So ist zum Beispiel denkbar, dass die Webanwendung mit Nutzerverwaltung und Bibliographie wiederverwendet wird, jedoch der Bereich Sprachdaten nicht genutzt wird oder verändert werden soll. Auch kann das Datenbankprojekt vollkommen

[8] Spring Boot Framework ist in The Spring Team (2022) beschrieben und die Programmiersprache Java ist in Oracle (2021) dokumentiert.

unabhängig von der Webapplikation genutzt werden, zum Beispiel um Datenbanken aus Tabellendokumenten zu erstellen und, wenn gewünscht, in einem Docker Container auszuliefern. Die Datenbanken und der Container können wiederum unabhängig vom UV2-Annotationstool mit anderen Anwendungen verwendet werden.

2. Die FAIR-Prinzipien für Forschungssoftware

Die Diskussion der FAIR-Prinzipien für Forschungssoftware beginnt mit den Prinzipien der Wiederverwendbarkeit (*reusable*) und der Interoperabilität (*interoperable*), weil beide zentral für die Entwicklung des UV2-Annotationstools sind. Anschließend werden die Prinzipien der Auffindbarkeit (*findable*) und der Zugänglichkeit (*accessible*) diskutiert. Letztere sind vor allem für die Veröffentlichung von Softwareanwendungen relevant.

2.1. Wiederverwendbarkeit

Das Prinzip der Wiederverwendbarkeit (s. (2) unten) expliziert die Zweigestaltigkeit von Softwareanwendungen: Software ist nutzbar im Sinne eines ausführbaren Artefakts und wiederverwendbar. Wiederverwendbarkeit betrifft vor allem, aber nicht ausschließlich, den Quellcode und beinhaltet Kriterien wie die Verständlichkeit und die Anforderungen, dass Softwareelemente verändert, erweitert und in andere Softwareanwendungen integriert werden können. Wie in Abschnitt 1.2. zur Softwarearchitektur dargelegt, ist die Wiederverwendung einzelner Softwarekomponenten des UV2-Annotationstools, aber auch die Wiederverwendung der ausführbaren Artefakte möglich. Die Verwendung von Maven erleichtert zusätzlich die Lesbarkeit und Verständlichkeit des Quellcodes.

(2) R. Reusable

The software is both usable (it can be executed) and reusable (it can be understood, modified, built upon, or incorporated into other software).

R1. Software is described with a plurality of accurate and relevant attributes.

R1.1. Software must have a clear and accessible license.

R1.2. Software is associated with detailed provenance.

R2. Software includes qualified references to other software.

R3. Software meets domain-relevant community standards.

(Chue Hong et al. 2021, 13-14, Hervorhebung im Original)

Das Prinzip der Wiederverwendbarkeit enthält die drei Unterprinzipien R1-R3, die technische, rechtliche und organisatorische Aspekte betreffen. Das Kriterium R1 besagt allgemein, dass eine Softwareanwendung mit Metadaten ausgezeichnet werden soll, die sowohl technische als auch forschungsrelevante Details beschreiben. Dies kann als Teil des Quellcodes oder in Form einer Dokumentation geschehen. Explizit wird die Softwarelizenz genannt (R 1.1), weil die Art der Lizenz entscheidend ist, inwieweit und mit welchen Bedingungen der Quellcode wiederverwendet werden darf. Das UV2-Annotationstool wird unter der European Public

License 2.0 veröffentlicht, einer Open Source Lizenz mit einer starken *copy-left* Klausel. Die Software darf daher ganz oder in Teilen wiederverwendet werden, jedoch mit der Einschränkung, dass die neu entstandenen Anwendungen auch unter einer kompatiblen Open Source Lizenz veröffentlicht werden müssen. Des Weiteren werden für das UV2-Annotationstools nur freie und Open Source Tools und Software verwendet. Auf diese Weise entstehen keine lizenzbedingten Einschränkungen für eine Wiederverwendung.

Das Kriterium R1.2 bezieht sich auf technische und organisatorische Informationen, wie zum Beispiel die Autorenschaft und die Entwicklungshistorie, aber auch auf die genutzten Tools und Techniken sowie die Designentscheidungen. Gerade bei größeren Open Source Projekten ist die Autorenschaft einzelner Softwareelemente häufig wesentlich komplexer, als bei Forschungsdaten, da mehrere Akteure die Software über den Entwicklungszeitraum ergänzen und verändern. Mit der Versionsverwaltung Git, die auch für das UV2 Annotationstool verwendet wird, lassen sich die Entwicklungshistorie und die genaue Autorenschaft einzelner Softwareelemente dokumentieren. Jede Änderung des Quellcodes, ein sogenannter *commit*, trägt einen Bezeichner und ist eindeutig einem Zeitpunkt und einem Entwickler zuzuordnen. Bei Softwareprojekten, die Maven verwenden, sind die Angaben zur Autorenschaft, zu verwendeten Tools und den in R2 genannten Abhängigkeiten zu anderen Softwarebibliotheken in der *project object model* Datei (pom.xml) aufgeführt. Diese Datei steuert die Automatisierung, ist aber auch mit einem Inhaltsverzeichnis zu vergleichen und gibt einen strukturierten Überblick zu den genannten Angaben.

Das Kriterium R3 ist dahingehend zu verstehen, dass Software dann wiederverwendbar ist, wenn die in der Forschungsrichtung üblichen Techniken (Programmiersprache, Teststandards), aber auch Dateiformate verwendet werden. In Bezug auf übliche Dateiformate können laut Umfrage der AG Digitale Romanistik von 2015 tabellenförmige Formate als *community standard* in der Romanistik bezeichnet werden. Das UV2 Annotationstool ermöglicht es Tabellendokumente als CSV[9] Dateien direkt zu importieren und zu exportieren. Bezüglich der Programmiersprachen oder der Teststandards gibt es unseres Wissens nach keine *community standards*, weshalb für das UV2-Annotationstool die weitverbreiteten Programmiersprachen Java 11, HTML5 und TypeScript verwendet werden.

2.2. Interoperabilität

Interoperabilität bezeichnet die Zusammenarbeit von unabhängig voneinander ausführbaren Systemen, wobei Chue Hong et al. (2021) hier Zusammenarbeit auf die Möglichkeit des Datenaustausches beschränken. Fragen zur Kompatibilität von Softwareelementen und Softwaresystemen sind im Geltungsbereich der Wiederverwendbarkeit angesiedelt.

[9] CSV steht für *Comma-Separated Values* und ist ein Dateiformat zur Darstellung von tabellenförmigen Daten (cf. Shafranovich 2005).

(3) I. Interoperable

The software interoperates with other software through exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs).

I1. Software reads, writes and exchanges data in a way that meets domain-relevant community standards.

I2. Software includes qualified references to other objects.

(Chue Hong et al. 2021, 12-13, Hervorhebung im Original)

Das UV2 Annotationstool hat zwei Schnittstellen, an denen Daten ausgetauscht werden. Dies sind zum einen die RESTful Webservices und die Schnittstelle zwischen Applikation und Datenbank. Zur Übertragung der Daten, zum Beispiel der Sprachdaten, zwischen Client und Applikation, wird das JSON Format genutzt. JSON ist ein Internetstandard zum Austausch hierarchisch strukturierter Daten (cf. Bray 2017). Die Applikation wandelt die JSON Daten in Java Einheiten um (und umgekehrt), so dass diese mittels der Spring Data JPA, der Schnittstelle zwischen Datenbank und Applikation, in der Datenbank gespeichert oder von dort gelesen werden können. Das Kriterium I1 nimmt Bezug auf sogenannte *community standards.* Für (halb-)automatisch annotierte Korpora existieren verschiedene Modelle und Dateiformate, die von globaleren Standards wie XML[10] oder TSV[11] abgeleitet sind. So zum Beispiel die XML-basierten Formate TEI[12] oder WebLicht TCF[13], die TSV Formate der CoNLL Familie oder WebAnno TSV[14]. Des Weiteren werden auch OWL oder graphenbasierte Formate wie Salt verwendet.[15]

Für das UV2-Annotationstool wurde aus den folgenden Gründen der Internetstandard JSON und eine relationale Datenbank gewählt: Zum einen ist eine Konvertierung in oben genannte Formate grundsätzlich möglich, die integriert werden kann. Zum anderen verfolgt das UV2-Annotationstool eine andere Zielsetzung als Werkzeuge zur Erstellung eines automatisch annotierten Korpus, das als Endergebnis zur Veröffentlichung vorliegt. Vielmehr dient das UV2-Annotationstool der Erstellung einer textbasierten Sprachdatensammlung, die vollständige Texte, aber auch einzelne Sätze enthält und je nach Forschungsinteresse (re-)annotiert und erweitert werden kann. Ziel des Tools ist es, die linguistische Analyse zu unterstützen, indem eine personalisierte, aber FAIRe Datenbasis erstellt, annotiert, durchsucht und veröffentlicht werden kann.

[10] XML steht für *Extensible Markup Language* und ist in Bray et. al. (2008) dokumentiert.
[11] Das TSV Format (Tab Separated Values) ist ein Datenaustauschformat für tabellenförmige Daten, deren Spalten mittels des tab-Zeichens (U+0009) getrennt werden (cf. Korpela 2005).
[12] TEI ist ein XML-basiertes Format zur Auszeichnung von Texten, das von der Text Encoding Initiative kuratiert wird (cf. TEI Consortium 2021).
[13] WebLicht TCF (Text Corpus Format) ist ein XML-basiertes Format, das in CLARIN-D/SfS-Universität Tübingen (2012) dokumentiert ist.
[14] Das WebAnno TSV Format ist in Eckart de Castilho et al. (2016) und The WebAnno Team (2021) beschrieben.
[15] Für eine Dokumentation von OWL siehe W3C OWL Working Group (2012) und für Salt Zipser & Romary (2010).

Das Kriterium I2 bezieht sich auf Datenobjekte, wie zum Beispiel Konfigurationsdateien, die zur Ausführung der Software nötig sind. Im Falle des UV2-Annotationstools benötigt der Applikationscontainer eine Parameterdatei, um sich mit dem Datenbankcontainer zu verbinden. Diese Dateien sollten mindestens mit einem Bezeichner, aber idealerweise einer auflösbaren Referenz verknüpft sein, wie zum Beispiel einer URL oder einem DOI.

2.3. Auffindbarkeit

Das Prinzip *findable* gibt vier Kriterien an, wie eine Software in Form von Quellcode oder als ausführbares Artefakt sowie dazugehörige Metadaten leicht gefunden werden kann.

(4) F. Findable

> *The software, and its associated metadata, should be easy to find for both humans and machines.*
>
> *F1. Software is assigned a globally unique and persistent identifier.*
>
> *F1.1. Different components of the software must be assigned distinct identifiers representing different levels of granularity.*
>
> *F1.2. Different versions of the same software must be assigned distinct identifiers.*
>
> *F2. Software is described with rich metadata.*
>
> *F3. Metadata clearly and explicitly include the identifier of the software they describe.*
>
> *F4. Metadata are FAIR and are searchable and indexable.*
>
> (Chue Hong et al. 2021, 9-11, Hervorhebung im Original)

Das wichtigste Kriterium ist sicher F1. Die Software sollte mit einem eindeutigen und persistenten Bezeichner versehen werden. Diese Bezeichner können Softwareelemente unterschiedlicher Komplexität (F1.1) und verschiedene Versionen einer Software (F1.2) kennzeichnen. Das Arbeitspapier der Research Data Alliance/ FORCE11 SSCID WG et al. (2020) gibt eine Übersicht, welche Bezeichner sich für Softwareelemente verschiedener Komplexität eignen. So wird zum Beispiel Wikidata für eine Auszeichnung auf Projektebene angeführt, weil dort eine ausführliche Beschreibung mit Metadaten und Verweisen vorhanden ist. Ein DOI bietet sich für die Identifikation von Softwareversionen und Projektordnern an. Ein softwarespezifischer Bezeichner ist der Software Heritage Persistent Identifier (SWHID), der auch mit Bezeichnern der Versionsverwaltung Git kompatibel ist (cf. Di Cosmo & Zacchiroli 2017, Software Heritage 2021). Dieser erlaubt eine Identifikation von Dateien, aber auch von *commits* und Programmcodeelementen. Das UV2-Annotationstool ist auf Projektebene (Quellcode und Container) über die Gitlab URL auffindbar. Eine URL ist zwar ein eindeutiger, aber kein persistenter Bezeichner, weshalb eine spätere Veröffentlichung mit einem DOI geplant ist. Die Kriterien F2-F4 beziehen sich auf Metadaten, die auch über Suchmaschinen gefunden werden sollen. Bezüglich ihrer Umsetzung gibt es keine Unterschiede

zwischen Softwareanwendungen und Forschungsdaten, sodass hier nicht näher auf diese Kriterien eingegangen wird.

2.4. Zugänglichkeit

Eine Softwareanwendung sollte nicht nur auffindbar, sondern auch zugänglich bzw. abrufbar sein. Zugänglichkeit meint hier in erster Linie technische Erreichbarkeit und nicht barrierefreie Nutzung der Softwareanwendung. Zugänglichkeit oder Abrufbarkeit ist garantiert, wenn ein standardisiertes Kommunikationsprotokoll verwendet wird, wie zum Beispiel HTTP. HTTP erfüllt die Kriterien A1.1 und A1.2.

(5) A. Accessible

The software, and its metadata, must be retrievable via standardized protocols.

A1. Software is retrievable by its identifier using a standardized communications protocol.

A1.1. The protocol is open, free, and universally implementable.

A1.2. The protocol allows for an authentication and authorization procedure, where necessary.

A2. Metadata are accessible, even when the software is no longer available.

(Chue Hong et al. 2021, 11-12, Hervorhebung im Original)

In Bezug auf das Problem des Langzeitbetriebs bzw. der Langzeitarchivierung von Softwareanwendungen ist das Kriterium A2 ein Kompromiss. So wird vorgeschlagen, dass die Metadaten, die eine Software beschreiben, zugänglich bleiben sollen, auch wenn die Software selbst nicht mehr abrufbar ist. Der Quellcode und die ausführbaren Dateien des UV2-Annotationstools sind über HTTP zugänglich. Das Tool kann als Webservice auf einem lokalen Rechner oder in einem Rechenzentrum betrieben werden.

3. Ausblick

Mit den FAIR-Prinzipien für Forschungssoftware geben Chue Hong et al. (2021) einen Leitfaden für die Entwicklung und Evaluation von Softwareanwendungen vor.

Wie im Beitrag aufgezeigt, werden die Prinzipien bei der Entwicklung des UV2-Annotationstools mit etablierten Verfahren der Softwaretechnik umgesetzt. Die Nutzung der Versionsverwaltung Git und eines Anbieters wie Gitlab oder Github ermöglicht es, schon zu Beginn der Entwicklung, die Prinzipien der Auffindbarkeit und Zugänglichkeit bzw. Abrufbarkeit umzusetzen. Wesentlich vielschichtiger ist die Umsetzung der Prinzipien der Interoperabilität und Wiederverwendbarkeit, weil hierbei technische, organisatorische und rechtliche Faktoren eine Rolle spielen. Um Wiederverwendbarkeit auf rechtlicher Ebene zu sichern, werden für das UV2-Annotationstool eine Open Source Lizenz gewählt und nur freie Software und offene Standards verwendet. Der Aufbau des Quellcodes ermöglicht eine Wiederverwendung auf der Ebene der Schichten (Client, Applikation, Datenbank)

sowie auf der Ebene der Bereiche Datenspeicherung, Sprachdaten, Nutzerverwaltung und der Bibliographie. Die Wiederverwendung auf der Ebene der Schichten wird durch die Implementierung der RESTful Webservices und die Verwendung der Spring Data JPA erleichtert. Die komponentenbasierte Architektur des Applikationsprojekts erlaubt die getrennte Wiederverwendung der oben genannten Bereiche. Hinsichtlich der zu verwendenden Datenformate und Programmiersprachen werden in den FAIR-Prinzipien geltende *community standards* erwähnt. Innerhalb der Romanistik können tabellenförmige Datenformate als Standard bezeichnet werden, der im UV2-Annotationstool berücksichtigt wird. Weitere fachspezifische Anforderungen an Forschungssoftware zu erörtern und zu formulieren, ist sicherlich ein längerer Prozess, den dieser Beitrag mit anstoßen möchte.

Bibliografie

AG DIGITALE ROMANISTIK. 2015. „Ergebnisse der Umfrage der AG Digitale Romanistik zur Langzeitarchivierung von digitalen Forschungsdaten für die Romanistik.“ *Mitteilungsheft des Deutschen Romanistenverbands e.V., Frühjahr 2015*, 36–40.
<http://deutscher-romanistenverband.de/wp-content/uploads/sites/14/Auswertung_Forschungsdaten-Umfrage.pdf>, 17.2.2022.

BECK, Kent. 2003. *Test-driven development: by example*. Boston: Addison-Wesley.

BRAY, Tim. 2017. „The JavaScript Object Notation (JSON) Data Interchange Format.“ STD 90, RFC 8259.
<https://doi.org/10.17487/RFC8259>, 17.2.2022.

BRAY, Tim, et al. (eds.). 2008. *Extensible Markup Language (XML) 1.0*. W3C Recommendation 26 November 2008.
<https://www.w3.org/TR/xml/>, 17.2.2022.

BRÜGGE, Bernd & Allen H. Dutoit. 2014. *Object-oriented software engineering using UML, Patterns, and Java.* Harlow, Essex: Pearson.

CHACON, Scott & Ben Straub. 2014. *Pro Git*. New York. Apress.

CHUE HONG, Neil P. et al. 2021. „FAIR Principles for Research Software (FAIR4RS Principles).“ Research Data Alliance.
<https://doi.org/10.15497/RDA00065>

CLARIN-D/SfS-Universität Tübingen. 2012. *WebLicht: Developer Manual.*
<https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format>, 17.2.2022.

DI COSMO, Roberto & Stefano Zacchiroli. 2017. „Software Heritage: Why and How to Preserve Software Source Code.“ *iPRES 2017: 14th International Conference on Digital Preservation*, Kyoto, Japan.
<https://hal.archives-ouvertes.fr/hal-01590958>, 24.9.2022

ECKHART Arnold et al. 2021. „Einfach FAIR. Geisteswissenschaftliches Arbeiten und nachhaltiges Publizieren von Forschungsdaten mit Git.“ Vortrag auf der Konferenz Forschungsdaten in den Geisteswissenschaften (FORGE 2021), Universität zu Köln, 08.-10.09.2021.

ECKART DE CASTILHO, R. et al. (2016): „A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures.“ *Proceedings of the LT4DH workshop at COLING 2016*, Osaka, Japan.
<https://aclanthology.org/W16-4011/>, 24.9.2022.

FIELDING, Roy Thomas. 2000. *Architectural Styles and the Design of Network-based Software Architectures*. PhD Dissertation. University of California, Irvine.

HICKSON, Ian et al. 2022. *HTML Living Standard*.

<https://html.spec.whatwg.org/>, 17.2.2022.

JUNIT-TEAM. 2015-2022. *JUnit5*.
<https://junit.org/junit5/>, 17.2.2022.

KORPELA, Jukka. 2005. „Tab Separated Values (TSV): a format for tabular data exchange."
<https://jkorpela.fi/TSV.html>, 17.2.2022.

KREFELD, Thomas & Stephan Lücke. 2020. „54 Monate VerbaAlpina – auf dem Weg zur FAIRness." *Ladinia* XLIII, 139-156.

LESTER, Brent. 2018. *Continuous Integration Vs. Continuous Delivery Vs. Continuous Deployment: Distinguishing Terms and Understanding how their Implementation Methods and Tools Differ*. Sebastopol, CA: O'Reilly Media.

MCILROY, Malcolm Douglas. 1969. „Mass Produced Software Components." In *Software Engineering Concepts and Techniques*, ed. Naur, Peter, Brian Randell & Friedrich Ludwig Bauer, 88–98, Brussels: Scientific Affairs Division, NATO.

MEYER, Bertrand. 1997. *Object-oriented software construction*. Upper Saddle River, NJ: Prentice Hall PTR.

MICROSOFT 2012-2022. *TypeScript*.
<https://www.typescriptlang.org/>, 17.2.2022.

NORTH, Richard & Sergei Egorov et al. 2015-2022. *Testcontainers*.
<https://www.testcontainers.org/>, 17.2.2022.

ORACLE. 2021. *Java*.
<https://dev.java/>, 17.2.2022.

RESEARCH Data Alliance/FORCE11 Software Source Code Identification WG et al. 2020. „Use cases and identifier schemes for persistent software source code identification (V1.1)." Research Data Alliance.
<https://doi.org/10.15497/RDA00053>, 24.9.2022.

RICHARDS, Mark & Neal Ford. 2020. *Handbuch moderner Softwarearchitektur* (Übersetzung von Jorgen W. Lang). Heidelberg: dpunkt.

SHAFRANOVICH, Yakov. 2005. „Common Format and MIME Type for Comma-Separated Values (CSV) Files. RFC Editor, RFC 4180.
<https://doi.org/10.17487/RFC4180>, 17.2.2022.

SOFTWARE HERITAGE. 2021. *Development Documentation: SoftWare Heritage persistent IDentifiers (SWHIDs).*
<https://docs.softwareheritage.org/devel/swh-model/persistent-identifiers.html>, 17.2.2022.

TEI CONSORTIUM. 2021. *TEI P5: Guidelines for Electronic Text Encoding and Interchange* (v4.3.0). Zenodo.
<https://doi.org/10.5281/zenodo.5347789>, 17.2.2022.

THE APACHE SOFTWARE FOUNDATION. 2022. *Apache Maven Project*.
<https://maven.apache.org/index.html>, 17.02.2022.

THE SPRING TEAM. 2022. *Spring Boot Framework*.
<https://spring.io/projects/spring-boot>, 17.2.2022.

THE WEBANNO TEAM. 2021. *WebAnno User Guide*, version 3.6.11.
<https://webanno.github.io/webanno/releases/3.6.11/docs/user-guide.html#sect_webannotsv>, 17.2.2022.

WILKINSON, Mark D. et al. 2016: „The FAIR Guiding Principles for scientific data management and stewardship." *Science Data* 3 (160018).
<https://doi.org/10.1038/sdata.2016.18>, 24.9.2022.

W3C OWL WORKING GROUP. 2012. *OWL 2 Web Ontology Language. Document Overview* (Second Edition). W3C Recommendation 11 December 2012.
<https://www.w3.org/TR/owl-overview/>, 17.2.2022.

ZIPSER, Florian & Laurant Romary. 2010. „A model oriented approach to the mapping of annotation formats using standards." *Proceedings of the*

Workshop on Language Resource and Language Technology Standards, LREC 2010. Malta.
<http://hal.archives-ouvertes.fr/inria-00527799/en/>, 17.2.2022.

Zusammenfassung

Der Beitrag stellt die FAIR-Prinzipien für Forschungssoftware nach Chue Hong et al. (2021) vor und diskutiert ihre Umsetzung am Beispiel des UV2-Annotationstools. Hierbei werden die umgesetzten Entwicklungsstrategien des Continuous Integration/Continuous Delivery und des test-driven development, die komponentenbasierte Architektur und die thematische Strukturierung des Quellcodes sowie die Art des ausführbaren Artefakts (Docker Container) vorgestellt und bezüglich der FAIR-Prinzipien bewertet.

Abstract

The article introduces the FAIR principles for research software, as proposed in Chue Hong et al. (2021), and discusses their realisation by means of the UV2 annotation tool. We examine the following aspects of the tool's implementation: the component-based architecture, the structuring of the source code, the type of the executable artefact (docker container), and the way the development strategies Continuous Integration/Continuous Delivery and test-driven development are realized.

2022, n°9
pp. 254-275
doi: https://doi.org/10.15460/apropos.9.1895

Florian Zacherl

Linguistische Online-Ressourcen auf Basis traditioneller Werke

Anforderungen und digitale Möglichkeiten am Beispiel des *Romanischen Etymologischen Wörterbuchs*

Florian Zacherl
ist wissenschaftlicher Mitarbeiter in der IT-Gruppe Geisteswissenschaften an der Ludwig-Maximilians-Universität München.
Florian.Zacherl@itg.uni-muenchen.de

Keywords

Digitale Lexikographie – Wörterbücher – Webportale

1. Einleitung

Die manuelle Aggregation von Informationen aus gedruckten linguistischen Quellen stellt oftmals einen aufwendigen Prozess dar. Eine digitale Online-Präsentation hat das Potential diesen deutlich zu beschleunigen. Dieser Artikel analysiert am Beispiel des Romanischen Etymologischen Wörterbuchs (REW, entspricht Meyer-Lübke 1935) die Konzeption eines Webangebots auf Basis einer solchen traditionellen Quelle, das die digitalen Möglichkeiten umfassend nutzt. Der Fokus dieses Artikels liegt dabei auf den Anforderungen, die eine solche Online-Ressource erfüllen sollte, und der Frage, welche zusätzlichen Möglichkeiten sie im Gegensatz zur gedruckten Vorlage bieten kann. Primär wird die Funktionalität betrachtet, die menschlichen Nutzenden den direkten Zugriff ermöglicht (im Gegensatz zur maschinellen Nutzung).

