

***Digital Humanities* auf dem Prüfstand**

Analysemethoden für digitale Korpora von E-Mails über Internetseiten bis zu Wikipedia

Ursula Reutner

Abstract

The digital revolution has changed our lives in many ways. In academics, this change manifests itself in the development of new disciplines, as well as in extended research areas and new analytical methods for those already established. In the humanities, the emergence of digital humanities has attracted much attention and led to a debate about the value and explanatory power of this new field. Are digital technologies no more than analytical tools for more easily and reliably generating knowledge that could in the past also be achieved, only with greater effort? Or do these methods and procedures instead allow results that are simply not conceivable otherwise? These questions are reason enough to consider some research projects in Romance linguistics with regard to the knowledge gained through digitalization and the findings as to its consequences, thus testing the impact of combining *digital* and *humanities*. Due to the fuzziness of the term *digital humanities*, it is necessary to first clarify what it actually means. After looking at current definitions and efforts to explain the term, we present some research methods involving different ways of analyzing digital text corpora such as e-mails, webpages, and Wikipedia. We outline the methods applied and the insights achieved, and we subsequently critically examine each approach. The result is a comprehensive overview of a clearly defined field of study in the humanities that demonstrates the opportunities and limits of digital humanities in this specific area.

Keywords: digital humanities, corpus-linguistics, discourse analysis, methodology, social media, Wikipedia, e-mail, web pages

Keywords: *Digital Humanities*, Korpuslinguistik, Diskursanalyse, Methodik, Social Media, Wikipedia, E-Mail, Internetseiten

1 Einleitung

Die digitale Revolution hat unser Leben in vielerlei Hinsicht verändert. In den Wissenschaften zeigt sich dies in der Entstehung neuer Wissenschaftsdisziplinen und in erweiterten Untersuchungsgebieten sowie Analysemethoden für bereits etablierte Disziplinen. Letztere werden gerne unter dem Schlagwort *Digital Humanities* zusammengefasst, das zugleich emotional aufgeladen ist. Verfechter erhoffen sich von der digitalen Auswertung einen besonderen Erkenntnisgewinn, Kritiker sehen einen im Vergleich zum Ertrag ungerechtfertigten Aufwand. Befürworter halten den Wert der Analyse großer Datenmengen hoch, Gegner den Mehrwert der genauen Analyse von Einzeldaten. Apologeten führen die Objektivierung durch maschinelles Zählen, Berechnen und Visualisieren an, Kontrahenten vermissen hermeneutisches Deuten sowie die intensive und zugleich intuitive Auseinandersetzung mit den einzelnen Daten. Wie so oft liegt die Wahrheit irgendwo dazwischen und lässt sich die Entscheidung über die Methode nur abhängig vom jeweiligen Erkenntnisziel treffen.

Die aufgeworfenen Fragen sind dennoch Anlass genug, einmal romanistische Forschung im Hinblick auf den Erkenntnisgewinn durch Digitalisierung zu betrachten und zugleich die Verbindung aus Geisteswissenschaften und Digitalem in Form von *Digital Humanities* auf den Prüfstand zu stellen. Dabei ist erst einmal zu klären, was unter den vagen Begriff der *Digital Humanities* eigentlich zu fassen ist. Dies erfordert zunächst einen Blick auf kursierende Definitionen und Erklärungsversuche und erlaubt im Anschluss Anmerkungen zur Tradition der *Digital Humanities* in der Sprachwissenschaft sowie zu Fragen der Abgrenzung. Auf dieser Basis werden exemplarisch einige Forschungsarbeiten zu originär digitalen Korpora vorgestellt: Arbeiten zur sprachlichen und inhaltlichen Analyse von E-Mails, zur Bildlichkeit und Farbgestaltung von Internetseiten und zu verschiedenen Aspekten der Online-Enzyklopädie Wikipedia, darunter die Sprache, Bildlichkeit und der Aufbau der Artikelkörper, die Formulierung der Artikeltitel und der Grad der sprachlichen Gewalt auf den Diskussionsseiten. In jedem Fall werden das Korpus, das Erkenntnisziel, die Methode und der Erkenntnisgewinn skizziert und das Vorgehen kritisch hinterfragt. Dadurch ergibt sich ein umfassendes Bild zu einem klar definierten Forschungsgebiet der Geisteswissenschaften, das einige Chancen und Grenzen der *Digital Humanities* in diesem Bereich aufzeigt.

2 Digital Humanities

2.1 Von der Vagheit der Definition

Worüber sprechen wir überhaupt, wenn wir von *Digital Humanities* reden? So leicht zu beantworten ist diese zunächst einfach klingende Frage nicht, denn eine klare Definition ist schwer zu finden. Nicht umsonst setzt sich ein ganzes Werk mit dem sprechenden Titel *Defining Digital Humanities* mit dem Thema auseinander (Terras/Nyhan/Vanhoutte 2013). „Answering the question ‚What is digital humanities?‘ continues to be a rich source of intellectual debate for scholars“, halten die Herausgeber einleitend fest (Nyhan/Terras/Vanhoutte 2013, 6) und stellen zugleich die Sinnhaftigkeit einer Definition in Frage. Eine solche sei nicht nur unmöglich, sondern eventuell auch unproduktiv, da sie das gerade erst aufkommende Feld unnötig begrenze:

Indeed, at the current time, not only does a comprehensive definition appear to be impossible to formulate, when the breadth of work that is covered by a number of recent and forthcoming companions is considered [...], it might ultimately prove unproductive, by fossilising an emerging field and constraining new, boundary-pushing work. (Nyhan/Terras/Vanhoutte 2013, 6)

So überrascht es kaum, dass auch die einschlägigen Handbücher eher das Tätigkeitsfeld umreißen als eine klare Definition liefern (cf. Schreibman/Siemens/Unsworth 2004; McCarty 2005; Unsworth/Siemens/Schreibman 2016; Jannidis/Kohle/Rehbein 2017). Definitiorische Einigkeit besteht lediglich darin, dass es sich um eine Verbindung aus Geisteswissenschaften und Informatik handelt (cf. Definitionen 1–6). Darüber hinausgehend bleiben die Deutungen vage und liefern ein „ungemein breites Bild“ (2), das teils auf geisteswissenschaftliche Forschung reduziert wird, deren Ergebnisse anderweitig „nicht zu erzielen wären, oder nur auf einer niedrigeren Ebene intersubjektiver Wahrnehmbarkeit“ (3). Sie erwähnen die Anwendung, Entwicklung und Erforschung computergestützter Verfahren (4–5) sowie die mögliche Konsequenz eines generellen Wandels in den Geisteswissenschaften (6).

1. Im weitesten Sinne handelt es sich dabei um die Beantwortung geisteswissenschaftlicher Fragestellungen mithilfe digitaler Methoden. (DARIAH-DE 2015, 8)
2. Verstehen wir die *Digital Humanities* als die Summe aller Versuche, die Informationstechniken auf den Gegenstandsbereich der Geisteswissenschaften anzuwenden, ergibt sich ein ungemein breites Bild. (Thaller 2017, 13)
3. Unter *Digital Humanities* verstehen wir alle Arten geisteswissenschaftlicher Forschung, die versuchen, durch den Einsatz moderner Informationstechnologien oder aus der Informatik abgeleiteter Instrumente inhaltliche Ergebnisse zu

erzielen, die ohne den Einsatz dieser Instrumente weder gar nicht zu erzielen wären, oder nur auf einer niedrigen Ebene intersubjektiver Nachprüfbarkeit. (Thaller 2014)

4. [...] I propose a twofold definition: First, DH encompasses all kinds of research in the Humanities that partly gains its findings from applying computer-based procedures, practices, and tools. In this understanding, *Digital Humanities* is pure Humanities scholarship [...]. Second, DH encompasses the design, development, and generalization of these computer based procedures, practices and tools, as well as the study of their underlying theories and models. In this understanding, Digital Humanities is rather an auxiliary science [...]. (Rehbein 2020, 252)
5. Die Forscherinnen und Forscher in diesem Feld beschäftigen sich damit, neue Entwicklungen in der Informatik auf ihre Verwendbarkeit in den Geisteswissenschaften zu prüfen oder eigenständig geeignete Verfahren zu entwickeln, und sie erforschen die Algorithmen und Datenstrukturen, die sich als geeignet erwiesen haben. (Jannidis/Kohle/Rehbein 2017, XI)
6. [...] harnessing computing power to facilitate, improve, expand and perhaps even change the way humanists work. (Gardiner/Musto 2015, 4–5)

Weitergehende Erklärungen weisen zudem auf Beteiligte wie Kommunikations-, Grafik- und Bilddesigner (7) sowie interdisziplinäre Fragestellungen (8) hin.

7. Digital Humanities projects most closely involve communication/graphic/visual designers who are concerned with the symbolic representation of language, the graphical expression of concepts, and questions of style and identity. (Burdick/Drucker/Lunenfeld/Presner/Schnapp 2012, 12)
8. [Digital humanities] asks what it means to be a human being in the networked information age and to participate in fluid communities of practice, asking and answering research questions that cannot be reduced to a single genre, medium, discipline, or institution. (Burdick/Drucker/Lunenfeld/Presner/Schnapp 2012 xii–xiii)

Die Definitionen der weltweit am häufigsten konsultierten Enzyklopädie sprechen ähnlich vage von „a variety of topics“ (9), eröffnen die Spanne zwischen dem niedrigschwelligen Einsatz digitaler Ressourcen bis hin zu großen Data-Mining-Projekten (9–10) und schließen zudem die „Reflexion“ über die Methoden und ihre Anwendung ein (10).

