ROMANISTISCHES

RK

KOLLOQUIUM XXXIV

Digitale romanistische Sprachwissenschaft: Stand und Perspektiven

Lidia Becker, Julia Kuhn,
Christina Ossenkop,
Claudia Polzin-Haumann,
Elton Prifti (eds.)

narr\f
ranck
e\atte
mpto

Digitale romanistische Sprachwissenschaft: Stand und Perspektiven

ROMANISTISCHES

RK

KOLLOQUIUM XXXIV

Herausgegeben von Lidia Becker, Julia Kuhn, Christina Ossenkop, Claudia Polzin-Haumann und Elton Prifti

Band 34

Lidia Becker, Julia Kuhn, Christina Ossenkop,
Claudia Polzin-Haumann, Elton Prifti (eds.)

Digitale romanistische Sprachwissenschaft: Stand und Perspektiven

narr\f
ranck
e\atte
mpto

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

DOI: https://www.doi.org/10.24053/9783823395065

Internet: www.narr.de
eMail: info@narr.de

CPI books GmbH, Leck

ISSN 2750-042X
ISBN 978-3-8233-8506-6 (Print)
ISBN 978-3-8233-9506-5 (ePDF)
ISBN 978-3-8233-0309-1 (ePub)

Inhalt

Einleitung

Die ersten Versuche des Einbezugs von Informatik in sprachwissenschaftliche Analyseverfahren, die bis in die frühen 1960er Jahre zurückgehen, trugen aufgrund ihres Erfolgs zu einer zunehmenden Öffnung der Linguistik in Bezug auf die digitale Welt bei. Dies erlebt einen substanziellen Qualitätssprung insbesondere seit der Jahrtausendwende, der auch mit der Konsolidierung der sogenannten *Digital Humanities* in Verbindung steht. Die Digitalisierung erreichte auch die traditionsreiche romanistische Sprachwissenschaft, in der mittlerweile etwa bei der Konzeption neuer Forschungsprojekte digitale Komponenten nahezu ein Muss geworden sind. Bei länger laufenden Projekten wird intensiv, manchmal sogar etwas übereifrig der digitale Anschluss gesucht; bei abgeschlossenen Projekten wird versucht, die Ergebnisse in digitalisierter Form zugänglich und weiter verwertbar zu machen. Durch die tiefgreifenden Veränderungen, die eine automatisierte Gewinnung, Verarbeitung, Darstellung und Nutzung von Forschungsdaten und -ergebnissen mit sich bringen, befindet sich auch die romanistische Sprachwissenschaft an einem methodischen Wendepunkt, über den es zu diskutieren gilt.

Diese zentrale Fragestellung stand im Fokus der 34. Edition des *Romanistischen Kolloquiums*, das Ende 2019 an der Universität Wien stattfand. Der Schwerpunkt wurde dabei über die Beschreibung des aktuellen Standes der digitalen romanistischen Sprachwissenschaft hinaus auf zentrale theoretische und methodische Fragen, Probleme und Herausforderungen sowie auf konkrete, methodisch innovative und zukunftsträchtige Praxisbeispiele gelegt. Im vorliegenden Band sind ausgewählte Beiträge des Kolloquiums in drei inhaltlichen Blöcken versammelt, in denen allgemeine methodologische Fragen behandelt und Facetten der Digitalisierung in der historisch-etymologischen Lexikographie sowie in der Korpuslinguistik vorgestellt werden. Es zeichnet sich dabei eine substanzielle Erweiterung des Spektrums der Digitalisierung von der Gewinnung, Verwaltung und eindimensionalen Nutzung digitaler Daten zur zusätzlichen Digitalisierung der Forschungsmethoden aus. Sowohl im Hinblick auf die romanischen Sprach- und Kulturräume als auch auf die sprachgeschichtlichen Perioden sowie die sprachwissenschaftlichen Disziplinen decken die Beiträge des Bandes ein breites Spektrum ab.

Der erste Block wird mit dem Beitrag *Digital Humanities auf dem Prüfstand. Analysemethoden für digitale Korpora von E-Mails über Internetseiten bis zu Wiki-*

pedia von Ursula Reutner eröffnet, in dem sich die Autorin mit dem Aufkommen und mit der graduellen Konsolidierung der digitalen Geisteswissenschaften in methodischer Hinsicht auseinandersetzt. Dies wird anhand einiger Forschungsprojekte in der romanistischen Linguistik unternommen, wobei die durch die Digitalisierung gewonnenen Erkenntnisse und die daraus resultierenden Konsequenzen betrachtet und so die Wirkung der Verbindung von Digital- und Geisteswissenschaften überprüft werden. Nach einem Blick auf aktuelle Definitionen und Erklärungsversuche des Begriffs *Digital Humanities* stellt der Beitrag einige Forschungsmethoden vor, die verschiedene Möglichkeiten der Analyse digitaler Textkorpora wie E-Mails, Webseiten und Wikipedia beinhalten. Die verschiedenen Methoden und die damit erzielten Erkenntnisse werden kritisch abgewogen. So sollen insgesamt die Möglichkeiten und Grenzen der *Digital Humanities* als geisteswissenschaftliches Fachgebiet aufgezeigt werden.