Das REW erschien erstmals 1911, wobei als Grundlage hier die dritte und finale (neubearbeitete) Auflage dient. Diese enthält anhand von 10701[1] (vorwiegenden lateinischen) Lemmata im wesentlichen Listen der zugehörigen „romanischen Vertreter“ (Meyer-Lübcke 1935, XI) sowie vor allem in umstrittenen oder unklaren Fällen weiterführende Informationen und vom Autor abgelehnte Etymologien. Aufgrund seiner Anlage als Sammlung und kritische Einordnung der „wichtigeren der ungemein zahlreichen und vielfach weit zerstreuten etymologischen Untersuchungen auf dem Gebiete der romanischen Sprachen“ (Meyer-Lübcke 1935, VIII)

[1] Die Lemmata sind von 1 bis 9721 nummeriert. Die Anzahl ergibt sich unter Berücksichtigung von Einfügungen und Entfernungen im Vergleich zu den beiden früheren Auflagen.

ist das REW trotz seines Alters weiterhin ein unverzichtbares Hilfsmittel in der romanischen Etymologie. Das einzige vergleichbare neuere Werk, das denn gesamtromanischen Sprachraum abdeckt, ist das „Dictionnaire Étymologique Roman“, das aber bei weiten nicht den Umfang des REW hat und methodisch durchaus umstritten ist (vgl. z.B. Vàrvaro 2011).

Dieser Artikel ist im Kontext eines Dissertationsprojekts entstanden, welches sich primär mit der feingranulierten Erschließung von strukturierten Daten aus Wörterbuchtexten befasst und zu diesem Zweck das REW als Beispiel systematisch auswertet und digital publiziert (vgl. Zacherl in Vorb. a). Das in diesem Rahmen entstandene Webangebot ist als Entwurf bereits zugänglich (vgl. REWOnline), zum jetzigen Stand (14.10.2022) aber noch nicht vollständig funktional.

Das folgende Kapitel untersucht zunächst, welche Anforderungen eine Online-Ressource in Gegenüberstellung mit einer gedruckten Ressource sinnvollerweise erfüllen sollte, während im Weiteren daraus abgeleitete Anforderungen an die Datenmodellierung (Kapitel 3) und den Aufbau der Oberfläche eines Webangebots (Kapitel 4) betrachtet werden. Schließlich wird ein kurzer Ausblick auf die zusätzlichen Möglichkeiten einer maschinellen Nutzung gegeben (Kapitel 5).

2. Anforderungen an linguistische Online-Ressourcen

Die behandelten Funktionen werden hier grob in drei Kategorien eingeteilt, die allerdings nicht immer trennscharf unterschieden werden können: *Grundsätzliche Anforderungen* stehen für elementare Eigenschaften, die eine Online-Ressource erfüllen muss, um die Aufgabe der gedruckten Vorlage weiterhin zu erfüllen. Dies schließt auch ein, dass spezifische Nachteile von Online-Publikationen im Vergleich zu traditionellen Werken soweit wie möglich ausgeglichen werden. Mit *Arbeitserleichterungen* sind spezifisch digitale Werkzeuge gemeint, die klassische linguistische oder allgemein wissenschaftliche Arbeitsweisen vereinfachen und Aufgaben beschleunigen. Unter *erweiterte Möglichkeiten* werden Funktionalitäten eingeordnet, die sich nur durch die Umwandlung in die digitale Form ergeben und die nicht (oder nicht mit realistischem Aufwand) mit dem gedruckten Werk erreicht werden können.

2.1. Grundsätzliche Anforderungen

Zwei Basisbedingungen, die eine wissenschaftliche Publikation und insbesondere ein Wörterbuch erfüllen muss, sind der Zugriff auf die vollständige relevante Information einschließlich Verweisen auf andere Literatur und die Möglichkeit der kleinteiligen Zitation. Der Zugriff selbst erscheint trivial, aber gerade bei Wörterbüchern (oder ähnlichen stärker strukturierten Texten) bietet eine Online-Ressource die Möglichkeit die Inhalte aufzubereiten und auch in anderer Struktur (oder auch in verschiedenen wählbaren Formaten) darzustellen. Je größer allerdings der Unterschied zwischen der digitalen Darstellung und dem ursprünglichen Quellenmaterial ist, desto hilfreicher ist eine zusätzliche Zugriffsmöglichkeit auf den Originaltext, um das Ausgangsmaterial transparent zu machen und auch Fehler, die

im Prozess der technischen Transformation eventuell aufgetreten sind, auf einfache Weise erkennbar zu machen. Bei einem Wörterbuch bedeutet das weiterhin, dass nicht nur die eigentlichen Einträge, sondern auch andere, weniger strukturierte Abschnitte wie Vorwörter oder ähnliches online zugreifbar sind.

Für die Zitation werden traditionell Seitenzahlen und eventuell zusätzlich Zeilennummern verwendet, die in bei Print-Publikationen eine formatabhängige Notwendigkeit sind, bei digitalen Publikationen aber weder erforderlich noch sinnvoll sind (cf. z.B. Prätor 2011). Dies erzeugt eine eigene Problematik, die allerdings in Bezug auf Wörterbücher weniger relevant ist, da die dort übliche Referenzierung auf einzelne Lemmata weiterhin möglich ist. Relevant bleibt allerdings die mangelnde Stabilität (im inhaltlichen Sinne) von Online-Publikationen, die bei einem gedruckten Werk innerhalb einer Auflage zwingend gegeben ist. Eine Form von Versionierung, die eine statische Ausgabe von Texten oder Textbestandteilen erzeugt, ist also notwendig. Diese sollte bei einer vollständigen Nutzung der digitalen Möglichkeiten unmittelbar und kleinteilig sein. Das bedeutet, dass Korrekturen nach einer Änderung sofort zitierbar sind und die Version sich auf kleinere Sinnabschnitte bezieht, die im Falle eines Wörterbuchs die einzelnen Einträge sein könnten. Gerade die Unmittelbarkeit stellt einen großen Vorteil gegenüber traditionellen Publikationen dar, bei denen Änderungen nur in sehr großen Zeitintervallen eingepflegt werden können. Somit wäre die Möglichkeit wünschenswert solche Änderungen möglichst direkt und einfach vornehmen zu können (cf. Kapitel 2.3).

2.2. Arbeitserleichterungen

Um die Arbeit mit wissenschaftlichem Material zu erleichtern, bestehen vor allem die Möglichkeiten, erstens die Textbasis an sich aufzubereiten und/oder anzureichern und zweitens die Schaffung verbesserter Such- und Zugriffsfunktionalitäten. Das primäre Ziel einer Aufbereitung sollte eine verbesserte Lesbarkeit sein. Zudem soll der Wechsel zu anderen Teilen des Werks vereinfacht bzw. durch möglichst vollständige Einträge ggf. unnötig gemacht werden. In gedruckten Publikationen besteht grundsätzlich ein gewisser Zwang, Platz einzusparen, was gerade bei älteren Werken oftmals zu einer Vielzahl von abkürzenden Schreibweisen, Auslassungen unter bestimmten Bedingungen und sehr engen blockartigen Texten führt. Ein digitales Format hat diese Einschränkungen nicht[2] und kann somit die Darstellung an der optimalen Lesbarkeit und der Präferenz der Nutzenden ausrichten. Bestehende Abkürzungen im Quellenmaterial können entweder komplett ersetzt oder auf geeignete Weise aufgelöst werden. Dies erspart den Wechsel zu Abkürzungsverzeichnissen, Bibliographien oder Ähnlichem. Interne Referenzen (z.B. auf ein bestimmtes Lemma) sollten mit Links hinterlegt werden. Auch für externe Referenzen (z.B. Literaturangaben) bietet sich dies an, falls die entsprechenden Ressourcen digital zugreifbar sind und die notwendigen

[2] Grundsätzlich besteht natürlich ein begrenztes Datenvolumen. Diese Einschränkung ist allerdings meistens (und insbesondere im gegebenen Fall) vernachlässigbar. Die Darstellung (also insbesondere Abstände, Schriftgrößen etc.) wird nur durch die jeweilige Bildschirmgröße beschränkt, wobei auch dies durch Zoom und „unendliches“ Scrollen kaum eine realistische Begrenzung ist.

technischen Möglichkeiten (wie beispielsweise eine seitengenaue Verlinkung) aufweisen. Auch bei Entitäten, die in der Quelle keine Verknüpfung im eigentlichen Sinne aufweisen, kann eine solche aufgebaut werden. So können beispielsweise Bedeutungen oder sprachliche Formen, die an verschiedenen Stellen vorkommen, passend verknüpft werden, sodass eine zusätzliche Vernetzung der Artikel untereinander erzeugt wird (cf. Kapitel 4.2).

Der Einstieg in die Ressource kann ebenfalls variabler gestaltet werden als bei einem gedruckten Werk. In einem solchen gibt es zum Teil ein Inhalts- bzw. Lemmaverzeichnis, in anderen Fällen hilft nur die (beispielsweise alphabetische) Ordnung beim Auffinden der einzelnen Einträge. Zusätzlich sind in den meisten Fällen Wortverzeichnisse vorhanden, die aber nicht zwangsläufig vollständig sind. Das folgende Zitat illustriert dieses Problem, das wiederum primär mit den Platzbeschränkungen von gedruckten Werken zusammenhängt:

> Die Wortverzeichnisse der anderen Sprachen sind möglichst vollständig, das deutsch-romanische bietet naturgemäß nur eine Auswahl, ist gegen die erste Ausgabe in den Stichwörtern kaum erweitert worden, erschöpft auch nicht den im Texte enthaltenen Stoff, da eine noch weitere Ausdehnung des Raumes ausgeschlossen war […] (Meyer-Lübke 1935, 815)

Eine digitale Ressource kann diese Möglichkeiten ungemein erweitern, sei es durch Volltextsuche bzw. spezialisierte Suchmöglichkeiten (z.B. nach bestimmten Formen, Bedeutungen, Sprachen, Literaturangaben etc.) oder auch durch vollständige automatisiert aus den tatsächlichen Vorkommen generierte Verzeichnisse der entsprechenden Entitäten. So ist insbesondere bei passender Datenstrukturierung grundsätzlich sowohl ein semasiologischer als auch ein onomasiologischer Zugang möglich, unabhängig davon, wie dies im Quellenmaterial der Fall war.

2.3. Erweiterte Möglichkeiten

Eine originär digitale Option ist die direkte Einbindung von Nutzenden, wenn ihnen die Möglichkeit gegebenen wird, auf bestimmte Weise einen eigenen Beitrag zu liefern. In der Sprachwissenschaft werden zum Teil Formen von Crowdsourcing für Transkriptionsaufgaben oder die Erhebung neuer Daten eingesetzt. Dies kann über externe Portale wie Zooniverse (cf. *Zooniverse* 2009) geschehen, wie etwa beim ISTROX-Projekt (cf. ISTROX 2020), oder über eigene Tools, wie beispielsweise beim *Atlas der deutschen Alltagssprache* (cf. z.B. Möller/ Elspaß 2014) oder dem Projekt *VerbaAlpina* (cf. Krefeld/Lücke 2021). Die entsprechende Plattform dient in diesem Fall gleichzeitig der Erhebung und der Publikation. Obwohl im Falle der Digitalisierung eines bestehenden Werks keine neuen Daten erhoben werden, ist durchaus eine Einbindung von Nutzenden denkbar. Die direkte Möglichkeit von z.B. Fehlerkorrekturen (wie es beispielsweise in den Wikimedia-Projekten üblich ist) ist im wissenschaftlichen Bereich allerdings nicht verbreitet. Wenn diese berücksichtigt wird, dann maximal über Kontaktformulare oder Ähnliches. Ein Beispiel liefert das Wörterbuchnetz des Kompetenzzentrums – Trier Center for Digital Humanities der Universität Trier:

„Falls Sie einen Erfassungsfehler entdecken, dann schreiben Sie uns bitte unter Angabe der Wörterbuchsigle und der betreffenden Kontextstelle." (FAQ Wörterbuchnetz, <https://woerterbuchnetz.de/>)

Die Nachteile eines solchen Ansatzes sind ein gewisser zusätzlicher Arbeitsaufwand für Nutzende und vor allem, dass die zeitliche Dauer bis zum Erscheinen der Änderung von Außenstehenden nicht eingeschätzt werden kann,[3] was gerade im Fall von Zitationen in Publikationen mit festen Abgabefristen problematisch ist.[4]

Um diese Probleme zu lösen, ist eine niedrigschwellige und direkte Möglichkeit zur Eingabe von Änderungen nötig, die einem bestimmten Muster entsprechen.[5] Diese können bei Bedarf bzw. in bestimmten Spezialfällen durch einen Moderationsmechanismus ergänzt werden, wobei dieser zumindest den Vorteil der unmittelbaren Einpflegung der Änderung zunichtemacht. Für eine solche Fehlerkorrektur ist auch die in Kapitel 2.3. angesprochene Einbindung des originalen Quellenmaterials entscheidend, um Nutzenden einfach und intuitiv den Abgleich zwischen der gedruckten Passage und der digitalen Repräsentation zu ermöglichen. Des Weiteren sind auch andere Anwendungsfälle möglich, wie die Verknüpfung mit externen Ressourcen, in Fällen in denen dies nicht automatisiert möglich ist.

Um eine Einbindung von externen Personen ohne übermäßigen Aufwand möglich zu machen, kann es sinnvoll sein, bereits vorhandene interne Tools für z.B. Mitarbeitende in einem Projekt auch für Außenstehende (zumindest mit eingeschränkter Funktion bzw. Komplexität) verwendbar zu machen. Voraussetzung hierfür ist die konsequente Nutzung von Webtools und eine gewisse Intuitivität der Oberfläche sowie eine ausführliche Dokumentation.

Ein weiterer spezifischer Vorteil bei der digitalen Darstellung der Informationen aus der Quelle ist die zusammenführende Nutzung der Daten, nicht nur, um Zusammenhänge wie in Kapitel 2.2. beschrieben leichter aufzufinden, sondern auch um die Daten zu aggregieren und beispielsweise statistische Auswertungen des vollständigen Werks zu erstellen. Dies ermöglicht Visualisierungen, in denen die Quelle aus den verschiedensten Gesichtspunkten beleuchtet werden kann (cf. Kapitel 4.4), wobei sich die dafür notwendigen Daten direkt aus den Anforderungen aus Kapitel 2.2 ergeben.

3. Modellierung der Wörterbuchinhalte

Aus den im vorherigen Kapitel aufgestellten Anforderungen an eine Online-Ressource lassen sich wiederum bestimmte Bedingungen für die digitale Darstellung des Quellenmaterials ableiten. Vor allem die in den Kapitel 2.2 und 2.3

[3] Intern kann der Arbeitsablauf auch dadurch verzögert werden, dass mehrere Instanzen involviert sind, beispielsweise wenn die Änderungswünsche erst von wissenschaftlichem Personal verifiziert und dann von technischem Personal eingepflegt werden müssen.

[4] Ob umgekehrt die Möglichkeit des selbständigen Einspeisens von Änderungen zu Manipulationen führt, muss in der Praxis beobachtet und analysiert werden.

[5] Hiermit sind vor allem Änderungen an der Textbasis (Einfügungen, Löschungen und Ersetzungen) sowie bestimmte regelbasierte Eingriffe in den Prozessablauf (beispielsweise die Auflösung einer abkürzenden Schreibweise) gemeint (vgl. Kapitel 3.5).

genannten Funktionen erfordern eine weitreichende Extraktion der in den Wörterbuchartikeln enthaltenen Information. Dieses Kapitel analysiert diese Bedingungen und stellt exemplarisch dar, wie die linguistischen Daten in einer relationalen Datenbank (also in tabellarischer Form) dargestellt werden können.

Folgende grundsätzlichen Bedingungen können aufgestellt werden:

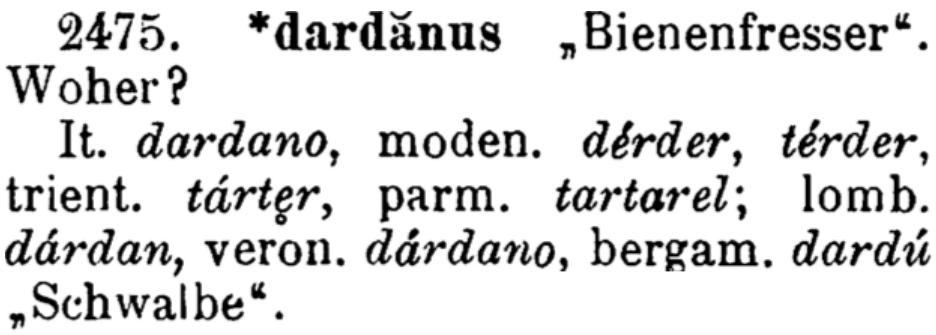
1. Für jeden Artikel (und für weitere textuelle Bestandteile der Quelle) müssen eine oder mehrere Repräsentationen erstellt werden können, die zumindest die quellentreue Transkription beinhalten.
2. Die einzelnen Grundbestandteile der Artikel, z.B. sprachliche Formen, Bedeutungen, bibliographische Angaben etc. (cf. hierzu auch Renders 2011, 118–121), müssen in geeigneter Form abgebildet werden.
3. Die Relationen zwischen diesen Bestandteilen (z.B. welche Form welche Bedeutung hat oder welche Form von welcher Form abstammt) müssen explizit dargestellt werden.
4. Änderungen bzw. Versionen müssen explizit abgespeichert werden können.
5. Datenstrukturen, die bestimmte Bestandteile mit externen Entsprechungen verknüpfen, müssen vorhanden sein.

Besonders die dritte Bedingung hat einige nicht-triviale Konsequenzen, weswegen sie im Weiteren gesondert betrachtet wird, bevor eine Möglichkeit für ein grundlegendes Datenmodell beschrieben wird (3.2 – 3.4). Dieses erfüllt die Bedingungen 1-3. Die letzten beiden Abschnitte gehen einzeln auf Bedingung 4 (3.5) und Bedingung 5 ein (3.6).

3.1. Explizite Darstellung von Wörterbuchrelationen

Die Information innerhalb von Wörterbüchern kann explizit oder implizit kodiert sein. Mit *impliziten Informationen* sind hierbei solche gemeint, die aus allgemeinen oder quellenspezifischen Konventionen hergeleitet werden können, aber nicht direkt im Text abgebildet sind. Ein einfaches Beispiel besteht in der Zuordnung von sprachlichen Formen zu Bedeutungen. Abb. 1 zeigt einen Teil eines Wörterbuchartikels des REW, der dies illustriert. Der Eintrag enthält (das Lemma eingeschlossen) neun verschiedene Formen, eine explizite Bedeutung wird aber nur bei zwei davon angegeben. Die Zuordnung zu den anderen erfolgt über die Konventionen der Quelle, die im Vorwort angegeben sind:

> […] die romanische Bedeutung wird nur dann gegeben, wenn sie von der des Stichwortes abweicht. […] Bei den Ableitungen und Zusammensetzungen gilt eine Bedeutung für sämtliche ihr vorangehenden Formen. (Meyer-Lübke 1935, XI)



2475. *dardănus „Bienenfresser“. Woher?

It. *dardano*, moden. *dérder*, *térder*, trient. *tártęr*, parm. *tartarel*; lomb. *dárdan*, veron. *dárdano*, bergam. *dardú* „Schwalbe“.

1 | Ausschnitt REW Lemma 2475

Für eine technische Verarbeitung der Daten muss diese Information also zuerst inferiert und dann in strukturierter Form abgelegt werden. Tab. 1 illustriert dies in tabellarischer Darstellung.[6]

Sprachabkürzung	Form	Bedeutung
lat.	dardănus	Bienenfresser
it.	dardano	Bienenfresser
moden.	dérder	Bienenfresser
moden.	térder	Bienenfresser
trient.	tártęr	Bienenfresser
parm.	tartarel	Bienenfresser
lomb.	dárdan	Schwalbe
veron.	dárdano	Schwalbe
bergam.	dardú	Schwalbe

Tab. 1 | Explizite Darstellung der Bedeutungszuordnungen, die im Text von Abb. 3 enthalten sind

3.2. Grundlegende Datenmodellierung

Die entscheidende Frage ist nun, in welchem Format diese Informationen abgelegt werden. Ein weit verbreiteter Ansatz in der digitalen Sprachwissenschaft und den Digital Humanities im Allgemeinen ist die Repräsentation als annotierter Text. Dabei kommt meistens XML zum Einsatz, entweder nach einem individuellen Schema (cf. z.B. Renders 2011) oder mit Hilfe des Standardformats TEI[7] (cf. z.B. Tasovac 2020). Das zentrale Element ist hierbei der (eventuell angepasste und erweiterte) Text eines Wörterbuchartikels. Bei den genannten impliziten Daten stößt dieser Ansatz allerdings an seine Grenzen, da diese Information nicht oder nur teilweise im Ursprungstext enthalten ist. Im besonderen Maße ist dies beispielsweise bei etymologischen Relationen zwischen verschiedenen sprachlichen Formen der Fall, die überhaupt nicht explizit als solche enthalten sind, sondern nur aus der Anordnung der Formen im Artikel erschlossen werden können. Hier soll ein alternativer, datenorientierter Ansatz vorgeschlagen werden, der explizite, tabellarische Daten vorsieht, wie sie beispielsweise in Tab. 1 illustriert

[6] Die Zuordnung der Bedeutungen ist bei weitem nicht die einzige Information, die inferiert werden muss, wichtig sind im vorliegenden Beispiel auch die Herkunftsrelation zwischen der lateinischen und den romanischen Formen. Die Zuordnung der lateinischen Sprache zum Lemma wird ebenfalls aus einer Konvention hergeleitet.

[7] Das Format wird von der gleichlautenden *T*ext *E*ncoding *I*nitiative entwickelt (cf. TEI 1994).

werden.[8] Gerade für artikelübergreifende, analytische Abfragen ist dies hilfreich, da sie ohne Durchsuchen von potentiell allen Artikeltexten durchgeführt werden können.

Den Kern stellt dabei eine grundlegende Einteilung in zwei Klassen von Daten dar: *Lexikalische Daten* und *einordnende Daten*. Dies liegt in der Natur eines Wörterbuchs begründet, dass als semi-strukturierter Text aufgefasst werden kann. Weite Teile folgen einer fixen Struktur, wie die Listen mit romanischen Formen im REW, die aber immer wieder von diskursiven Abschnitten unterbrochen werden. Dieses Modell versucht diesen Voraussetzungen gerecht zu werden. Mit *lexikalischen Daten* wird die abstrakte sprachwissenschaftliche Information bezeichnet, die aus dem Wörterbuchartikel extrahiert wird. Diese Information ist vor allem für die technische Verarbeitung von Suchanfragen und Ähnlichem nötig, bleibt aber den Nutzenden weitgehend verborgen. Der Begriff *einordnende Daten* bezeichnet hier Informationen, die für die Einbettung der strukturierten lexikalischen Daten in den Kontext der Wörterbuchartikel (also beispielsweise die Anordnung und Reihenfolge der Formen im Text) nötig sind, und diskursive Elemente, die nicht oder nur teilweise ausgewertet werden können. Sie sind rein auf den Artikel bezogen und dienen zusammen mit den lexikalischen Daten dazu, den vollständigen Artikelinhalt zu rekonstruieren. Ein großer Vorteil einer solchen hybriden Modellierung ist, dass die strukturierten Daten sehr effizient für technische Anwendungen verwendet werden können, da unabhängig vom Zugriffsweg (beispielsweise semasiologisch vs. onomasiologisch) ein direkter Zugriff auf die entsprechenden Daten möglich ist, während ein annotationsbasiertes Modell weiterhin in der durch die Lemmatisierung des Quellenmaterials vorgegebenen Perspektive verbleibt, sodass Anfragen, die nicht dieser Perspektive entsprechen, zumindest deutlich aufwendiger sind. Das gilt sowohl für interne Funktionen eines Webangebots als auch für die maschinelle Datenverarbeitung von externer Seite, wenn über eine technische Schnittstelle (cf. Kapitel 5) auf die Daten zugegriffen wird. Gleichzeitig bleibt die Darstellung des zugrundeliegenden Artikels uneingeschränkt möglich.