9. Digital Humanities is an area of scholarly activity at the intersection of computing or digital technologies and the disciplines of the humanities. Developing from the fields of humanities computing, humanistic computing, and digital

humanities praxis, [It] developed out of humanities computing and has become associated with other fields, such as humanistic computing, social computing, and media studies. [It] embraces a variety of topics, from curating online collections of primary sources (primarily textual) to the data mining of large cultural data sets. (Wikipedia EN 2023)

10. systematische Nutzung computergestützter Verfahren und digitaler Ressourcen in den Geistes- und Kulturwissenschaften sowie die Reflexion über deren Anwendung. (Wikipedia DE 2023)

Die Liste an Definitionen ließe sich beliebig fortsetzen (cf. z. B. Gibbs 2013, 290), ohne das vage Ergebnis zu konkretisieren. Daran ändert auch eine mögliche Unterscheidung zwischen der Untersuchung von Digitalisierungsphänomenen mit geisteswissenschaftlichen Methoden, *Humanities for Digitalization*, kurz H4D, und der Anwendung digitaler Methoden auf geisteswissenschaftliche Fragestellungen, *Digitalization for Humanities*, kurz D4H, wenig. Der Versuch einer Annäherung an die Begriffsdefinition mithilfe struktureller Semantik liefe daher ins Leere, sodass allein ein prototypensemantischer Ansatz Klärung bringen dürfte, der lediglich Klarheit bezüglich der idealtypischen Ausprägung des Faches erfordert, wozu am Ende dieses Beitrags ein Vorschlag formuliert wird.

2.2 Traditionen in der Sprachwissenschaft

Beschränken wir uns an dieser Stelle auf den unstrittigen Aspekt der Verbindung aus Geisteswissenschaften und Informatik bzw. auf die auch im Terminus *Digital Humanities* selbst enthaltene Begrifflichkeit aus Digitalem und Geisteswissenschaften, so lässt sich zunächst festhalten, dass eine solche in der Sprachwissenschaft intensiv praktiziert wurde, bevor das Schlagwort selbst in aller Munde war. Beide Kernelemente aus der Definition (10), „computergestützte Analysemethoden [Verfahren]“ und „digitale Ressourcen“, finden sich seit ihrem Aufkommen auch in der sprachwissenschaftlichen Forschung, die in beiden Richtungen der Kooperation zwischen Geisteswissenschaften und Informatik, also sowohl in D4H, als auch in H4D, präsent ist.

Computergestützte Analysemethoden werden in der Sprachwissenschaft traditionell in der Korpuslinguistik eingesetzt, die damit einen Teilbereich der *Digital Humanities avant la lettre* darstellt. Gegenüber manuellen Auswertungsverfahren bietet die automatisierte Auswertung einige Vorteile: den Einbezug größeren Datenmaterials, die Reduzierung von menschlichem Versehen, die Anwendung statistischer Verfahren mit der Möglichkeit, Muster zu erkennen und Ergebnisse auf statistische Signifikanz zu prüfen. Mit der Etablierung

der modernen Korpuslinguistik ist die Sprachwissenschaft damit bereits seit längerem im Bereich der *Digitalization for Humanities* (D4H) verortet.

Im Hinblick auf digitale Ressourcen ist zwischen digitalisierten und originär digitalen Texten zu unterscheiden. Die Digitalisierung von Texten aus der nicht digitalen Welt erlaubt zum einen, Kulturgüter zu bewahren und einer breiteren Öffentlichkeit zugänglich zu machen. Zum anderen ist sie eine Voraussetzung für die computergestützte Analyse und eine automatisierte Verknüpfung von Daten. Umfangreiche Digitalisierungsprojekte gelten im Bereich der Romanistik derzeit zum Beispiel Sprachatlanten und etymologischen Wörterbüchern wie dem *Romanischen Etymologischen Wörterbuch*, dem *Französischen Etymologischen Wörterbuch* oder dem *Lessico etimologico italiano* (↑Zacherl; ↑Husson/Kremer und ↑Nannini) und profitieren zweifellos vom Aufschwung der *Digital Humanities*.

Digitale Ressourcen im Sinne originär digitaler Texte sind wiederum auch ohne diesen Aufschwung ein Forschungsgebiet der Sprachwissenschaft. Dabei sind zweierlei Arten von Texten zu unterscheiden: Die einen entstanden früher oder entstehen auch heute noch parallel in nicht digitaler Form. Sie haben damit immer Entsprechungen in der nicht digitalen Welt, mit denen sie verglichen werden können, wodurch sich die sprachlichen Neuerungen durch die Digitalisierung beschreiben lassen (zu den entsprechenden Parametern, cf. Reutner 2013b). Andere Texte gäbe es ohne die Digitalisierung vermutlich nicht im jeweiligen Ausmaß. Zu ihnen zählen etwa umfangreiche Enzyklopädien in kleineren Minderheitensprachen wie sie mit entsprechenden Wikipediaversionen aufkommen und dabei zum Beispiel Fragen der Normierung neu aufwerfen (cf. z. B. Reutner 2020, 784, 794). Da Sprachwissenschaft traditionell jeglicher Realisierungsform von Sprache gilt, sei sie nun schriftlich, mündlich oder seit Jüngstem eben auch digital produziert, stehen beide Typen digitaler Manifestationsformen von Sprache automatisch im Interesse der Sprachwissenschaft und belegen völlig unabhängig von der Existenz einer Disziplin *Digital Humanities* zugleich einen Beitrag der Sprachwissenschaften im Bereich *Humanities for Digitalization* (H4D).

2.3 Fragen der Abgrenzung

Wo also ist sinnvollerweise die Grenze zu ziehen zwischen ureigenen Bereichen und Verfahren einer bestimmten Disziplin und neuen Bereichen und Verfahren, mit denen diese in die *Digital Humanities* fällt? Verkürzt ließe sich fragen: Ab wann werden *Humanities* zu *Digital Humanities*? Werfen wir einen Blick auf die Methoden, so stellt sich die Frage, ob der Einsatz jedweder computergestützten Methode aus einer geisteswissenschaftlichen Arbeit ein Werk der *Digital Huma-*

nities entstehen lässt, oder ob ein bestimmter Anteil oder Komplexitätsgrad des Digitalen erreicht sein muss, damit in Verbindung mit geisteswissenschaftlichen Fragestellungen von *Digital Humanities* gesprochen werden kann. Zugespitzt ließe sich die Frage formulieren: Beginnen *Digital Humanities* bereits, wenn der Geisteswissenschaftler den Computer anschaltet und eine Exceltabelle erstellt? Selbst wenn einer solch ironischen Anmerkung ein eindeutiges „nein“ entgegenzusetzen ist, bleibt angesichts der vagen Definition des Fachgebiets die Grenzziehung zwischen einem zu geringen Einsatz des Digitalen und einem genügenden Anteil im Zweifelsfall schwer. Werden *Digital Humanities* als reine *community of practice* verstanden, so gehören ihr ohnehin jegliche Forschungsarbeiten an, deren Urheberinnen oder Urheber glauben dazugehören bzw. dazugehören möchten, was durchaus legitime Abgrenzungsversuche natürlich ad absurdum führt.

Des Weiteren besteht keine Übereinkunft, ob die Nutzung eines bereits existierenden Softwareprogramms ausreicht, damit von *Digital Humanities* gesprochen werden kann, oder ob eine Forschungsarbeit ihnen nur dann angehören sollte, wenn speziell für die aufgeworfene Fragestellung ein neues Werkzeug (Tool) entwickelt wurde oder zumindest ein vorhandenes Werkzeug spezialisierte Antworten gibt. Manches spricht dafür, dass *Digital Humanities* einen höheren Anteil des Digitalen umfassen sollte, als es die wiederholte Anwendung etablierter Programme zu leisten vermag. Zugleich besteht die eigentliche Herausforderung häufig weniger in der Entwicklung des jeweiligen Werkzeugs, als in der passenden Formulierung der Forschungsfrage sowie der durchdachten Vorstrukturierung, Aufbereitung und Interpretation der Daten, was die Frage nach der neuartigen Programmierung wiederum in den Hintergrund rücken lässt.

Wie kann eine bestimmte geisteswissenschaftliche Fragestellung mit den Methoden der Informatik beantwortet werden? So lautet eine der zentralen Fragen bei der Verbindung aus Geisteswissenschaften und Informatik, und ihre Lösung hängt entscheidend von der gelungenen Übersetzung des Forschungsinteresses auf ‚Digitalisch‘ ab. Idealerweise sind Fragestellung und Methodenwahl oder -entwicklung eng verzahnt und entstehen interdisziplinär im Dialog zwischen Vertretern aus der Informatik und den Geisteswissenschaften. Eine enge Definition der *Digital Humanities* könnte die Frage, wie hoch und wie komplex der Anteil des Digitalen in den Geisteswissenschaften sein sollte, daher etwa auch lösen, indem sie eine Interdisziplinarität im Vorgehen voraussetzt.

3 Erkenntnisziel und Korpus

Das Erkenntnisziel für diesen Beitrag ist die Frage nach dem zusätzlichen Erkenntnisgewinn bei der mit digitalen Mitteln erfolgten Auswertung: Sind die Methoden der *Digital Humanities* einfach nur Hilfsmittel, um zuvor mühevoll ermitteltes Wissen einfacher und sicherer zu generieren? Oder werden dank dieser Methoden auch Ergebnisse erzielt, die ohne sie nicht denkbar wären? Diese Fragen können auch an dieser Stelle weder generell noch abschließend beantwortet werden. Wohl aber lassen sie sich anhand einiger Beispiele vertiefen, die eine Annäherung an die Antwort erlauben. Sie sind demnach Anlass genug, einmal den potentiell einschlägigen Teil der Forschung am eigenen Lehrstuhl im Hinblick auf den Erkenntnisgewinn durch *Digital Humanities* zu betrachten.