Im Mittelpunkt des zweiten Beitrags *Digitale romanistische Linguistik als Ort methodologischer Paradigmendiskussion? Kritische Überlegungen, Bedarfe und Potenziale* von Vera Mathieu und Julia Montemayor steht die Diskussion über den *mixed methods*-Ansatz bzw. über die Kombination von qualitativen und quantitativen Methoden in romanistischen Studien, in denen computergestützte Analyseansätze zunehmend an Relevanz gewinnen und die neue, viel versprechende Perspektiven eröffnen. Ferner werden dabei die Möglichkeiten der softwarevermittelten qualitativen Kategorisierung von Sprachdaten sowie die partielle Adaption korpuslinguistischer Verfahren bei der Analyse von Sprachdaten anhand exemplarischer Einblicke in die Praxis vorgestellt.

Im zweiten Block, in dem drei Beiträge vereint sind, werden einzelne Aspekte der sich immer weiter konsolidierenden Digitalisierung in der historisch-etymologischen Lexikographie beleuchtet. Gegenstand der Abhandlungen sind drei monumentale romanistische Werke: das *Romanische Etymologische Wörterbuch* (REW) in seiner 3. Edition, das *Lessico Etimologico Italiano* (LEI) sowie das *Französische Etymologische Wörterbuch* (FEW). Florian Zacherl setzt sich in seinem Beitrag *Automatisierte Erschließung von strukturierten Daten aus Wörterbuchtexten* mit den nicht wenigen inhaltlichen und technischen Herausforderungen auseinander, die die komplexe und besonders aufwändige Umwandlung der Werke von der papiernen Version in strukturierte digitale Versionen mit zahlreichen neuen und sehr nützlichen Verwendungsmöglichkeiten mit sich bringt. Nach einigen grundsätzlichen Überlegungen zur Darstellung lexikalischer Daten, insbesondere in einer relationalen Datenbank, wird am Beispiel des REW_3 eine Methode des besagten Umwandlungsprozesses vorgestellt, die aus vier Arbeitsschritten besteht: Extrahierung des Originaltextes aus den gescannten Seiten mittels optischer Zeichenerkennung und Speicherung in

einer relationalen Datenbank; Extraktion der einzelnen Artikel und strukturierte Hierarchisierung ihrer Bestandteile durch eine formale Grammatik und Darstellung in einer baumartigen Struktur; Umwandlung dieser in tabellarische Daten, wobei implizite Konventionen, die von den Konventionen des jeweiligen Quellenmaterials abhängen, aufgelöst werden, sowie abschließend die (digitale) Veröffentlichung der extrahierten und neu organisierten Daten, die zudem stets verbessert, korrigiert und mit anderen Online-Ressourcen und dem *Semantic Web* verbunden werden können.

Ein ähnlicher Weg wurde auch beim FEW eingeschlagen. Es befindet sich seit mehreren Jahren in einem (Retro)digitalisierungsprozess, der vom französischen Forschungslabor *Analyse et Traitement Informatique de la Langue Française* (ATILF) durchgeführt wird. Benjamin Husson und Sarah Kremer stellen in ihrem Beitrag *Les outils de l'informatisation du FEW* die verschiedenen informatischen und typografischen Werkzeuge und Techniken vor, die dabei verwendet werden. Die Herausforderungen und die groben prozeduralen Abläufe sind ähnlich wie im Falle des REW_3. Im Rahmen des Beitrags werden verschiedene Teilprojekte vorgestellt, deren Ziel die komplexe Umwandlung des Textes in digitale Daten ist. Es werden dabei die wichtigsten technischen Hürden des besagten Prozesses präsentiert, die mit Hilfe moderner Technologien und Standards bewältigt werden mussten und noch müssen. Ein Schwerpunkt wurde auch im Bereich der digitalen Typographie bzw. in der soliden Planung, Entwicklung, eleganten Gestaltung und erfolgreichen Umsetzung einer speziellen digitalen Schriftart gelegt, die eine *online*-Visualisierung des in typographischer Hinsicht komplexen Wörterbuchs ermöglicht.

