„Ja, jetzt ist das langweilig. Aber in zwanzig Jahren!“
Bereitstellung, Zugang und Analyse literarischer Blogs am Beispiel des Techniktagebuchs

Blessing, André
andre.blessing@ims.uni-stuttgart.de
Universität Stuttgart, Institut für maschinelle Sprachverarbeitung

Hess, Jan
jan.hess@dla-marbach.de
Deutsches Literaturarchiv Marbach, Germany

Jung, Kerstin
kerstin.jung@ims.uni-stuttgart.de
Universität Stuttgart, Institut für maschinelle Sprachverarbeitung

„Ja, jetzt ist das langweilig. Aber in zwanzig Jahren!“ – Wie der Untertitel des *Techniktagebuchs*[1] bereits suggeriert, versteht sich das 2014 von der Schriftstellerin Kathrin Passig ins Leben gerufene literarische Blog als Teil des kulturellen Gedächtnisses. Mehr als sechzig regelmäßige Autor:innen bloggen hier oftmals humorvoll-pointiert über ihre Erlebnisse, Erfahrungen, Erfolge und Misserfolge im Zusammenhang mit technischen Themen und Gegenständen. Bereits seit 2008 werden literarische Blogs wie das *Techniktagebuch* am Deutschen Literaturarchiv Marbach gesammelt und archiviert. Im Rahmen des Projekts SDC4Lit – Science Data Center for Literature[2] werden nicht nur der Workflow zur Archivierung von Literatur im Netz aktualisiert, sondern auch Methoden und Werkzeuge zur (explorativen) Analyse von elektronischer Literatur getestet und (weiter-)entwickelt, um sie in Form von Analyse-Pipelines über die SDC4Lit-Plattform zur Verfügung zu stellen (Schlesinger 2021).

Am Beispiel des *Techniktagebuchs* soll der hier vorgeschlagene Poster-Beitrag einerseits die Herausforderungen bei der wissenschaftlichen Arbeit mit diesen bislang wenig erforschten „born-digital“ Literaturformen[3] aufzeigen sowie andererseits vor allem auch (explorative) Zugangs- und Analysemöglichkeiten literarischer Blogs präsentieren. Als Standard für die Archivierung von Webinhalten hat sich das WARC-Format (IPCC) etabliert. Resultierend aus der Blog-Software enthält der im Falle des *Techniktagebuchs* ca. neun Gigabytes große Blog-Crawl[4] neben den 7677 Blogbeiträgen eine Vielzahl an Metaseiten wie dem aus Tags bestehenden Stichwortverzeichnis oder den Blog-Archivseiten, die für die reine Inhaltsanalyse störend sind und entsprechend vorab identifiziert werden müssen. Eine Besonderheit des *Techniktagebuchs* und zugleich Grund für die Auswahl von Tumblr als Blogging-Plattform[5] ist zudem die Möglichkeit zur Rückdatierung einzelner Blogposts. Neben den aus aktuellen (technischen) Erlebnissen der Autor:innen resultierenden und entsprechend auf den tatsächlichen Schreibzeitpunkt datierten Beiträgen enthält das *Techniktagebuch* zahlreiche retroperspektivisch verfasste und entsprechend zurückdatierte, sodass das eigentliche Publikationsdatum zum Teil über die in der jeweiligen URL enthaltene Tumblr-ID ermittelt werden muss. Bereits aus der Diskrepanz zwischen dem eigentlichen Publikationsdatum und dem von den verschiedenen Autor:innen deklarierten Datum ergeben sich verschiedene Fragestellungen: Welche Autor:innen schreiben eher über Aktuelles, welche eher über Vergangenes? Lassen sich dabei jeweils thematische Schwerpunkte ermitteln? Wie verändern sich diese im zeitlichen Verlauf? Gibt es thematische "Trendsetter" unter den Autor:innen? Inwiefern beeinflussen sie sich gegenseitig?

Um sich Fragestellungen wie diesen nähern zu können, wurden die in den WARC-Records enthaltenen Daten zunächst mithilfe von WARC-Bibliotheken wie WARCIO[6] und JWARC[7] aufbereitet und die zu analysierenden Texte mithilfe von trafilatura (Barbaresi 2019) extrahiert. Mittels regelbasierter Informationsextraktion (Blessing 2014) wurden hieraus u. a. die in den Blogposts im Normalfall nur in Klammern am Textende genannten Autor:innen ermittelt. Basierend auf den so gewonnenen Informationen wurden verschiedene Netzwerke zur Visualisierung verschiedener Zusammenhänge erstellt. Analysiert wurden so beispielsweise die direkten Referenzierungen der Autor:innen innerhalb der Blogposts unter Berücksichtigung der deklarierten Publikationsdaten. Dabei wurde u. a. sichtbar, dass die Links zwar vorwiegend genutzt werden, um auf auch hinsichtlich des deklarierten Datums früher entstandene Blogposts zurück zu verweisen oder gar Fortsetzungsgeschichten in Form mehrerer Blogbeiträge aufzubauen. Die hier dennoch zahlreich zutage tretenden, vermeintlich anachronistischen Verweise auf künftige Beiträge verdeutlichen aber zugleich, dass die Autor:innen die Rückdatierungsoption durchaus kreativ zu nutzen wissen.

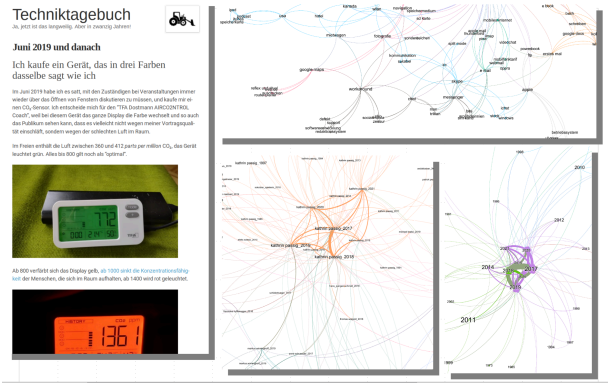

Abb. 1: Collage von Gephi-Visualisierungen zur Analyse des Techniktagebuchs.

