DisKo: Zur Einbindung von Citizen Humanities beim Aufbau eines Diversitäts-Korpus

Schumacher, Mareike
schumacher@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Marie, Flüh
marie.flueh@uni-hamburg.de
Universität Hamburg

Peter, Leinen
P.Leinen@dnb.de
Deutsche Nationalbibliothek

Abstract

Für die Korpuskonstituierung im Projekt DisKo (Diversitäts-Korpus) haben wir ein Konzept entwickelt, das sowohl auf Ansätzen aus den Digital als auch den Public Humanities aufbaut und diese zu einer Citizen-Humanities-Komponente zusammenfügt (vgl. Heinisch 2020). Dieses Konzept dient dazu, ein Erzähltextkorpus aufzubauen, in dem Gender nicht nur binär dargestellt wird. DisKo wird zur Grundlage eines Gender-Classifiers 2.0 (aufbauend auf dem Gender-Classifier 1.0 von Schumacher und Flüh 2021), den wir mithilfe von Verfahren des überwachten maschinellen Lernens so trainieren, dass diverse Gender-Zuschreibungen automatisch klassifiziert werden. Den Grundsätzen von Offenheit, Transparenz und Empowerment folgend, die für Citizen-Humanities-Projekte zentral sind (vgl. Heinisch 2020, Dunn und Hedges 2012: 19), werden Vertreter:innen unterschiedlicher Communities angesprochen und in die Korpuskonstituierung einbezogen. Dabei steht das Community-Building zwischen Offenheit und Zielgruppenspezifik.

DisKo (Diversitäts-Korpus)

DisKo ist ein Kooperationsprojekt zwischen der Deutschen Nationalbibliothek (DNB), der Technischen Universität Darmstadt und der Universität Hamburg. Die DNB sammelt im gesetzlichen Auftrag seit 1913 u.a. alle in Deutschland veröffentlichten Medienwerke; seit 2006 schließt dies auch die sogenannten Netzpublikationen, also genuin digitale Werke, mit ein. Die Digitalisierungsstrategie zielt auf eine systematische und auch projekt- und anlassbezogene Digitalisierung der physischen Bestände sowie die Vernetzung zur Wissenschaft ab und schreibt seit 2020 mit dem jährlichen DH-Call ein Unterstützungsangebot für Forschende aus, die mit den Daten der DNB arbeiten möchten. Grundlage dieser Aktivität bildet die Reform des Urheberrechts-Wissensgesellschafts-Gesetzes (UrhWissG) im Jahr 2018. Die konkrete Förderung durch die DNB besteht in der Bereitstellung von Metadaten, digitalen bzw. digitalisierten Objekten und einer passenden Infrastruktur zur Bearbeitung und Analyse der teilweise sensiblen Daten. Das Projekt DisKo ist seit Frühjahr 2022 Teil dieser Förderlinie und kann darum nicht nur auf genuin digitale Medien zugreifen, sondern auch Texte retrodigitalisieren lassen, die bisher von keinem Digitalisierungsprojekt erfasst wurden. Darüber hinaus bietet die Zusammenarbeit mit der DNB die Möglichkeit, bei der Korpusbildung eine quantitative Einschätzung der Grundgesamtheit (Calvo Tello 2021: 96–97, Schöch 2017: 225) aller in einer definierten Zeitspanne in Deutschland erschienen Romane zu berücksichtigen, da jedes mit einer deutschen ISBN erscheinende Buch hier gemeldet und in doppelter Ausführung abgegeben oder als digitales Objekt eingereicht werden muss.

Im Projekt m*w werden seit 2019 Genderrollen und -stereotype erforscht. Der Gender-Classifier 1.0 erkennt und klassifiziert weibliche, männliche und neutrale Genderrollen durchschnittlich zu 78% (F1-Score) in Novellen, Romanen und Dramen (vgl. Flüh/Lemke/Schumacher 2022). Bei der Annotation des Trainingskorpus und Fallstudien mit diesem Classifier (vgl. Schumacher und Flüh 2020; Flüh und Schumacher 2021/2022; Flüh/Horstmann/Schumacher 2022) zeigt sich, dass Brüche mit stereotypen Genderzuschreibungen in älteren Texten selten, in zeitgenössischen aber häufiger vorkommen. Aus dem Desiderat, einen Classifier zu trainieren, der diverse, nicht nur binäre Genderrollen erkennen und klassifizieren kann, resultierte das Projekt DisKo: die Erschließung eines Trainingskorpus aus den Beständen der DNB aus zeitgenössischen Romanen mit nicht-binären Genderdarstellungen.

