Merkmale registrieren oder textuelle Phänomene identifizieren? Zur Vereinbarkeit von automatischer und manueller Textsortenanalyse

Thielert, Frauke
frauke.thielert@upb.de
Universität Paderborn, Deutschland

Haaf, Susanne
haaf@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Schuster, Britt-Marie
brittms@mail.upb.de
Universität Paderborn, Deutschland

Georgi, Christopher
christopher.georgi@upb.de
Universität Paderborn, Deutschland

Kategorien wie „Textsorte“, „(kommunikative) Gattung“ oder „Genre“ gehören zu einem disziplinenübergreifenden Bestand und werden entsprechend in Sprach-, Literatur-, Kultur- sowie den Sozialwissenschaften verwendet. Allgemein lässt sich die Frage stellen, ob und inwieweit die genannten Kategorien in die Digital Humanities eingehen und inwieweit sie methodologisch reflektiert werden.

Textanalysen in der Text- und Korpuslinguistik

Die Auseinandersetzung mit einer Kategorie wie „Textsorte“ kann auf eine jahrzehntelange Fachgeschichte, besonders in der Sprachwissenschaft, zurückblicken, in der zwar keine Einigkeit hinsichtlich des Verständnisses des Konzepts „Textsorte“ erzielt worden ist, jedoch deutlich geworden ist, dass in diesem Zusammenhang die Unterschiede zwischen Texten unter Einbezug unterschiedlichster Textebenen zu modellieren sind. Textsorten zeigen sich – grosso modo – nicht nur anhand von musterhaften Ausprägungen auf textgrammatischer, -semantischer und -pragmatischer Ebene, sondern berühren auch die Materialität, Kodalität (Nutzung unterschiedlicher Zeichenressourcen) und ggf. eine spezifische Ortsgebundenheit, die Lokalität. Heutige Textsortenmodelle sind Mehrebenen-Modelle, was mit der Annahme verknüpft ist, dass die einzelnen Ebenen in einem wechselseitigen Abhängigkeitsverhältnis stehen (vgl. Adamzik [2] 2016; Heinemann/Heinemann 2002). Um eine „Textsorte“ oder ein „Textmuster“ zu erfassen, ist eine umfassende Nutzung des linguistischen Beschreibungsinstrumentariums erforderlich. Die Attraktivität von Kategorien wie „Textsorte“ ist v.a. darin gesehen worden, dass sie Einblick in den ‚kommunikativen Haushalt‘, also in spezifische Ordnungsleistungen einer Gesellschaft und Kultur ermöglichen (vgl. Fix 2006). I.d.R. wird die Ausprägung von Textmustern auf rekurrente Aufgaben und deren Lösung zurückgeführt, die wiederum einen Einblick in gesellschaftliche Relevanzen bieten. Gerade der in den letzten zwei Jahrzehnten geführte (text)linguistische Diskurs hat zudem erbracht, dass zunächst als dem Text äußerlich gedachte Faktoren wie Kontext, einschließlich der Beziehung zwischen Textproduzent und -rezipient, nichts dem Text Äußerliches sind, sondern durch den Text hergestellt werden. Zudem ist eine Kategorie wie „Stil“ die etwa auch Dialogizität oder Perspektivität umfasst, verstärkt als Textsortenstil verstanden worden, der sich aus der Sichtung aller Textebenen im Zusammenspiel ergibt (vgl. Sandig [2] 2006). Eine wichtige Neuorientierung in der textlinguistischen Betrachtung stellen Modelle dar, die konsequent von der textlichen Oberfläche ausgehend, ohne sich allerdings auf Syntax und Lexik zu beschränken, thematische, situative und funktionale Hinweise und damit zentrale Textdimensionen erschließen (vgl. Hausendorf et al. 2017, historisch: Schuster 2019).