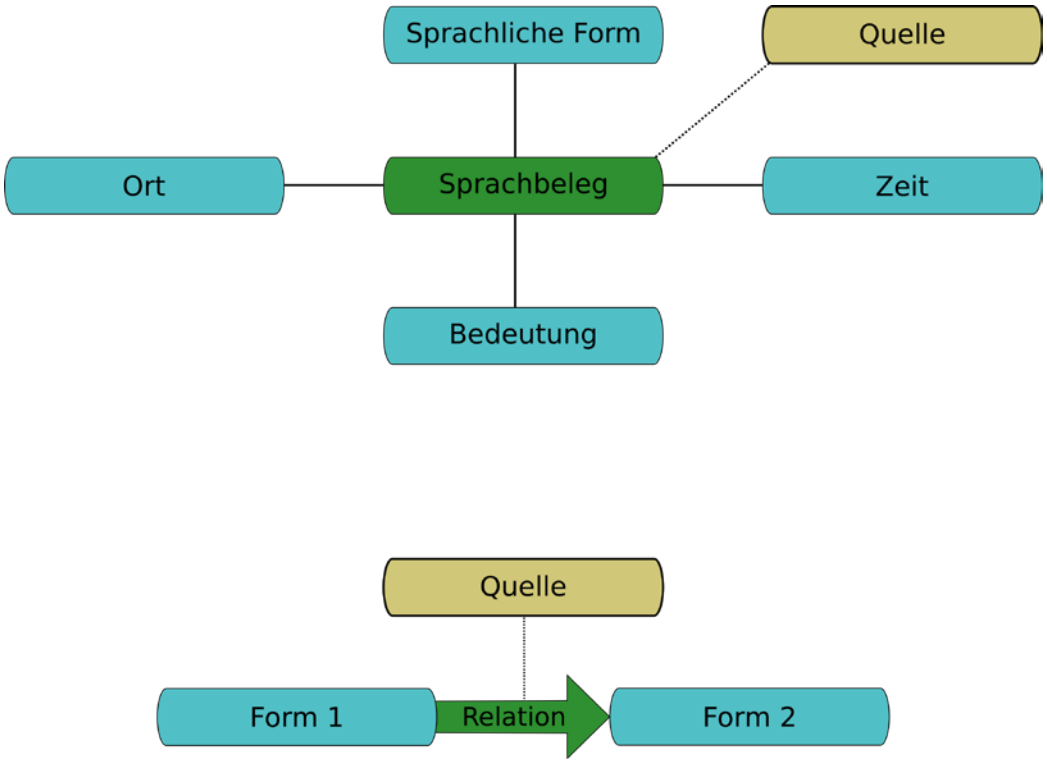
3.3. Lexikalische Daten

Die strukturierten lexikalischen Daten bestehen in diesem Modell aus sogenannten *Sprachbelegen* und Relationen zwischen sprachlichen Formen. Ein *Sprachbeleg* ordnet eine Form ein, indem er sie mit einer Bedeutung und (falls vorhanden) einem Ort sowie einem Zeitpunkt bzw. -intervall in Verbindung setzt. Relationen beschreiben verschiedene Formen von Verbindungen zwischen zwei sprachlichen Formen.[9] Im Fall eines etymologischen Wörterbuchs sind das vor allem Herkunftsrelationen wie Vorgänger oder Entlehnungen, es sind aber auch andere möglich, um beispielsweise verschiedene Flexionsformen des gleichen Lexems miteinander zu verknüpfen. Alle Datensätze werden als Teil eines globalen Informationspools verstanden. Die Beziehung zum Artikel wird als Quelle aufgefasst, aus der die

[8] Die Annotation von Text wird allerdings „im Kleinen“ verwendet, indem in Textbestandteilen, die nicht vollständig strukturiert erfasst werden können, bestimmte Entitäten annotiert werden (cf. Kapitel 3.4).

[9] Diese Art der Modellierung ist nicht direkt mit dem sogenannten *lemon*-Modell (cf. McCrae et al. 2012, 8–15) kompatibel, das für die Darstellung von lexikalischen Daten für das Semantic Web erstellt wurde und oft als Standard verwendet wird. Zacherl (in Vorb. b) führt die Gründe dafür auf.

jeweilige Information stammt. Abb. 2 veranschaulicht die beiden Informationstypen. Eine detailliertere Analyse der grundsätzlichen Darstellung von linguistischen Rohdaten und der Vorschlag einer konkreten Methodik zur Überführung von Wörterbuchtexten in relationale Daten, die nach einem solchen Datenmodell strukturiert sind, findet sich in Zacherl (in Vorb. b).



2 | Schematische Darstellung der beiden Grundelemente für das lexikalische Datenmodell

3.4. Einordnende Daten

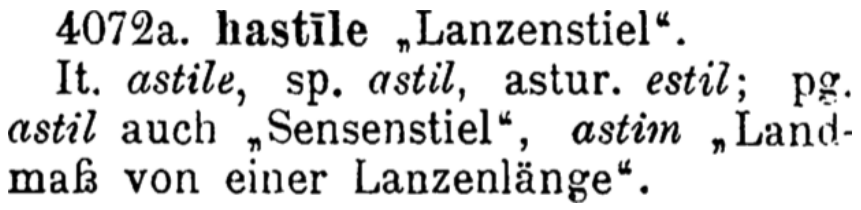
Die Grobstruktur eines Wörterbuchartikels wird über Beleglisten dargestellt. Jeder Artikel besteht aus mindestens zwei dieser Listen; eine repräsentiert die Kopfzeile mit der Angabe der Lemmata, alle weiteren beschreiben den eigentlichen Artikelinhalt. Die Listen werden grundsätzlich aufsteigend nummeriert, eine Ausnahme bilden nur Entlehnungen, die sich immer auf konkrete Sprachbelege beziehen. Diese stehen außerhalb der Nummerierung und werden den entsprechenden Belegen zugeordnet.

Die Sprachbelege selbst müssen für eine spätere Darstellung ebenfalls entsprechend ihrer Position in der Liste nummeriert werden. Grundsätzlich wäre eine einfache aufsteigende Nummerierung ausreichend, um die Belege entsprechend ihrer ursprünglichen Anordnung wiederzugeben. Um die logische Struktur besser abzubilden und auch bis zu einem gewissen Maß abweichende Darstellungen der Wörterbucheinträge zu ermöglichen (cf. Kapitel 4.1), bietet sich allerdings die Verwendung von mehreren hierarchisch funktionierenden Indizes an:

- Index 1 (Subliste): Listen von Belegen werden häufig durch Semikola oder andere Trennzeichen in unterschiedliche Gruppen aufgeteilt. Die dort

angegebenen Formen haben beispielsweise ähnliche Bedeutungen, die gleiche Wortart etc. Der erste Index nummeriert diese Einteilung.

- Index 2 (Position innerhalb der Subliste): Dieser Index wird über die Sprache[10] definiert. Wenn für eine Sprache mehrere Formen bzw. Bedeutungen gegeben werden, werden diese auf dieser Ebene noch zusammengefasst. So besteht das Beispiel aus Abb. 3 aus zwei verschiedenen Sublisten, wobei die erste aus drei Elementen und die zweite aus nur einem Element besteht. Kurzschreibweisen werden behandelt, als wären sie ausgeschrieben (d.h. Angaben wie pg., sp. *astil* werden wie pg. *astil*, sp. *astil* als zwei Elemente indiziert).
- Index 3 (Varianten innerhalb einer Sprache): Falls mehrere Formen bzw. Varianten einer Form angegeben werden, werden sie hier unterschieden (beispielsweise die portugiesischen Formen *astil* und *astim* in Abb. 3). Dabei wird wiederum nicht zwischen abkürzenden Schreibweisen und expliziter Trennung durch Kommata im Originaltext unterschieden (also pg. *(f)ata* und pg. *ata*, *fata* würden beispielsweise identisch nummeriert).
- Index 4 (Bedeutung): Falls eine Form mehrere Bedeutungen hat, werden diese durch den letzten Index nummeriert. Dabei spielt es keine Rolle, ob die Bedeutungen in der Quelle explizit angegeben sind oder inferiert werden müssen.


4072a. **hastīle** „Lanzenstiel".
It. *astile*, sp. *astil*, astur. *estil*; pg. *astil* auch „Sensenstiel", *astim* „Landmaß von einer Lanzenlänge".

3 | REW Lemma 4072a

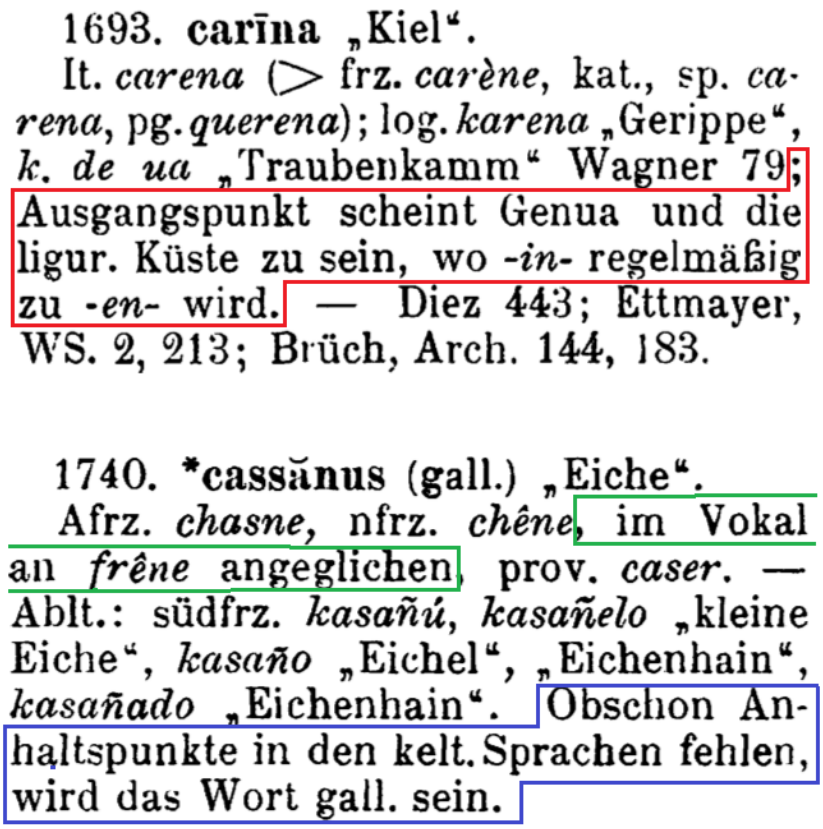
Tab. 2 illustriert am bereits erwähnten Beispiel die verschiedenen Indizes. Zusätzlich zur Indizierung werden den Sprachbelegen außerdem alle Literaturangaben zugeordnet, die sich auf diese beziehen.

10 Hier und im Folgenden wird der Begriff *Sprache* stellvertretend für eine Sprache oder einen Dialekt verwendet. Das dient nur der Vereinfachung, da beides im Normalfall durch entsprechende Abkürzungen angegeben und syntaktisch nicht unterschieden wird.

Sprachabkürzung	*Form*	*Bedeutung*	*Subliste*	*Position*	*Variante*	*Bedeutung*
it.	astile	Lanzenstiehl	0	0	0	0
sp.	astil	Lanzenstiehl	0	1	0	0
astur.	estil	Lanzenstiehl	0	2	0	0
pg.	astil	Lanzenstiehl	1	0	0	0
pg.	astil	Sensenstiehl	1	0	0	1
pg.	astim	Landmaß von einer Lanzenlänge	1	0	1	0

Tab. 2 | Indizierung der sprachlichen Formen aus dem Eintrag aus Abb. 3

Diskursive Elemente sind Texteinschübe, die keiner festen Struktur unterliegen. Die häufigsten Vertreter sind Einschübe, die sich auf einen einzelnen Beleg beziehen, Einschübe, die sich auf eine komplette Belegliste beziehen und vollständige Sätze, die strukturell in keinem Bezug zu einer Belegliste stehen (cf. Abb. 4). Die ersten beiden Typen werden als Kommentar zum jeweiligen Beleg bzw. zur jeweiligen Liste aufgefasst und diesen zugeordnet.[11] Der letzte Typ wird als „entartete" Belegliste modelliert, d.h. als eine Belegliste, die nur aus einem Kommentar besteht, aber keine Belege enthält. Somit kann jeder Artikel als eine geordnete Menge von Beleglisten aufgefasst werden.[12]



1693. **carīna** „Kiel".
It. *carena* (> frz. *carène*, kat., sp. *carena*, pg. *querena*); log. *karena* „Gerippe", *k. de ua* „Traubenkamm" Wagner 79; Ausgangspunkt scheint Genua und die ligur. Küste zu sein, wo *-in-* regelmäßig zu *-en-* wird. — Diez 443; Ettmayer, WS. 2, 213; Brüch, Arch. 144, 183.

1740. ***cassănus** (gall.) „Eiche".
Afrz. *chasne*, nfrz. *chêne*, im Vokal an *frêne* angeglichen, prov. *caser*. — Ablt.: südfrz. *kasañú*, *kasañelo* „kleine Eiche", *kasaño* „Eichel", „Eichenhain", *kasañado* „Eichenhain". Obschon Anhaltspunkte in den kelt. Sprachen fehlen, wird das Wort gall. sein.

4 | Verschiedene Arten von Texteinschüben im REW (rot: auf alle vorangegangenen Formen bezogen, grün: auf den direkten Vorgänger bezogen, blau: eigenständiger Satz)

[11] Alle anderen selteneren Formen von Texteinschüben können ebenfalls bestimmten Entitäten zugeordnet werden und werden somit analog behandelt.

[12] Ein Eintrag besteht zusätzlich zu den Beleglisten auch aus dem unbearbeiteten Originaltext, so dass dieser alternativ zu einer angereicherten Version ebenfalls dargestellt werden kann. Dies ist prinzipiell redundant, weil der Text auch rekonstruiert werden könnte, vereinfacht die Prozesse aber deutlich.

Diskursive Elemente werden nicht systematisch ausgewertet,[13] die darin enthaltenen relevanten Entitäten, die über die Struktur erkennbar sind, werden aber erkannt und mit XML annotiert, sodass sie an der Oberfläche entsprechend dargestellt werden können.

3.5. Korrekturen und Ausnahmen

Alle Änderungen werden explizit als Datensätze abgelegt. Es können zwei grundlegende Typen unterschieden werden: Fehler an der Textbasis und *Ausnahmen*, die in den Ablauf des Transformationsprozesses eingreifen, der aus dem ursprünglichen Text die entsprechenden Daten erzeugt. Beim ersten Typus können zusätzlich Korrekturen von Digitalisierungsfehlern, die beispielsweise bei einer automatischen Texterkennung oder auch bei manueller Transkription entstanden sind, unterschieden werden von Fehlern in der Quelle an sich. Letztere sollten nur in sehr offensichtlichen Fällen (fehlende schließende Klammern, offensichtliche Tipp- oder Druckfehler, etc.) korrigiert werden. Trotzdem kann eine solche Korrektur nicht immer völlig ohne Interpretation stattfinden, wenn beispielsweise unklar ist, an welcher Stelle eine schließende Klammer vergessen wurde. An der Oberfläche der Web-Anwendung können solche editorischen Eingriffe dann beispielsweise gesondert markiert werden. Unabhängig vom Typ werden sämtliche Änderungen mit einem Zeitstempel und dem entsprechenden User-Login markiert, so dass die Änderungshistorie jederzeit nachvollziehbar bleibt.[14]

Je nach Präferenz können die Korrekturen nach Erstellung sofort umgesetzt werden, wodurch eine neue Version des/der betroffenen Artikel erstellt wird oder es kann für alle oder bestimmte Änderungen (je nach Login und Art der Änderung) ein Moderationsprozess vorgeschaltet werden. Die Versionen der Artikel werden dabei explizit in der Datenbank abgelegt, d.h. nach jeder Änderung wird dort eine Kopie der Artikelrepräsentation erstellt, die diese Änderung enthält. Grundsätzlich wäre dies nicht nötig, da aufgrund der zeitlichen Markierung aus der Textbasis und den jeweiligen Korrekturen die zum damaligen Zeitpunkt gültige Version rekonstruiert werden könnte. Dahinter steckt aber die wenig realistische Annahme, dass sich der Programmcode für den Transformationsprozess zu keinem Zeitpunkt ändert. Ein weiterer Vorteil der expliziten Darstellung der Artikelversionen ist, dass deren Erstellung weniger rechenintensiv ist.

3.6. Datenanreicherung

Bei der Betrachtung der Möglichkeiten für eine externe Vernetzung ist die offensichtlichste die Verknüpfung der Literaturangaben, die im Text genannt werden, mit potentiell vorhandenen digitalen Versionen der jeweiligen Quelle, die ohne Zugangsbeschränkung verfügbar sind. Für Quellen, die über Seitenzahlen referenziert werden, ist dabei eine einfache Abbildung von Quellenabkürzung und eventuell vorhandener Bandnummer auf entsprechende Links nötig. Etwas

[13] Dies wäre nur mit sehr weit fortgeschrittenen Methoden des *Natural Language Processings* möglich.

[14] Die exakte Darstellung beider Arten von Änderungen hängt stark von der algorithmischen Umsetzung des Transformationsprozesses ab und wird deshalb an dieser Stelle nicht genauer spezifiziert. Entsprechende Beispiele finden sich in Zacherl (in Vorb. b).

komplizierter ist der Fall bei Wörterbüchern, die über Lemmanummern referenziert werden. In einem solchen Fall ist zusätzlich eine Abbildung von Seitenzahlen auf die jeweils dort gelisteten Lemmata nötig.

In Bezug auf die Bedeutungen bietet sich eine Verknüpfung mit einem kontrollierten Vokabular oder einer Ontologie an. Das hat den Vorteil, dass einerseits intern doppelt vorhandene Konzepte, die durch unterschiedliche sprachliche Ausdrücke beschrieben werden, verknüpft werden können, andererseits hat es einen Normierungseffekt, der gerade beim Zusammenführen von Daten aus verschiedenen (potentiell in unterschiedlichen Sprachen verfassten) Werken wichtig ist. Ein Beispiel für eine solches externes Portal ist die Wissensdatenbank Wikidata (cf. Wikidata 2012). Durch die Zuordnung zum entsprechenden Konzept in dieser Datenbank können z.B. die im REW vorkommenden synonymen Bedeutungsangaben „Abfall“, „Unrat“ und „Müll“ zusammengefasst werden, was einen Abgleich der entsprechend zugeordneten sprachlichen Formen deutlich erleichtert.

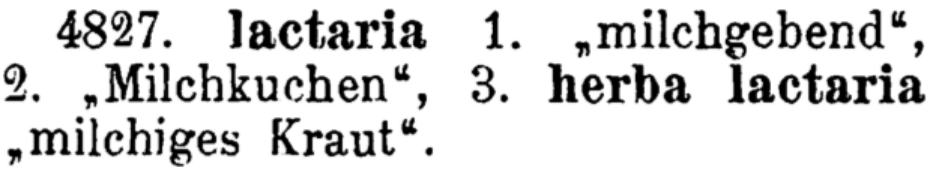
Für eine geographische Visualisierung ist außerdem eine Zuordnung von Sprachen bzw. Dialekten zu Verbreitungsgebieten notwendig. Ob eine solche Visualisierung für eine gegebene Quelle in Frage kommt, ist sehr spezifisch und ergibt tendenziell nur bei Werken Sinn, die eine große Anzahl kleinräumiger, dialektaler Formen enthalten.

4. Gestaltung der Oberfläche

Im Folgenden wird exemplarisch eine mögliche Umsetzung der Konzepte aus Kapitel 1 beschrieben und der daraus entstandene Entwurf eines Webauftritts vorgestellt.

4.1. Darstellung der Artikel

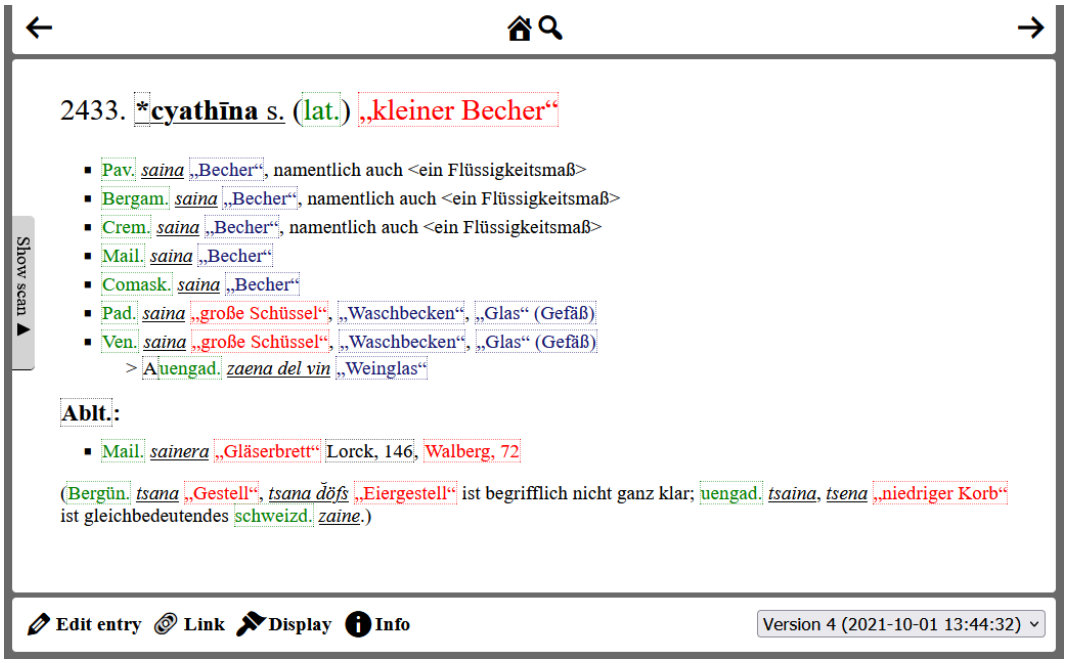
Der Kern eines jeden Online-Wörterbuchs ist die Darstellung der einzelnen Artikel. Hierbei stellt sich die grundlegende Frage, ob die Artikel einzeln (cf. z.B. TLIO) oder im Kontext der vorangegangenen und nachfolgenden Artikel dargestellt werden (cf. z.B. Wörterbuchnetz). Eine Entscheidung hängt u.a. davon ab, wie häufig einzelne Einträge Bezug auf ihre direkte Umgebung nehmen bzw. ob ähnliche Lemmata (Varianten, Komposita, etc.) in einem Artikel zusammengefasst werden oder nicht. Das REW referenziert beispielsweise grundsätzlich über Lemmanummern und fasst verhältnismäßig viele Lemmata in einem Artikel zusammen (cf. Abb. 5). Eine einzelne Darstellung der Artikel, die oftmals übersichtlicher ist, bietet sich also an.


4827. **lactaria** 1. „milchgebend“, 2. „Milchkuchen“, 3. **herba lactaria** „milchiges Kraut“.

5 | Kopfzeile REW Lemma 4827

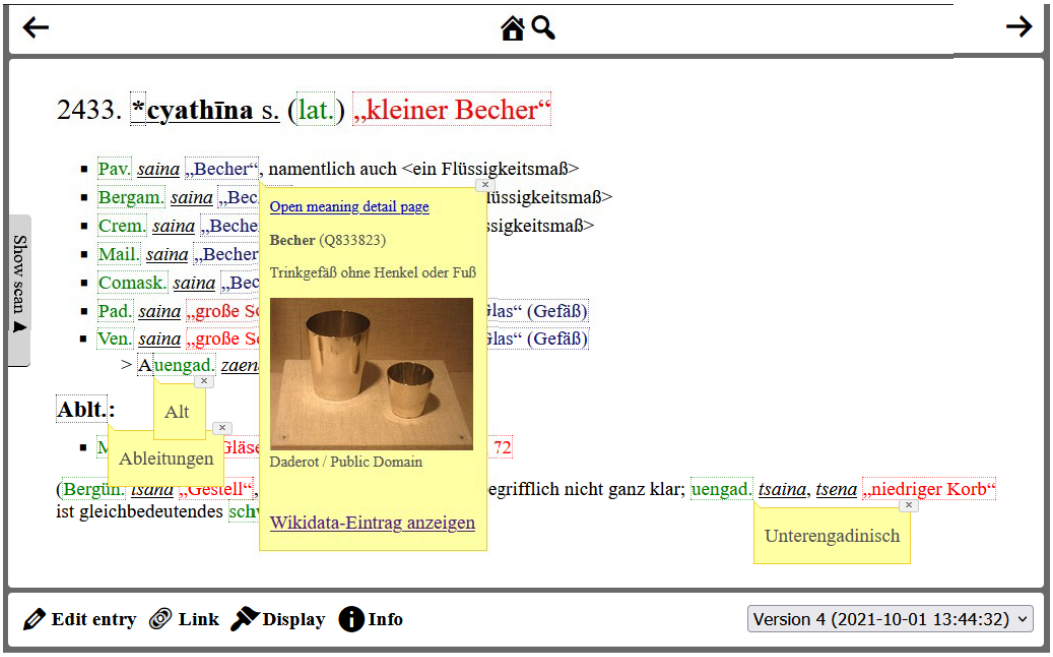
Ein großer Unterschied der standardmäßigen Darstellung (cf. Abb. 6) eines Artikels ist, dass die Beleglisten als Aufzählungen (also vertikal) angezeigt werden. Entlehnungen, die im Originaltext in Klammern hinter dem jeweiligen Beleg aufgeführt

werden, stellen eingerückte Unterlisten dar. Außerdem werden verschiedene Bestandteile eines Artikels durch verschiedene Absätze stärker visuell voneinander abgegrenzt. Einleitende Angaben, die eine Belegliste genauer spezifizieren (Ableitungen, Zusammensetzungen, etc.) werden prominenter dargestellt. All dies ist Konsequenz des bereits erwähnten Wegfalls der Platzbeschränkungen eines gedruckten Werks und hat die Absicht die Übersichtlichkeit und Lesbarkeit zu erhöhen. Diskursive Elemente behalten ihre grundsätzliche Darstellung.