Erfasst wird ein Zeitrahmen der letzten rund 70 Jahre; in diesem Fall beinhaltet die Grundgesamtheit also alle in Deutschland zwischen 1950 und 2022 erschienen belletristischen Werke. Übertragen auf den Gesamtbestand der DNB bedeutet das, dass prinzipiell ca. 450.000 physische Objekte und ca. 435.000 digitale Objekte mit dem Erschließungsmerkmal “Belletristik” in Frage kommen. Die Überschneidungsmenge der Bestände ist leider nicht bekannt, kann jedoch über Algorithmen des Werkclustering durch die DNB eingegrenzt werden. Angesichts des Umfangs des als Basis für das maschinelle Lernen potentiell geeigneten Datensatzes sind wir mit grundlegenden Herausforderungen der Korpuskonstituierung konfrontiert, wie sie auch Gius et al. (2019) beschreiben: Die Menge (digital) vorliegender Texte ist so groß, dass der naheliegendste Weg der Korpusbildung darin bestünde, sich dabei auf eine Auswertung der Metadaten zu beschränken. Weil es sich bei “Figurengender” um einen textimmanenten Aspekt handelt, der in den Metadaten nicht erfasst wird, ist diese Vorgehensweise hier nicht möglich. Für den Aufbau des Korpus kommen standardisierte Methoden wie *Random Sampling* und *Stratified Sampling* (Calvo Tello 2021: 107; Schöch 2017: 226) ebenfalls nur bedingt in Frage. *Random Sampling* ist ungeeignet, weil wir Texte benötigen,

in denen zuverlässig Figuren diverser Gender-Kategorien vorkommen. Beim Aufbau einer balancierten Sammlung (*Stratified Sampling*), in der "für alle Kombinationen wesentlicher Merkmale eine Mindestanzahl von Datensätzen" (Schöch 2017: 226) vorkommen müsste, ist problematisch, dass es keine endliche Liste wesentlicher Merkmale der Genderthematik gibt. Ob eine Figur im Hinblick auf Gender stereotyp oder ungewöhnlich ist und welche Kategorien es jenseits der traditionellen Einteilung in "männlich" und "weiblich" gibt, ist nicht klar definiert. Darum nutzen wir eine dritte Methode: die Verkleinerung der Population durch ergänzende Kriterien (Calvo Tello 2021: 108–109), mit Zügen einer opportunistischen Auswahl (Schöch 2017: 226) unter Einbezug von geisteswissenschaftlichem Crowd-Sourcing (vgl. Dunn und Hedges 2012). Vor dem Hintergrund, dass große Teile der Grundgesamtheit aktuell ausschließlich physisch vorliegen und somit für digitale Analysemethoden vorerst nicht in Frage kommen, scheint dies der einzig mögliche Weg zu sein. Die drei Kriterien, die wir bei der Korpuszusammenstellung berücksichtigen, sind:

1. Kriterium der Ausbalanciertheit
   - aus jedem Jahr wird zunächst nur ein Roman übernommen
   - von jedem Autor/jeder Autorin wird nur ein Roman übernommen
2. Kriterium der Heterogenität in Bezug auf
   - literarische Genres
   - Autor*innengender
3. Kriterium der thematischen Relevanz: Nur Romane werden übernommen, in denen Figuren vorkommen, die
   - mit stereotypen Genderrollen brechen
   - sich nicht klar in ein binäres Gendersystem fügen

Während Parameter der Kriterien I. und II. aus den in der DNB erfassten Metadaten abgeleitet werden können, handelt es sich bei III. um ein Kriterium, das nicht erfasst wird. Erschwerend kommt hinzu, dass die Parameter der Kategorie III. nicht klar definiert sind, sondern von Interpretationen und (unbewussten) Vorannahmen abhängen. Um sowohl im Hinblick auf die Interpretation von Figurengender als auch auf versteckte Vorannahmen eine möglichst große Heterogenität zu erreichen und auf diese Weise den Aspekt des Representation-Bias (Suresh und Guttag 2019: 4) mit einzubeziehen, haben wir für die Korpuskonstituierung eine dreiteilige Citizen-Humanities-Komponente konzipiert.