Die Beschränkung auf die eigene Forschung erklärt sich aus drei Gründen: Erstens lassen sich die Möglichkeiten und Grenzen der angewandten Methoden besser einschätzen, als es der zwangsweise oberflächliche Blick auf fremde Forschung erlaubt, da der Methodeneinsatz in der Praxis des Forschungsprozesses bereits hinterfragt und in seinen Alternativen abgewogen wurden und die Gründe, warum welche Methode gewählt wurde, gut bekannt sind. Zweitens muss vermieden werden, die Forschung anderer ohne tiefergehende Einsicht in die jeweiligen Hintergründe, das vollständige Korpus und die angestrebten Überlegungen zu bewerten, was in einem kurzen Beitrag nicht solide möglich wäre und den Autoren damit gegebenenfalls nicht gerecht werden würde. Drittens wäre die Frage der Auswahl der zu untersuchenden Studien vor dem Hintergrund der vagen Definition des Feldes ohnehin kaum repräsentativ lösbar, sodass ein Einbezug fremder Studien auch nur zu impressionistischen Ergebnissen führen könnte. Aus der eigenen Forschung lässt sich hingegen nach klaren Kriterien ein quantitativ überschaubares Korpus erstellen, dessen qualitative Hintergründe bekannt sind und das ohne potenzielle Verletzung Dritter kritisch betrachtet werden kann. Es erlaubt damit zwar keine Verallgemeinerung der Ergebnisse, wohl aber Einsichten in einige grundsätzliche Fragen bei Studien zu einem bestimmten Forschungsthema.

In Betracht kommen prinzipiell Untersuchungen, die auf einem digitalen Korpus basieren und/oder ein Korpus mit digitalen Methoden niederschwelliger bis anspruchsvollerer Natur auswerten. Die Verwendung digitaler Ressourcen ist bei der statistischen Auswertung von Phänomenen in größeren Textkorpora (cf. Chalier/Eiber/Reutner 2020) oder der Arbeit mit digitalisierten Lexika hilfreich, die die systematische Abfrage nach bestimmten, eventuell auch kombinierten Markierungsangaben ermöglichen. Diese erleichtert z. B. Studien zu italienischen Lehnwörtern im Französischen (cf. Reutner 2008) oder zu Euphe-

mismen im Italienischen (cf. Reutner 2009; 2014a), Französischen (cf. Reutner 2009; 2013a) und Spanischen (cf. Reutner 2011; 2012a). Die Auswertung der Lexika wäre theoretisch auch durch manuelle Durchsicht möglich, in der Praxis aber eine äußerst zeitaufwändige Sisyphusaufgabe. Handelt es sich so nun um zusätzlichen Erkenntnisgewinn durch Digitalisierung oder nicht? Da die Erkenntnisse bis zu einem gewissen Grad auch ohne Digitalisierung erzielbar wären, kann die Frage zunächst verneint werden. Da die Erkenntnisse aufgrund des enormen Zeitaufwands vermutlich nicht oder nur selten manuell ermittelt werden würden und zugleich systematischer vorgegangen werden konnte und mehr Hypothesen untersucht werden konnten, ist sie zugleich zu bejahen und damit ein weiteres Beispiel dafür, wie schwer der zusätzliche Erkenntnisgewinn in der Realität zu bestimmen ist.

Grundlage für den vorliegenden Beitrag sind aber nicht Studien auf der Basis digitalisierter Ressourcen und ihrer Funktionalitäten, sondern Analysen von originär digitalem Sprach- und Bildmaterial, Material also, das digital entstanden ist. Untersucht werden Studien zu E-Mails (↑4.1), Webauftritten (↑4.2) und Wikipediaseiten (↑4.3). Alle verfolgen als übergeordnetes Forschungsinteresse die Frage, ob Sprach- und Kulturunterschiede der nicht virtuellen Welt im digitalen Raum bewahrt oder eher homogenisiert werden. Sie beinhalten damit bis zu drei Vergleichsdimensionen: Im kulturellen Vergleich werden Unterschiede zwischen einzelnen Sprachkulturen bestimmt (cf. Reutner 2012b). Im medialen Vergleich werden die digitalen Realisierungsformen den ihnen zugrundeliegenden traditionellen Textsortenmustern gegenübergestellt (E-Mail vs. Brief, Webauftritt vs. Printkatalog von Firmen, Online-Enzyklopädie vs. Printenzyklopädie). Der fachspezifische Vergleich gilt Unterschieden, die sich aus Sparten und Themen ergeben.

Das Interesse des vorliegenden Beitrags an diesen Studien liegt in der Frage nach den Chancen und Grenzen der Methodenwahl. Hierfür werden die einzelnen Projekte in jeweils gleicher Anordnung behandelt. Zunächst wird das jeweilige Korpus aus E-Mails, Internetseiten oder Wikipediabeiträgen vorgestellt. Im Folgenden werden das Erkenntnisziel formuliert, die angewandten Analysemethoden skizziert und der Erkenntnisgewinn resümiert. Abschließend werden das gewählte Vorgehen sowie alternative Vorgehensweisen kritisch hinterfragt und dabei die oben aufgeworfenen Fragen aufgegriffen.

4 Originär digitale Korpora und ihre Analyse

4.1 E-Mails und Nachrichten: Sprache und Bild

Korpus: Grundlage der Analyse sind je 100 französische und spanische Beleggleichschreiben, mit denen ein zuvor versandter Fragebogen als E-Mail-Anhang zurückgesandt wurde (cf. Reutner 2010).

Erkenntnisziel: E-Mail-Schreibtraditionen werden im medialen Vergleich zum traditionellen Brief und im kulturellen Vergleich zwischen Frankreich und Spanien herausgearbeitet.

Methode: Die E-Mails wurden manuell im Hinblick auf die Existenz und Art der Anredeformel, Schlussformel und Unterschrift sowie die Form und den Inhalt des Nachrichtenkörpers untersucht. Die einzelnen Kategorien wurden in Exceltabellen aufbereitet, dort mit Parametern wie Alter und Herkunft verknüpft und in Diagrammen visualisiert. Insgesamt wurde ein originär digitales Korpus überwiegend manuell analysiert. Einfache digitale Methoden unterstützten die Auswertung und erlaubten die grafische Darstellung der Ergebnisse.

Erkenntnisgewinn: Es ergeben sich einige Unterschiede zum herkömmlichen Brief, die bei den Franzosen besonders ausgeprägt sind. Diese orientieren sich weniger stark an den Normen des klassischen Briefs als die Spanier, die wiederum durch den überwiegenden Gebrauch von Anrede- und Schlussformeln, Unterschriften sowie meist vollständigen Sätzen auffallen. Auch sind die spanischen E-Mails häufig länger, da sie neben den notwendigen Fakten meist noch zusätzliche Aussagen im Sinne des Beziehungsaufbaus enthalten. Während mediale Unterschiede zum klassischen Brief damit nur teilweise erkennbar sind, stechen französisch-spanische Kulturunterschiede deutlich hervor.

Reflexion: Das Korpus ist relativ klein, dafür im Hinblick auf Verfasser und Inhalt ausgesprochen homogen und damit für einen bestimmten Bereich der E-Mail-Kommunikation aussagekräftig. Für die Ermittlung genereller Schreibtraditionen in E-Mails bräuhete es eine Vielzahl solcher Korpora, die in vergleichsweise ähnlich homogener Qualität unter Berücksichtigung des Datenschutzes schwer zu beschaffen sind. Stünden sie zur Verfügung, würde ihre Auswertung von digitalen Analysemethoden profitieren. Bei dem relativ kleinen Korpus ist deren Einsatz im Hinblick auf das Erkenntnisziel nicht hilfreich, da zunächst ohnehin die einzelnen Kategorien manuell zu bestimmen sind und die automatisierte Auswertung erst bei einem größeren Korpus ihre volle Kraft entfalten könnte. Letztendlich benötigen auch statistische Verfahren große Stichproben, um valide und sinnvoll interpretierbare Aussagen liefern zu

können, sodass hier eine sorgfältige manuelle Auswertung sicherlich das beste zur Verfügung stehende Mittel war.

Eine quantitative Analyse großer Datenmengen erfolgt häufig beispielsweise bei Kurznachrichten, die in Sozialen Netzwerken versandt und von den Anbietern ausgewertet werden. So reagiert zum Beispiel Facebook sensibel auf bestimmte im Messenger fallende Stichwörter, wenn es auf den Seiten des Senders im Anschluss tatsächlich oder vermeintlich passende Werbeanzeigen schaltet. Die automatisierte Stichwortsuche führt teils zu guten Ergebnissen, geht zugleich aber mit der fehlerhaften Interpretation einiger Stichwörter einher, die die beschränkte qualitative Aussagekraft einer einfachen Stichwortsuche schnell offenbart. Manuell wäre eine Stichwortsuche in solch großen Datenmengen wiederum überhaupt nicht zu leisten, während die automatisierte Datenauswertung immerhin auch eine Verknüpfung der Stichwörter mit personenbezogenen Daten erlaubt. Die Qualität der Ergebnisse lässt sich durch Methoden des *Natural Language Processing* (NLP) verbessern, die Kontexte, Inhalte oder auch die Grundstimmung der Autoren teilweise gut einzuordnen erlauben. Ein komplett automatisches Erkennen etwa von Straftätern, Straftaten oder sprachlicher Gewalt ist aufgrund der Komplexität der natürlichen Sprache, die weit über den Gebrauch einzelner Stichwörter oder kontextualisierte Inhalte hinausgeht, bislang nicht möglich. Polizei, Nachrichtendienste und Soziale Netzwerke, die die Integrität der auf ihren Plattformen geäußerten Aussagen im Blick haben, arbeiten daran und stoßen immer wieder an natürliche Grenzen. Diese sind bislang auch noch der automatischen Bilderkennung gesetzt, die manchmal zum Beispiel vergleichsweise harmlose Bilder zensieren und demgegenüber pornographische Inhalte ungefiltert erscheinen lässt. Hinzu kommt, dass Bilder aufgrund der komplexen Wahrnehmungsebenen selbst bei hermeneutischer Analyse oft schwierig in der Interpretation sind.