Mehrebenen-Modelle zur Beschreibung von Textsorten sind fast ausnahmslos Produkt von Annahmen, die ebenso aus Sprach- und Kommunikationstheorien wie aus einzelnen Textexemplaren hergeleitet werden. Diese werden zumeist nur an geringen Textmengen überprüft. Da wichtige Untersuchungsebenen ‚vorgegeben‘ sind, verfährt die Methode top down. Wie korpuslinguistische Untersuchungen mit kulturanalytischen Interesse – also nicht im engeren Sinne textlinguistische Studien – deutlich gemacht haben, ließen sich einige auch in der Textlinguistik für wichtig erachtete Ausdrucksmuster durch die Berechnung von Kollokationen, n-Grammen auf Wort und Phrasenebene oder Keywords ermitteln (vgl. Bubenhofer/Scharloth 2016). Dabei handelt es sich um Bottom-Up-Verfahren, die zu neuen Hypothesen und Annahmen führen können.

Innerhalb der Diskussionen um Textsortenklassifikation und Texttypologie ist deutlich geworden, dass „Textsorten“ keine starren Entitäten sind; sie sind nicht vollständig festgelegt und erlauben Veränderungen. Aus dieser Variabilität ergibt sich das generelle Potential zum Wandel von Textsorten, der durch die Nutzung und Grenzen von Spielräumen bestimmt wird. Die entsprechenden Konventionalisierungsprozesse sind jedoch bisher kaum betrachtet worden.

Textanalysen in den Digital Humanities

Den bisher skizzierten Textauffassungen stehen Zugriffe auf die Kategorie „Text“ gegenüber, die in den Digital Humanities bevorzugt werden. Grundsätzlich scheint die Kategorie „Textsorte“ eine Hilfskategorie zu sein, mit der größere Datenmengen (z.B. Referenzkorpora) geordnet werden. Fragen der Textstrukturiertheit werden im Zusammenhang mit dem Text-Encoding z.B. in digitalen Editionen aufgeworfen (vgl. z.B. TEI-P5 Guidelines 2019), wobei die Ergebnisse nur selten Niederschlag in quantitativen Analysen finden. Texte werden zudem für das Training von Methoden ganz unterschiedlicher Anwendungen (z.B. Sentiment-Analysis, Stilometrie oder Topic Modelling) verwendet. Der Text(sorten)begriff bleibt dabei unterspezifiziert, indem „Text“ mit Dokumenten, Sätzen oder Mengen sinntragender Struktureinheiten gleichgesetzt (vgl. z.B. Ravi/Ravi 2015: 16; de Rose et al. 1997: 6) oder nach Alltagsverständnis differenziert wird (vgl. z.B. Medhat et al. 2014: 1096). Einschlägige Kategorien der DH sind daneben die des (Gattungs)Stils, Autorenstils oder Registers. Dabei deckt sich das Stilverständnis nicht mit dem holistischen Verständnis von „Stil“ als einer alle Textebenen durchwirkenden Kategorie, mit der sozialer Sinn erzeugt wird. Das Text- und insbesondere auch das Stil- und Registerverständnis der DH ist wesentlich an Merkmalen orientiert, wie dies etwa in der folgenden Äußerung zum Tragen kommt, die hinsichtlich des Verständnisses hochaggregierter geisteswissenschaftlicher Kategorien in den DH charakteristisch ist: „Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively.“ (Hermann et al. 2015: 44).

Merkmale bei Untersuchungen zu Textgattungen und Diskursen sind etwa Frequenzen von Inhalts- und Funktionswörtern, der Variantenreichtum des Wortschatzes, Satzlängen, n-Gramme oder mit Parsern ermittelte syntaktische Strukturen; die Auswahl wird in der Regel nicht begründet und scheint durch ihre Operationalisierbarkeit selbst gerechtfertigt. Exemplarisch hierfür steht das Topic-Modeling (Fankhauser et al. 2016, Viehhauser 2017). Dabei wird Text im Sinne des „bag-of-words“-Ansatzes als „Behältnis“ von Wörtern verstanden, wobei die grammatikalische Struktur und selbst die Wortfolge unberücksichtigt bleiben (vgl. Blei et al. 2003). Bibers (1988) und Bibers/Finegans (2014) multidimensionale Analysen (Schöch/Pielström 2014: S. 2f.), die sich am Genre- und Registerbegriff orientieren, fassen eine Vielzahl von Merkmalen zu Merkmalbündeln zusammen, berücksichtigen jedoch kaum die Funktionalität bzw. pragmatische Dimension von Texten. Auffällig ist, dass in diesen und anderen Studien Merkmalen wie der Satzlänge oder Komplexität von Sätzen eine Bedeutsamkeit für Stil, Register oder Genre zugeschrieben wird, die in qualitativen Studien randständig ist. Dass „formal features“ auch durchaus auf interpretierbaren Kategorien basieren, rückt ebenfalls wenig ins Bild. Zusammenfassend darf behauptet werden, dass bei Text- und Stilklassifizierungen in den DH Merkmale der Textoberfläche bevorzugt behandelt werden.