Citizen Humanities in DisKo

Der Aufbau eines Diversitäts-Korpus bringt drei Herausforderungen mit sich. Erstens können die Texte in der Grundgesamtheit nicht algorithmisch erschlossen werden, da große Datenmengen nur physisch vorliegen. Zweitens ist die Auswahl von Texten, in denen Gender nicht (nur) binär dargestellt wird, bereits eine interpretatorische Leistung. Darüber hinaus muss drittens der sog. Representation-Bias mit einbezogen werden, der personengebunden funktioniert (vgl. D'Ignazio und Klein 2020: 53; Suresh und Guttag 2019). Im Falle von DisKo ist einerseits die Frage zentral, was genau unter nicht-binären Genderdarstellungen zu verstehen ist. Reicht es aus, wenn eine literarische Figur einer (binären) Genderkategorie mit einer Reihe von Eigenschaften beschrieben wird, die traditionell eher der anderen (binären) Genderkategorie zugeschrieben werden? Oder muss eine Figur explizit mit Begriffen wie "queer" oder "gender-fluid" charakterisiert werden? Wie explizit oder implizit müssen nicht-binäre Genderdarstellungen angelegt sein, damit sie als solche erkannt werden? Zu Beginn der Korpusgestaltung steht also eine interpretatorische Leistung, deren Ziel die Auslegung des Verständnisses von nicht-binären Genderkategorisierungen ist. Hinzu kommt, dass bei dieser Thematik Aspekte von Macht und Suppression eine Rolle spielen. D'Ignazio und Klein weisen in *Data Feminism* darauf hin, dass es wichtig ist, Daten zu sammeln, die marginalisierte Gruppen sichtbar machen (D'Ignazio und Klein 2020: 119). Darüber hinaus sollte die betreffende Community beim Sammeln der Daten einbezogen werden, um einen Empowerment-Effekt zu erreichen (vgl. D'Ignazio und Klein 202: 120, Heinisch 2020: 164). Außerdem ist für den Aufbau des Diversitäts-Korpus *Embodied Knowledge* (Christie et al. 2020) von Bedeutung; gerade bei Projekten, die Aspekte des Feminismus und der queer Community umfassen, ist die körperliche soziale Erfahrung Grundbestandteil einer Verstehensleistung, die zu einem kulturellen Wissen beiträgt (vgl. Christie et al. 2020). Im Sinne eines geisteswissenschaftlichen Crowd-Sourcing (Dunn und Hedges 2012) betrachten wir es darüber hinaus als Gelingensbedingung des Projektes DisKo, möglichst viele Personen an der Korpuszusammenstellung zu beteiligen, die im Hinblick auf ihre Genderzugehörigkeit und ihren beruflichen Hintergrund divers sind. Über einen eigens entwickelten Fragebogen werden laufend Vorschläge für DisKo eingereicht. Fakultativ können gleichzeitig Daten zu Genderzugehörigkeit und beruflichem bzw. interessengeleitetem Hintergrund angegeben werden. Von Beginn an werden auf der Webseite des m*w-Projektes auf einer eigenen Seite die Liste der für DisKo eingereichten Buchtitel sowie auch die dabei angegebenen Metadaten zu den Einreichenden offen einsehbar zugänglich gemacht (vgl. Schumacher und Flüh 2022). So werden die ethischen Grundsätze von Citizen Humanities *Transparenz* in Bezug auf beteiligte Interessengruppen (vgl. Heinisch 2020: 15), *Offenheit* der Ergebnisse des Crowd-Sourcing (vgl. Dunn und Hedges 2012: 19) sowie ein informativer *Mehrwert für Beteiligte* (vgl. Heinisch 2020: 15) gewährleistet.

Der Fragebogen wird in drei Phasen in unterschiedlichen Communities verbreitet, deren Auswahl auf Basis der drei Dimensionen der partizipatorischen Wissenschaft – wissenschaftlicher Impact und Output, Lernen, Involviertheit und Empowerment der Teilnehmenden und gesellschaftlicher Impact und Awareness in Bezug auf die Thematik – getroffen wurde (vgl. Heinisch 2020: 155). Der Idee Wodwards folgend, dass Communities sich um verbindende Fragestellungen herum bilden, wird jede Phase von einer Frage geleitet (Wodward 2007: 117). Ergänzend wird eine Disseminationsstrategie mit digitalen und analogen Anteilen umgesetzt, die auf Erkenntnisse aus dem Disseminationsprojekt forTEXT zurückgreift (Gius et al. 2021, Schumacher und Gius 2022, Schumacher und Horstmann 2019). Die drei Phasen sind non-exklusiv, d.h., wenn z.B. Mitglieder der primären Zielgruppe

aus Phase II schon in Phase I in den Community-Diskurs eintreten, so sind sie willkommen. Seit Beginn des Projektes und über alle Phasen hinweg wird der Blog des m*w-Projektes als Herzstück der Kommunikation genutzt.