6 | Standardansicht eines Artikels

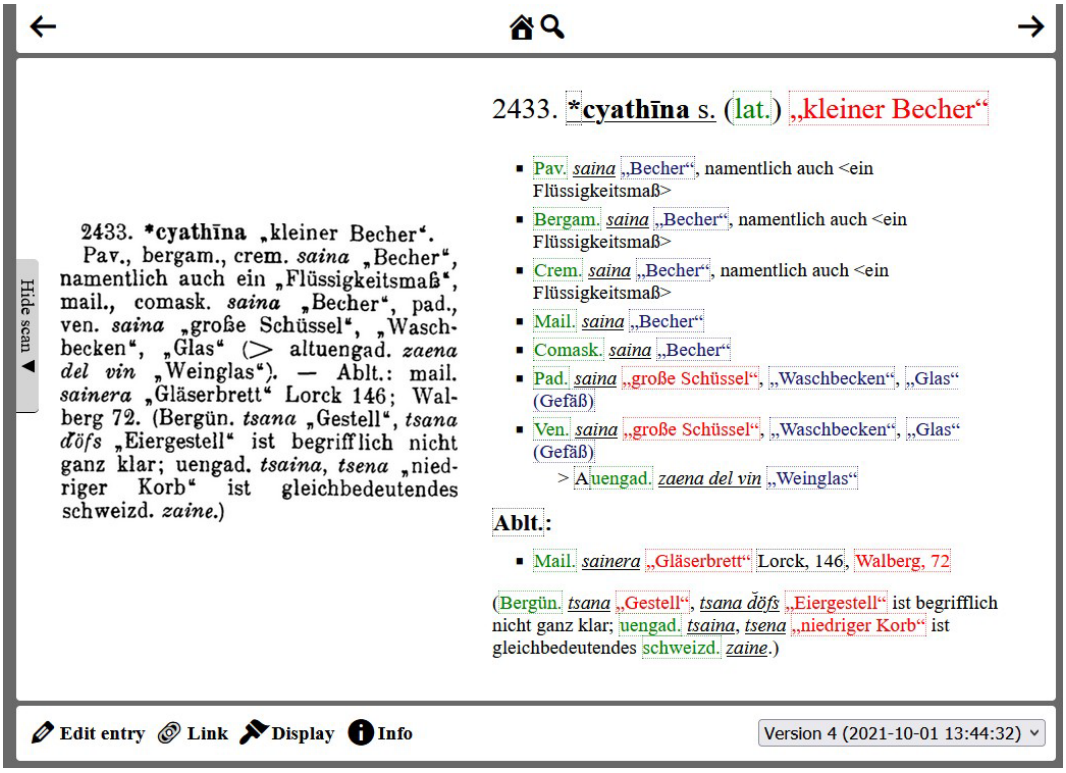
Die einzelnen Bestandteile der Artikel werden unterschiedlich hervorgehoben. Die grundsätzlichen Konventionen aus der Quelle (Lemmata sind fett markiert, andere sprachliche Formen kursiv bzw. in Kapitälchen und Bedeutungen in Anführungszeichen) werden so übernommen. Alle interaktiven Elemente, die beim Überfahren mit der Maus und Klick (bzw. Antippen auf mobilen Endgeräten) zusätzliche Informationen anzeigen, werden durch einen gestrichelten Rahmen markiert. Vor allem trifft das auf alle Formen von Abkürzungen zu, die so aufgelöst werden können. Bedeutungen und bibliographische Referenzen werden außerdem unterschiedlich formatiert, je nachdem ob sie mit externen Daten verknüpft sind oder nicht. In beiden Fällen wird (falls vorhanden) eine Verlinkung auf die externe Entsprechung angegeben, im Fall der Bedeutung werden zusätzlich noch Basisinformationen und eventuell ein Bild zur Illustration des Wikidata-Konzepts nachgeladen. Abb. 7 zeigt einige Beispiele.



7| Auflösung von Abkürzungen und Anzeige zugeordneter Konzepte aus Wikidata

Für beliebige Artikel ist jederzeit eine Anzeige des Ausschnitts aus dem Scan des Wörterbuchs über einen Button auf der linken Seite möglich.[15] Dieser dient einerseits zum Nachweis der Quelle, bietet aber auch eine unaufwendige Möglichkeit aufgefundene (potentielle) Fehler, die im Verlauf der Verarbeitung entstanden sind, im Abgleich mit dem Quellenmaterial als solche zu verifizieren und im Anschluss entsprechend zu korrigieren (vgl. hierzu Kapitel 4.3). Abb. 8 zeigt die gemeinsame Ansicht von Scan und Artikel.

[15] Zusätzlich zu den in Kapitel 3 behandelten Daten muss dafür das Bildmaterial zur Verfügung stehen, sowie eine Zuordnung von Pixelkoordinaten zu den jeweiligen Artikeltexten. Texterkennungssysteme können zum Teil beispielsweise das HOCR-Format erzeugen, in dem eine Zuordnung von Zeilen zu Pixelkoordinaten vorhanden ist, die für diese Zwecke genutzt werden kann.



8 | Artikeldarstellung mit Ursprungstext

Die weiteren Bedienelemente sind in zwei Gruppen eingeteilt. In der oberen Hälfte finden sich Navigations- und Suchfunktionen. Es kann zum vorherigen und nächsten Artikel gewechselt werden, sowie auf die Startseite, die verschiedene Einstiegsmöglichkeiten sowie ein vollständiges Inhaltsverzeichnis umfasst, das auch das Vorwort sowie die Abkürzungsverzeichnisse enthält.[16] Das Lupensymbol erlaubt den Zugriff auf verschiedene Suchfunktionalitäten wie eine uneingeschränkte Volltextsuche und spezialisierte Suchen nach bestimmten Entitäten.

Im unteren Bereich finden sich Interaktionsmöglichkeiten und Dokumentation. Die ersten beiden Bedienelemente dort erlauben das Bearbeiten des Artikels (cf. Kapitel 4.3) und die Verlinkung bzw. Zitation. Bei letzterem können zwei verschiedene URLs[17] erzeugt werden, wobei eine auf die aktuelle Artikelversion verweist, also auf eine statische Darstellung, die für wissenschaftliches Zitieren genutzt werden kann, und die andere auf die jeweils neuste Version des jeweiligen Artikels. Weiterhin ist die Auswahl von unterschiedlichen Darstellungsvarianten möglich, die in verschiedenen Stufen die Ähnlichkeit zum originalen Text steigern, bis hin zur ursprünglichen Spaltendarstellung, wie sie auch der Scan zeigt. Der Info-Button enthält eine detaillierte Dokumentation der verschiedenen Notationskonventionen, was sowohl diejenigen, die aus dem ursprünglichen Wörterbuch übernommen wurden, als auch die spezifischen der digitalen Variante umfasst. Rechts unten kann schließlich zwischen den verschiedenen Artikelversionen

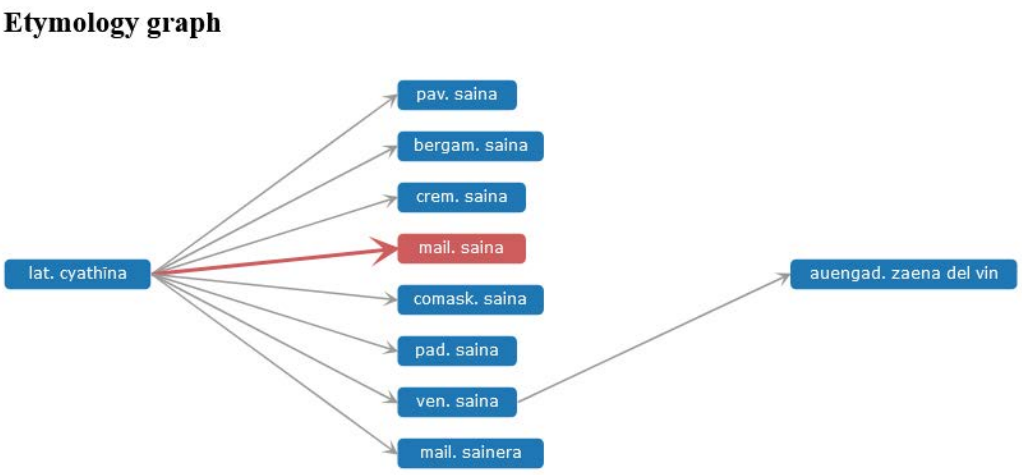
[16] Letztere sind grundsätzlich nicht nötig, da alle Abkürzungen an den Stellen aufgelöst werden, an denen sie vorkommen. Zur Übersicht und um der Struktur des Originalwerks möglichst zu entsprechen, werden sie trotzdem aufgeführt.

[17] Die URLs sollten möglichst auf persistenten Identifikatoren wie DOIs basieren.

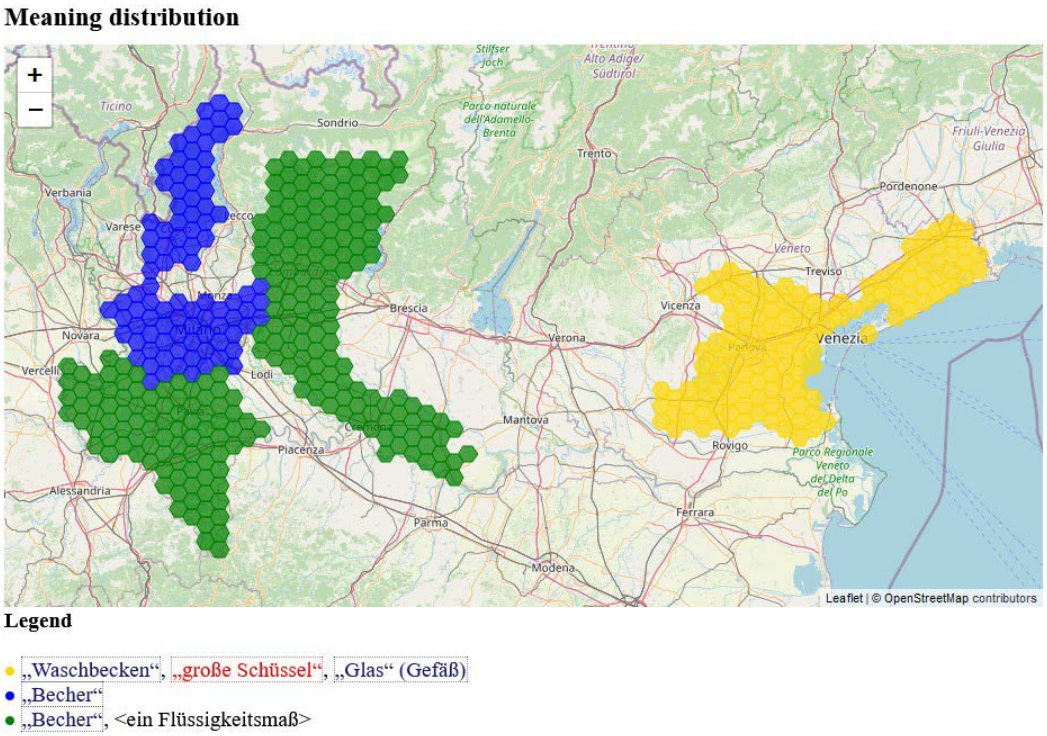
gewechselt werden (falls mehrere vorhanden sind). Wenn eine veraltete Version ausgewählt wird, erscheint im oberen Teil eine entsprechende Warnung, um die versehentliche Nutzung von Informationen, die bereits korrigierte Fehler enthalten, möglichst zu vermeiden.

4.2. Detailseiten für Wörterbuchbestandteile

Für die drei wichtigsten Entitäten im Kontext des Wörterbuches, nämlich sprachliche Formen, Bedeutungen und referenzierte Literatur, werden weiterhin sogenannte *Detailseiten* angeboten, die über die Vorkommen der jeweiligen Entitäten innerhalb der Artikel verlinkt werden und artikelübergreifend Informationen zu diesen aggregieren. So enthält beispielsweise die Detailseite einer Form alle Vorkommen derselben in den verschiedenen Artikeln mit entsprechenden Bedeutungen. Umgekehrt enthält die Detailseite einer Bedeutung eine Liste aller Vorkommen und der entsprechenden sprachlichen Formen. Weiterhin sind verschiedene Visualisierungen möglich. Abb. 9 zeigt eine Visualisierung, in der eine einzelne Form mit Hilfe einer Graphdarstellung in den Kontext der etymologischen Relationen eingebettet wird, während Abb. 10 ein Beispiel für eine geographische Visualisierung zeigt, die die Verteilung der Bedeutungen einer bestimmten Form darstellt.



9 | Visualisierung von Herkunftsrelationen

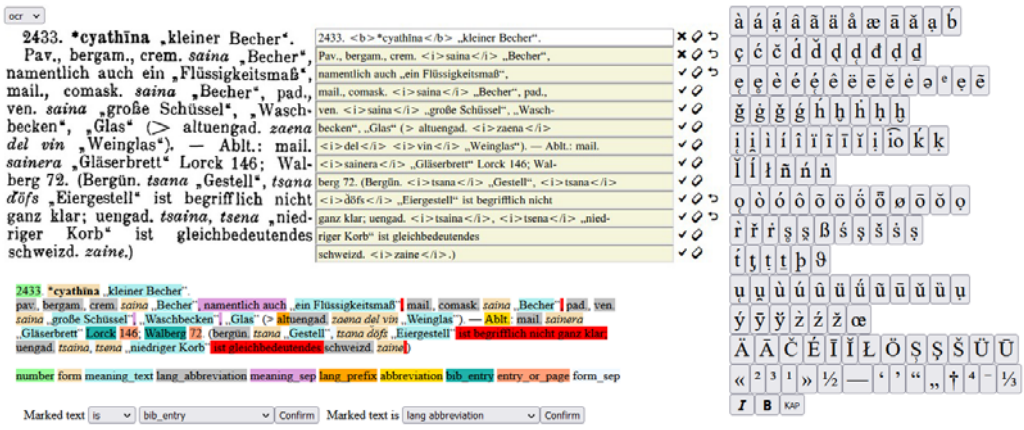


10 | Visualisierung der geographischen Verteilung von Bedeutungen einer Form

Die Detailseiten haben grundsätzlich eigene URLs, die die ID der Entität enthalten, somit können diese auch einzeln verlinkt werden.

3.3. Interaktionsmöglichkeiten

Für Nutzende der Online-Ressource sind zwei Möglichkeiten vorgesehen, wie sie selbst dazu beitragen können, die Datenbasis zu verbessern bzw. zu erweitern. Die wichtigste Möglichkeit ist wohl die zur selbstständigen Korrektur von Fehlern im digitalisierten Quellentext. Abb. 11 zeigt einen Entwurf für eine entsprechende Eingabemaske. Den Kern bildet der Ausschnitt des Scans zum jeweiligen Artikel mit zugehörigen Texteingabefeldern, die den Zeilen im Scan entsprechen und passend zu diesem angeordnet werden. Absätze im Text werden über zwei verschiedene Hintergrundfarben markiert, die sich abwechseln. Im rechten Teil findet sich eine Art virtueller Tastatur, die häufig benötigte Zeichen enthält, die mit vielen Tastatur-layouts nicht direkt eingegeben werden können. Unten findet sich schließlich eine Darstellung des Verarbeitungsergebnisses, in dem die unterschiedlichen erkannten Bestandteile farblich markiert werden. Bei jeder Änderung in einem der Textfelder wird ein Korrektur-Datensatz angelegt und die Darstellung aktualisiert, so dass vor allem die Korrektur von strukturellen Fehlern (z.B. Satzzeichen, Klammern, etc.) besonders deutlich im Ergebnis sichtbar wird.



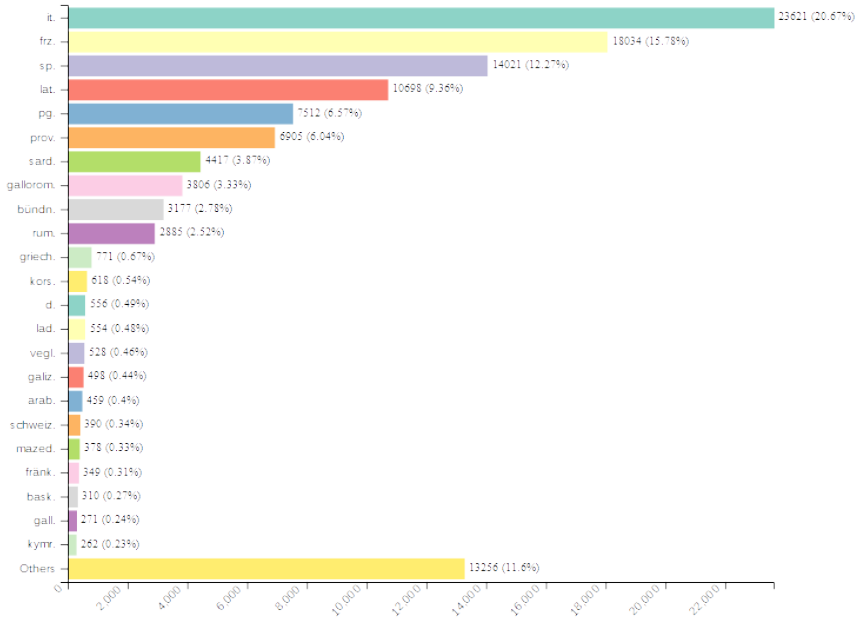
11 | Entwurf einer Oberfläche zur Korrektur

Eine weitere Interaktionsmöglichkeit bietet das Gebiet der Anreicherung und Vernetzung. Die in Kapitel 3.6 genannten Daten können nur zum Teil automatisiert erstellt werden. In den übrigen Fällen, sei es bei der Zuordnung von Bedeutungen zu Wikidata-Einträgen, von Sprachabkürzungen zu Verbreitungsgebieten oder von literarischen Quellen zu entsprechenden Online-Repräsentationen, wird bei einem fehlenden Datum stattdessen ein Oberflächenelement angezeigt, das Nutzenden das Nachtragen ermöglicht. So kann zum Beispiel beim Fehlen einer Wikidata-Verknüpfung eine Abfrage an die Wikidata-Server geschickt werden, die aus der Bedeutungsbeschreibung nach Möglichkeit verschiedene Kandidaten liefert, aus denen der richtige ausgewählt werden kann. Falls keiner der Kandidaten sinnvoll ist, kann auch manuell die ID eines Eintrags angegeben werden, wobei grundsätzlich auch immer die Möglichkeit besteht, bisher nicht vorhandene Konzepte in Wikidata neu zu erstellen.

4.4. Statistische Auswertung

Abb. 12 zeigt exemplarisch eine der statistischen Visualisierungen, die aus den aus dem REW extrahierten Daten erzeugt werden können. In diesem Fall wird der prozentuale Anteil der verschiedenen Sprachen dargestellt. Dafür werden alle Formen verwendet, die strukturiert erfasst wurden, also nicht ausschließlich innerhalb von diskursiven Elementen vorkommen (cf. Kapitel 3.4). Dialekte werden unter den übergeordneten Sprachen zusammengefasst.[18] Da die Darstellung direkt aus dem aktuellen Datenbestand generiert wird, sind die zugrundeliegenden Zahlen nicht statisch, sondern können grundsätzlich bei jeder Änderung oder Korrektur leichten Schwankungen unterliegen.

[18] Diese Einschränkungen dienen der Übersichtlichkeit und sind nicht obligatorisch.



12 | Visualisierung des Anteils der Sprachen bezogen auf alle Formen, die aus dem REW strukturiert erfasst wurden (Stand 31.01.2022)

Weitere statistische Auswertungen, beispielsweise die Häufigkeitsverteilung der literarischen Quellen oder die Aufgliederung der Lemmata nach Sprachen ist in ähnlicher Weise möglich.

5. Ausblick

Dieser Beitrag konzentriert sich auf Möglichkeiten und Methodiken, um die digitalisierten Informationen aus einem Wörterbuch für die direkte menschliche Nutzung zu verwenden. Gerade das sehr kleinteilige Datenmodell, das in Kapitel 3 beschrieben wurde, eignet sich allerdings auch gut für eine maschinelle Nutzung der Daten. So ist sowohl der Zugriff über eine technische Schnittstelle der Online-Ressource als auch die Umwandlung der sprachlichen Kerndaten, d.h. der lexikalischen Daten wie sie in Kapitel 3.3. definiert wurden, in Formate des Semantic Webs wie RDF (cf. RDF 2014) ohne weiteres möglich. Dies eröffnet weitreichende Möglichkeiten wie individuelle Abfragen auf dem Datenbestand oder auch zusätzliche Visualisierungen auf Basis externer Tools.

Auch ein vollständiger Export kann vorgesehen werden. Das bezieht sich nicht nur auf die Kerndaten (wie beispielsweise eine Liste der Lemmata oder die vollständigen Artikeltexte), sondern auch auf sekundäre Daten, wie sie in Kapitel 3.6 für die Anreicherung des Ursprungsmaterials beschrieben werden. Beides kann somit in anderen Projekten wiederverwendet werden.[19] Sinnvollerweise sollten die Exportdaten (in gewissen Zeitschnitten) ebenfalls in passenden Repositorien archiviert werden.

[19] Voraussetzung hierfür ist die Veröffentlichung unter Verwendung einer entsprechend offenen Lizenzierung. Im vorliegenden Beispiel werden alle Daten unter Creative Commons BY-SA zur Verfügung gestellt.

Bibliografie

ISTROX. 2020. „Launch of ISTROX on Zooniverse: Press Release.“ <https://istrox.ling-phil.ox.ac.uk/news/2020/08/04/launch-istrox-zooniverse-press-release> 30.01.2022.

KOMPETENZZENTRUM – „Trier Center for Digital Humanities: Wörterbuchnetz“. <https://woerterbuchnetz.de>. 05.02.2022.

KOMPETENZZENTRUM – „Trier Center for Digital Humanities: „FAQ Wörterbuchnetz.“ <https://woerterbuchnetz.de>. 31.01.2022.

KREFELD, Thomas & Stephan Lücke. 2021. s.v. „Crowdsourcing” *VerbaAlpina-de* 21/2 (Erstellt: 16/1, letzte Änderung: 21/1), Methodologie. <https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D212%26letter%3DC%2312>.

MCCRAE, John et al. 2012. „Interchanging lexical resources on the Semantic Web.“ *Lang Resources & Evaluation* 46, 701–719. <https://doi.org/10.1007/s10579-012-9182-3>.

MEYER-LÜBKE, Wilhelm. 1935. *Romanisches etymologisches Wörterbuch* 3., vollst. neubearb. Aufl. Heidelberg: Winter. <https://nbn-resolving.org/urn:nbn:de:bvb:355-ubr07799-0>.

MÖLLER, Robert & Stephan Elspaß. 2014. „Zur Erhebung und kartographischen Darstellung von Daten zur deutschen Alltagssprache online: Möglichkeiten und Grenzen.“ In *20 Jahre digitale Sprachgeographie*, ed. Tosques, Fabio, 121–131, Berlin: Humboldt-Universität, Institut für Romanistik. <https://www2.hu-berlin.de/vivaldi/tagung/beitraege/pdf/20_jahre_web_version.pdf>.

PRÄTOR, Klaus. 2011. „Zur Zukunft des Zitierens. Identität, Referenz und Granularität digitaler Dokumente.“ *Editio* 25 Heft 2011, 170-183. <https://doi.org/10.1515/9783110239362.170>.

RDF. 2014. „Resource Description Framework.“ <https://www.w3.org/RDF/>. 19.02.2022.

RENDERS, Pascale. 2011. *Modélisation d'un discours étymologique. Prolégomènes à l'informatisation du Französisches Etymologisches Wörterbuch*. <https://orbi.uliege.be/handle/2268/94407>.

REWONLINE = Zacherl, Florian. 2021-. *Digitale Aufbereitung des Romanischen etymologischen Wörterbuches von Wilhelm Meyer-Lübke*. <https://www.rew-online.gwi.uni-muenchen.de>.

TASOVAC, Toma. 2020. *The Historical Dictionary as an Exploratory Tool: A Digital Edition of Vuk Stefanovic Karadzic's Lexicon Serbico-Germanico-Latinum*. <http://hdl.handle.net/2262/92750>.

TEI. 1994. „Text Encoding Initiative.“ <https://tei-c.org/>. 19.02.2022.

TLIO = *Tesoro della Lingua Italiana delle Origini*. 1997. <http://tlio.ovi.cnr.it/TLIO/>. 05.02.2022.

VÀRVARO, Alberto. 2011. „Il DÉRom: un nuovo REW?“ *Revue de linguistique romane* 75, 297-304.

WIKIDATA. 2012. <https://www.wikidata.org>. 19.02.2022.