Im Hinblick auf den vergleichsweise einfachen Ausgangspunkt des medialen und kulturellen Vergleichs von E-Mail-Korpora lässt sich festhalten, dass die nicht maschinelle Auswertung eines kleinen homogenen Korpus relativ sichere Aussagen über einen klar definierten Bereich ermöglicht und von komplexeren Methoden nur bedingt profitieren könnte.

4.2 Internetseiten: Bildlichkeit und Farbgestaltung

Korpus: Untersucht werden drei Subkorpora: Das erste Subkorpus umfasst die Startseiten der Internetauftritte von 66 deutschen und französischen Unternehmen, die in den jeweiligen Leitindizes gelistet sind (cf. Reutner/Schubach 2012). Da die einzelnen Sparten in den Indizes beider Länder unterschiedlich gewichtet sind und die Sparte Einfluss auf die Seitengestaltung haben kann,

werden zudem die Seiten von Unternehmen aus derselben Sparte verglichen. Ein zweites Subkorpus besteht daher aus jeweils fünf deutschen und französischen Banken und Automobilherstellern aus dem DAX und CAC40 (cf. Reutner 2014c), ein drittes aus acht spanischen und vier deutschen Banken und Versicherern aus dem DAX und IBEX35 (cf. Reutner 2015).

Erkenntnisziel: Die Forschungsleitfrage gilt der Dimension kultureller und branchenbedingter Unterschiede bei der Gestaltung von Webseiten.

Methode: Vergleichsparameter sind die Typografie, Farbwahl und Seitenanordnung sowie der Einsatz von Bildern. Für eine Aussage zur Typografie wurden manuell Hervorhebungen wie Fettdruck, Versalien, Kapitälchen, Kursivierungen und Unterstreichungen ausgezählt und mit Hilfe gängiger Software die Schriftart und -größe ermittelt. Zur Bestimmung von Farbwahl und Bildeinsatz wurden manuell die Position des Firmenlogos und die Anzahl der Bilder pro Seite sowie ihrer Überlappung durch Texte oder Textboxen ausgezählt; mithilfe eines Softwareprogramms wurde eine Farbraumanalyse aller Seiten vorgenommen und die Seitenaufteilung unterschiedlichen Rastern zugeordnet. Die teils manuell, teils mit Hilfe bereits existierender Programme gewonnenen Ergebnisse wurden in eine Exceltabelle aufgenommen und auf Korrelationen untersucht.

Erkenntnisgewinn: Die Ergebnisse belegen eine auffallende Neigung der französischen und einiger spanischer Seiten zur Kombination von Schriftarten, zum Einsatz auffälliger typographischer Elemente und zu einer relativ freien Seitengestaltung, während auf den deutschen Seiten klassische Schriften und eine klare Anordnung der einzelnen Elemente dominieren. Die romanischsprachigen Seiten zeigen zudem deutlich mehr Mut zur Farbe als die deutschen und fallen darüber hinaus durch einen stärkeren Einsatz von Bildern auf, die sich oder den Text teilweise überlappen. Vielfalt und Kreativität kennzeichnen die Seiten vor allem französischer Unternehmen, Übersichtlichkeit und strukturelle Klarheit die deutscher Firmen. Die spanischen Seiten nehmen in der Ästhetik ihrer Gestaltung eine mittlere Position zwischen den deutschen und den französischen ein.

Reflexion: Insbesondere für die Ermittlung von Farben und Schriftgrößen ist die computergestützte Analyse notwendig und erlaubt eine allein durch das Auge des Betrachters nicht erzielbare Genauigkeit der Bestimmung. Die Vergleichsparameter lassen sich für diese ersten Studien nur durch die intensive Auseinandersetzung mit den Subkorpora ermitteln. Auf ihrer Basis aber könnte ein größeres Korpus von Unternehmensseiten automatisiert untersucht werden. Die subtilere Analyse der Botschaft einzelner Bilder ist wiederum nur durch den menschlichen Betrachter möglich und erfordert somit ein hermeneutisches

Herangehen, das einerseits genauer, zugleich aber wiederum anfälliger für subjektive Präferenzen oder Voreingenommenheiten ist.

4.3 Wikipedia

4.3.1 Sprache und Inhalt der Artikelkörper

Korpus: (i) Kleinere Untersuchungen basieren auf den deutschen, englischen, italienischen, französischen und spanischen Wikipediaartikeln zum Thema *Euro* sowie Währungsartikeln in Printenzyklopädien (cf. Reutner 2013b; 2014b). (ii) Eine größere Studie analysiert 120 Artikel zu jeweils fünf Stichwörtern aus den vier Bereichen Geographie, Chemie, Medizin und Wirtschaft in je drei französischen und italienischen Enzyklopädien, darunter je zwei Printenzyklopädien und Wikipedia (cf. Eiber 2020).

Erkenntnisziel: Das Forschungsinteresse beider Studien gilt der Frage nach kulturellen Unterschieden zwischen den einzelnen Wikipediaversionen und medialen Unterschieden zwischen Wikipedia und Printenzyklopädien. Die größere Studie berücksichtigt zudem fachspezifische Besonderheiten.

Methode: (i) Inhaltlich wurden die Wikipediaartikel durch vollständige Lektüre (*close reading*) in thematische Teilbereiche untergliedert, die im Anschluss ausgezählt, gewichtet und quantitativ verglichen wurden, was Computer optimal zu leisten vermögen. Sprachlich wurden die französischen und italienischen Wikipediaartikel einer strukturellen Analyse unterzogen und die Ergebnisse mit den Sprachstrukturen der Währungsartikel traditioneller Printenzyklopädien verglichen. Das kleinere Korpus wurde damit traditionell ausgewertet. (ii) Auch die größere Studie hat einen traditionellen Anteil: Manuell ermittelt wurden Kulturbezüge und behandelte Themen sowie Fachausdrücke, wertende Ausdrücke und Abweichungen von der Standardsprache. Automatisierte Auswertungsverfahren ermöglichten die Extraktion statistisch signifikanter Schlüsselwörter und Kookkurrenzen, die Berechnung der durchschnittlichen Wort- und Satzlänge sowie des lexikalischen Reichtums durch das MTL-D-Maß. Das Textmaterial wurde hierfür durch *Tree Tagger* mit Annotationen zu Wortarten und Lemmata angereichert und konnte über CQPweb nach Medium (Print oder Wiki), Sprache (Französisch oder Italienisch) und Fach (Geographie, Chemie, Medizin, Wirtschaft) abgefragt werden.

Erkenntnisgewinn: (i) Der Vergleich ergibt sowohl kulturelle als auch intermediale Unterschiede. Kulturunterschiede treten in der jeweiligen thematischen Schwerpunktsetzung deutlich hervor. Die italienische Version legt zum Beispiel viel Wert auf die Ästhetik von Münzen und Scheinen. Nur in ihr werden diese detailliert mit Bildern und Erklärungen vorgestellt, während die deutsche Version in besonderer Ausführlichkeit technische Fragen zum Funktionieren

des Euro behandelt. Der mediale Vergleich zeigt, dass klassische Ideale der Enzyklopädiensprache teilweise beibehalten, zugleich aber durch Elemente der konzeptionellen Mündlichkeit ergänzt werden. Hierzu zählen eine geringere Informationsdichte, syntaktische Komplexität und lexikalische Elaboriertheit als es in Printenzyklopädien üblich ist, was häufig die Verständlichkeit, Klarheit, kurzum Leserfreundlichkeit von Wikipedia fördert.

(ii) Die größere Studie zeigt im Medienvergleich, dass Wikipediaartikel tendenziell länger sind als gedruckte Artikel und dabei nicht nur einzelne Aspekte ausführlicher behandeln, sondern insgesamt mehr Aspekte anschneiden, dabei aber auch widersprüchliche und unvollständige Informationen liefern. Der Grad der Fachsprachlichkeit von Wikipediaartikeln ist etwas niedriger als der gedruckter Artikel, wobei die durchschnittliche Wortlänge in etwa vergleichbar und nur der Anteil fachsprachlicher Ausdrücke geringfügig niedriger ist. Auffällig sind zudem jüngere Fachausdrücke aus dem Bereich der Informatik, die in Wikipedia bereits erscheinen und in den Printenzyklopädien noch fehlen. Im Hinblick auf den Neutralitätsgrad stechen in Wikipedia positive Bewertungen hervor, die in Printenzyklopädien unüblich sind und sich insbesondere in einer starken Frequenz von Hochwertwörtern wie fr. *célèbre*/it. *celebre* und fr. *fameux*/it. *famoso* manifestieren. Abgesichert werden die Informationen in Wikipedia wiederum durch Verweise auf Experten und Studien, die in Printenzyklopädien ebenso ausbleiben. Zudem zeigen sich in Wikipedia einige Performanz- und Kompetenzfehler (z. B. Tippfehler, unvollständige Sätze) sowie eine geringere lexikalische Variation (niedrigerer MTLD-Wert) als in Printenzyklopädien.