Anders als das FEW befindet sich das Langzeitprojekt LEI noch im Entstehungsprozess. In klassischer, papierner Form wurden 23 großformatige Bände publiziert, während der Rest des LEI (in etwa die Buchstabenstrecken G-Z) in genuin digitaler Form verfasst und publiziert werden. Demnach ist der Digitalisierungsprozess, der für das LEI im Jahr 2018 begonnen hat, entsprechend komplexer und mit größeren Herausforderungen verbunden. Ein Aspekt des umfangreichen Projekts *LEI digitale* (Prifti 2022) ist das semantische *mapping* des LEI, das im Grunde in der Verknüpfung von lexikalischen Einträgen und Konzepten besteht, die die außersprachliche Realität repräsentieren. Das ist der Gegenstand des Beitrags *La mappatura semantica del Lessico Etimologico Italiano (LEI). Possibilità, metodi e prospettive* von Alessandro Alfredo Nannini, der sich damit seit einigen Jahren befasst. Er stellt dabei die semantischen Strukturen des LEI, den darauf basierten *mapping*-Prozess sowie das zu Grunde liegende Begriffssystem dar. Ferner werden einige Perspektiven vorgestellt, die das semantische *mapping* für das LEI und für die Lexikographie im Allgemeinen öffnet.

Der dritte inhaltliche Block, der vier Beiträge mit korpuslinguistischem Bezug vereint, wird mit dem Beitrag *Kulturerbe, historische Linguistik und Semantic Web: Eine Einführung mit Fallbeispiel zu französischen linguistischen Ressourcen* eröffnet, in dem sich Sabine Tittel mit der digital gestützten Bearbeitung historischer Sprachressourcen auseinandersetzt, die als schriftliche Zeugnisse alle Aspekte des historischen Lebens erfassen und damit der Generierung und Bewahrung von kulturhistorischem Wissen dienen. Die Modellierung und Veröffentlichung historischer Sprachressourcen nach dem *Linked Open Data* (LOD)-Paradigma des *Semantic Web* ist eine Möglichkeit, dieses Wissen zugänglich zu machen. Der LOD-Ansatz ermöglicht einen Zugang, der weit über die derzeitigen Suchfunktionen des *World Wide Web* mit ihren Defiziten hinausgeht. Neben einer kurzen Einführung in LOD werden im Beitrag die syntaktische Struktur des Datenformats *Resource Description Framework* beschrieben und die Prinzipien der semantischen Abbildung auf Ontologien erläutert. Anhand eines Anwendungsfalls von altfranzösischen, mittelfranzösischen und modernen regionalen französischen Wörterbüchern wird dann gezeigt, wie historische linguistische Daten mit dem *OntoLex-Lemon*-Modell modelliert werden können.

Im Artikel *Digitale Phraseologie* von Elmar Schafroth werden drei digitale Projekte vorgestellt, die mit Phraseologie zu tun haben und sich auf das Italienische oder auf mehrere Sprachen gleichzeitig beziehen. Es handelt sich dabei um das Projekt FRAME (*Fraseologia Multilingue Elettronica*), in dem Satzglieder in sieben Sprachen (Chinesisch, Deutsch, Englisch, Französisch, Italienisch, Russisch, Spanisch) nach den Prinzipien der Konstruktionsgrammatik beschrieben werden. Das zweite Projekt, welches sich auf audiovisuelles Material stützt und in dem auf aktuelle Forschungsfragen wie das Verhältnis zwischen Phraseologie und Konstruktionsgrammatik eingegangen wird, richtet sich an Studierende der romanischen Sprachen, des Deutschen und des Englischen, die sich für die Phraseologie interessieren. Im Rahmen des dritten Projekts, GEPHRI (*Gebrauchsbasierte Phraseologie des Italienischen*), werden die 500 häufigsten verbalen Idiome des Italienischen in einer Datenbank, hauptsächlich nach den Prinzipien der Konstruktionsgrammatik, teilweise auch nach der Frame-Semantik beschrieben und zur Nutzung bereitgestellt.

Die automatische Identifizierung von Mehrwortausdrücken ist eine entscheidende Komponente für die Verarbeitung natürlicher Sprache, stellt aber neben der Erkennung von Idiomatizität auch Herausforderungen wie Variabilität, Mehrdeutigkeit und Unstimmigkeit dar. Um die Lösung dieses Problems bemüht sich im Beitrag *Construction, exploitation et exploration de ressources linguistiques pour le traitement automatique des expressions polylexicales en français: le projet PARSEME-FR* die Autorengruppe Mathieu Constant, Marie

Candito, Yannick Parmentier, Carlos Ramisch und Agata Savary. Dafür wurden im Rahmen des Projekts PARSEME-FR neue Modelle und Algorithmen sowie neue linguistische Ressourcen entwickelt, wie etwa Annotationsrichtlinien für Mehrwortausdrücke und entsprechend annotierte Korpora sowie Werkzeuge zur Strukturierung und Vervollständigung lexikalischer Ressourcen.