Unvereinbare Traditionen? Ein Fallbeispiel

Man kann mit Blick auf diese unterschiedlichen Forschungstraditionen, die hier bewusst pointiert gegenübergestellt wurden, konstatieren, dass herkömmliche qualitativ-linguistische Studien, obgleich sie stark mit dem Begriff „Muster“ operieren, sich bisher kaum für statistische Signifikanzen u.ä. interessiert haben, während wiederum stilo- und textometrische Studien mit einem „unterkomplexen Textbegriff“ arbeiten und nach Bubenhofer/Scharloth es bisher versäumt haben, „Texte als komplexes Gewebe zu operationalisieren“ (2015: 13). Grundsätzlich gilt: Während merkmalsorientierte Zugänge auf der textlichen bzw. sprachlichen Oberfläche operieren, gehen phänomenorientierte Modelle von textlichen Dimensionen (z.B. der Beziehungsdimension) aus, die in ihrer Relevanz für die textliche Kommunikation erkannt worden sind und auf ihre sprachliche Gestaltung hin befragt werden. Zwar mehren sich in den letzten Jahren die Versuche, im Sinne der „mixed methods“ quantitative und qualitative Methoden miteinander zu verbinden, jedoch ist im Hinblick auf den Text- und Textsortenbegriff bisher nicht deutlich, ob sich diese komplementär zueinander verhalten oder zu möglicherweise sich widersprechenden Befunden führen.

In unserem Beitrag möchten wir ein Mehrebenen-Modell vorstellen, das in dem DFG-Projekt: „Die Evolution von komplexen Textmustern: Entwicklung und Anwendung eines korpuslinguistischen Analyseverfahrens zur Erfassung der Mehrdimensionalität des Textmusterwandels“ entstanden ist. Es verbindet unterschiedliche Zugriffe auf die Kategorie „Text“ und bezieht quantitative und qualitative Text(sorten)analyse spiralförmig aufeinander. Am Beispiel der Verwendung personaldeiktischer Ausdrücke (*ich – du – wir – ihr*) und entsprechender Possessiva sowie Indefinitpronomen wie *man* , die in unterschiedlichen historischen Textgruppen leicht identifizierbar sind, möchten wir auf Basis eines Pilotkorpus von Zeitungstextsorten des Zeitraums 1830 bis 1930 sowie mehrerer Vergleichskorpora aus dem Deutschen Textarchiv (DTA) zeigen:

1. welche Texteigenschaften (allein) durch die automatische, korpusbasierte Textanalyse, insbesondere durch die Nutzung von Part-of-Speech- und Lemma-Informationen, auch in Bezug auf verschiedene Binnentextsorten, zutage treten und hinsichtlich welcher Forschungsfragen dies

aufschlussreich ist. So werden durch diachrone Längsschnittuntersuchungen Frequenz, Signifikanz und Typizität entsprechender Ausdrücke, letzteres insbesondere durch Bezugnahme auf Vergleichskorpora, jedoch auch eine hohe Varianz der Ausdrücke sichtbar. Eine derartige Zugriffsweise erlaubt, ergänzt durch POS-sensitive Suchen, einen Einblick in Konstanz und Wandel von Verfasserreferenz und Rezipientenansprache. Sie bieten durch ihre Irritationsmomente einen Ansatzpunkt, um Hypothesen zu Zeiträumen, die für Wandelphänomene interessant sind, zu bilden. Sie dienen ferner zum Abgleich mit auf schmalen Korpora generierten Ergebnissen (vgl. Lefévre 2017: 150), die durch eine solche Zugangsweise relativiert werden. So zeigt sich – gemessen an der vorliegenden Forschungsliteratur und an Vergleichskorpora – ein erstaunlicher Anstieg von *ich* -, *du* -, *-wir* und *ihr* - Verwendungen.