Der Sprach- und Kulturvergleich ergibt, dass französische Artikel eine größere thematische Breite aufweisen als italienische und damit auch durchschnittlich länger sind. Die italienische Wikipedia ist insgesamt stärker als die französische am Modell italienischer Printenzyklopädien ausgerichtet als die französische Wikipedia an ihren Printentsprechungen. Ein Beispiel sind Länderartikel, die in der italienischen Wikipediaversion ebenso wie in italienischen Printenzyklopädien ein Kapitel zu Traditionen und Folklore enthalten. Die französische Wikipedia entfernt sich weiter vom Modell der zeitgenössischen französischen Printenzyklopädie, was sich unter anderem dann zeigt, wenn im Artikel *Banque* kritische Reflexionen zum Bankensystem aufscheinen, die wiederum an entsprechende Passagen in der *Encyclopédie* von Diderot und D'Alembert erinnern. Sowohl die französisch- als auch die italienischsprachige Fassung von Wikipedia sind durch Bezugnahmen auf Frankreich bzw. Italien geprägt, was in Form von expliziten Vergleichen oder implizit zum Beispiel durch die Erwähnung der *Lettres persanes* im französischsprachigen Artikel

zu Afghanistan oder des Vatikanstaats im italienischsprachigen Artikel zu Saudi-Arabien erfolgt. Der Grad der Fachsprachlichkeit ist in der französischen Wikipedia nur etwas höher als in der italienischen, in der wiederum mehr Anglizismen verwendet werden. Die französische Wikipedia lässt zudem ein verstärktes Bemühen um sprachliche Rücksichtnahme gegenüber gesellschaftlichen Gruppierungen erkennen, was für die italienische Wikipedia nicht gleichermaßen gilt. Erscheinungen konzeptioneller Mündlichkeit treten sprachspezifisch auf. In der französischen Wikipedia fallen beispielsweise die Tilgung stummer Buchstaben oder auch Linksdislokationen auf, in der italienischen eher Kongruenzschwächen oder der Rückgang schriftsprachlicher Pronomina.

Der fachliche Vergleich ergibt gemessen an der Wortlänge und dem Anteil von Fachausdrücken einen höheren Grad der Fachsprachlichkeit in Chemie- und Medizinartikeln als in Artikeln aus den Bereichen Wirtschaft und Geographie. Kulturbezüge erscheinen in beiden Versionen unabhängig vom Fach, sodass selbst zunächst kulturunspezifische Themen wie Alkohol in den Artikeln Bezüge zu den Ländern Frankreich bzw. Italien aufweisen. Sprechsprachliche Elemente treten in Artikeln aller Fächer auf, was die Schlussfolgerung nahelegt, dass diese medienbedingt sind. Insgesamt lässt sich sagen, dass Wikipediaartikel fachliche, sprachliche und kulturelle Spezifika aufweisen, die sich ebenso bei gedruckten Enzyklopädieartikeln nachweisen lassen und somit Charakteristika der Diskurstradition im digitalen Medium fortsetzen. Zu diesen treten Erscheinungen konzeptioneller Mündlichkeit, die durch die multiplen Textvergleiche auf die Produktionsbedingungen im Wiki zurückgeführt werden konnten.

Reflexion: (i) Einzelne Artikel sind nicht repräsentativ für eine gesamte Enzyklopädie. Ihre Analyse erlaubte es, einige Aspekte festzuhalten, die in einem größeren Korpus untersucht werden können. Diese gilt es zunächst zu ermitteln, eine automatisierte Auswertung ist an dieser Stelle kaum sinnvoll. (ii) Das größere Korpus ist aussagekräftiger und kann vom Einsatz digitaler Methoden stark profitieren. Die automatische Auswertung setzt zwar einen relativ hohen Aufwand bei der Digitalisierung und Annotation der Daten voraus, erlaubt dann aber die automatische Berechnung von Wortlängen, Satzlängen, lexikalischer Varianz und das Erkennen sprachlicher Muster, die manuell so nicht zu Tage treten würden und zugleich nicht so leicht statistisch zu verifizieren wären. Schwierig erwies sich die digitale Analyse bei der genaueren Ermittlung der behandelten Themen und kulturellen Bezüge, der Frequenz von Fachausdrücken, wertenden Ausdrücken und manchen Verstößen gegen die sprachliche Norm. Automatisierbar wäre eventuell die Termextraktion. Doch sind entsprechende Programme bislang meist für das Englische gut trainiert, was nur eines der Beispiele für einen Bedarf an besseren Werkzeugen für die

Analyse von Daten in romanischen Sprachen und anderen Sprachen jenseits des Englischen ist.

4.3.2 Aufbau und Bildersprache der Artikelkörper

Korpus: Für die Auswertung wurde ein Korpus mit 557 Lemmata aus Wikipedia erstellt und durch die Artikelvarianten, so vorhanden, aus zehn Sprachversionen ergänzt: der deutschen, englischen, französischen, italienischen, spanischen, portugiesischen, katalanischen, galicischen, baskischen und okzitanischen. Das Korpus wurde nach der inhaltlichen Ausrichtung der Artikel in drei Kategorien eingeteilt: kulturunspezifische, intern-kulturspezifische und extern-kulturspezifische Artikelthemen. Das kulturunspezifische Korpus dient als Referenzkorpus und umfasst Artikel zu naturwissenschaftlichen Inhalten aus dem medizinischen, chemischen und physikalischen Bereich. Das intern-kulturspezifische Korpus umfasst Artikel zu kulturell geprägten Themenbereichen wie zum Beispiel Literatur, Geschichte, Kunst oder Philosophie der Gegenden, in denen die oben genannten Sprachen gesprochen werden. Das extern-kulturspezifische Korpus enthält Artikel, die kulturelle Aspekte aus anderen Ländern beschreiben (laufendes Projekt).

Erkenntnisziel: Forschungsleitend ist die Frage, wie sich die einzelnen Sprach- und Kulturräume im Hinblick auf Aufbau und Bildlichkeit der Enzyklopädieartikel unterscheiden.

Methode: Für die computergestützte Analyse wurde eine ikonografische Auswertung aller 557 Lemmata in den zehn verschiedenen Wikipediaversionen vorgenommen. Bei der Auswertung wurden Symbole ignoriert und lediglich aussagekräftige Bilder untersucht, die eine gewisse Größe bei der Darstellung auf der Seite überschreiten. Dabei wurde der konkrete Inhalt der Bilder in der automatischen Analyse außer Acht gelassen und der Fokus auf die einfache Frage gelegt, ob die Bilder identisch sind oder nicht. Da sämtliche Bilder von Wikipedia zentral in Wikimedia abgespeichert werden, tragen gleiche Bilder den gleichen Dateinamen. Abweichende Dateinamen geben somit einen Hinweis auf eine unterschiedliche Bebilderung. Für die ikonografische Betrachtung wurde ein Algorithmus programmiert, der das Bildmaterial in den Artikelkörpern analysiert. Der Algorithmus generiert Matrizen, die den prozentualen Anteil gleicher Bebilderung in verschiedenen Sprachversionen wiedergeben. Für die qualitative Analyse wurden insgesamt 1.985 Artikel aus neun verschiedenen Wikipediaversionen manuell ausgewertet. Als Vergleichsparameter wurden zunächst die Basiselemente eines Artikels wie Artikellänge, Gliederung und Bilder herangezogen. Ferner wurden die Infobox und deren farbliche Gestaltung

und das Vorkommen übersetzter Textelemente in die qualitative Analyse einbezogen.

Erkenntnisgewinn: Die Auswertung ergibt in allen drei Subkorpora Unterschiede zwischen den Sprachversionen. Selbst im vermeintlich kulturunspezifischen Korpus zeigt sich eine unterschiedliche Schwerpunktsetzung. Während beispielsweise die italienischsprachige Wikipedia den geschichtlichen Aspekt in medizinischen Artikeln dominant zu Artikelbeginn beschreibt, erscheint dieser in der englischsprachigen Wikipedia nur am Rande zu Artikelende. Ferner zeigen sich Zusammenhänge zwischen verschiedenen Sprachversionen. So finden sich auffallende Ähnlichkeiten zwischen den iberoromanischen Versionen in Spanisch, Katalanisch und Galicisch. Die okzitanische Version hingegen orientiert sich vermehrt an der französischen. In den kulturunspezifischen und extern-kulturspezifischen Korpora hingegen erfolgt eine Orientierung insbesondere des Portugiesischen am Englischen.

Reflexion: Für die ikonografische Auswertung ist die computergestützte Analyse zum Teil hilfreich. Durch die Verwendung eines Algorithmus können im Vorfeld zur manuellen Auswertung ein großes Korpus untersucht und erste Tendenzen kultureller Einflüsse identifiziert werden. Eine Analyse der Bildpositionierung in Bezug auf den Textinhalt und vor allem eine genauere Bestimmung der Bildbotschaft ist jedoch nur durch eine manuelle Auswertung möglich. Computergestützt kann schon ein einzelnes aus unterschiedlichen Perspektiven fotografiertes Element (z. B. Eiffelturm, Kolosseum, Brandenburger Tor) gegebenenfalls nicht hundertprozentig zuverlässig als identisch erkannt werden. Noch weniger lässt sich maschinell eine mögliche Botschaft interpretieren, die durch die unterschiedlichen Perspektiven vermittelt wird. Die ergänzende manuelle Analyse ist daher unerlässlich und erlaubt, die groben Ergebnisse der computergestützten Analyse näher zu betrachten.