Netzwerkvisualisierungen zu den von den Autor:innen vergebenen Tags in Kombination mit dem Einsatz von Topic-Model-Analysen (Blei 2003) zeigen demgegenüber, dass diese manuell definierten Schlagwörter der Autor:innen nicht konsistent, mitunter eher humoristisch verwendet werden. Um das *Techniktagebuch* korpuslinguistisch und stilometrisch auszuwerten, kamen bislang CQPWeb (Hardie 2012) sowie Stylo (Eder 2016) zum Einsatz, sodass auch hinsichtlich des Vokabulars und des individuellen Blog-Schreibstils Unterschiede aufgezeigt werden können. Interessant und überraschend ist beispielsweise bereits die später auf dem Poster zu findende Antwort auf die Frage nach den fünf Wörtern, die Kathrin Passig wesentlich häufiger nutzt als alle anderen Autor:innen.

Mit Blick auf die bereits erwähnte SDC4Lit-Plattform werden die genannten Methoden und Werkzeuge nicht nur hinsichtlich ihres konkreten Nutzens bei der beispielhaften, explorativen Ana-

lyse des *Techniktagebuchs* getestet, (weiter-)entwickelt und in Form von Analysepipelines kombiniert, sondern auch hinsichtlich ihrer Verwendbarkeit für die über das SDC4Lit-Repositorium zur Verfügung gestellten Materialien von Literatur im Netz. Letztlich sollen die erstellten Analyse-Pipelines demnach nicht nur auf anderen (literarischen) Blogs, sondern auch auf Werken der Netzliteratur Anwendung finden können, um möglichst niedrigschwellige Zugangsmöglichkeiten zu diesen trotz ihrer genuinen Digitalität eher vernachlässigten Literaturformen zu schaffen, die keinesfalls erst „in zwanzig Jahren“ in den (wissenschaftlichen) Fokus rücken sollten.

Fußnoten

1. Techniktagebuch, https://techniktagebuch.tumblr.com/ , Abruf: 15.7.2021.
2. SDC4Lit, https://sdc4lit.de/ , letzter Zugriff: 15.7.2021.
3. In Untersuchungen literarischer Blogs finden computergestützte Analysemöglichkeiten bislang kaum Verwendung, vgl. Fassio (2021), Knapp (2012, 2014) und Ainetter (2006).
4. Der Blogcrawl wurde erstellt mithilfe von Brozzler (https://github.com/internetarchive/brozzler , letzter Zugriff: 30.11.2021) und Heritrix (https://github.com/internetarchive/heritrix3 , letzter Zugriff: 30.11.2021).
5. Techniktagebuch, https://techniktagebuch.tumblr.com/post/76963766003/20140218 , letzter Zugriff : 15.07.2021.
6. WARCIO: WARC (and ARC) Streaming Library, https://github.com/webrecorder/warcio , letzter Zugriff: 30.11.2021.
7. JWARC, Java library for reading and writing WARC files with a typed API, https://github.com/iipc/jwarc , letzter Zugriff: 30.11.2021.

Bibliographie

Ainetter, Sylvia (2006): Blogs - literarische Aspekte eines neuen Mediums. Eine Analyse am Beispiel des Weblogs Miagolare (= Innsbrucker Studien zur Alltagsrezeption 5). Wien: Lit Verlag.

Barbaresi, Adrien (2019): “Generic Web Content Extraction with Open-Source Software”, in: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), 267–268.

Blei, David M. / Ng, Andrew Y. / Jordan, Michael I. (2003): “Latent dirichlet allocation”, in: Journal of Machine Learning Research 3, 993–1022.

Blessing, André / Kuhn, Jonas (2014): “Textual Emigration Analysis (TEA)”, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC' 14) European Language Resources Association (ELRA), Reykjavik, Iceland, 2089–2093.

Eder, Maciej / Rybicki, Jan / Kestemont, Mike (2016): “Stylometry with R: a package for computational text analysis”, in: R Journal, 8(1), 107–121.

Fassio, Marcella (2021): Das literarische Weblog. Praktiken, Poetiken, Autorschaften (= Praktiken der Subjektivierung 21). Bielefeld: Transcript.

Hardie, Andrew (2012): “CQPweb – combining power, flexibility and usability in a corpus analysis tool”, in: International Journal of Corpus Linguistics, 17(3), 380–409.

IIPC: The WARC Format. https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/, letzter Zugriff: 15.07.2021.

Knapp, Lore (2012): “Christoph Schlingensiefs Blog. Multimediale Autofiktion im Künstlerblog”, in: Ansgar Nünning und Jan Rupp (Hrsg.), Narrative Genres im Internet: Theoretische Bezugsrahmen, Mediengattungstypologie und Funktionen. Trier: Wissenschaftlicher Verlag, 117-132.

Knapp, Lore (2014): Künstlerblogs. Zum Einfluss der Digitalisierung auf literarische Schreibprozesse (Goetz, Schlingensief, Herrndorf), Berlin: Ripperger & Kremers.

Schlesinger, Claus-Michael / Ulrich, Mona / Hein, Pascal / Blessing, André (2021): Networks of Net Literature - Modelling, Extracting and Visualizing Link-Based Networks in the DLA corpus of net literature. Bergen: ELMCIP 2021, https://elmcip.net/node/16380, letzter Zugriff: 30.11.2021.