ZACHERL, Florian. In Vorb. a. *Digitale Tiefenerschließung traditioneller Lexikographie – am Beispiel des Romanischen Etymologischen Wörterbuchs*, univ. Diss. Ludwig-Maximilians-Universität München.

ZACHERL, Florian. In Vorb. b. „Automatisierte Erschließung von strukturierten Daten aus Wörterbuchtexten.“ In *Digitale romanistische Sprachwissenschaft: Stand und Perspektiven*, ed. Becker, Lidia et al., Tübingen: Narr

Francke Attempto.
ZOONIVERSE. 2009. <https://www.zooniverse.org/>. 19.02.2022.

Zusammenfassung

Das Zusammenführen von Informationen aus verschiedenen Quellen im Rahmen der linguistischen Forschung kann einen nicht zu unterschätzenden Aufwand darstellen. Webportale, die diese in digitalisierter Form enthalten, bieten eine mögliche Lösung für dieses Problem, müssen aber dazu bestimmte Anforderungen erfüllen. Dieser Beitrag analysiert diese Anforderungen und untersucht weitere originär digitale Möglichkeiten, die sich in diesem Kontext ergeben. Darauf aufbauend ermittelt er, was dies für Struktur und Format der zugrundeliegenden Daten bedeutet und zeigt am Beispiel eine konkrete Umsetzung der aufgestellten Prinzipien.

Abstract

Combining information from different sources in the context of linguistic research can lead to an effort that should not be underestimated. Web portals that contain a digital representation of that information offer a possible solution to this problem, but need to fulfil certain criteria to achieve this goal. The article analyses these criteria and examines further possibilities, exclusive to the digital form, that arise in this context. On this basis, the contribution identifies requirements for suitably structured and formatted data and presents an example that illustrates a concrete implementation of the principles set out.

Bild: © Alexandre Melay, *Soul has weight* (2016) | www.alexandremelay.com

Premiers travaux

apropos

[Perspektiven auf die Romania]

Winter 2022

9

www.apropos-romania.de

ISSN: 2627-3446

2022, n°9
pp. 277-305
doi: https://doi.org/10.15460/apropos.9.1939

Alejandro Cienfuegos Pérez

Atribución de autoría y humanidades digitales en el Siglo de Oro español

Alejandro Cienfuegos Pérez
estudió lenguas romances (español y francés) en la Universidad de Göttingen (Alemania).
alejandro.cienfuegos.perez@gmail.com

Palabras clave

Literatura – lingüística – humanidades digitales – siglo de oro español – verificación de autoría – atribución de autoría

Este artículo se basa en una tesis de maestría que se realizó en el programa "TransRomania-Studien: Romanische Sprachen, Literaturen und Kulturen" en la Georg-August-Universität Göttingen bajo la supervisión de PD Dr. Nanette Rissler-Pipka. El trabajo ha sido abreviado y revisado para su publicación en la sección Premiers Travaux de *apropos*.

1. Introducción

Partimos de la hipótesis de que los nuevos métodos de la estilometría, en particular el empleo de la herramienta *stylo* de R (Eder 2018; Eder y Rybicki 2012; Eder y Rybicki 2011) puede si no dar una solución definitiva al menos aportar nuevas perspectivas a una discusión literaria largamente debatida: la de la autoría de *La tía fingida* atribuida por muchos a Cervantes. Así mismo, también consideramos relevante no obviar la perspectiva literaria cuando la determinación de autoría se ocupa del estilo literario de un autor [1].

Proseguiremos con una de las partes que consideramos más complejas en el entorno del Siglo de Oro que es llevar a cabo la compilación de un corpus suficientemente representativo y fiable de textos digitales. El camino desde las fuentes, pasando por la problemática y limitaciones hasta la cosecha de textos y configuración del corpus será expuesto aquí profusamente. También es ineludible la relación que esto tiene con el texto dubitado. De él se comentará su relevancia subrayando los avatares desde la aparición del manuscrito original hasta llegar a las múltiples versiones que existen hoy en día del mismo.

[1] A pesar de que consideramos la importancia del uso del lenguaje inclusivo y la distinción autor/autora y similares, evitaremos su empleo a fin de garantizar una mejor legibilidad del texto. En cualquier caso, entiéndase que utilizamos un masculino genérico para referirnos a ambos sexos.

Llegados a este punto, se empleará la herramienta *stylo* para las pruebas estilométricas utilizando diferentes funciones. En base a los datos obtenidos en los experimentos podremos pasar al análisis de *La tía fingida* y aquella/s novela/s que presenten una mayor similitud estilométrica desde una perspectiva literaria. Esta última parte del trabajo se omite en este artículo, aunque se puede consultar en Cienfuegos-Pérez (2022).

2. Corpus

2.1 Problemática de los textos del Siglo de Oro

El Siglo de Oro español es de gran interés para los trabajos de atribución de autoría debido a diversos motivos que tienen que ver sobre todo con las circunstancias de transmisión de los mismos. Sin embargo, esas mismas circunstancias de transmisión también dan lugar numerosos problemas para la investigación como aquí expondremos.

A pesar de que actualmente están a disposición de la investigación numerosos textos y ediciones digitales, como comenta Rißler-Pipka (2018) para el análisis estilométrico es necesario disponer de los textos digitales, pero también los metadatos deben estar completos. Estos contienen informaciones extralingüísticas de los textos: autor, fecha y lugar de publicación, etc. Su importancia se debe al hecho de que la calidad de las ediciones de los textos de este periodo es difícilmente comprobable. Juola (2008, 247) apunta en este sentido a otro inconveniente que puede existir en las ediciones digitales como las de Google Scholar, JSTOR, o el Project Gutenberg que podrían haber sido corrompidas debido al proceso de escaneado o reescritura.

Otros de los aspectos problemáticos para la atribución de autoría es la determinación del corpus como abierto o cerrado (*open/closed set*) que vimos en Juola (2008). En este sentido, la primera pregunta que habría que responder es ¿sabemos a ciencia cierta que el posible autor se encuentra entre los seleccionados? Para nuestro objetivo un análisis exhaustivo conllevaría reunir todo el corpus de obras de los posibles autores, que aun suponiendo que pudiera ser Cervantes y existiendo determinadas hipótesis podría ser cualquiera que, por ejemplo, hubiera realizado trabajos del mismo género (novela breve) en torno a la fecha de publicación de la obra cuya autoría queremos determinar. Como veremos, existen limitaciones para poder reunir este corpus ideal.

Siguiendo con los manuscritos es necesario reseñar que los textos que mejores resultados estadísticos dan en cuanto a fiabilidad son aquellos cuyo contenido es el expresamente creado por el autor. Pues bien, incluso tomando como punto de partida los manuscritos del Siglo de Oro, estos presentan una serie de problemas dado que los procesos de impresión y el concepto de propiedad intelectual no eran tal y como los conocemos hoy en día. Francisco Rico (2000) publicó un trabajo coordinado al respecto de esta problemática con el nombre *Imprenta y crítica textual en el Siglo de Oro*. En este libro, podemos no solamente hacernos una idea

de las ediciones que se realizaban por el propio impresor de los textos originales sino también del origen del problema que nos ocupa: la atribución de autoría.

Por otra parte, Arellano-Ayuso (1997, 41-42) comenta las vicisitudes de la transmisión literaria en el Siglo de Oro observando la trayectoria de la difusión de los textos desde el manuscrito pasando por la edición y la posterior venta del libro. Arellano-Ayuso (1997, 42) divide su artículo distinguiendo y explicando los avatares de los textos según su tipo. Dentro de los textos impresos en prosa, Arellano-Ayuso (1997, 51) da muestra de algunos problemas en el proceso de impresión que dificultaron el paso del manuscrito al libro. En cuanto a las *Novelas ejemplares* de Cervantes que se publican en 1613 comentan que algunas de ellas eran anteriores, pues en *El Quijote* de 1605 se cita al *Rinconete y Cortadillo.* Entre 1604-1606 fue compilada una colección miscelánea de obras para el cardenal Fernando Niño de Guevara por parte de Francisco Porras de la Cámara. En ella incluye sendas versiones con variantes de *Rinconete y Cortadillo* y de *El celoso extremeño,* además de *La tía fingida*. El manuscrito de Porras de la Cámara, desapareció cuando estaba en manos de Bartolomé José Gallardo (1835), que había realizado algunas copias. Hoy tenemos la copia manuscrita ("basada" en el códice Porras de 1604-1606) y la versión impresa de la novela *El celoso extremeño (1613)*. Lo curioso de estas dos versiones es que el final difiere. Esto ha llevado a discrepancias entre los estudiosos: Maregalli (1992) achaca el final de 1613 a la censura, Canavaggio (1992), sin embargo, considera que el mismo Cervantes habría llevado a cabo la reescritura para dar a la trama y los personajes mayor complejidad (cfr. Arellano-Ayuso 1997, 55). En cualquiera de los casos, nuestra intención no es en este punto entrar en este debate, sino exponer algunos ejemplos de las vicisitudes y los cambios realizados en los textos que son otro de los variados problemas inherentes a la literatura del Siglo de Oro.

Una vez expuestos los aspectos limitantes más generales pasaremos al siguiente apartado donde expondremos las fuentes de las que disponemos para poder compilar un corpus de trabajo.

2.2 Fuentes para la recopilación de textos del corpus

En otro orden de aspectos para un trabajo de estilometría es indispensable reunir un amplio corpus que sea fiable y representativo. Para este fin, existen en la actualidad algunos recursos que son de gran ayuda pero que sin embargo también presentan en muchos casos ciertas limitaciones. José Calvo Tello en Github[2] pone a disposición una lista de portales que aportan recursos digitales en español. Entre ellos podemos destacar la base de datos de la Biblioteca Virtual Miguel Cervantes (BVMC) que da acceso a un inmenso catálogo de libros entre los que se hayan una gran cantidad del Siglo de Oro. Sin embargo, como comenta Rißler-Pipka (2018) muchos de estos libros son ediciones antiguas por cuestiones de derechos de autor lo cual es un perjuicio a la hora de realizar un escaneado además de que su lenguaje puede no corresponder con el español moderno. Otras bases de datos reseñables que pueden servir de fuente para la recopilación de textos son: AHCT (Association

[2] <https://github.com/morethanbooks/Atlas-de-Datos/blob/master/atlas%20de%20datos.csv>.

for Hispanic Classical Theater); CORDE (Corpus diacrónico del español) que no pone a disposición ningún texto completo permitiendo solamente la búsqueda de títulos y autores en un periodo determinado.

Todos los aspectos problemáticos y limitantes expuestos en este apartado podrían hacer pensar que un trabajo como el que nos ocupa podría no presentar la suficiente fiabilidad y demasiados escollos para poder ser conclusivo. Pues bien, tampoco lo pretende, recuperemos en este punto las palabras de Holmes y Kardos (2003) que afirmaban que "la estilometría […] no pretende anular la escolástica tradicional de los expertos en literatura e historia, más bien trata de complementar su trabajo proveyendo significados alternativos a los trabajos de investigación sobre proveniencia dudosa"[3] (Holmes y Kardos 2003, 1, traducción propia) y adjuntaban una insistencia final donde enuncian que "la evidencia estilométrica tiene que ser contrastada con la provista por estudios más convencionales hechos por los estudios literarios"[4] (Holmes y Kardos 2003, 5, traducción propia). Por ende, el objetivo de aportar nuevas perspectivas sigue siendo legítimo y posible pese a las limitaciones que la época impone al corpus. Además, aunque debemos de tener en cuenta todos los aspectos condicionantes aquí señalados, estos no han sido óbice para obtener resultados satisfactorios en otros trabajos en los cuales nos basamos para aplicar sus métodos.

Por último, consideramos que la utilización de las MFW[5] como marcador de estilo puede en parte hacer frente a algunos de las restricciones como la alteración que pueden producir en la estadística las modificaciones y ediciones del texto. Pensamos que siendo estas a menudo las denominadas "palabras funcionales" son menos susceptibles de ser editadas pues no varían el significado o el mensaje de un determinado texto. Por otra parte, en algunos casos pueden faltar partes del texto para lo cual no existe a nuestro saber ninguna solución. No obstante, esta consideración es una mera hipótesis. Una respuesta a esta conllevaría un estudio y análisis profundos de las vicisitudes y los cambios más frecuentes en los textos del Siglo de Oro que estableciera exactamente cómo y en qué medida se modificaban los textos.

Vistas las fuentes, pasaremos pues en el próximo apartado al proceso de compilación de nuestro corpus y la explicación de su configuración.

2.3 Cosecha de los textos y configuración del corpus

Para compilar el corpus con el que vamos a trabajar llevamos a cabo en primer lugar una consulta en el CORDE[6]. Dentro de este, seleccionamos todos los autores entre

[3] Original: "Stylometry – the statistical analysis of literary style – does not seek to overturn traditional scholarship by literary experts and historians, rather it seeks to complement their work by providing an alternative means of investigating works of doubtful provenance" (Holmes y Kardos 2003, 1)

[4] Original: "stylometric evidence must always be weighed in the balance along with that provided by more conventional studies made by literary scholars" (Holmes y Kardos 2003, 5)

[5] Most frequent words: se refiere a las palabras que más frecuencia presentan en una determinada lengua. Suelen corresponder con las denominadas palabras funcionales (preposiciones p. ej.). A partir de este punto utilizaremos la abreviatura MFW.

[6] REAL ACADEMIA ESPAÑOLA: Banco de datos (CORDE) [en línea]. Corpus diacrónico del español. <http://www.rae.es> [30.10.21]

1547 y 1620, periodo que coincide prácticamente con la vida de Miguel de Cervantes (1547-1616). Es una selección amplia que pretende abarcar todas las hipótesis posibles por lejanas que parezcan. Dentro de este periodo nos interesa saber que autores realizaron prosa narrativa breve en España, tanto culta como tradicional. El buscador arroja un total de 63 documentos con 677554 palabras. De estos resultados eliminaremos los numerosos anónimos que aparecen, pues no nos interesa incluir más textos dubitados. También se excluirán cartas, memoriales, premáticas y otros textos que por su extensión excesivamente breve o su tipología pueden ser problemáticas al compararlas con la TF.

En un análisis ideal dispondríamos de suficientes textos de prosa breve similares a las *Novelas Ejemplares* de diferentes autores de los que además tendríamos numerosos ejemplos, no obstante, no es el caso y no hay nada que podamos hacer para remediarlo. Una nueva búsqueda incluyendo la narrativa extensa da en el CORDE como resultado 39 documentos con un total de 4061193 palabras. Las *Novelas Ejemplares* tienen alrededor de 5000 hasta ca. 23000 palabras y como vimos, para el análisis estilométrico, necesitamos textos que sean lo más similares en cuanto a su extensión, sin embargo, dados los límites del corpus de la prosa breve no estamos en condiciones de rechazar muestras.

A este corpus debemos adjuntar otros textos que tienen algunas similitudes con las *Novelas Ejemplares* pero que pertenecen a una época posterior a Cervantes. Los incluiremos para añadir más fuentes de autores que están presentes en los corpus anteriores y que ayudarán a identificar mejor su estilo. Creemos que es poco probable que alguno de los presentes de este subcorpus hayan sido los creadores de la TF, pues esta precede a las primeras publicaciones que realizaron como escritores. Estos textos tienen también un argumento a favor para su inclusión que es su disponibilidad en formato txt. a través del *Github* de Fradejas Rueda[7]. Por otra parte, existe un problema y es que no disponemos de todos los metadatos acerca del origen de la edición de los textos más allá de la explicación de su propietario[8] por este motivo trataremos de contrastar los textos y completar estos datos en nuestro corpus.

Es necesario señalar que los autores que están representados por una sola obra son problemáticos para la clasificación utilizando *stylo* pues el algoritmo va a tratar de encontrar los textos que presentan una mayor similitud. Al no existir más ejemplos del mismo autor, es probable que no se puedan observar las características estadísticas de forma clara y el texto sea situado en algún lugar erróneo dando prioridad a otras cuestiones como el género, la extensión o la temática y genere de este modo irregularidades. Por este motivo y para aplacar el obstáculo que supone la diferente extensión de los textos, vamos a fragmentar los documentos del corpus igualando su longitud entorno a las 6000-7000 palabras, lo cual también nos posibilita multiplicar las muestras de un determinado autor e

[7] https://github.com/7PartidasDigital

[8] "Los textos relacionados con M.ª de Zayas y Alonso Castillo Solórzano proceden de los ficheros de Alejandro García-Reydi (USAL), los del entorno del Quijote de Avellaneda los he cosechado en la Cervantes Virtual y algunos me los ha pasado Javier Blasco Pascual (UVa)". Fradejas Rueda en: <https://github.com/7PartidasDigital/NovelaBarroca>.

incorporarlo incluso disponiendo de un solo texto de base. Estas muestras podrían ser útiles a la hora de configurar los parámetros de la herramienta donde servirán quizás para tratar de anular algunas de las señales que se desvelan tras las pruebas estilométricas[9]

Tras este trabajo de compilación, hemos realizado varios (sub)corpora: un primer corpus de la prosa narrativa breve contemporánea a Cervantes, muy reducido; un segundo corpus con la prosa narrativa extensa también contemporánea a Cervantes que presenta un número de palabras por texto muy superior al de las *Novelas Ejemplares*; y un tercer corpus de obras del Siglo de Oro de autores que suponemos menos susceptibles de guardar relación con la TF.

Al corpus añadiríamos por último la versión de la obra dubitada. Encontramos diferentes ediciones en la BVMC por lo que surge la cuestión en torno a cuál de ellas seleccionar. José Manuel Lucía Megías (2018) hace un comentario que nos parece relevante con respecto a las ediciones:

> La tía fingida, […] a pesar de los esfuerzos editoriales de los últimos años, aún carece de la edición crítica que dé cuenta de la complejidad de sus materiales textuales y el hecho de contar con testimonios de muy diversa naturaleza —copias antiguas y modernas— que dan cuenta de dos redacciones de la obra. (Lucía Megías 2018, 346)

De entre los testimonios que disponemos seleccionaremos la versión de Porras de la Cámara realizada por Francenson y Wolf (Berlín, 1818) en la edición realizada por Florencio Sevilla Arrollo que se encuentra disponible en formato digital en la BVMC. Esta edición, la berlinesa, es en palabras de Lucía Megías (2018, 344-345) "*muy correcta y fiel a su testimonio base, en que solo se aleja de los tres grandes grupos de cambios lingüísticos y ortográficos*" que como indica y presenta profusamente en el apéndice son: cambio de grafías, cambio de mayúsculas y minúsculas y cambios en la puntuación. Para nuestro análisis la modernización de las grafías es un aspecto positivo que hemos buscado también en el resto de textos para poder disponer de cierta homogeneidad. Los cambios en la puntuación no son relevantes pues no son un elemento fiable en los textos de esta época por lo que no serán tenidos en cuenta en el análisis cuantitativo. Los datos de la edición seleccionada son los siguientes:

- Autor: desconocido ¿Cervantes?
- Título: La tía fingida
- Fecha de publicación: 16??
- Edición: Novela de La tía fingida [versión Porras de la Cámara por Francenson/Wolf]
- Versión digital: <http://www.cervantesvirtual.com/nd/ark:/59851/bmc1z4h>.

No se nos escapa el hecho de que el corpus que hemos recopilado parte de los datos que nos da el CORDE. Como ya comentamos, nos basamos en aspectos sincrónicos seleccionando todas aquellas obras disponibles que correspondieran

[9] que, como resumen Cerezo Soler y Calvo Tello (2019, 237) citando algunos ejemplos de trabajos al respecto, son: el género literario (Kestemont 2011), la época de composición (Jockers 2014) y otras.

con la categoría de prosa narrativa. Sin embargo, el CORDE tiene en cuenta aspectos lingüísticos, no literarios y por lo tanto da como resultado una serie de textos de diferentes géneros literarios. Aunque en base a los resultados del buscador del CORDE hay una cierta homogeneidad en realidad desde un punto de vista literario nos encontramos con un corpus más bien heterogéneo.

3. *Stylo* de R: justificación de la herramienta.

En este punto ya tenemos listo un corpus considerable de textos en prosa del Siglo de Oro. Así pues, el siguiente paso antes de poner en marcha el análisis es presentar la herramienta que vamos a utilizar para llevarlo a cabo para así poder también comprender los resultados que arrojará y su índice de fiabilidad.

R es un entorno y lenguaje de programación que está enfocado en el análisis estadístico pero que como apunta Rueda (2019) y muestran los numerosos estudios que se han servido de este entorno tiene un gran potencial para su uso en los estudios filológicos. Como ejemplo de los trabajos realizados utilizando R, Rueda (2019) hace referencia a estudios del ámbito de la lingüística (Gries 2013, Gries 2016, Desagulier 2017, Levshina 2015, Winter 2019) y de la literatura (Jockers 2014). Si bien esta lista, se puede ampliar inmensamente.

Dentro de este entorno, para llevar a cabo el análisis estilométrico existen múltiples herramientas. Una de las más recientes y que ha dado buenos resultados en varias investigaciones es el paquete *stylo* basado en R (Eder, Rybicki y Kestemont 2016) mantenido y desarrollado por el grupo de estilística computacional[10]. Este paquete aporta diferentes funciones para el análisis estilométrico además de una interfaz de uso sencillo y diagramas de gran calidad.

Recordemos en este punto que Juola (2008) comentaba respecto a las herramientas a utilizar para el análisis estilométrico que lo que es más importante para el usuario casual es la capacidad de seleccionar los algoritmos y las características a utilizar dinámicamente, en función del idioma, el género, el tamaño, de los documentos disponibles. En este sentido *stylo,* nos parece una herramienta que cumple con todas estas necesidades. Con respecto a la precisión nos remitiremos al trabajo de Blasco Pascual y Ruiz Urbón (2009) que realizan una evaluación y cuantificación de algunas técnicas de atribución de autoría en textos españoles. En su trabajo, muestran que el *perfil de palabra simple* es el más efectivo para los textos del Siglo de Oro con un 70% de precisión en un corpus de 10 autores que se eleva al 90% entre dos autores. Además, dan cuenta de que el perfil más efectivo es el de *marca de puntuación simple1*[11] que presenta un 85% de fiabilidad en un corpus de 10 autores; sin embargo este método no es aplicable al Siglo de Oro dado que, como ya comentamos en el apartado sobre la problemática de los textos de esta época, en particular su puntuación no es suficientemente fiable[12].

[10] <https://computationalstylistics.github.io/resources/>

[11] Este perfil es calculado mediante la división de la frecuencia de una serie de signos de puntuación entre el número total de caracteres contenidos en el texto.

[12] Pascual y Urbón (2009) comentan en sus conclusiones que hay dos razones que suponen un problema para la atribución de autoría y que tienen que ver una con la fase de creación y otra con la fase de edición de los textos. Son las siguientes: *a) Los hábitos de escritura en los Siglos de Oro* b) Los correctores, componedores y

De este modo, de los diferentes métodos posibles para el análisis solo queda uno efectivo: "Desgraciadamente del *top ten* obtenido para el español sólo resulta operativo el *perfil de palabra simple*" (Blasco Pascual y Ruiz Urbón 2009, 44).

Así pues, llegados a este punto queda explicado el motivo del uso de *stylo*: medir una de las realidades textuales que es la frecuencia de palabras y dentro de estas *el perfil de palabra simple* que parece ser teóricamente el más adecuado y efectivo para este conjunto de textos[13]. Con el corpus compilado y la herramienta presentada podemos pasar a su empleo práctico.

4. Pruebas estilométricas

El primer paso a realizar es una serie de experimentos con el fin de determinar si nuestro corpus es organizado de manera lógica (por autores) por parte de la herramienta. Por el momento dejaremos la obra dubitada a un lado. Vamos a realizar un *Cluster Analysis* de cada uno de los subcorpora para crear un corpus final con el que trabajaremos. Eder (2015) considera que la extensión del texto que mejor define la señal autorial está en torno a las 5000 palabras y Hernández Lorenzo (2019, 192) comenta que el estilo en la narrativa suele estar más diseminado que por ejemplo en la poesía. Sin embargo, la novela breve tiende a concentrar el estilo. Además, como explicaremos a continuación hemos dividido las novelas extensas en fragmentos de unas 6000 palabras para lograr cierta homogeneidad de los fragmentos, así pues, utilizaremos el rango de las 500 MFW para nuestras pruebas estilométricas.