2. was durch eine flankierende manuelle Annotation mit einem vordefinierten Tagset ins Blickfeld rückt. Es wird deutlich, dass die personaldeiktischen Verwendungen sich nicht gleichmäßig über alle Textsorten verteilen, sondern sich besonderen Textsorten wie dem Erfahrungs- und Erlebnisbericht verdanken. Ferner wird deutlich, dass sich relativ von Textkotext und -kontext bestimmte Lesarten (z.B. das Verfasserkollektive oder Rezipienten umschließende, inklusive *wir*) herausbilden, die weiterführende Analysen zu sprachlicher Inklusion und Exklusion erlauben und damit die Beziehungsdimension von Texten erschließen sowie die Beantwortung von Fragestellungen zu Funktionalität und Sprachhandlungsprofilen der vorliegenden Textsorten ermöglichen.

Somit stehen einerseits die Wandelbarkeit der Verteilung von sprachlichen Einheiten vor dem weiten Horizont von Textgruppen, andererseits die Funktionalität von sprachlichen Einheiten für die Konstitution bestimmter Textsorten im Vordergrund. Sowohl die unterschiedliche Verteilung von personaldeiktischen Formen als auch die spezifische Funktionalität von sprachlichen Einheiten, wie wir diskutieren möchten, ist nicht selbsterklärend, sondern gleichermaßen von Forschungshypothesen und -interessen abhängig. Abschließend möchten wir deshalb Überlegungen zu den folgenden Fragen bieten: Ist die „Bricolage“ (Bubenhofer/Dreesen 2018) aus Ansätzen und Methoden sehr unterschiedlicher Forschungstraditionen überhaupt sinnvoll? Lassen sich komplexe, kontextbasierte deiktische Kategorien messen, aber auch: Lassen sich damit verknüpfte Handlungsmuster überhaupt operationalisieren und in einem Tagset darstellen?

Bibliographie

Adamzik, Kerstin (2016): *Textlinguistik. Grundlagen, Kontroversen, Perspektiven.* 2., völlig neu bearbeitete, aktualisierte und erweiterte Neuauflage. Berlin, Boston: De Gruyter.

Biber, Douglas / Finegan, Edward (1994): *Multi-Dimensional Analyses of Authors' Styles:* Some Case Studies from the Eigtheenth Century. Oxford: Oxford University Press.

Biber, Douglas (1988): *Variation across Speech and Writing* . Cambridge: Cambridge University Press.

Blei, David M. / Ng, Andrew Y. / Jordan, Michael I. (2003): "Latent Dirichlet Allocation", in: *Journal of Machine Learning Research* 3: 993–1022.

Bubenhofer, Noah / Dreesen, Philipp (2018): "Linguistik als antifragile Disziplin? Optionen in der digitalen Transformation", in: *Digital Classics Online* 4: 63–75.

Bubenhofer, Noah / Scharloth, Joachim (2016): "Kulturwissenschaftliche Orientierung in der Computer- und Korpuslinguistik", in: *Sprache – Kultur – Kommunikation / Language – Culture – Communication* . Ein internationales Handbuch zu Linguistik als Kulturwissenschaft / An International Handbook of Linguistics as a Cultural Discipline. Bd. 43. Berlin, Boston: De Gruyter: 924–933.

Bubenhofer, Noah / Scharloth, Joachim (2015): "Maschinelle Textanalyse im Zeichen von Big Data und Data-driven Turn – Überblick und Desiderate", in: *Zeitschrift für Germanistische Linguistik* 43: 1–26.

DeRose, Steven J. / Durand David G. / Mylonas, Elli / Renear, Allen H. (1997): "What is text, really?", in: *SIGDOC Asterisk Journal of Computer Documentation* 21.3: 1–24.