Phase I: Wie wird ein literaturwissenschaftliches Korpus aufgebaut, das zur Basis automatisierter Gender-Klassifikation verwendet werden soll?

Kernzielgruppe dieser Phase ist die Digital-Humanities-Community, die hauptsächlich aufgrund der Methodik Interesse an dem Projekt zeigt. Als etabliertes Medium der Wissenschaftskommunikation innerhalb dieser Community, das darüber hinaus auch erhebliches Potential für geisteswissenschaftliche Wissenschaftskommunikation allgemein bietet (vgl. Geier und Gottschling 2019), wird ein Twitter-Account aufgebaut. Dabei setzen wir auf ein organisches Wachstum, bei dem Interesse an dem Projekt über die getwitterten Inhalte ausgelöst wird. Außerdem werden etablierte Informationskanäle genutzt, wie z.B. der Discord-Server DHall, die DHd-Mailingliste, der Fachinformationsdienst für Allgemeine und Vergleichende Literaturwissenschaft oder das Projektschaufenster der Webseite des DHd-Verbandes. Auch die Teilnahme an Fachkonferenzen stellt einen wichtigen Teil der ersten Phase dar.

Phase II: Was macht non-binäre Genderdarstellung aus?

In dieser Phase geht es darum, den ersten Outreach des Projektes zu generieren, indem Mitglieder der LGBTIQ+-Community und deren sogenannte Allies, die Interesse an der Genderthematik aufweisen, angesprochen werden. Zentral ist dabei die zielgruppengenaue Ausrichtung. Auch Kenntnis literarischer Texte ist vonnöten, sodass eine Community an der Schnittstelle zwischen LGBTIQ+-Themen und Interesse für Literatur gefunden werden muss. Diese Phase ist eng mit der ersten Phase des Community-Buildings verzahnt. Beteiligte der DH-Community dienen als Multiplikator*innen, indem sie z.B. im Rahmen ihrer Lehre auf das Projekt DisKo aufmerksam machen. Seitens des Projektes werden Gastvorträge in universitären Lehrveranstaltungen durchgeführt. Außerdem werden Flyer in Bibliotheken ausgelegt, die mittels QR-Code auf die DisKo-Umfrage verweisen. Darüber hinaus wird DisKo bei der Plattform *Bürger schaffen Wissen* eingereicht.

Phase III: Wie bedeutsam ist non-binäre Genderdarstellung für unsere Gesellschaft?

Das Thema Gender und insbesondere (non-)Binarität wird aktuell in unterschiedlichen Bereichen des öffentlichen Lebens diskutiert. Die gesellschaftliche Brisanz der Gender-Thematik und die Relevanz für den alltäglichen Umgang miteinander ist aber nicht nur derzeit ein wichtiges Thema. Mit dem Projekt m*w, in dessen Rahmen DisKo aufgebaut wird, möchten wir offenlegen, dass auch in der Literaturgeschichte immer wieder Figuren eine Rolle spielen, die sich nicht in ein binäres Gendersystem fügen lassen. Wir möchten zeigen, dass die aktuelle Debatte also nicht neu ist, sondern Gender-Diversität schon lange ein gesellschaftliches Thema ist, das u.a. in Kulturprodukten wie literarischen Texten eine wichtige Position einnimmt. In dieser Phase setzen wir die in Citizen-Science-Ansätzen häufig sehr stark verankerte Idee einer *Third Mission* von Forschungsprojekten (vgl. Heinisch 2020: 152–153) um, indem wir Ergebnisse mit einer möglichst breiten Öffentlichkeit teilen und in die aktuellen Debatten einfließen lassen. Darum suchen wir in dieser Phase weitere Wege der Wissenschaftskommunikation, wie z.B. über die Videoplattform TikTok oder den Bilder-Sharing-Dienst Instagram.