4.3.3 Formulierung der Artikeltitle

Korpus: Das Korpus besteht aus insgesamt 1.176 möglichen Titeln für 36 Artikelkörper der französischsprachigen Fassung von Wikipedia, wobei pro Inhalt sowohl die dauerhaft als Haupttitel verbleibende Bezeichnung als auch alternative Benennungen aus Definitionen, Umleitungen und Umbenennungen aufgenommen wurden. Ausgewählt wurden einerseits Inhalte, deren Benennung strittig sein könnte: historische und jüngste Ereignissen (von der Oktoberrevolution in Russland bis hin zu den Attentaten auf *Charlie Hebdo* in Frankreich), körperliche und geistige Einschränkungen (von Blindheit über Körperbehinderung zu Trisomie 21) sowie Hautfarben. Andererseits wurden Inhalte ausgewählt, bei denen zwar der Inhalt der Benennung unumstritten

ist, nicht aber deren Form: Artikel zu US-amerikanischen Sportvereinen sowie Titel mit Lexemen, deren Schreibung von der französischen Orthografiereform betroffen ist (cf. Reutner/Eiber 2019).

Erkenntnisziel: Von Interesse sind die Gründe für die Durchsetzung oder Ablehnung einzelner Titelvarianten. Über diese wird in der Online-Enzyklopädie Wikipedia gemeinschaftlich entschieden, was sie von traditionellen Print-encyklopädien unterscheidet, in denen die Formulierung der Lemmata meist zentral vorgegeben ist.

Methode: Für die vorliegende Untersuchung wurden die Inhalte, bei denen Abänderungen zu erwarten sind, intuitiv ausgewählt. Im Anschluss galt es, alle alternativen Titel zu einem bestimmten Inhalt zu ermitteln. Dies erfolgte über den Link *informations sur la page*, der ausgehend von der Webversion von Wikipedia zu einer Seite führt, die als weiteren Link *nombre de redirections* enthält und dort nicht nur die Anzahl, sondern auch die Titel aufzeigt, die sich aus Umleitungen oder Umbenennungen ergeben. Die alternativen Titel sowie die Umbenennungsaktionen und teils auch deren Begründungen konnten also mit Hilfe der Wikifunktionalitäten recherchiert werden. Im Anschluss wurde versucht, die Gründe für die Bevorzugung eines bestimmten Titels nachzuvollziehen. Gelesen wurden so die entsprechenden Diskussionsseiten und Versionsgeschichten, die manchmal erklärende Kommentare enthalten. Automatisch wurden damit nur die alternativen Titel ermittelt, manuell wurden die untersuchten Inhalte ausgewählt und die genauen Abläufe rekonstruiert, die dann im Anschluss linguistisch-hermeneutisch ausgewertet wurden.

Erkenntnisgewinn: Bei der Ereignisdenomination ist die Wahrung der Neutralität und im Falle jüngster Ereignisse zusätzlich die Passung des Titels zum jeweils aktuellen Kenntnisstand ausschlaggebend. Bei den Titeln, die auf soziale Gruppen Bezug nehmen, wird größtmögliche sprachliche Rücksichtnahme ihnen gegenüber angestrebt, solange diese nicht die Verständlichkeit kompromittiert. Bei den Titeln mit ausländischen Eigennamen und von der Orthografiereform betroffenen Ausdrücken treten sprachliche Vorlieben Einzelner zutage, die geduldet werden, solange die Systematizität gewahrt bleibt. Insgesamt ergibt die Studie ein Streben nach möglichst neutralen, ideologisch unbelasteten, ökonomischen, sprachsensiblen und zugleich eindeutigen Titeln, die inhaltlich zum Informationsstand und sprachlich zu anderen Titeln passen.

Reflexion: Die Wikifunktionalitäten erlauben das Auffinden der alternativen Titel zu einem bestimmten Inhalt. Der zu untersuchende Inhalt selbst kann bislang nur intuitiv bestimmt werden. Wünschenswert wären Funktionen oder Programme, die eine gezielte Suche nach einer bestimmten Art von Abänderung (z. B. „Titel wurde geändert“) oder einer bestimmten Art von Kommentaren

(z. B. „zeige mir alle Kommentare, in denen Gründe für eine Änderung genannt werden“) ermöglichen. Wikipedia gibt zudem nur die Änderungen zwischen zwei zuvor auf Verdacht ausgewählten Versionen an. Eine Funktion: «zeige mir die Versionen an, in denen eine bestimmte Änderung vorgenommen wird» gibt es noch nicht. Für das Auffinden solcher Versionen, in denen relevante Änderungen erfolgen, könnten im Rahmen der Digital Humanities sinnvolle Programme entwickelt werden, um die Datenbanksuche in diesem Bereich zu verfeinern.

4.3.4 Sprachliche Gewalt der Diskussionsseiten

Korpus: Der Studie liegt das Untersuchungskorpus «wuf 15» zugrunde, das über das Korpusabfragesystem Cosmas des Instituts für Deutsche Sprache in Mannheim zugänglich ist und sämtliche Nutzerdiskussionen der französischsprachigen Wikipedia aus dem Jahre 2015 enthält (cf. Eiber/Reutner 2020).

Erkenntnisziel: Forschungsleitend ist die Frage, welche Rolle sprachliche Gewalt bei den Aushandlungsprozessen auf den Diskussionsseiten von Wikipedia spielt, die die Online-Enzyklopädie wiederum von traditionellen Printenzyklopädien unterscheiden.

Methode: Zur Ermittlung von Textstellen, die von sprachlicher Gewalt geprägt sein könnten, wurden Ausdrücke gesucht, die konflikträchtige Verhaltensweisen bezeichnen: *troll* (1.649 Okkurrenzen), *faux-nez* (3.091 Okkurrenzen), *révocation(s)* (39.979 Okkurrenzen) und *vandalisme* (145.064 Okkurrenzen). Zum anderen wurden Ausdrücke ermittelt, die in ihrer Funktion als Schimpfwort einen Hinweis auf bestehende Konflikte darstellen könnten: *connard* (2.306 Okkurrenzen), *fils de pute* (28 Okkurrenzen), *merde* (1.051 Okkurrenzen), *salaud* (78 Okkurrenzen) und *salope* (72 Okkurrenzen). Die Okkurrenzen wurden mithilfe eines Log-Likelihood-Quotienten auf die statistisch signifikanten Kookkurrenzen der Ausdrücke *troll* (204 Kookkurrenzen), *faux-nez* (170 Kookkurrenzen), *révocation(s)* (268 Kookkurrenzen) und *vandalisme* (562 Kookkurrenzen) reduziert bzw. auf die Stellen, in denen *connard* (34), *fils de pute* (3), *merde* (12), *salaud* (13) und *salope* (4) als Schimpfwort verwendet werden. Von den konflikträchtigen Verhaltensweisen konnten für *révocation(s)* und *vandalisme* signifikante Muster der Typen *révocation(s)/vandalisme* + Verb¹ und *révocation(s)/vandalisme* + Adjektiv² ermittelt werden, deren Kontexte genauer analysiert wurden. Bei den Schimpfwörtern wurden ihre Frequenz, ihr Adressat (Gesprächspartner oder Dritter), die Wahl der Anredeform (*vous* oder *tu*) und die Verwendung sprachlicher Mittel zur Intensivierung für die Bestimmung des Grades sprachlicher Gewalt herangezogen.

Erkenntnisgewinn: Die Studie zeigt ein insgesamt geringes Ausmaß sprachlicher Gewalt in der Nutzerdiskussion von Wikipedia. Die potentiell konfliktträchtigen Verhaltensweisen *révocation(s)* oder *vandalisme* erscheinen in Verbindung mit Verben meist als höfliche Aufforderung und kombiniert mit Adjektiven als höfliche Kritik. Die Frequenz von Schimpfwörtern ist niedrig. Sind sie Zeichen verbaler Aggression, so richten sie sich meist direkt an den Gesprächspartner und weniger häufig an abwesende Dritte. Die Anredeformen variieren je nach Sprechakt: Handelt es sich um eine höfliche Aufforderung, wird eher gesiezt, liegt eine Beschimpfung vor, wird eher geduzt. In nur wenigen Fällen werden Schimpfwörter durch weitere sprachliche Mittel intensiviert. Insgesamt fällt die Häufigkeit konstruktiver Sprechakte auf, während der Einsatz von Schimpfwörtern selten ist und Passagen mit einem erheblichen Maß an sprachlicher Gewalt nur punktuell zu verzeichnen sind.

Reflexion: Die Studie kann zumindest für die untersuchten Merkmale eine gewisse Repräsentativität beanspruchen, da ihr ein ausreichend großes Korpus zugrunde liegt. Die automatische Abfrage und Quantifizierung sprachlicher Formen erleichtert die Belegsammlung. Die statistische Kookkurrenzanalyse erlaubt zudem das Auffinden von Mustern, die manuell nicht auf diese Weise gewonnen werden könnten. Allerdings muss die Bestimmung der Suchwörter intuitiv erfolgen, sodass der automatischen Analyse subjektive Voraussetzungen zugrunde liegen. Nach der Extraktion müssen zudem zahlreiche Fehltreffer aussortiert werden, die kaum vermeidbar sind und damit die Grenzen der Automatisierung aufzeigen. Ob ein Schimpfwort zum Beispiel tatsächlich als Schimpfwort zum Einsatz kommt oder metasprachlich verwendet wird, indem darüber gesprochen wird, ist durch automatische Programmierung nicht immer sicher zu erkennen.

5 Schlussbemerkungen

Im Ergebnis zeigt sich die Verwendung digitaler Ressourcen zweifelsohne ertragreich: Digital vorliegende Texte bieten ein ganz eigenes Bild von Sprache, das linguistisch untersucht werden kann und damit einen Beitrag der Geisteswissenschaften zur Digitalisierungsdebatte darstellt (H4D). Digitale Texte können zudem breit rezipiert werden und sind leichter systematisch auszuwerten als Printtexte, was zugleich einen Mehrwert der Digitalisierung für die Geisteswissenschaften (D4H) beinhaltet.