4.1 Análisis de *clúster* para ajustar el corpus

Corpus 1: Prosa narrativa breve

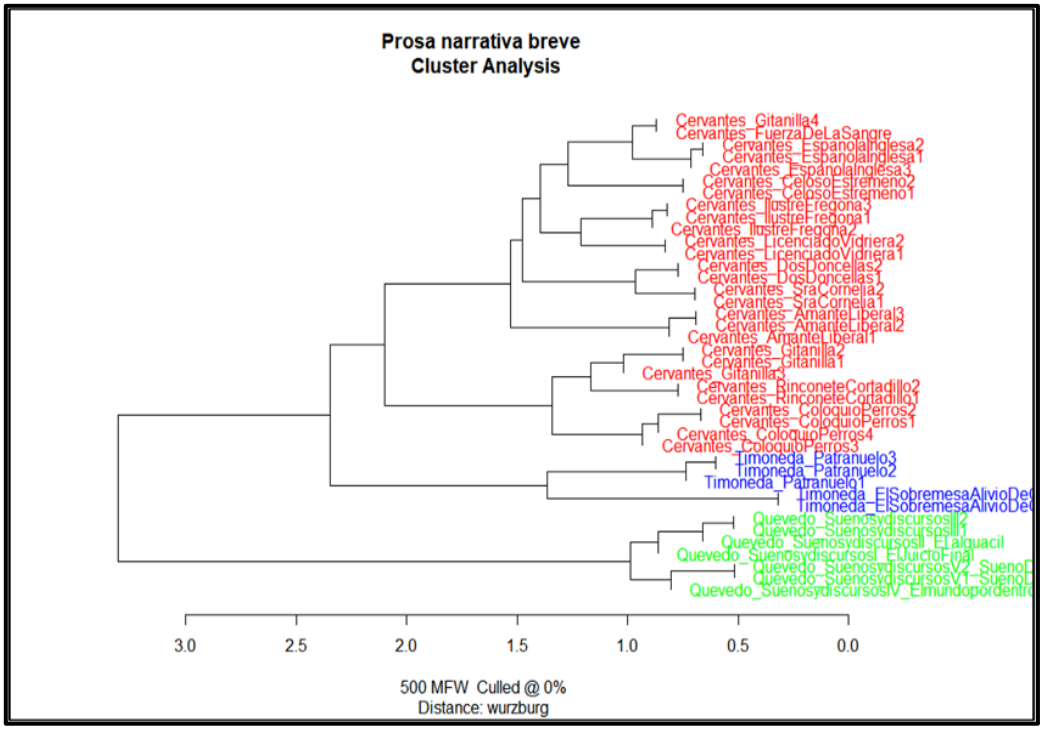
Tras varias pruebas con la función *stylo ()*, encontramos que la organización de los textos es eficiente utilizando de las 100 a las 500 MFW y *Cosine Delta*. También se observó en los primeros experimentos que existían algunas irregularidades que podrían ser debidas a la distinta extensión de los textos. Esta heterogeneidad de los textos está presente con mayor o menor medida en todo el corpus. Es por ese motivo por el cual hemos decidido fragmentar las obras en partes de aproximadamente 6000 palabras con el fin de homogeneizar lo más posible nuestros experimentos. La organización de los textos con 500 MFW se puede ver en la siguiente imagen. La herramienta reconoce sin problemas la obra de los autores y los organiza como se esperaba.

Si observamos el eje vertical, vemos que la herramienta reconoce dos Cervantes diferentes y que algunos de los textos no están organizados con sus correspondientes partes. Esto muestra que hay una cierta heterogeneidad en este

cajistas de imprenta podrían ser en alta medida los responsables de muchas de las marcas que nuestros procedimientos de medición actuales contemplan. (Blasco Pascual y Ruiz Urbón 2009, 44)

[13] Sería interesante en este punto presentar la funcionalidad y el uso de la herramienta estilo. Sin embargo, esto extendería mucho este artículo por lo que referimos al lector/a las publicaciones de sus creadores Maciej Eder, Jan Rybicki y Mike Kestemont (2016) y los tutoriales que se encuentran disponibles en la red, por ejemplo en: <https://computationalstylistics.github.io/stylo_nutshell/>.

corpus cervantino y pone quizás de manifiesto la relación entre ciertas partes de diferentes obras (por ejemplo, la cuarta parte de *La gitanilla* y *La fuerza de la sangre*). En comparación a otras pruebas hechas utilizando los textos completos, la división parece ayudar a que los textos se organicen mejor. Así mismo esto nos permitirá más adelante incluir algunos autores de los que solo disponemos de un texto que de otra manera sería imposible por no ser suficientemente representativo del "estilo" ya que el sistema trataría de asociarlo a un texto cercano y causaría imprecisiones. El corpus siguiente va a incluir por lo tanto los textos del corpus 1 y del corpus 2.

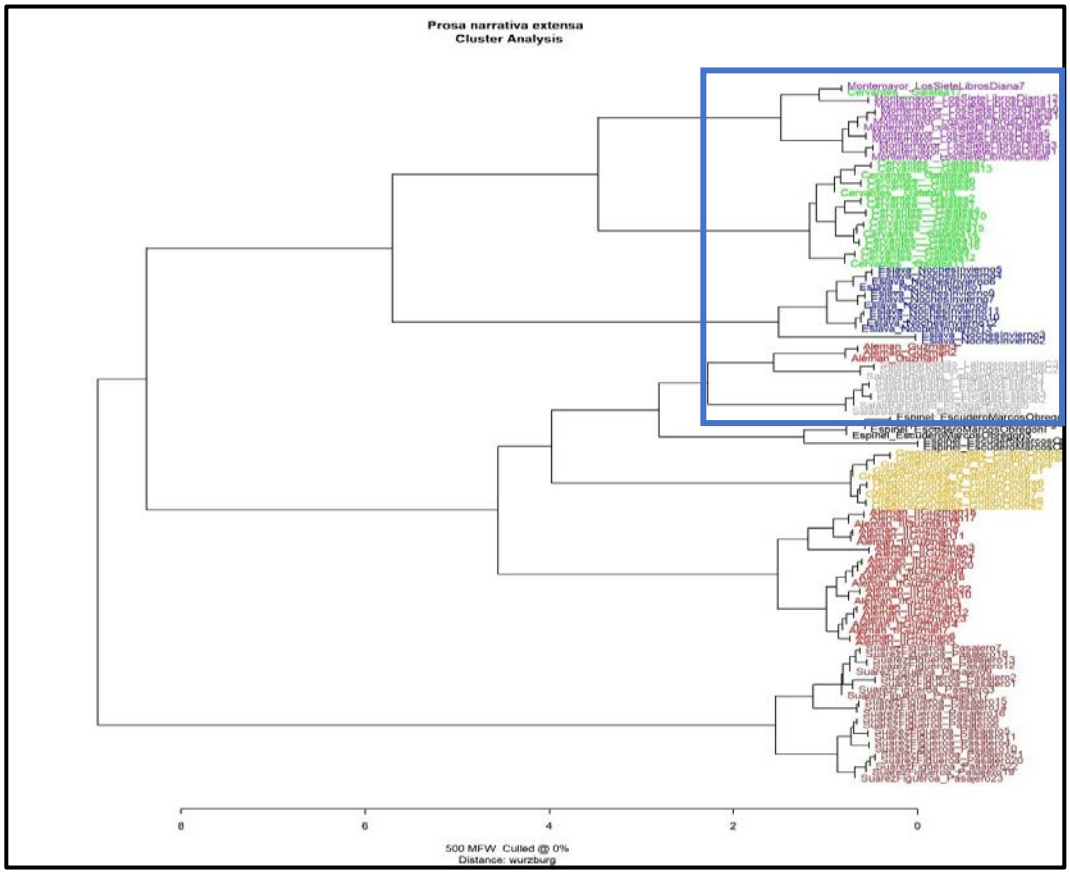


1 | prosa narrativa breve, análisis de *clúster*.

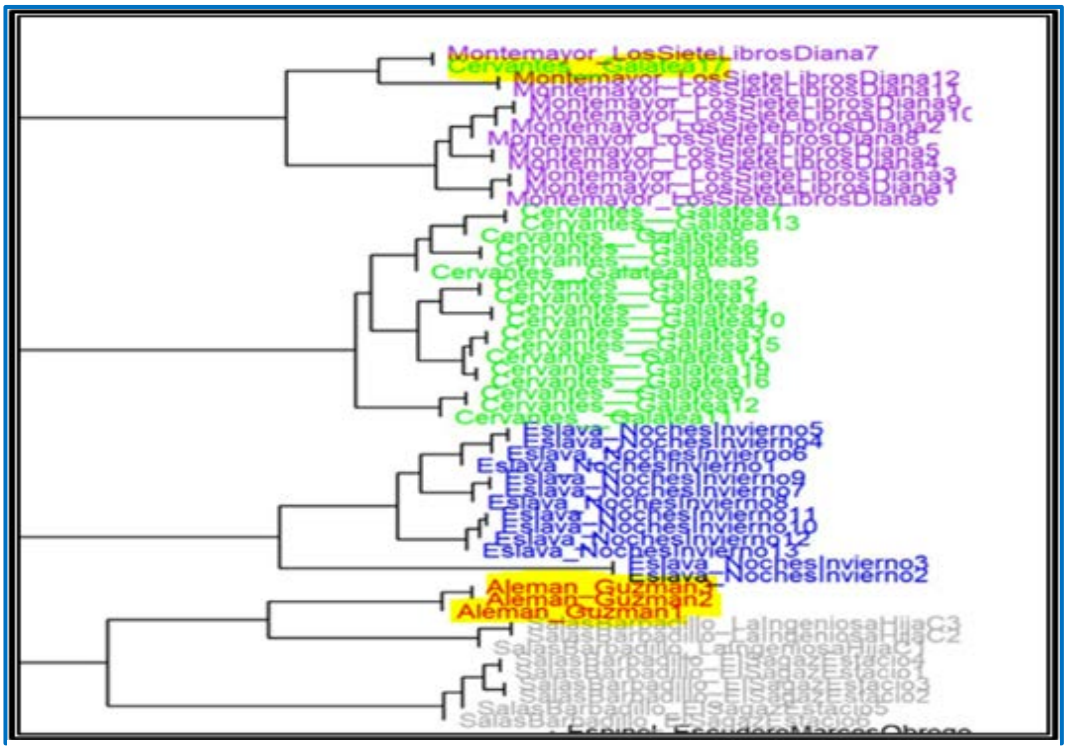
Corpus 2: Prosa narrativa extensa

Continuamos el experimento manteniendo la misma configuración de 500 MFW y *Cosine Delta*. Esta vez con nuestro segundo (sub)corpus que contiene las obras que corresponden a la prosa narrativa extensa fragmentada en partes de aprox. 6000 palabras. Vemos que en los resultados la organización también es la esperada (imagen 2). Solamente sorprende la parte diecisiete de *La Galatea* que aparece junto a la séptima parte *Los siete libros de Diana* de Montemayor. Igualmente, las tres primeras partes del *Guzmán de Alfarache* de Alemán aparecen separadas del resto de la obra en un subcluster en el que encontramos los textos de Salas Barbadillo. Aunque este dato podría ser interesante para el análisis literario, no es el tema que nos ocupa. Nuestro objetivo es conseguir un entorno libre de interferencias en el que podamos incorporar un texto dubitado. Por este motivo y considerando que las obras están suficientemente representadas, vamos a prescindir de esas partes de los textos.

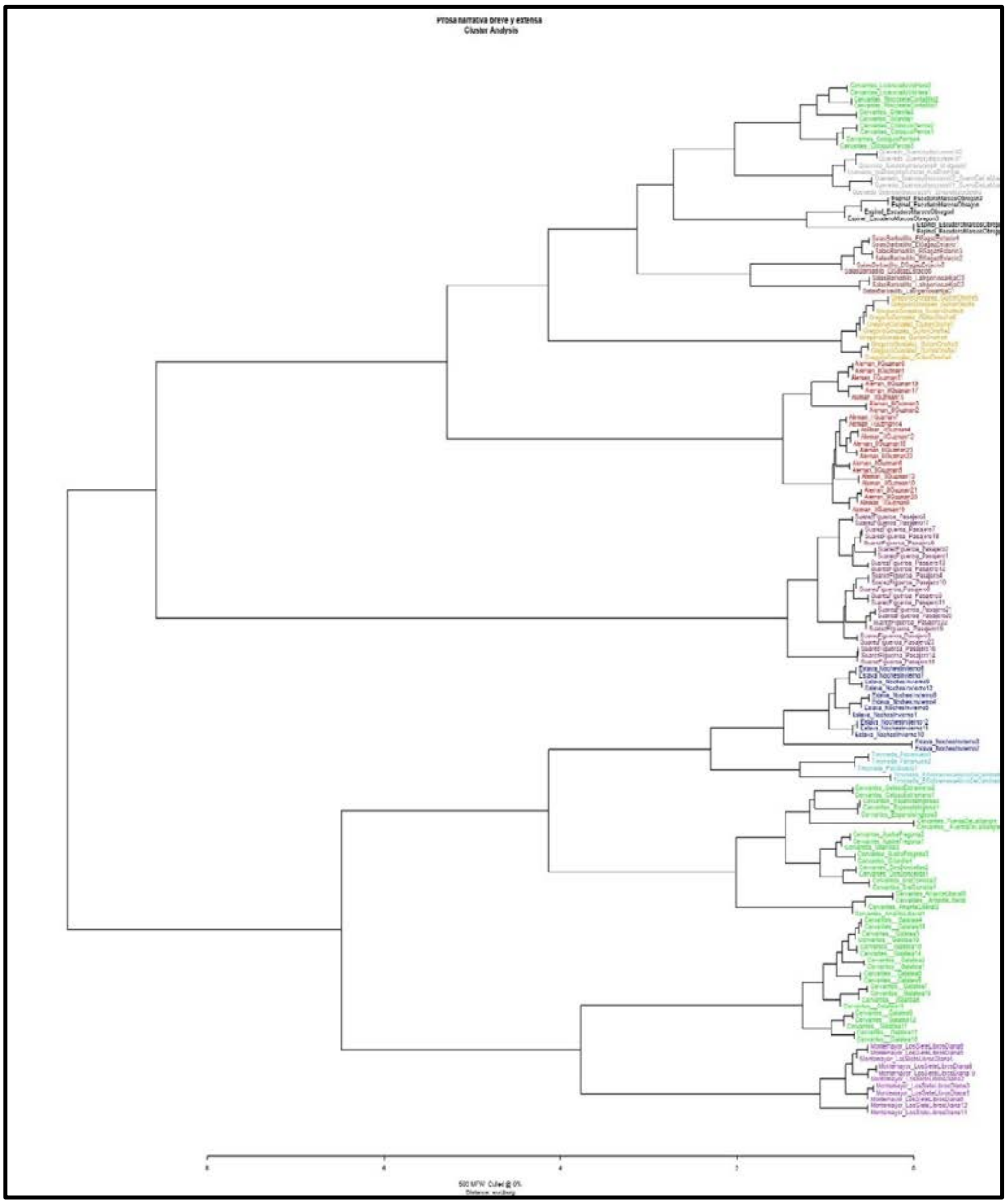
Al repetir la prueba sin estas partes, la organización de las obras por autores es satisfactoria. En la imagen se puede ver el dendrograma con los primeros resultados:



2 | prosa narrativa extensa, análisis de *clúster*, 500 MFW, *Cosine Delta*.

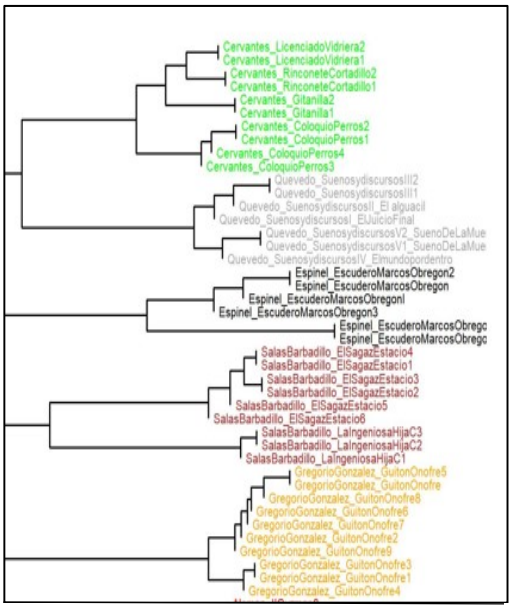


3 | detalle del dendrograma (Imagen 2), partes del texto problemáticas.

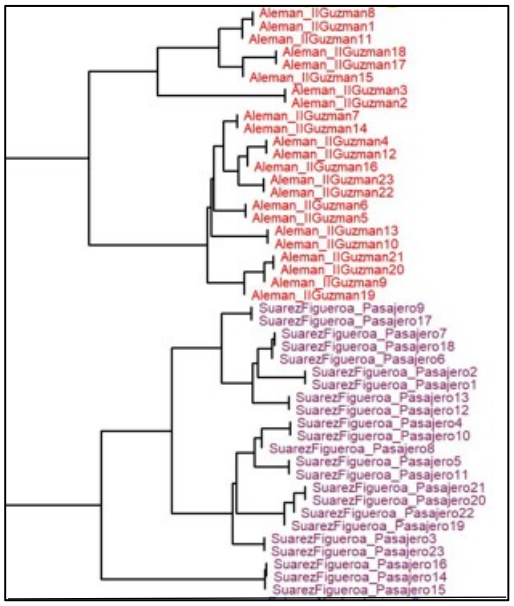


4 | prosa narrativa breve y extensa, análisis de *clúster*, 500 MFW, *Cosine Delta*.

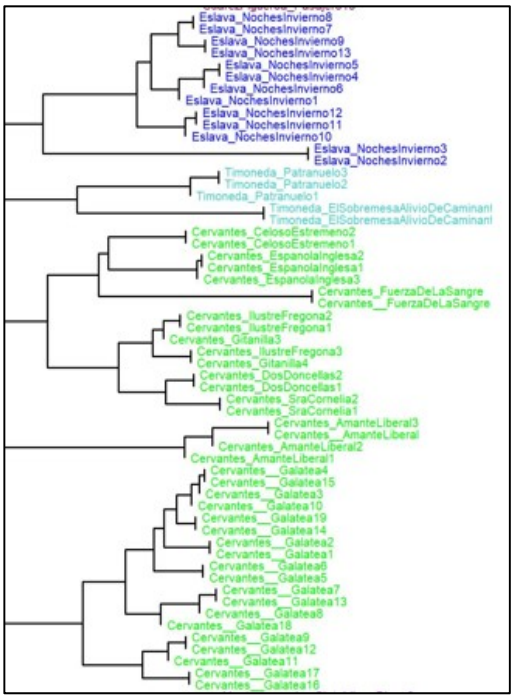
Llegados a este punto, vemos que una vez excluidos los fragmentos de los textos que causaban irregularidades, la herramienta es capaz de organizar las obras por autores en los (sub)corpus por separado por lo que ya podemos proceder a juntar ambos (sub)corpus para ver cómo se comportan en conjunto. La imagen de la página siguiente muestra que la organización de los autores y sus obras en *clusters* es correcta utilizando las 500 MFW y *Cosine Delta*. El amplio número de textos hace que la imagen del dendrograma se vea con dificultad por lo que incluimos imágenes con los detalles de las diferentes partes:



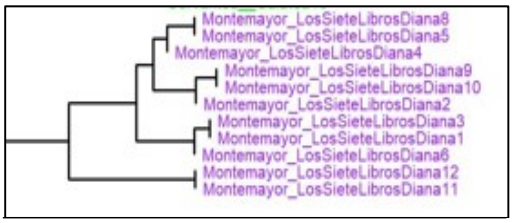
5 | sección 1 del dendograma (Imagen 4)



6 | sección 2 del dendograma (Imagen 4)



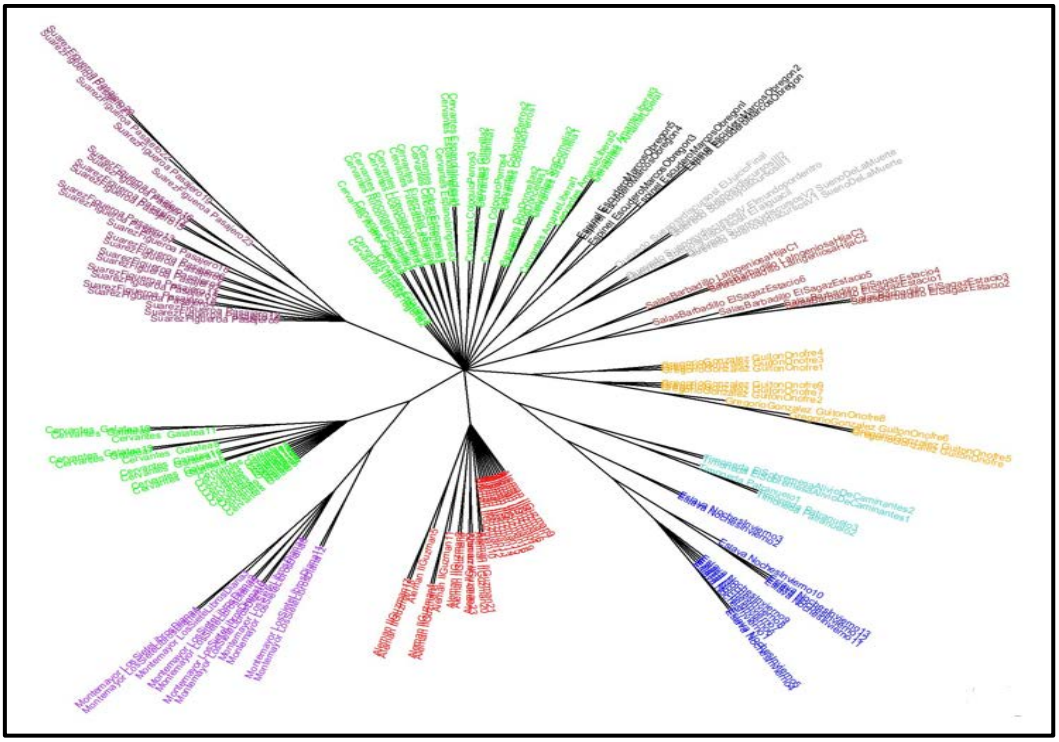
7 | sección 3 del dendograma (Imagen 4)



8 | sección 4 del dendograma (Imagen 4)

Algunos resultados dignos de mención son la clasificación de dos de las partes de *El escudero Marcos Obregón* de Espinel (imagen 5), de *Noches de Invierno de Eslava* (imagen 7) o de *La fuerza de la sangre* de Cervantes (imagen 7). En los dos primeros casos entendemos que son partes de los textos que se diferencian del resto por múltiples posibles motivos. Podrían ser una muestra de la señal del género literario o temática, por ejemplo. Esto no obstante no va a ocupar nuestro trabajo, aunque resulta interesante y podría analizarse en un trabajo específico. Lo que es para nosotros fundamental es que dichas partes fueron agrupadas correctamente en base a sus autores.

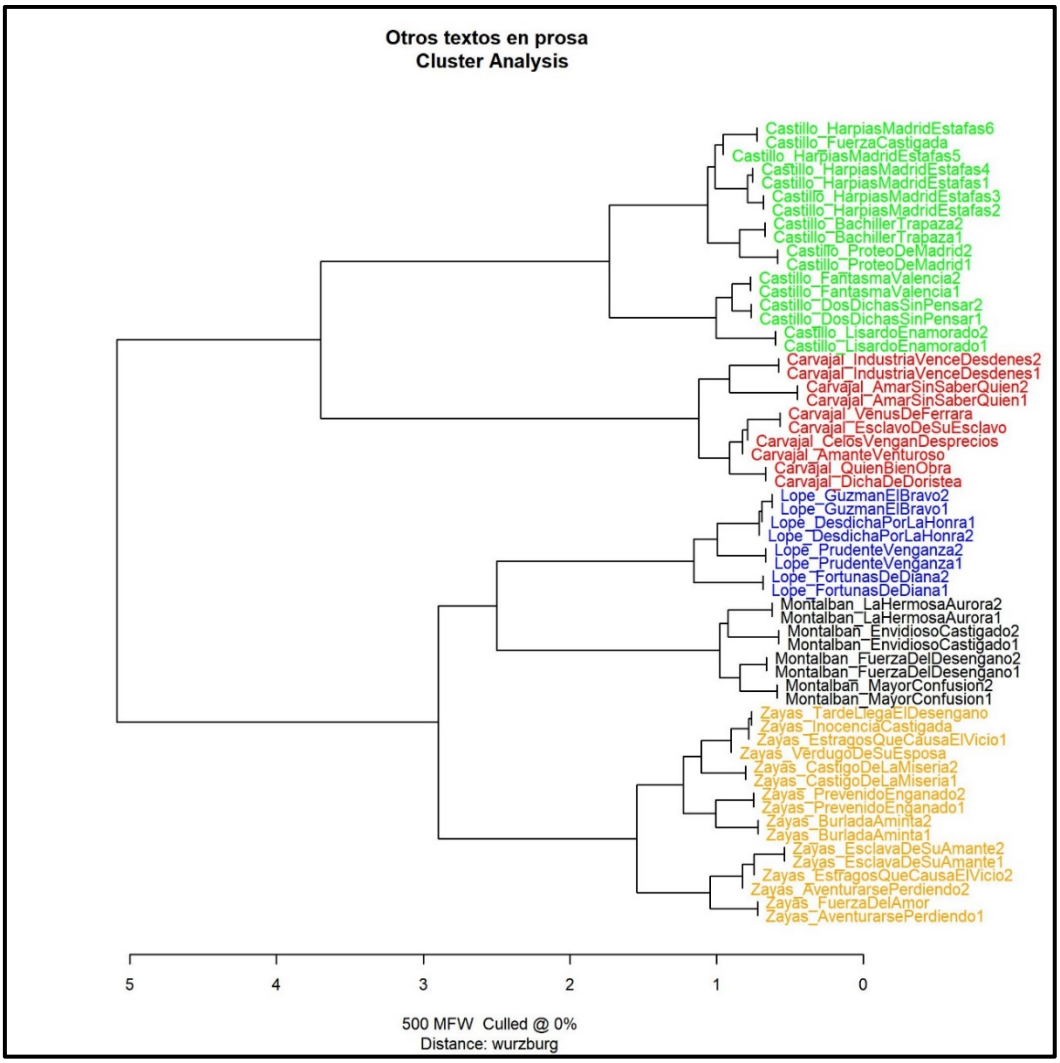
Ahora vamos a utilizar el árbol de consenso (BCT) que reúne los resultados de los dendrogramas que van desde las 100 MFW hasta las 500 MFW con el incremento especificado de 100 MFW. Esta nueva visualización nos permite observar de forma más evidente la relación entre los textos del corpus. Vemos que la clasificación de autores y obras es consistente en el sentido de que no se mezclan que como ya comentamos es lo que nos interesa en este punto. Sin embargo, también se puede observar que las ramificaciones correspondientes a Cervantes (verde claro) que salen del eje central corresponden a cada uno de los textos lo cual sugiere una gran heterogeneidad en su obra. Se podría decir que el programa reconoce a doce Cervantes. Por otra parte, la aparición de *La Galatea* a la izquierda separada del resto de su obra responde a la diferencia del género y su cercanía a la novela *Los siete libros de la Diana* de Montemayor lo respalda, pues esta novela pertenece también al género pastoril.