Erste Ergebnisse

Zum jetzigen Zeitpunkt (Stand Dezember 2022) befinden wir uns in Phase I des Citizen-Humanities-Projektes DisKo. Maßnahmen zur Verbreitung der Umfrage wurden auf Twitter, Mastodon, der DHd-Webseite, internen Kanälen der DNB und über Flyer umgesetzt. Außerdem wurde ein Gastvortrag im Rahmen der Ringvorlesung “Einführung in die Digital Humanities” an der Universität Hamburg gehalten. Der Rücklauf ist noch nicht sehr hoch. Insgesamt gab es 17 Teilnehmende der Umfrage. Allerdings wurden von diesen insgesamt 31 Titel angegeben, was bedeutet, dass jede*r Teilnehmende t durchschnittlich 1,8 Titel eingereicht hat. Die tatsächliche Verteilung ist recht heterogen, es wurden bis zu sechs Titel pro Person angegeben. Die freiwilligen Angaben zum eigenen Hintergrund wurden meist beantwortet, wie aus Abb. 1 ersichtlich wird.

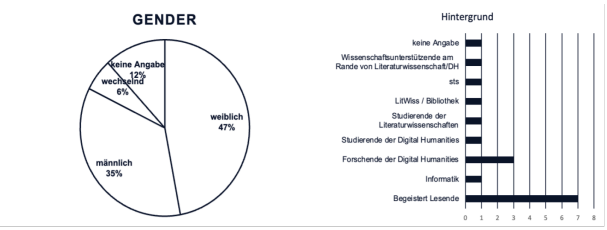

Abb. 1: Metadaten der Einreichenden der DisKo-Umfrage

Die eingereichten Titel umfassen derzeit eine Zeitspanne, die von 1928–2022 reicht, also 94 Jahre abdeckt. Der Schwerpunkt liegt mit 22 Erzähltexten auf Titeln, die nach 2000 erschienen sind. Nur acht der für DisKo vorgeschlagenen Texte sind vor dem Jahr 2000 erschienen. Es handelt sich bei den Texten sowohl um ursprünglich deutschsprachige Erzähltexte als auch um Übersetzungen. Zwar birgt die Integration von Übersetzungen für das Machine-Learning-Training die Gefahr, eine unkontrollierbare Variable einzubauen, da unklar ist, inwiefern die Ergebnisse der automatischen Erkennung davon beeinflusst werden. Dafür können aber Titel der Weltliteratur aufgenommen werden, die tatsächlich gelesen werden.

In Bezug auf die Genderdarstellungen reichen die Texte von der Erwähnung der nicht-stereotypen Genderindentität einer Figur (wie in Murakamis *Kafka am Strand*) bis hin zu einer Fülle nicht-stereotyper Genderidentitäten, die zum Hauptthema des Erzähltextes werden. Letzteres ist z.B. bei Evaristos *Mädchen, Frau, etc.* der Fall, sodass wir diesen Roman für unser Machine-Learning-Training als Testtext gewählt haben. Eine linguistische Besonderheit zeigt Leckies *Maschinen-Trilogie* , die durchgehend im generischen Femininum geschrieben wurde. Neben Murakami und Leckie wurden bisher acht weitere Texte ins Trainingskorpus übernommen. Von jedem der zehn Texte wurden zwei 2.000 Tokens umfassende Passagen ins Trainingskorpus integriert – eine vom Beginn und eine vom Ende des Textes (um eventuelle Transitionen berücksichtigen zu können). Erste Tests mit einem auf diesem 40.000 Tokens umfassenden Trainingskorpus trainierten Modell zeigen noch keine zufriedenstellenden Ergebnisse. Der F1-Score liegt insgesamt bei 0,35, die Erkennung der Kategorie "Divers" bei 0. Ein Blick auf die annotierten Beispiele im Trainingskorpus zeigt, dass für diese Kategorie nur 60 Vorkommnisse annotiert wurden, während die Kategorien "Frau" und "Neutral" jeweils rund 500 Vorkommnisse aufweisen, "Mann" sogar 902. Um hier zu einem ausgewogeneren Verhältnis zu kommen, könnten statt Anfangs- und Endpassagen, Ausschnitte aus den Texten ausfindig gemacht werden, in denen sich Genderzuschreibungen der Kategorie "Divers" häufen. Eine andere Möglichkeit wäre die Annotation kompletter Erzähltexte, in denen non-binäre Genderdarstellungen zum Hauptthema gemacht werden. Nach Abschluss der Pilot-Trainingsphase werden wir darum ein erneutes Training mit relevanteren Samples oder Volltexten anschließen.