Letzterer besteht vor allem aber im Einsatz computergestützter Verfahren, deren Anwendung bei der Auswertung großer Datenmengen aus mindestens drei Gründen sinnvoll ist: Erstens ist eine replizierbare und damit aussage-

kräftige quantitative Analyse von Datenmaterial größeren Umfangs manuell realistischerweise in vielen Fällen überhaupt nicht möglich. Zweitens liefert die automatisierte Analyse solide Daten, die für sich genommen und im Vergleich von Subkorpora auf statistische Relevanz geprüft und zudem in ihrer Aussagekraft mit unterschiedlichen Parametern verknüpft werden können. Vor allem aber ergeben sich drittens durch die digitale Analyse Ergebnisse, die mit bloßem Auge nicht unbedingt auffallen würden und in der vorliegenden Präzision lediglich von der Maschine ermittelt werden können, die zugleich unvoreingenommen nach bestimmten Parametern sucht. Die Extraktion signifikanter Kookkurrenzen lässt zum Beispiel sprachliche Muster erkennen, die vorher so noch nicht ermittelt wurden. Einfache Stichwortsuchen erlauben das grobe Durchforsten großer Datenmengen auf bestimmte Inhalte, und komplexere Methoden des *Natural Language Processing* zudem die Erfassung von Wortzusammenhängen und inhaltlichen Kontexten. Die Berechnung lexikalischer Vielfalt sowie einfacher Wort- und Satzlängen eröffnet zudem neue Einsichten in Stilfragen. Die Farbraumanalyse und automatische Identifikation identischer Bilder ermöglicht darüber hinaus neue Erkenntnisse zum kulturspezifischen Farb- und Bildgebrauch.

Durch den Einbezug größeren Datenmaterials werden manchmal zwar tatsächlich nur Tendenzen kleinerer Voruntersuchungen bestätigt. In vielen Fällen aber erbringt die digitale Analyse Ergebnisse, die bei einer manuellen Ermittlung nicht in gleichem Maße zu erzielen wären. Zudem zwingen computergestützte Methoden dazu, Kriterien und Fragestellungen abstrakt zu formulieren und dann rein datenbasiert und strikt objektiv zu verifizieren oder zu falsifizieren. Ob und wenn ja welcher Einsatz digitaler Methoden aus den Geisteswissenschaften *Digital Humanities* macht, muss an dieser Stelle offenbleiben, da die Grenzen des neuen Fachgebiets bislang noch nicht klar definiert sind. Die Frage aber, ob mit Hilfe digitaler Methoden Informationen nicht nur einfacher und qualitativ hochwertiger ermittelt werden, sondern darüber hinaus auch inhaltlich neuartige Ergebnisse erzielt werden können, ist damit klar zu bejahen.

Gewonnen werden zunächst aber eben immer nur statistisch belastbare Informationen, die bis zu einem gewissen Grad oberflächlich bleiben und gegebenenfalls hermeneutisch hinterfragt werden müssen. Bei der Bildanalyse zeigt sich besonders deutlich, dass subtilere Botschaften und kulturelle Bezüge bislang kaum zufriedenstellend in automatisierter Weise ermittelt werden können, was zum einen an der Komplexität visueller Informationsverarbeitung liegt und zum anderen an der subjektiven, kultur- und kontextabhängigen Interpretation von Bildern. Auch bei der Analyse sprachlicher Daten sind komplexere sprachliche Phänomene, die sich aus dem ganz konkreten Gebrauch

der Wörter in einem bestimmten Kontext ergeben, kaum maschinell greifbar. Denken wir nur an Ironie, für die nur schwer durchwegs verlässliche Parameter der automatischen Erkennung zu definieren sind, und ebenso wenig für die Frage, ob ein bestimmtes Wort als solches oder metasprachlich gebraucht wird. Die automatisierte Datenanalyse ist bei entsprechend großer Datenmenge und aufgrund der Reduktion von menschlichem Versehen oder der Neigungen, Texte mit einer bestimmten Brille zu lesen, aussagekräftiger und objektiver als ein hermeneutisches Herangehen, kann zugleich aber einige Fragen, die Wissenschaftler an Sprache und Bilder richten, überhaupt nicht greifen. Damit stellen manuelle und automatisierte Verfahren insgesamt keinen Widerspruch dar, sondern ergänzen sich in fruchtbarer Weise. Die adäquate Formulierung der geisteswissenschaftlichen Fragestellung für die Anwendung digitaler Methoden erfolgt zudem am besten im Austausch zwischen Geisteswissenschaften und Informatik, der den Fachbereich der *Digital Humanities* wiederum generell prägen sollte und insgesamt mehr interdisziplinäre Ansätze wünschenswert erscheinen lässt.

Bei Fragen, für deren Beantwortung die digitale Analyse dann sinnvollerweise herangezogen wird, läuft ihr Einsatz aber auch Gefahr, der Illusion der Präzision zu erliegen, das heißt, eine besonders hohe Aussagekraft und Objektivität zu suggerieren und dabei die subjektive Beeinflussung der Endergebnisse auszublenden. Statistische Verfahren und Methoden des maschinellen Lernens vermindern zwar die Subjektivität und stellen hohe Ansprüche an die Daten, die vernünftig balanciert und der Fragestellung angemessen zusammengestellt sein müssen. Entsprechende Kenntnisse und Wissenschaftsethos vorausgesetzt, wird auch das Verfahren auf der Basis objektiver Kriterien ausgewählt und nicht bewusst oder unbewusst danach, welches die Ergebnisse liefert, die den eigenen Wunschvorstellungen am besten entsprechen. Dennoch ist eine subjektive Beeinflussung in den meisten Fällen unumgänglich, denn schließlich muss das digital auszuwertende Korpus meist auf bestimmte Hypothesen hin überprüft werden, deren adäquate Formulierung das Hauptqualitätsmerkmal der Forschung darstellt. Abgesehen von manchen vollautomatisierten Untersuchungen, die zum Beispiel die Anwendung künstlicher Intelligenz oder maschinellen Lernens erlauben, ist vor den meisten geisteswissenschaftlichen Studien klar zu entscheiden, welche Einzelphänomene im Hinblick auf bestimmte Parameter herangezogen werden. An welchen sprachlichen Formen zeigt sich zum Beispiel Fachsprachlichkeit? An welchen Neutralität? Und an welchen sprachliche Gewalt? Die Operationalisierung einer Fragestellung in konkret auffindbare Formen beeinflusst das Ergebnis massiv, ebenso wie die Entscheidung zwischen alternativen Methoden der statistischen Prüfung. Zudem nehmen

Forschende nicht nur darauf Einfluss, welche Parameter ausgewählt werden, um davon ausgehend bestimmte Schlüsse ableiten zu können, sondern auch, wie die automatisiert generierten Ergebnisse letztendlich zu werten sind. Damit kann die Digitalisierung die Objektivierung der Forschung im Sinne eines reduzierten Einflusses des Forschenden auf die Ergebnisse zwar unterstützen, nicht aber automatisch objektive Ergebnisse garantieren. Denn die wahre Aussagekraft der Daten hängt in vielen Fällen entscheidend von den jeweiligen Vorannahmen, der Art der Fragestellung und ihrer Interpretation ab. Demnach sind *Digital Humanities* durch ihre ausgeprägte Präzision und den Einbezug großer Datenmengen also besonders anfällig dafür, eine hohe Verlässlichkeit der eigenen Ergebnisse nahezulegen. Doch liefern sie einerseits zwar objektive Informationen aus vorhandenen Daten, die formulierten Endergebnisse tragen andererseits aber immer einen subjektiven Kern in sich, sobald sie sinnvollerweise über bloßen Positivismus hinausgehen.

Insgesamt erscheinen *Digital Humanities* zumindest auf der Basis der vorliegenden Daten als klare Hilfe, die die Arbeit der Geisteswissenschaften zu erleichtern, ihre Reichweite zu vergrößern und ihre Ergebnisse besser zu visualisieren vermag. Häufig wird dabei zugleich neues Wissen gewonnen, das den teils enormen Aufwand rechtfertigt. Sicher stellen Geisteswissenschaftler zudem Fragen, bei deren Beantwortung Algorithmen unterstützen oder sogar Ergebnisse generieren helfen, die manuell nicht in dieser Weise erzielt werden können. Letztendlich aber sollten Geisteswissenschaftler ihre Fragen nicht danach ausrichten, ob sich davon abgeleitet ein interessantes Werkzeug konstruieren lässt, sondern diejenigen Fragen aufwerfen, die es zu beantworten gilt. Nicht immer sind dies die Fragen, bei deren Beantwortung große Datenmengen, bestimmte Werkzeuge oder Algorithmen hilfreich sind.

Zugleich können viele Fragestellungen natürlich auch neue Methoden und Werkzeuge notwendig machen und so in interdisziplinärer Kooperation zu interessanten Forschungsfragen in der Informatik führen, deren Lösungen wiederum relevante Erkenntnisse in den Geisteswissenschaften nach sich ziehen. Auf den Prüfstand gestellt erweisen sich *Digital Humanities* damit als vernünftige Ergänzung der traditionellen Geisteswissenschaften. Manche Forschungsfragen können von digitalen Methoden und Ressourcen profitieren oder werden sogar erst durch sie aufgeworfen, andere wiederum nicht. Die Digitalisierung befruchtet Teilbereiche aller geisteswissenschaftlichen Disziplinen damit höchst sinnvoll. Nicht sinnvoll wäre nur eine Vernachlässigung der Teilbereiche, bei denen dies nicht der Fall ist, und im Extremfall ein vollständiges Aufgehen der Geisteswissenschaften (Humanities) in *Digital Humanities*.