Der darauffolgende Beitrag *Erste Schritte zu einer digitalen Chrestomathie des Altrumänischen* stammt von Sam Mersch und bietet einen Einblick in das digitale Editionsprojekt der altrumänischen Chrestomathie von Moses Gaster. Ziel der Abhandlung ist die Darstellung der Probleme und der entsprechenden Lösungsüberlegungen, die sich bei der Durchführung der digitalen Editionsarbeit ergeben, wobei der Schwerpunkt auf den technischen Aspekten liegt.

Der Artikel *Digital Humanities und Sprachgeschichte am Beispiel des USTC. Zum Wert digitaler Kataloge frühneuzeitlicher Drucke für die sprachgeschichtliche Forschung* von Thomas Scharinger schließt den Sammelband ab. Darin wird die Eignung des *Universal Short Title Catalogue* (USTC) für die sprachhistorische Forschung zu den romanischen Sprachen erörtert, der als digitale Datenbank mit detaillierten Informationen (z. B. Autorschaft, Ort, Region, Sprache, Thema) zu mehr als 740.000 im frühneuzeitlichen Europa gedruckten Ausgaben die Rekonstruktion des Gebrauchs einer bestimmten Sprache in einem bestimmten Gebiet zu einer bestimmten Zeit ermöglichen kann. Anhand von drei Fallstudien wird gezeigt, dass die mit dem USTC generierten Daten genutzt werden können, um etwa die Beziehung zwischen Latein und den romanischen Volkssprachen, die Rivalität zwischen zwei konkurrierenden romanischen Sprachen wie Katalanisch und Spanisch sowie die Verbreitung einer romanischen Sprache über ihr ursprüngliches Gebiet hinaus zu untersuchen.

Die Studien in diesem Band, die räumlich weitgehend die gesamte Romania abdecken und eine Vielzahl von verschiedenartigen Aspekten und Betrachtungsperspektiven behandeln, bezeugen die stete Konsolidierung eines digitalen Wandels auch in der romanistischen Sprachwissenschaft, der sich zunehmend im Bereich der Methode entfaltet und eine immer stärker werdende interdisziplinäre Orientierung aufweist. Dies eröffnet neue, viel versprechende Forschungsperspektiven; innovative methodische Wege zeichnen sich immer deutlicher ab. Die wissenschaftliche Diskussion über die Fortschritte, vor allem aber über die Entwicklungsperspektiven, allen voran im methodischen Bereich, muss allerdings ununterbrochen fortgeführt werden und diese Prozesse begleiten. Dabei muss auch dem rasanten Rhythmus der Entwicklung der Digitalität Rechnung getragen werden.

Wir sind Clara Comas Valls, Charlotte Siemeling, Magnus Fischer, Giulia Agnello-Steil und Valentina Fabris für ihre Unterstützung bei der Erstellung

der Druckvorlage sowie Kathrin Heyng (Narr Francke Attempto Verlag) für die Betreuung dieses Bandes zu Dank verpflichtet.

1 Bibliographie

FEW = Wartburg, Walther von et al. (1922-2002): *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes*, 25 vols., Bonn et al., Klopp et al.

LEI = Prifti, Elton/Schweickard, Wolfgang (eds.) (1979–): *Lessico Etimologico Italiano (LEI), fondato da Max Pfister*, Wiesbaden, Reichert.

Prifti, Elton (2022): „Il LEI digitale. Un resoconto, con particolare attenzione alla dialettologia", in: Cortelazzo, Michele A./Morgana, Silvia/Prada, Massimo (eds.): *Lessicografia storica dialettale e regionale*, Firenze, Franco Cesati, 293–314.

REW_3 = Meyer-Lübke, Wilhelm (^{3}1935): *Romanisches etymologisches Wörterbuch*, Heidelberg, Winter.

Lidia Becker
Julia Kuhn
Christina Ossenkop
Claudia Polzin-Haumann
Elton Prifti

Methodologie

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). Definitorische 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 dazuzugehören 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 angestrengten 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 Begleitschreiben, 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 Sätze 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 bräuchte 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 Werbeannoncen 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 MTLD-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 Enzyklopädiesprache 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