9 | prosa narrativa breve y extensa, *Bootstrap Consensus Tree,* 500 MFW, *Cosine Delta*.

Corpus 3: Otros textos en prosa

El último (sub)corpus que queda por organizar es el correspondiente a aquellos textos que incluyen a los autores que en un principio consideramos menos susceptibles de ser los creadores de la TF por una cuestión de cronología. Aquí realizamos nuevamente un análisis de *clúster*. Los resultados son los esperados utilizando nuevamente las 500 MFW con *Cosine Delta*:



10 | otros textos en prosa, análisis de *clúster*, 500 MFW, *Cosine Delta*.

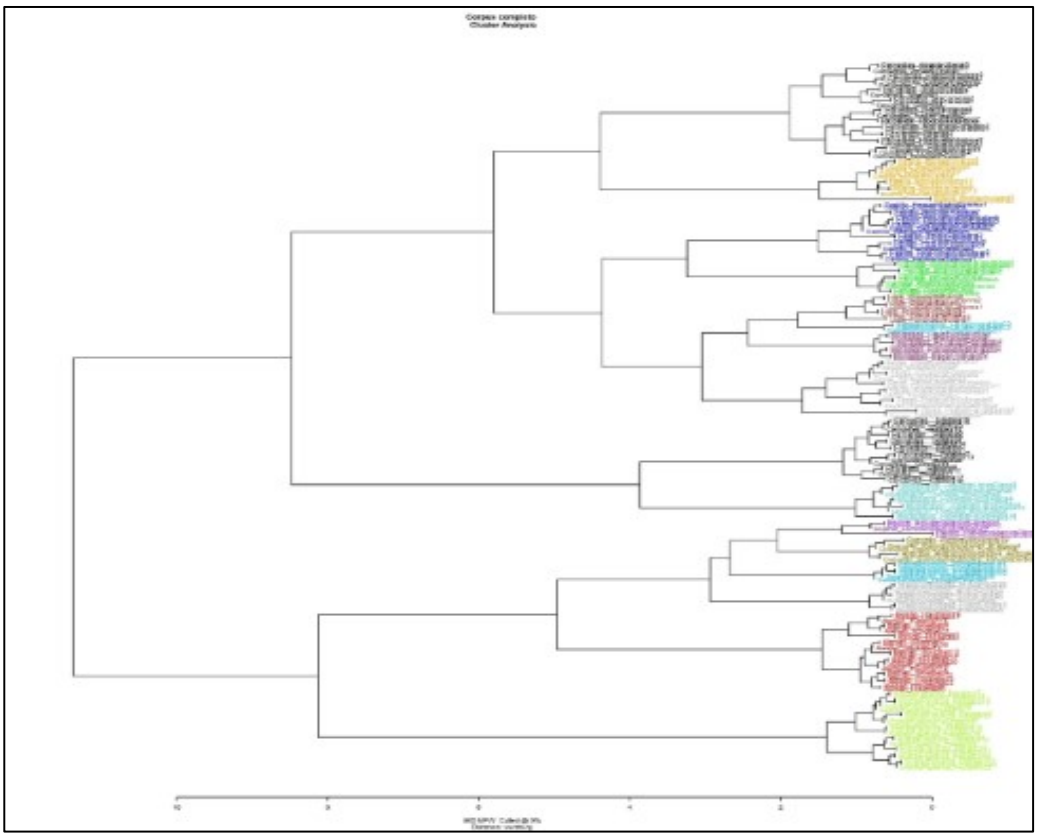
En este caso, no hay grandes sorpresas pues todos los textos son organizados en *clústeres* en los que se incluye un solo autor.

Corpus Completo

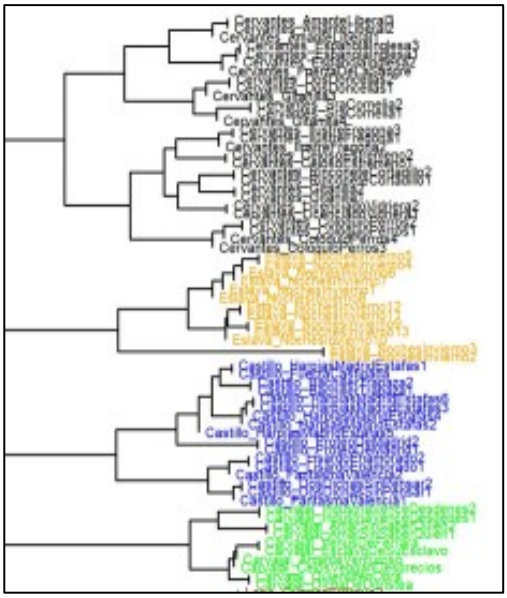
Tras esta progresiva comprobación de la organización de nuestro (sub)corpus por parte de *stylo* llegamos al que hemos llamado "corpus completo". En él incluimos todos los textos y autores de las pruebas anteriores para tratar de comprobar si la organización sigue siendo la esperada. Los resultados se vuelven más difícilmente visibles pero la organización de los textos por autores es correcta, aunque sigue existiendo cierta heterogeneidad en el corpus cervantino. Esto da lugar a que se cuelen dentro de un mismo *cluster* aunque separado en los subclusters *El patrañuelo* y *El sobremesa o alivio de caminantes* de Timoneda junto a las *Novelas Ejemplares*. También cercano encontramos el texto de Eslava, *Noches de Invierno*, pero este está algo más alejado en el eje horizontal donde se muestra que la relación aparece con respecto al *cluster* en el que están las *Novelas Ejemplares* y el texto de Eslava. Por otra parte, *La Galatea* se desmarca también y aparece en un

clúster en el que también está la obra de Montemayor. Sabemos a que se puede deber esto: se trata de dos novelas (prosa extensa) que pertenecen al género pastoril.

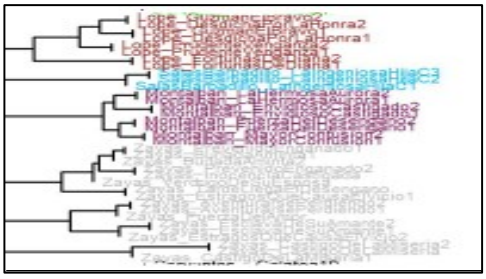
En este punto podemos plantearnos retirar la obra de Timoneda. *El sobremesa y alivio de caminantes* y *El patrañuelo*. Al hacer esto los resultados son idénticos, pero desaparece Timoneda del dendrograma:



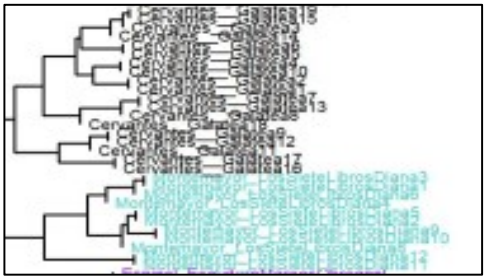
11 | corpus completo, análisis de clúster, 500MFW, *Cosine Delta*



12 | detalle 1 del dendograma (img. 11)



13 | detalle 2 del dendograma (img. 11)



14 | detalle 3 del dendograma (img. 11)

En este punto ya hemos puesto a prueba la herramienta en su función más básica y hemos comprobado que la organización de los textos en base a la figura autorial funciona correctamente con los (sub)corpora por separado y también en conjunto eliminando algunas partes y algunos textos. En el siguiente apartado vamos a proceder a comprobar la efectividad de estos resultados para luego introducir nuestro texto dubitado.

4.2 Precisión a través de la validación cruzada y clasificación

Para llevar a cabo la organización de los textos según su autor hemos utilizado la configuración de las 500 MFW y la distancia *Cosine Delta*. Sin embargo, aún nos queda responder a una pregunta de gran relevancia ¿con qué precisión se están organizando los textos? En este sentido, Eder añadió una revisión al observar que el límite fiable de las MFW puede variar según el corpus utilizado en el análisis. En base a esto, Eder (2017) llegó a la conclusión de que el factor más determinante de un texto es la fuerza de la señal autorial: a algunos les basta un número reducido de palabras mientras que otros continúan siendo dudosos pese a una mayor extensión. En nuestro caso utilizamos aún solo las 500 MFW adaptándonos a las circunstancias que impone la brevedad y la heterogeneidad de las longitudes de los textos del corpus. Las observaciones llevadas a cabo por Eder tienen como consecuencia que para poder conocer la fiabilidad de atribución de autoría para un autor es preciso llevar a cabo un análisis supervisado para comprobar el grado de fiabilidad que puede esperarse de la organización de los diferentes textos por parte de *stylo* antes de utilizarlo sobre el texto dubitado (cfr. Hernández Lorenzo 2019, 193-194). El código de este experimento puede consultarse en el anexo de Cienfuegos-Pérez (2022).

Los resultados reconocen quince clases en base a quince textos que se han tomado como referencia. Estos representan a los 14 autores que hemos seleccionado (Cervantes está repetido). El porcentaje general de acierto es del 100% con una desviación estándar del 0% utilizando desde las 100 hasta las 500 MFW con un incremento de 100 MFW. Si accedemos a los datos sobre el acierto atendiendo al número de palabras frecuentes utilizadas vemos que ya con las 100 MFW este porcentaje es del 100% y se mantiene así hasta las 500 MFW.

Sabiendo estos datos, utilizamos la función *classify ()* para la validación cruzada cuyo código y explicación se encuentra nuevamente en el anexo "código" (2) de Cienfuegos-Pérez (2022). Los resultados de esta prueba son otra vez correctos. Los autores predichos para los textos dados son los siguientes:

Aleman	Carvajal	Castillo	Cervantes	Cervantes
1	1	1	1	1
Eslava	Espinel	GregorioGonzalez	Lope	Montalban
1	1	1	1	1
Montemayor	Quevedo	SalasBarbadillo	SuarezFigueroa	Zayas
1	1	1	1	1

Tab. 1 | resultados de la validación cruzada, 500 MFW, *Cosine Delta*.

Así pues, vemos que utilizando tan solo las 500 MFW el resultado de acierto general es del 100%. El sistema ha conseguido reconocer la existencia de 15 clases correspondientes a los 14 autores. Para poder llegar a estos resultados hemos tenido que dividir los textos de aquellos autores de los cuales disponíamos de un solo ejemplo y homogeneizar las longitudes de todo el corpus. También, como ya se indicó, tuvimos que prescindir de algunos textos que causaban irregularidades e imprecisión por lo que consideramos que no nos permitirían introducir un texto dubitado en el próximo paso.

Por último, solo nos queda introducir el texto dubitado en este proceso. Así pues, repetimos la última prueba, pero esta vez introduciendo el texto dubitado en la carpeta *secondary_set* que es la que contiene las muestras de diferentes autores que van a ser comparadas con el *primary_set* conteniendo numerosos textos de ejemplo que le sirven al programa para "aprender" el estilo. Los resultados de las clases esperadas y predichas se pueden ver en la siguiente tabla:

Esperado ("expected")	Predicho ("predicted")
[1] "Aleman"	[1] "Aleman"
[2] "Carvajal"	[2] "Carvajal"
[3] "Castillo"	[3] "Castillo"
[4] "Cervantes"	[4] "Cervantes"
[5] "Cervantes"	[5] "Cervantes"
[6] "Desconocido"	[6] "Cervantes"
[7] "Eslava"	[7] "Eslava"
[8] "Espinel"	[8] "Espinel"
[9] "GregorioGonzalez"	[9] "GregorioGonzalez"
[10] "Lope"	[10] "Lope"
[11] "Montalban"	[11] "Montalban"
[12] "Montemayor"	[12] "Montemayor"
[13] "Quevedo"	[13] "Quevedo"
[14] "SalasBarbadillo"	[14] "SalasBarbadillo"
[15] "SuarezFigueroa"	[15] "SuarezFigueroa"
[16] "Zayas"	[16] "Zayas"

Tab. 2 | resultados de la función *classify*, corpus + La TF, 500 MFW, *Cosine Delta*

Como se puede ver en la tabla, el texto *Desconocido* que no es otro que *La tía fingida*, ha sido clasificado como perteneciente a Cervantes. Por lo tanto, el resultado que arroja la prueba de atribución de autoría con la función *classify* es que entre los autores seleccionados el texto dubitado presenta mayor similaridad estilística con Cervantes y por lo tanto parece pertenecer a este autor.

4.3 Verificación de autoría

Llegados a este punto en el que hemos comprobado la efectividad y hemos visto que la TF es atribuida a Cervantes podemos continuar con las pruebas en este caso de verificación de autoría: suponemos que el autor es Cervantes, pero queremos ponerlo a prueba. Esto sirve también como validación cruzada de los resultados anteriores. Para realizar este experimento vamos a utilizar otra característica de

stylo denominada *General Imposters* (GI) [14] también conocida como segundo sistema de verificación (o2). Kestemont et al. (2016) apuntan que esta característica fue introducida por Koppel y Winter (2014) y aplicada en el estudio de los escritos disputados de Julio Cesar.

Para ello seguimos el *script* que propone Eder (2018) adaptándolo a nuestro corpus (ver anexo "corpus" en Cienfuegos-Pérez 2022). La primera de las pruebas consiste en comparar las frecuencias relativas de las MFW de la TF con todos los demás textos a través de la función *imposters()*[15]. Utilizamos la medida de distancia por defecto que es *Delta de Burrows* obteniendo los siguientes datos:

Alemán	Carvajal	Gregorio Gonzalez	Lope
0.00	0.00	0.00	0.00
Castillo	Cervantes	Montalban	Montemayor
0.02	0.98	0.00	0.00
Eslava	Espinel	Quevedo	Salas Barbadillo
0.00	0.02	0.00	0.01
Suarez Figueroa	Zayas		
0.02	0.05		

Tab. 3 | resultados de la función *imposters* con Delta de Burrows.

Para entender los coeficientes solo hace falta saber que la atribución se da en los valores cercanos a 1 y se rechaza en el 0. Los resultados son los que nos da el algoritmo al tratar de comparar un texto anónimo determinado con un conjunto de textos de candidatos entre los que se encuentra también el probable autor. Dado que no indicamos otra cosa, el método trata de comprobar uno tras otro todos los autores disponibles que considera como candidatos potenciales. Sin embargo, también podemos focalizarnos en un solo autor haciendo que el método compare el texto dubitado con todos los textos de un candidato. Así pues, le indicamos al programa que queremos hacer esto. Lo llevamos a cabo de manera iterativa y con todos los autores siguiendo el consejo de Eder (2018) de forma que los resultados nos sirvan también a modo de validación cruzada. Tras llevar a cabo este laborioso proceso, vemos que los valores de las ocho pruebas varían levemente con respecto a la anterior destacando Zayas que obtiene un 0,15 en vez de 0,05, el resto de resultados de esta validación corresponden con los dados en el primer experimento con una desviación de 0,01/ 0,02 (ver tabla 5).

[14] Lo que hace esta característica es en palabras de Kestemont et al. (2016, 88): "La [..] GI no consiste en evaluar si dos documentos son simplemente similares en cuanto al estilo de escritura, dado un vocabulario de características estático, sino que pretende evaluar si dos documentos son significativamente más similares entre sí que otros documentos, a través de una variedad de espacios de características estocásticas y en comparación con selecciones aleatorias de los llamados autores distractores (Juola, 2015), también llamados 'impostores'."

[15] El script con el código tras esta función puede consultarse en el Github del Computational Stylistics Group: <https://github.com/computationalstylistics/stylo/blob/master/R/imposters.optimize.R>.

Alemán	0	Eslava	0.01		
Carvajal	0	Espinel	0.01	Montemayor	0
Castillo	0.02	Gregorio González	0.01	Quevedo	0
Cervantes	0.97	Lope	0	Salas Barbadillo	0.03
Suarez F.	0.01	Zayas	0,15	Montalbán	0

Tab. 4 | resultados de la función imposters en las pruebas individuales para cada autor.

En este caso solo hay un valor que se acerque a 1.00 (atribución) y es Cervantes. El único candidato (candidata) que parece mostrar cierta similitud es Zayas. Sigamos pues con el experimento para ver si estos resultados son estables. Realizaremos de nuevo la primera prueba utilizando la función *imposters*, pero esta vez vamos a emplear *Wurzburg Delta* (*aka Cosine Delta*) que es la medida de distancia que hemos aplicado desde los primeros experimentos pues según Eder (2018) muestra una mejoría notable con respecto a las otras (*Burrow's Delta, Eder's Delta*). También incluiremos la medida *Eder Delta* para poder contrastar los resultados. Tras realizar este proceso obtenemos los siguientes datos:

Autores	Wurzburg (Cosine)	Eder	Burrows
Alemán	0.03	0.02	0.00
Carvajal	0.00	0.00	0.00
Castillo	0.00	0.02	0.02
Cervantes	0.96	0.95	0.98
Eslava	0.00	0.04	0.01
Espinel	0.00	0.02	0.02
Gregorio González	0.00	0.01	0.00
Lope	0.00	0.01	0.00
Montalbán	0.00	0.00	0.00
Montemayor	0.00	0.00	0.00
Quevedo	0.05	0.00	0.00
Salas Barbadillo	0.00	0.00	0.01
Suarez Figueroa	0.00	0.01	0.02
Zayas	0.04	0.13	0.05

Tab. 5 | resultados de imposters con diferentes distancias, *Cosine Delta*, *Eder's Delta y Burrow's Delta.*

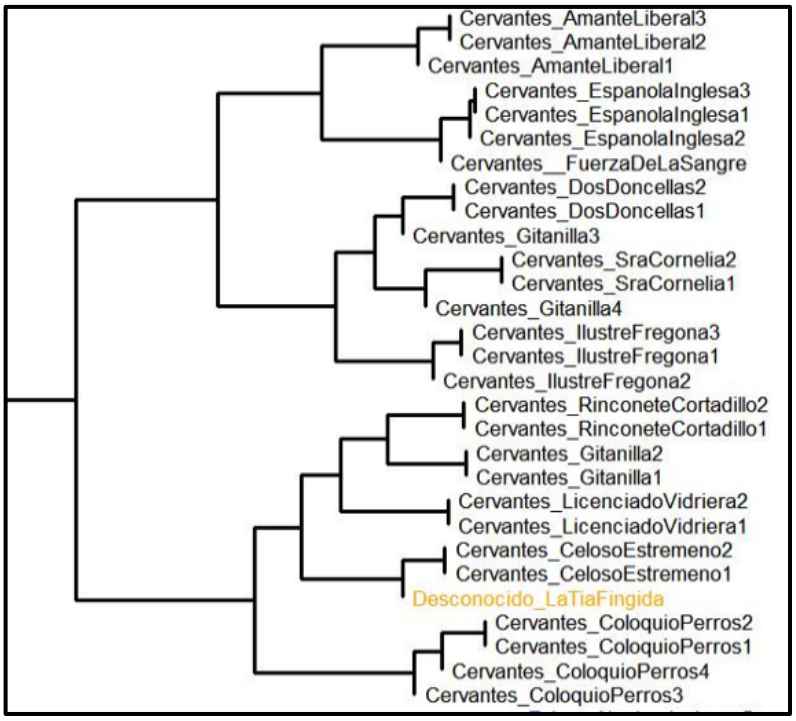
Para explicar los resultados, volvemos a la documentación de Eder (2018) quien afirma que cualquier resultado por encima de 0,5 puede ser indicativo de que la verificación de autoría para un candidato dado fue exitosa. Entonces, con los datos que obtuvimos podemos concluir que Cervantes es siempre reconocido como el autor más cercano a la obra dubitada con todas las distancias estadísticas utilizadas. No obstante, existen algunos autores que presentan un cierto grado de similitud. En un principio podríamos obviar este hecho dada la notable diferencia que presentan con respecto al alcalaíno, no obstante, hemos querido probar esta función para comprobar los resultados que nos da para Cervantes. Gracias a ella, ahora sabemos que utilizando diferentes distancias (*Eder's Delta, Burrow's Delta,*

Würzburg Delta) hay una sola atribución con una alta probabilidad pues la función *imposters()* apunta a que *La tía fingida* es de Cervantes. También vimos que la función *classify()* reconocía que el texto dubitado es del complutense. En definitiva, tanto la atribución como la verificación de autoría dan como resultado que el autor de esta novela es el Manco, aunque hay que insistir que en este corpus no están todos los posibles autores y textos que se podrían tener en consideración debido en parte a las limitaciones ya comentadas.

Por lo tanto, la única posible pega que vimos es que el programa parece tener algún problema para descartar de forma definitiva a varios de los autores destacando entre ellos Zayas que se queda siempre en la zona "de grises"[16]. Sin embargo, la atribución cervantina es tan constante y clara que no consideramos esto un impedimento. Además, en este punto conviene recordar el apunte que hace Eder (2018) con respecto a los resultados "cuando realizas el test nuevamente, el resultado final puede diferir levemente debido a la naturaleza estocástica del test". Por este motivo consideramos que debemos continuar indagando un poco. Además, nos interesa descubrir con qué obras se establecen los nexos entre *La tía fingida* y *Las Novelas Ejemplares* para poder realizar el análisis literario que nos permita comparar el estilo desde esa perspectiva.

En primer lugar, vamos a realizar un dendrograma utilizando el "corpus completo" y empleamos la función *stylo()* con las 500 MFW y la distancia *Cosine Delta*. Habíamos dejado para el final la comprobación de los resultados incluyendo a *La tía fingida*. La imagen que muestra el dendrograma es idéntica a la que se mostraba en el "corpus completo". La diferencia estriba en Cervantes, donde nos encontramos con *La tía fingida* que, aunque está sola en una ramificación, aparece asociada en un mismo *subcluster* con *El celoso extremeño*:

[16] Estas mismas pruebas fueron realizadas utilizando los textos completos (sin fragmentar) y en un rango de MFW más alto (3200). Esto nos obligaba a dejar fuera algunos autores de los cuales solo tenemos una obra, pero en ese caso la posible atribución a Zayas quedaba en todos los casos por debajo del rango inferior que daba la función *imposters.optimize()* y por lo tanto descartada.