Fazit

Die Korpuskonstituierung ist für Projekte, die Verfahren des maschinellen Lernens einsetzen, ein Dreh- und Angelpunkt. Alle weiteren Verfahrensschritte und Ergebnisse wie z.B. die Performanz eines Classifiers oder Analyseergebnisse, die durch dessen Einsatz erzielt werden, werden vom genutzten Korpus massiv beeinflusst. Unser Konzept der Korpuskonstituierung greift sowohl arbeitspraktische Ansätze der Digital Humanities wie das *Random Sampling* oder die opportunistische Auswahl als auch kritische Betrachtungen von Aspekten wie Representation-Bias und Empowerment auf. Mithilfe einer Citizen-Humanities-Komponente begegnen wir zentralen Herausforderungen beim Aufbau des Diversitäts-Korpus DisKo. Die strategische Korpuskonstituierung, die wir in unserem Vortrag präsentieren und zur Diskussion stellen wollen, ist dabei nicht nur ein optionaler Bestandteil, sondern wird zur Gelingensbedingung für eine möglichst diverse und ausgewogene Datenbasis.

Bibliographie

Calvo Tello, José. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning* . 1. Aufl. Bd. 4. Digital Humanities Research. Bielefeld, Germany: transcript Verlag / Bielefeld University Press. https://doi.org/10.14361/9783839459256 .

Chenier, Elise. 2014. „Oral History and Open Access: Fulfilling the Promise of Democratizing Knowledge" https://nanocrit.com/issues/issue5/notes-women-who-rock-making-scenes-building-communities-participatory-research-community-engagement-and-archival-practice [zugegriffen: 6. Juli 2021].

Christie, Alex, Jana Millar Usiskin, Jentery Sayers undKathryn Tanigawa. 2020."Introduction: Digital Humanities, Public Humanities - Nanocrit.Com." https://nanocrit.com/issues/issue5/introduction-digital-humanities-public-humanities [letzter Zugriff: 6. Juli 2021].

Dunn, Stuart E. und Mark Hedges. 2012. *Crowd-Sourcing Scoping Study: Engaging the Crowd with Humanities Research* . https://www.semanticscholar.org/paper/Crowd-Sourcing-Scoping-Study%3A-Engaging-the-Crowd-Hedges-Dunn/9940b0520332a6b0605559fd7c8c46672b3fb655 [zugegriffen: 6. Juli 2021].

Flüh, Marie und Mareike Schumacher. 2021. „Digitale Diachrone Korpusanalyse Am Beispiel Des Projekts ‚m*w – Gender Stereotype in Der Literatur'". *Digital Humanities and Gender History* . https://doi.org/10.22032/dbt.49173 .

Flüh, Marie, und Mareike Schumacher. 2022. „Jung, wild, emotional? Rollen und Emotionen Jugendlicher in zeitgenössischer Fantasy-Literatur". Gehalten auf der DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum" (DHd 2022), Potsdam, März 7. https://doi.org/10.5281/zenodo.6327983 .

Flüh, Marie, Jan Horstmann und Mareike Schumacher. 2022. "Distant Gender Reading

Genderaspekte in Fantasy-Jugendromanen von 2008 bis 2020". In: Weertje Willms (Hg.): *Gender in der deutschsprachigen Kinder- und Jugendliteratur: Vom Mittelalter bis zur Gegenwart . Gender in der deutschsprachigen Kinder- und Jugendliteratur* . Berlin: De Gruyter. https://www.degruyter.com/document/isbn/9783110726404/html?lang=de .

Flüh, Marie, Mark Lemke und Mareike Schumacher. 2022: The model of choice.Using pure CRF- and BERT-based classifiers for gender annotation in German fantasy fiction. In: Digital Humanities 2022 – Responding to Asian Diversity (DHTokyo).

Flüh, Marie und Mareike Schumacher (forthcoming): "Macht versus Emotion. Handlungstreibende Muster in Günderrodes Dramen digital, distant und scalable gelesen". In *Noch Zukunft haben. Das Werk Karoline von Günderrodes neu gelesen. Neue Romantikforschung, hg. von* Roland Borgards, Martina Wernli und Frederike Middelhoff Berlin: Springer.