Da für letztere noch keine abschließende Definition existiert, liegt es auf der Basis der obigen Erwägungen nahe, abschließend eine Annäherung an das Fach mithilfe der Prototypensemantik vorzuschlagen. Die idealtypische Ausprägung von *Digital Humanities* liegt unseres Erachtens in Forschungsarbeiten, die ein interdisziplinäres Zusammenspiel zwischen Informatik und Geisteswissenschaften erfordern, von dem beide Disziplinen profitieren, egal ob die Kooperation in Richtung *Humanities for Digitalization* (H4D) oder *Digitalization for Humanities* (D4H) verläuft.

Bibliografie

- Burdick, Anne/Drucker, Johanna/Lunenfeld, Peter/Presner, Todd/Schnapp, Jeffrey (2012): *Digital_Humanities*, Cambridge, MIT.
- Chalier, Marc/Eiber, Bettina/Reutner, Ursula (2020): „Sciences naturelles avares en mots et sciences humaines en étalant trop? Réponses statistiques à de vieux stéréotypes sur le discours scientifique“, SHS Web of Conferences 78, 06008 (Congrès Mondial de Linguistique Française – CMLF 2020), <https://doi.org/10.1051/shsconf/20207806008>.
- DARIAH-DE (2015): *Einführung: Projekte und Forschungsfragen in den Digital Humanities*, in: Digital Research Infrastructure for the Arts and Humanities-DE (ed.): *Handbuch für Digital Humanities. Anwendungen, Forschungsdaten und Projekte*, Berlin, ePubli, 6–14.
- Encyclopédie* = d'Alembert, Jean Le Rond/Diderot, Denis (eds.) (1751–1772): *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers, par une société de gens de lettres*, Paris/Neuchastel, Briasson/Faulche, <http://enccre.academie-sciences.fr/encyclopedia/> (23.10.2023).
- Eiber, Bettina (2020): *Diskurstraditionen im medialen Wandel. Eine korpuslinguistische Untersuchung französischer und italienischer Printenzyklopädie- und Wikipediaartikel*, Tübingen, Narr.
- Eiber, Bettina/Reutner, Ursula (2020): „Gardez votre calme: Wikipédia entre collaboration fructueuse et violence verbale“, in: Bernard-Barbeau, Geneviève/Meier, Franz/Schwarze, Sabine (eds.): *Conflits sur/dans la langue : perspectives linguistiques, argumentatives et discursives*, Frankfurt a. M./Berlin/Bern/Bruxelles/New York/Oxford/Warszawa/Wien, Lang, 191–213.
- Gardiner, Eileen/Musto, Ronald G. (2015): *The digital humanities. A primer for students and scholars*, New York, Cambridge University Press.
- Gibbs, Fred (2013): „Digital Humanities Definitions by Type“, in: Terras, Melissa/Nyhan, Julianne/Vanhoutte, Edward (eds.): *Defining Digital Humanities. A reader*, Surrey/Burlington, Ashgate, 289–297.

- Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (2017): „Warum ein Lehrbuch für Digital Humanities?“, in: Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (eds.): *Digital Humanities. Eine Einführung*, Stuttgart, Metzler, XI–XIII.
- McCarty, Willard (2005): *Humanities Computing*, Basingstoke/New York, Palgrave Macmillan.
- Nyhan, Julianne/Terras, Melissa/Vanhoutte, Edward (2013): *Introduction*, in: Terras, Melissa/Nyhan, Julianne/Vanhoutte, Edward (eds.): *Defining Digital Humanities. A reader*, Surrey/Burlington, Ashgate, 1–10.
- Rehbein, Malte (2020): „Historical Network Research, Digital History, and Digital Humanities“, in: Kerschbaumer, Florian/von Keyserlingk-Rehbein, Linda/Stark, Martin/Düring, Marten (eds.): *The Power of Networks. Prospects of Historical Network Research*, London, Routledge, 251–277.
- Reutner, Ursula (2008): „Les emprunts récents de l’italien au français“, in: Horiot, Brigitte (ed.): *Le français ailleurs et toujours : place et fonctions du français dans les autres langues*, Lyon, St. Joseph, 119–136.
- Reutner, Ursula (2009): *Interpretationen zu französischen und italienischen Euphemismen*, Tübingen, Niemeyer.
- Reutner, Ursula (2010): „E-Mail-Kulturen im Vergleich. Zum Sprachverhalten spanischer und französischer Linguisten“, in: *Romanistik in Geschichte und Gegenwart* 16 (2), 3–28.
- Reutner, Ursula (2011): „El eufemismo como fenómeno cultural y lexicográfico“, in: *Lingüística española actual* 33 (1), 55–74.
- Reutner, Ursula (2012): „La asignación de la marca de eufemismo. Una comparación de todas las formas acotadas en el DGLE, el DRAE y el DUE“, in: Botta, Patricia/Pastor, Sara (eds.): *Rumbos del hispanismo en el umbral del Cincuentenario de la AIH*, vol. 8: *Lengua*, Roma, Bagatto, 293–303 (= 2012a).
- Reutner, Ursula (2012): „Das interkulturelle Potential digitaler Medien im historischen Vergleich“, in: Reutner, Ursula (ed.): *Von der digitalen zur interkulturellen Revolution*, Baden-Baden, Nomos, 33–52 (= 2012b).
- Reutner, Ursula (2013): „*Nous, lexicographes, nous avons donc toujours tort ? Traitement de l’euphémisme dans le Petit Robert*“, in: *Cahiers de lexicologie* 103, 167–192 (= 2013a).
- Reutner, Ursula (2013): „Wikipedia und der Wandel der Wissenschaftssprache“, in: *Romanistik in Geschichte und Gegenwart* 19 (2), 231–249 (= 2013b).
- Reutner, Ursula (2014): „Eufemismo e lessicografia. L’esempio dello Zingarelli“, in: *Studi di lessicografia italiana* XXXI, 317–344 (= 2014a).
- Reutner, Ursula (2014): „L’enciclopedia digitale «Wikipedia». Linee di analisi interculturale e intermediale“, in: Suomela-Härmä, Elina (ed.), *Dal manoscritto al web: canali e modalità di trasmissione dell’italiano. Tecniche, materiali e usi nella storia della lingua*, Firenze, Cesati, 2014, 689–698 (= 2014b).

- Reutner, Ursula (2014): „Französisches Bilderspiel und deutsches Informationspaket. Ein Vergleich der Internetpräsenzen von Banken und Automobilherstellern“, in: Rentel, Nadine/Reutner, Ursula/Schröpf, Ramona (eds.): *Von der Zeitung zur Twitterdämmerung. Medientextsorten und neue Kommunikationsformen im deutsch-französischen Vergleich*, Berlin/Münster, Lit, 135–160 (= 2014c).
- Reutner, Ursula (2015): „El sitio web – ¿un espacio cultural? Un estudio comparativo germano-español de bancos y aseguradoras“, in: Rentel, Nadine/Reutner, Ursula/Schröpf, Ramona (eds.): *Lingüística mediática y traducción audiovisual*, Frankfurt a. M./Berlin/Bern/Bruxelles/New York/Oxford/Warszawa/Wien, Lang, 3–26.
- Reutner, Ursula (2020): „Minor‘ Gallo-Romance Languages“, in: Lebsanft, Franz/Tacke, Felix (eds.): *Manual of Standardization in the Romance Languages*, Berlin, De Gruyter, 773–807.
- Reutner, Ursula/Eiber, Bettina (2019): „Fusillade au siège de Charlie Hebdo ou Attentat contre Charlie Hebdo? Wikipédia et la co-construction des titres d’articles“, in: *Romanistik in Geschichte und Gegenwart* 25 (2), 149–175.
- Reutner, Ursula/Schubach, Sebastian (2012): „Kulturspezifische Ästhetik im Internet. Typografie und Bildlichkeit im deutsch-französischen Vergleich“, in: Reutner, Ursula (ed.): *Von der digitalen zur interkulturellen Revolution*, Baden-Baden, Nomos, 235–263.
- Schreibman, Susan/Siemens, Ray/Unsworth, John (eds.) (2004): *A Companion to Digital Humanities*, Oxford, Blackwell.
- Terras, Melissa/Nyhan, Julianne/Vanhoutte, Edward (eds.) (2013): *Defining Digital Humanities. A reader*, Surrey/Burlington, Ashgate.
- Thaller, Manfred (2014): „Grenzen und Gemeinsamkeiten: Die Beziehung zwischen der Computerlinguistik und den Digital Humanities“, Präsentation bei der Jahrestagung des Verbandes Digital Humanities im deutschsprachigen Raum (DHd) in Passau am 27.03.2014.
- Thaller, Manfred (2017): „Digital Humanities als Wissenschaft“, in: Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (eds.): *Digital Humanities. Eine Einführung*, Stuttgart, Metzler, 13–18.
- Unsworth, John/Siemens, Raymond/Schreibman, Susan (eds.) (2016): *A New Companion to Digital Humanities*, Chichester/Malden, Wiley & Sons.
- Wikipedia DE (2023): *Digital Humanities*, San Francisco, Wikimedia Foundation, https://de.wikipedia.org/wiki/Digital_Humanities (23.10.2023).
- Wikipedia EN (2023): *Digital humanities*, San Francisco, Wikimedia Foundation, https://en.wikipedia.org/wiki/Digital_humanities (23.10.2023).