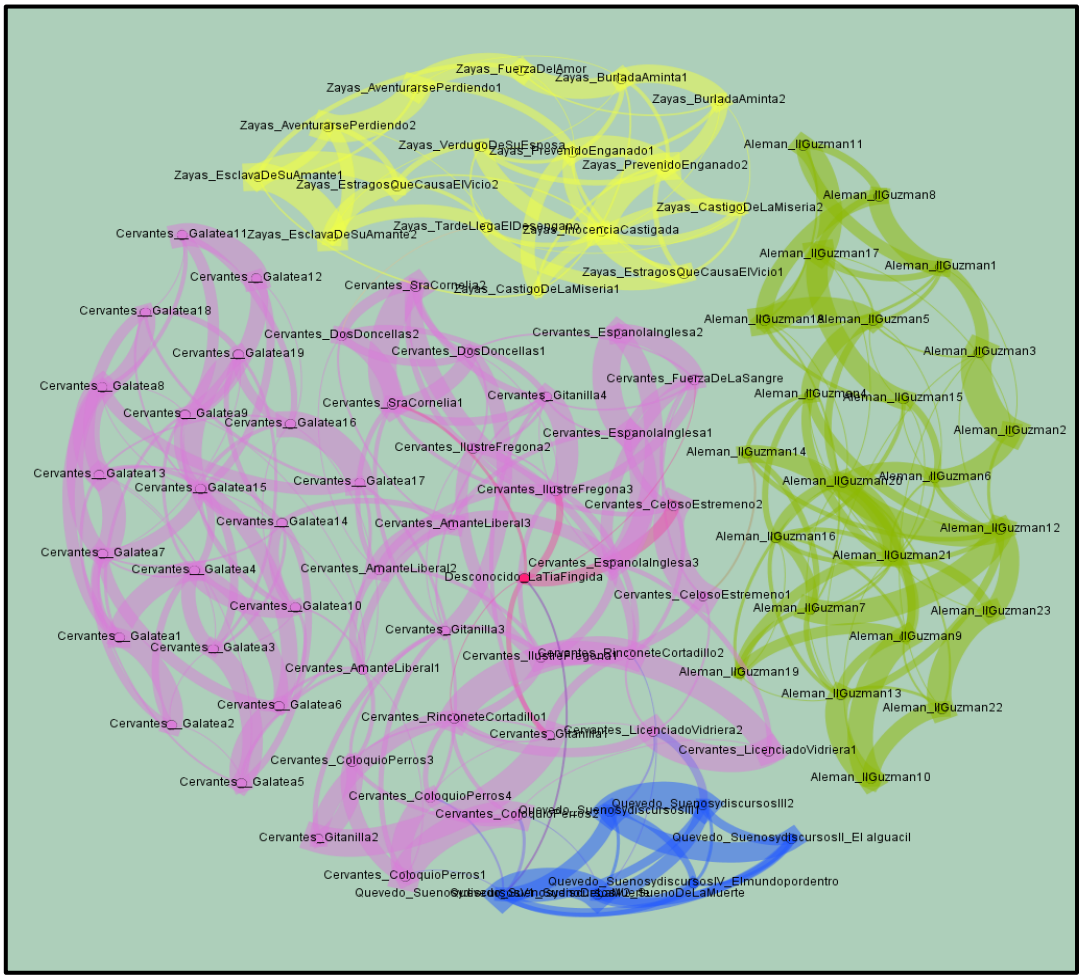


16 |detalle (obras de Cervantes) del análisis de clúster con el corpus "completo", 500 MFW, *Cosine Delta*.

Para seguir con la comprobación de los resultados podemos echar un vistazo a las relaciones entre todos los autores y el texto dubitado. Esto lo vamos a realizar en el siguiente apartado donde llevaremos a cabo un experimento que nos permitirá visualizar los parentescos entre textos.

4.4 Stylo.network () y Gephi

En la última prueba se puso de manifiesto que parece existir una relación entre la obra cervantina y *La tía fingida*. Además, sabemos que Zayas, Quevedo y Alemán parecen tener un cierto grado mínimo de similitud estilística con Cervantes. El problema es que no conocemos donde residen estas relaciones. Para obtener esta información ya hemos utilizado el dendrograma y este nos muestra como resultado un parentesco con *El celoso extremeño*. Ahora nos queda por comprobar un segundo método de visualización: el *Bootstrap Consensus Tree* (BCT) a partir de la frecuencia relativa de las 100 a las 500 MFW utilizando aquellos autores que la función *imposters()* no acababa de descartar del todo. La función *stylo()* nos va a servir para crear el gráfico. Vamos a utilizar la función *stylo.network()* con la que creamos un BCT cuyos resultados exportaremos a *Gephi* (cf. Bastian, Heymann y Jacomy 2009). Se trata de dos tablas de Excel en formato csv conteniendo los datos de los nodos y las aristas. Con ello empleamos el *layout* de Fruchterman y Reingold (1991) con el que, tras ajustar los parámetros de visualización, obtenemos las siguientes imágenes:



17 | gráfico *Gephi* del BCT de cuatro autores (Zayas, Cervantes, Quevedo, Alemán) con distancia *Cosine Delta* (500 MFW)

18 | detalle de la imagen 17, *La tía fingida* y sus nexos.

A través de este gráfico podemos ver con mayor claridad las relaciones que tiene la TF con otros textos en base a los parámetros que seleccionamos con la función *stylo()*. Vemos efectivamente que no tiene ningún nexo con Zayas, pero sí alguna conexión con el primer fragmento de la quinta parte de los *Sueños y discursos*, "el sueño de la muerte" de Quevedo. Sin embargo, las relaciones con las *Novelas Ejemplares* de Cervantes son mucho más notorias y variadas entre las que destaca una relación que sobresale por encima del resto: se trata de la segunda parte de *El celoso extremeño*.

En este punto dada la evidente conexión que aparenta tener *La tía fingida* con el de *El celoso extremeño* procedimos en este punto a llevar a cabo el análisis literario de ambas obras y su confrontación con el objetivo de poder descifrar algunos de los elementos que se esconden tras los datos que hemos visto hasta ahora. Los resultados de este análisis pueden ser consultados en Cienfuegos-Pérez (2022).

5. Conclusiones

En nuestros experimentos con *stylo* procuramos que la organización de textos indubitados fuera hecha por autor para anteponer la relevancia del empleo de determinadas palabras funcionales a cualquier otro aspecto como marcador del estilo autorial. A través de los experimentos vimos que la organización de los textos por autor de las obras no dubitadas era correcta[17]. Al introducir *La tía fingida* en el corpus después de realizar la validación y demostrar que el porcentaje de fiabilidad con las 500 MFW y la distancia *Cosine Delta* era del 100% vimos que esta obra era asociada a Cervantes y más particularmente a *El celoso extremeño*. Además, llevamos a cabo una serie de experimentos estilométricos utilizando diferentes funciones para determinar la autoría. El primero de ellos fue de atribución de autoría (función *classify*) seguido de la verificación de autoría (función *imposters*). En ambos casos la TF fue asociada a Cervantes sin que otro autor mostrara resultados tan evidentes. Asimismo, el empleo de diferentes distancias estadísticas como *Delta de Burrows* y *Delta de Eder* mostró que la opción de la autoría cervantina es constante. Para ver dónde se encontraban los nexos entre las obras de Cervantes y otros autores de forma más concreta se utilizaron los parámetros ya indicados, se empleó la función *stylo.network* con la que se creó un BCT (*Bootstrap Consensus Tree*) y se exportaron los datos a *Gephi* obteniendo así un gráfico que evidenció que la relación entre *La tía fingida* y *El celoso extremeño* era más notoria que otras. En este punto, decidimos analizar las dos obras que presentaban la interrelación más destacada.

En el análisis literario nos centramos sobre todo en aspectos técnicos y expresivos que no se pueden medir cuantitativamente como son los contenidos, y las figuras retóricas. Pudimos observar que en el plano de los contenidos ambas obras pertenecen a un género, tipo y temática similar y utilizan técnicas narrativas semejantes. En el análisis de la expresión se reconocieron también algunas

[17] Es necesario precisar que, aunque no se hayan incluido en el trabajo, se llevaron a cabo numerosas pruebas previas utilizando los textos sin fragmentar, con diferentes configuraciones de las MFW y utilizando diversas medidas de distancia. Es así como llegamos hasta los parámetros que se incluyen en los experimentos finales.

semejanzas a través del uso de figuras retóricas compartidas, pero se evidenciaron divergencias notorias que están recogidas en Cienfuegos-Pérez (2022) dentro del apartado "Síntesis del análisis literario" y que se podrían englobar en la afirmación: *El celoso extremeño* se sirve de recursos que presentan una mayor complejidad comparados con los de *La tía fingida*.

Llegados a este punto solo nos queda responder a la gran pregunta ¿quién fue el autor de *La tía fingida*? Pues bien, hay determinados aspectos a considerar antes de hacer cualquier afirmación terminante sobre su identidad. En primer lugar, la elección del corpus no incluye todos los textos que serían ideales para una investigación de este tipo ni tampoco todos los autores que se podrían considerar. Además, la fiabilidad de los textos es solo relativa pues no cumplen siempre la premisa de poseer muestras incorruptas que señalaba Juola (2008). La problemática de los textos del Siglo de Oro muestra la necesidad de seguir trabajando en la calidad, la digitalización y la puesta a disposición de los mismos para el ámbito académico.

Además de esto, la decisión de contrastar determinados textos y no otros en el experimento con *classify* no es 100% objetiva. En el conjunto de los textos a comprobar (*secondary_set*, o *test_set*) elegimos ciertas obras para ver si el algoritmo las reconocía y las asignaba a sus respectivos autores, pero ¿Qué pasa si probamos con textos que se consideren estilísticamente más lejanos a un autor? Desafortunadamente la herramienta no ofrece ninguna solución a esto más allá de realizar la prueba con diferentes conjuntos de textos en el *secondary_set*. [18]

Entendidos estos escollos, podemos entrar en consideraciones. En base a las pruebas estilométricas solo tenemos una respuesta posible: *La tía fingida* es con una alta probabilidad obra de Cervantes. Sin embargo, en vista de los límites ya comentados lo que podríamos afirmar es más bien que: de entre este conjunto de autores y con el corpus seleccionado el único autor posible es Cervantes. Por otra parte, desde la perspectiva literaria las dos obras comparadas mostraron convergencias y divergencias. En la obra no dubitada se pudo observar una obra técnicamente más elaborada ¿Se puede desde el análisis literario de ciertas figuras retóricas y algunos elementos morfológicos decidir la autoría de la obra? Consideramos que no, tampoco era nuestro objetivo, pero sí sirve en conjunto con los experimentos estilométricos para llegar a ciertas hipótesis.

La primera de estas hipótesis es que esta obra podría ser una novela que Cervantes compuso con anterioridad a sus otras novelas breves, un periodo en el cual el autor aún estaba puliendo su pluma en este género. No olvidemos que la fecha de escritura aproximada que manejamos es la de la compilación de los manuscritos que realizó Porras de la Cámara para el cardenal Niño de Guevara (1604-1606). Por lo tanto, no sabemos a ciencia cierta cuando fue redactada ni en qué orden cronológico con respecto a los otros textos aparecidos en el manuscrito Porras.

[18] Evidentemente, realizamos pruebas con diferentes "sets" sin embargo consideramos que la selección es siempre arbitraria. De todos modos, las diferentes configuraciones clasificaban los textos de autor conocido correctamente y atribuían la TF a Cervantes. En este contexto la utilización de textos *estilísticamente* lejanos al autor puede servir en la investigación para validar los resultados.

Esto explicaría las coincidencias a la vez que visto en perspectiva podría servir como ejemplo de análisis de la evolución del alcalaíno. Así pues, aunque compartimos aquí su hipótesis de la autoría cervantina diferimos de Marín (1944, 40) en que la novela sea excelente, aunque emplea un vocabulario variado, numerosas técnicas expresivas y representa una buena muestra del género de la novela breve primigenia. Consideramos que no tiene el grado de complejidad y excelencia en el uso de los recursos técnicos y expresivos de los que sí hace gala *El celoso extremeño*. Concordamos por lo tanto en parte con la opinión de Foulché-Delbosc (1899) al reconocer cierta distancia estilística con respecto a Cervantes, pero al contrario que el hispanista francés nos parece que esto no sirve como argumento para no considerarla del complutense, no vemos tanta distancia como para poder afirmar rotundamente que no es obra suya.

La segunda presunción es que la obra fuera de un temprano imitador como hipotetizaba Márquez Villanueva (1995). Sin embargo, aunque un imitador pudiera emplear el vocabulario y las expresiones más características de Cervantes resulta poco creíble que prestara atención a todas las palabras funcionales y su frecuencia relativa. Opinamos por lo tanto que la posibilidad de un imitador no se puede descartar utilizando solamente la comparación de determinadas expresiones, pasajes textuales e índices verbales. Estas, en nuestra opinión, no sirven tampoco como argumentos definitivos para desmentir u otorgar la autoría de la TF como hicieron por ejemplo Icaza (1916) y De Val (1953) para rechazarla como obra de Cervantes o Madrigal (2003) para adjudicársela.

Otro aspecto relevante que nos queda por comentar es el de justificar para qué serviría considerar la TF de Cervantes ¿qué aporta esto a la literatura? Aunque desde nuestro punto de vista la obra nos parezca más prosaica que el resto de novelas cortas, el conocimiento de sus parecidos y diferencias nos da cierta perspectiva sobre la evolución del autor. También vemos en sus contenidos crítica social, dicotomía de espacios casa-calle, una serie de ideas sobre la mujer, imagen del mundo estudiantil, entorno del hampa y otras múltiples posibilidades de análisis que podrían ser contrastadas con el resto de su corpus conformando en su totalidad el significado último de determinados propósitos e ideas, el empleo y desarrollo de ciertas técnicas narrativas, etc. No olvidemos que el genial Cervantes es considerado el creador de la novela breve española[19] por lo que saber más sobre el proceso y la evolución de este género es descubrir también las estrategias que se esconden tras su nacimiento.

Llegamos pues al final habiendo visto que nuestra hipótesis inicial no estaba desencaminada. La estilometría ha demostrado ser una herramienta útil en el proceso de atribución de autoría, que pese a sus límites nos ha ayudado a adoptar nuevas perspectivas literarias. Sin ella no hubiéramos encontrado con tanta rapidez paralelismos con *El celoso extremeño*. La hipótesis de la autoría cervantina de la TF a pesar de que haya sido desdeñada por algunos críticos que no reconocían en esta obra la genialidad de Cervantes parece tener bastante sustento al menos mientras no encontremos una alternativa con textos que podamos contrastar y que presente

[19] Por ejemplo, Díez-Echarri y Roca Franquesa (1972, 222) o Montero Reguera (2006, 166)

unos resultados semejantes a los arrojados por el complutense en nuestros experimentos. Hasta entonces, si nos decantamos por la primera de las sospechas, la de la novela cervantina de composición temprana, tendríamos que admitir que quizás la excelencia de los trabajos conocidos del ingenioso Miguel de Cervantes es fruto de la evolución y mejoría del mismo o dicho de otro modo que sin esfuerzo y trabajo ni el gran talento de Miguel de Cervantes Saavedra hubiera alcanzado la calidad de las *Novelas ejemplares* ni dado vida al ingenioso caballero y a su leal escudero que viven hoy en el mundo entero.

Bibliografía

ARELLANO-AYUSO, Ignacio. 1997. "Las aventuras del texto: del manuscrito al libro en el Siglo de Oro." En *Unum ET Diversum: Estudios en honor de Ángel-Raimundo Fernández González*, ed. Spang, Kurt, 41-66, Pamplona: Ediciones Universidad de Navarra.

BASTIAN, Mathieu, Sebastien Heymann & Mathieu Jacomy. 2009. "Gephi: an open-source software for exploring and manipulating networks." *Proceedings of the international AAAI conference on web and social media* 3 (1), 361-362.

BLASCO PASCUAL, Francisco & Cristina Ruiz Urbón. 2009. "Evaluación y cuantificación de algunas técnicas de atribución de autoría en textos españoles." *Castilla: Estudios de literatura* 0: 27-47. <https://doi.org/10.24197/cel.0.2009> (09.10.21).

CANAVAGGIO, Jean. 1992. *Cervantes: en busca del perfil perdido*. Besalú: Llibres Detot.

CEREZO SOLER, Juan & José Calvo Tello. 2019. "Autoría y estilo. Una atribución cervantina desde las humanidades digitales. El caso de La conquista de Jerusalén." *Anales Cervantinos* 51, 231-250.

CERVANTES SAAVEDRA, Miguel. 2018. *La tía fingida*. Ediciones Cátedra.

CERVANTES SAAVEDRA, Miguel. 1818. *La tia fingida: novela inédita*. Edición digital basada en C. F. Franceson y F. A. Wolf, Berlín: Librería de G. C. Nauck. <https://www.cervantesvirtual.com/nd/ark:/59851/bmc1z4h1>.

CERVANTES, Miguel. 2018. *Novelas ejemplares*. MVP.

CIENFUEGOS PÉREZ, Alejandro (2022). *Atribución de autoría y Humanidades Digitales: Métodos de estilometría y aplicación al Siglo de Oro*. Tesis de Master, Georg-August-Universität Göttingen. DARIAH-DE. <https://doi.org/10.20375/0000-000f-3246-a>.

DE ICAZA, Francisco A. 1916. *De cómo y por qué "La tía fingida" no es de Cervantes: y otros nuevos estudios cervánticos*. Madrid: Imprenta Clásica Española.

DE VAL, Manuel Criado. 1953. *Análisis verbal del estilo: índices verbales de Cervantes, de Avellaneda y del autor de "La tía fingida"*. Madrid: Consejo superior de investigaciones científicas.

DESAGULIER, Guillaume. 2017. *Corpus linguistics and statistics with R*. Berlin: Springer International Publishing.

DÍEZ ECHARRI, Emiliano & José María Roca Franquesa. 1972. *Historia de la literatura española e hispanoamericana*. Madid: Aguilar.

EDER, Maciej. 2018. "Authorship verification with the package *stylo*." *Computational Stylistics Group Blog*. <https://computationalstylistics.github.io/blog/imposters/> (18.11.2022).

EDER, Maciej. 2017. "Short Samples in Authorship Attribution: A New Approach." *Digital Humanities. Montreal 8.-11.8.2017.* <https://dh2017.adho.org/abstracts/341/341.pdf> (18.11.2022)

EDER, Maciej. 2015. "Does size matter? Authorship attribution, small samples, big problem." *Digital Scholarship in the Humanities* 30 (2), 167-182.
<https://doi.org/10.1093/llc/fqt066> (22.11.22)

EDER, Maciej & Jan Rybicki. 2011. "Stylometry with R." *Digital Humanities*, 308-310.

EDER, Maciej, Jan Rybicki & Mike Kestemont 2016. "Stylometry with R: a package for computational text analysis." *The R Journal* 8 (1), 107–121.
<https://journal.r-project.org/archive/2016/RJ-2016-007/index.html> (15.08.21).

FOULCHE-DELBOSC, Raymond. 1899."Etude sur 'La tía fingida': I. le manuscrit Porras.- II. le manuscrit de la Colombine.- III. le texte de La tía fingida.- IV. l'attribution a Cervantes." *Revue hispanique: recueil consacré à l'étude des langues, des littératures et de l'histoire des pays castillans, catalans et portugais* 6 (19), 256-306.

FRUCHTERMAN, Thomas MJ & Edward M. Reingold. 1991. "Graph drawing by force-directed placement." *Software: Practice and experience* 21 (11), 1129-1164.

GALLARDO, Bartolomé José. 1835. "La tía fingida ¿es novela de Cervantes?" *El Criticón, papel volante de Literatura y Bellas artes* 1, 1-43.

GRIES, Stefan Th. 2016. *Quantitative corpus linguistics with R: A practical introduction*. New York: Routledge.

GRIES, Stefan Th. 2013. *Statistics for Linguistics with R. A practical introduction*. Berlin/Boston: De Gruyter Mouton.

HERNÁNDEZ-LORENZO, Laura. 2019. "Poesía áurea, estilometría y fiabilidad: métodos supervisados de atribución de autoría atendiendo al tamaño de las muestras." *Caracteres: estudios culturales y críticos de la esfera digital* 8 (1), 189-228.

HOLMES, David I. & Judit Kardos."Who was the author? An introduction to stylometry." *Chance* 16 (2), 5-8.

JOCKERS, Matthew L. 2014. *Text Analysis with R for Students of Literature*. Cham: Springer International Publishing.

JUOLA, Patrick. 2015. "The Rowling case: a proposed standard analytic protocol for authorship questions." *Digital Scholarship in the Humanities* 30 (suppl_1), i100-i113.

JUOLA, Patrick. 2008. "Authorship attribution." *Foundations and Trends® in Information Retrieval* 1 (3), 233-334.

KESTEMONT, Mike et al. 2016. "Authenticating the writings of Julius Caesar." *Expert Systems with Applications* 63, 86-96.

KESTEMONT, Mike. 2011. "Een stylometrisch onderzoek naar Jan van Boendales auteurschap voor de Brabantse yeesten." *Revue belge de Philologie et d'Histoire* 89 (3), 1019-1048.

KOPPEL, Moshe & Yaron Winter. 2014. "Determining if two documents are written by the same author." *Journal of the Association for Information Science and Technology* 65 (1), 178-187.
<http://dx.doi.org/10.1002/asi.22954> (18.11.2022).

LEVSHINA, Natalia. 2015. *How to do linguistics with R. Data Exploration and Statistical Analysis*. Amsterdam/Philadelphia: Benjamins.

LUCÍA MEGÍAS, José Manuel. 2018. "El códice Porras (casi) recuperado (la copia del Cigarral del Carmen de" "La tía fingida"). " *Anales cervantinos* 50, 333-351.

MADRIGAL, José Luis. 2003. "De cómo y por qué La tía fingida es de Cervantes." *Artifara* 2, s. p.

MARÍN, Luis Astrana. 1944. "Sobre La tía fingida" en *Cervantinas: y otros ensayos*. Vol. 6. Afrodisio Aguado, sa, 1944. 37-48.

MÁRQUEZ VILLANUEVA, Francisco. 1995. *Trabajos y días cervantinos*. Vol. 2. Biblioteca Estudios Cervantinos.

MAREGALLI, Franco. 1992. *Introducción a Cervantes* Vol. 111. Barcelona: Ariel.
MONTERO REGUERA, José. 2006. "El nacimiento de la novela corta en España (la perspectiva de los editores)." *Lectura y signo* 1, 165-175.
RICO, Francisco. 2000. *Imprenta y crítica textual en el Siglo de Oro, ed. lit. Pablo Andrés Escapa y Sonia Garza,* Valladolid, Centro para la Edición de los Clásicos Españoles.
RIßLER-PIPKA, Nanette. 2018. "Die Digitalisierung des goldenen Zeitalters – Editionsproblematik und stilometrische Autorschaftsattribution am Beispiel des Quijote Abstracts." *Zeitschrift Für Digitale Geistes-wissenschaften* 2018, s.p.
<https://zfdg.de/2018_004_v1> (18.11.2022).
RUEDA, José Manuel. 2019 "Estilometría y la Edad Media castellana" En *Theorien von Autorschaft und Stil in Bewegung: Stilistik und Stilometrie in der Romania (=Romanische Studien*, Beiheft 6), ed. Rissler-Pipka, Nanette, 49-74.
RYBICKI, Jan and Maciej Eder. 2012. "Stylometry: Computer-Assisted Analysis of Literary Texts." Workshop *Culture & Technology, European Summer School in Digital Humanities*. Leipzig University.
WINTER, Bodo. 2019. *Statistics for linguists: An introduction using R*. New York: Routledge.

Páginas web y recursos en línea.

Nota: los enlaces a las versiones digitales de los libros empleados para el corpus se encuentran en el anexo "Corpus" en Cienfuegos-Pérez (2022).
BVMC (Biblioteca virtual Miguel de Cervantes), Guía de obras atribuidas a Miguel de Cervantes Saavedra (15.06.21). En línea:
<http://www.cervantesvirtual.com/portales/miguel_de_cervantes/estados_atribuciones/>.
Los gráficos Gephi fueron creados con el software de la página:
<https://gephi.org/>.
Los textos digitales del subcorpus *otros textos en prosa* han sido recuperados de Fradejas Rueda, 7 Partidas Digital, Github:
<https://github.com/7PartidasDigital>.
ForTEXT & CATMA. „Tutorial: Stylo zur Analyse des Autoren-Stils nutzen" forTEXT & CATMA. 1 de agosto de 2019. Video, 12m22s.
<https://youtu.be/LuJe67898z0>.
Tutorial para el empleo de Stylo: Rueda, José Manuel Fradejas. "*Cuentapalabras. Estilometría y análisis de texto con R para filólogos*." (2020) Recuperado el 08.08.21 de:
<http://www.aic.uva.es/cuentapalabras/>.

Resumen

El presente trabajo expone el empleo de algunas de las herramientas que ofrecen las humanidades digitales para tratar de aportar nuevas perspectivas a un caso de autoría largamente discutido: el de la novela breve *La tía fingida* atribuida a Cervantes. Para llevar a cabo dicha tarea se mostrará el proceso de compilación del corpus y algunos de los problemas de los textos del Siglo de Oro. A través del uso de la herramienta *stylo* se presentarán los diferentes métodos y experimentos llevados a cabo, tanto de atribución como de verificación de autoría. Los resultados de la estilometría mostraron que *El celoso extremeño* y *La tía fingida* parecen tener un parentesco estilístico. Para tratar de confirmarlo se contrastarán ambas obras desde la perspectiva estilométrica y la literaria.

Abstract

This paper presents a series of experiments using tools offered by the digital humanities to bring new perspectives to a long-discussed case of authorship attribution: regarding the short novel *La tía fingida* ('The Pretended Aunt') attributed to Cervantes. The process of compilation of the corpus and some of the general problems of authorship attribution in the Spanish Golden Age are also discussed. By trying out the method of stylometry and using *stylo* as a tool, the different functions and experiments carried out, both for attribution and authorship verification, are presented. The results show that *El celoso extremeño* and *La tía fingida* seem to have a stylistic relatedness. To try to confirm this, both works will be contrasted from a stylometric and literary criticism perspective.

hosted by Hamburg University Press

Nächste Nummer

Künste des Dazwischen

Graphische Literatur und visuelle Poesie der Romania als Genres 'en marge'

hrsg. von Julia Dettke & Jasmin Wrobel

Sommer 2023

10

www.apropos-romania.de

ISSN: 2627-3446