Geier, Andrea und Markus Gottschling. 2019. „Wissenschaftskommunikation auf Twitter? Eine Chance für die Geisteswissenschaften!" *Mitteilungen des Deutschen Germanistenverbandes* 66 (3): 282–91. https://doi.org/10.14220/mdge.2019.66.3.282 .

Gius, Evelyn, Mareike Schumacher, Dominik Gerstorfer, Malte Meister, Sandra Bläß, Marie Flüh, Jan Horstmann, Janina Jacke, Christian Bruck und Marco Petris (2021): forTEXT. Literatur digital erforschen. URL: https://fortext.net [zugegriffen: 6. Juli 2021].

Gius, Evelyn, Krüger Katharina und Carla Sökefeld. 2019. „Korpuserstellung als literaturwissenschaftliche Aufgabe". Gehalten auf der DHd 2019 Digital Humanities multimedial und multimodal. 6. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum" (DHd 2019), Frankfurt am Main und Mainz, März 16. https://doi.org/10.5281/zenodo.4622112 .

Habell-Pallán, Michelle, Sonnet Retman und Angelica Macklin. 2014. „Notes on Women Who Rock: Making Scenes, Building Communities: Participatory Research, Community Engagement, and Archival Practice - nanocrit.com", 2014. URL: https://nanocrit.com/issues/issue5/notes-women-who-rock-making-scenes-building-communities-participatory-research-community-engagement-and-archival-practice [zugegriffen: 3. August 2021].

Heinisch, Barbara. 2020. "Citizen Humanities as a Fusion of Digital and Public Humanities?" *Magazén* , no. 2 (December): JournalArticle_3442. https://doi.org/10.30687/mag/2724-3923/2020/02/001 .

Henny-Kramer, Ulrike und Frederike Neuber. 2017. „Criteria for Reviewing Digital Text Collections, version 1.0 |". *IDE – Institut für Dokumentologie und Editorik* (blog). URL: https://www.i-d-e.de/publikationen/weitereschriften/criteria-text-collections-version-1-0/ . [zugegriffen: 3. August 2021].

Lassner, David. 2020. *Bericht aus dem Workshop zu Bias in Datensätzen und ML-Modellen. Erkennung und Umgang in den DH.* URL: https://digitalintellectuals.hypotheses.org/3262 [zugegriffen: 27. Juli 2022].

Schöch, Christof. 2017. „Aufbau von Datensammlungen". In *Digital Humanities: Eine Einführung* , hg. von Fotis Jannidis, Hubertus Kohle, und Malte Rehbein, 223–33. Stuttgart: J.B. Metzler. https://doi.org/10.1007/978-3-476-05446-3_16 .

Schumacher, Mareike und Evelyn Gius. 2022. "forTEXT – Literatur digital erforschen". *Mitteilungen des Deutschen Germanistenverbandes* Jg. 69, Heft 2. 2022. Vandenhoeck & Ruprecht Verlage.

Schumacher, Mareike, und Flüh, Marie. 2020. „m*w Figurengender zwischen Stereotypisierung und literarischen und theoretischen Spielräumen. Gendersterotype und -bewertungen in der Literatur des 19. Jahrhunderts". In *DHd2020: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts* , hg. von Christof Schöch 162–167. https://doi.org/10.5281/ZENODO.4621892 .

Suresh, Harini und John V. Guttag. 2019. "A framework for understanding unintended consequences of machine learning." arXiv preprint arXiv:1901.10002.

Schumacher, Mareike und Marie Flüh. 2022. "Jung, wild, emotional? Rollen und Emotionen Jugendlicher in zeitgenössischer Fantasy-Literatur". In *DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum" (DHd 2022),* Potsdam. https://doi.org/10.5281/zenodo.5555952 .

Schumacher, Mareike, und Marie Flüh . 2022. „DisKo – das Diversitäts-Korpus". Projektwebseite*. m*w* (blog). 2022. URL: https://msternchenw.de/diversitaets-korpus/ . [zugegriffen: 3. August 2021].

Schumacher, Mareike, und Jan Horstmann. 2019. „Social Media, YouTube und Co: Multimediale, multimodale und multicodierte Dissemination von Forschungsmethoden in forTEXT". Frankfurt am Main und Mainz, März 16. https://doi.org/10.5281/zenodo.4622253 .

Woodward, Kathleen. 2009. "The Future of the Humanities- in the Present & in Public." *Daedalus* 138 (1): 110–23. https://doi.org/10.1162/daed.2009.138.1.110.