Fankhauser, Peter/ Knappen, Jörg / Teich, Elke (2016): "Topical Diversification Over Time In The Royal Society Corpus" in: Eder, Maciej / Rybick, Jan (eds.): *Digital Humanities* , 11.–16. Juli 2016, Krakow: Conference Abstracts.

Fix, Ulla (2006): "Was heißt Texte kulturell verstehen? Ein- und Zuordnungsprozesse beim Verstehen von Texten als kulturellen Entitäten", in: Blühdorn, Hardarik / Breindl, Eva / Waßner, Ulrich Hermann (eds.): *Text – Verstehen.* Grammatik und darüber hinaus. Berlin / Boston: De Gruyter: 254–276.

Hausendorf, Heiko / Kesselheim, Wolfgang / Kato, Hiloko / Breitenholz, Martina (2017). *Textkommunikation* : ein textlinguistischer Neuansatz zur Theorie und Empirie der Kommunikation mit und durch Schrift. Berlin / Boston: De Gruyter.

Heinemann, Wolfgang / Heinemann, Margot (2002): *Grundlagen der Textlinguistik* . Interaktion – Text – Diskurs. Berlin / Boston: De Gruyter.

Herrmann, Berenike J. / Dalen-Oskam, Karina van / Schöch, Christof (2015): "Revisiting Style, a Key Concept

in Literary Studies" in *: Journal of Literary Theory* 9: 25–52.

Hermanns, Fritz (2009): "Linguistische Hermeneutik. Überlegungen zur überfälligen Einrichtung eines in der Linguistik bislang fehlenden Teilfaches", in: Felder, Ekkehard (eds.): *Sprache* . Heidelberg: Springer: 179–214.

Jannidis, Fotis (2019): "Digitale Geisteswissenschaften – Offene Fragen, schöne Aussichten", in: *ZMK* 10: 63–70.

Lefévre, Michel (2017): "Von der "Berlinischen Privilegierten Zeitung" zur "Königlich Privilegierten Berlinischen Zeitung". Entwicklungstendenzen in der Äußerungsstruktur, Textgestaltung und Syntax", in: Pfefferkorn, Oliver / Riecke, Jörg / Schuster, Britt-Marie (eds.): *Die Zeitung als Medium* . Berlin / Boston: De Gruyter: 149–163.

Medhat, Walaa / Hassan, Ahmed / Korashy, Hoda (2014): "Sentiment analysis algorithms and applications. A survey", in: *Ain Shams Engineering Journal* 5: 1093–1113.

Ravi, Kumar / Ravi, Vadlamani (2015): "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications", in: *Knowledge-Based Systems* 89: 14–46.

Sandig, Barbara (2006): *Textstilistik der deutschen Sprache* . 2. völlig neu bearbeitete und erweiterte Auflage. Berlin / Boston: De Gruyter.

Schöch, Christof / Steffen Pielström (2014): *Für eine computergestützte literarische Gattungsstilistik, Jahrestagung der Digital Humanities im deutschsprachigen Raum.* http://dig-hum.de/sites/dig-hum.de/files/Schoch-Pielstrom_2014_Gattungsstilistik.pdf. [letzter Zugriff 23. September 2019]

Schuster, Britt-Marie (2019): "Sprachgeschichte als Geschichte von Texten", in: Bär Joachim / Lobenstein-Reichmann, Anja /Riecke, Jörg (eds.): *Handbuch Sprache in der Geschichte* . Berlin / Boston: De Gruyter: 219–240.

TEI Consortium (2019): TEI P5. Guidelines for Electronic Text Encoding and Interchange. Originally edited by C.M. Sperberg-McQueen and Lou Burnard for the ACH-ALLC-ACL Text Encoding Initiative, now entirely revised and expanded under the supervision of the Technical Council of the TEI Consortium. Version 3.6.0 (16. Juli 2019)

Viehhauser, Gabriel (2017): “Digitale Gattungsgeschichten. Minnesang zwischen generischer Konstanz und Wende.” In: Zeitschrift für digitale Geisteswissenschaften. 2017. PDF Format ohne Paginierung. DOI: 10.17175/2017_003.