Schott Campus

Elisabeth Klein

Big Data in den Geisteswissenschaften?

Ansätze und Perspektiven zur Bewältigung großer Datenmengen aus Sicht der Linguistik

Symposion »Stand und Perspektiven musikwissenschaftlicher Digital-Humanities-Projekte«

Beitragsarchiv des Internationalen Kongresses der Gesellschaft für Musikforschung, Mainz 2016 – »Wege der Musikwissenschaft«, hg. von Gabriele Buschmeier und Klaus Pietschmann, Mainz 2018

gfm GESELLSCHAFT FÜR MUSIKFORSCHUNG

SCHOTT

Elisabeth Klein

Big Data in den Geisteswissenschaften? Ansätze und Perspektiven zur Bewältigung großer Datenmengen aus Sicht der Linguistik

›Big Data‹ ist ein Modebegriff unserer Zeit und wird in der Regel in einem Atemzug mit Begriffen wie ›pattern recognition‹, ›machine learning‹ oder ›deep learning‹ genannt, die für probabilistische Verfahren stehen, die Zugriff und Analyse unvorstellbar großer Datenmengen ermöglichen sollen. Mit dieser Terminologie werden vor allem Naturwissenschaften, seit der Entschlüsselung des menschlichen Genoms insbesondere die Life Sciences, die Wirtschaftswissenschaften und die Informatik assoziiert. Im Kontext geisteswissenschaftlicher Disziplinen wirkt der Begriff ›Big Data‹ jedoch bislang eher befremdlich und muss daher zuerst ›entzaubert‹ werden. Dies geschieht zunächst, indem im Folgenden, wo immer möglich, der im deutschen Sprachgebrauch weniger diffuse und kaum mit Assoziationen beladene Terminus ›Massendaten‹ verwendet wird. In einem weiteren Schritt sollen Wesen und Nutzen von Massendaten in den Geisteswissenschaften fokussiert werden, um dann ihr Potenzial zur Inspiration neuer Forschungsideen zu diskutieren.

An dieser Stelle lohnt es sich, das Thema Massendaten aus der Perspektive angrenzender Disziplinen wie Informationswissenschaft, Informatik, empirische Sozialforschung und (Computer-/Korpus-) Linguistik zu beleuchten. Aufgrund der Schnittmengen und Gemeinsamkeiten mit den Forschungsinteressen klassischer Geisteswissenschaften bieten sich diese Disziplinen an, um ihnen Methoden zu entleihen, die Zugänge zu Massendaten bieten. Dabei stellen sich drei grundsätzliche Fragen, die die Auseinandersetzung mit dem Thema Massendaten in Bezug auf die Geisteswissenschaften im Folgenden strukturieren sollen:

1. Was sind Massendaten und was bedeutet dieser Begriff in Bezug auf die Geisteswissenschaften?
2. Wie können Massendaten technisch zugänglich gemacht werden?
3. Ist ein interpretativer Zugang zu Massendaten möglich und falls ja, wie?

Wann immer möglich soll hier der Versuch unternommen werden, Anknüpfungspunkte speziell für die Musikforschung zu finden.

1. Was bedeutet Big Data für die Geisteswissenschaften?

Die Beschäftigung mit dem Thema Massendaten erfordert zunächst eine Klärung, was Big Data ist und was der Begriff in den und vor allem für die Geisteswissenschaften bedeutet. Doug Laney[1] bestimmt Big Data, d. h. Massendaten, durch drei Eigenschaften: Datenvolumen (volume), Produktions- und Analysegeschwindigkeit (velocity) und Heterogenität (variety).

[1] Doug Laney, »3D-Data Management: Controlling Data: Volume, Velocity and Variety«, https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf, 29.4.2017.

Naheliegenderweise zeichnen sich Massendaten durch ein hohes Datenvolumen aus, das sich oft im Tera- und Petabyte-Bereich bewegt, wobei der Big-Data-Begriff selbst bisher keine numerische Definition des Volumens beinhaltet. Daran zeigt sich, dass selbst die gesamte deutschsprachige *Wikipedia*, deren Texte (d. h. ohne Bilder und Mediendateien) als 6 GB großer sogenannter ›text dump‹ zum Download zur Verfügung stehen, noch nicht unbedingt als Big Data gelten muss.[2] Ihr Gehalt für geistes-, kultur- und sozialwissenschaftliche Forschung ist dabei ungleich höher und qualifiziert *Wikipedia* allein durch die enthaltene Informationsmenge als forschungsrelevante Massendaten. Jedoch ist es nicht unbedingt diese Eigenschaft, die den Umgang schwierig macht, sondern vor allem die hohe Produktionsgeschwindigkeit, mit der solche Daten entstehen, und die Notwendigkeit von Analysen in Echtzeit wie etwa bei Sensordaten. Dabei weisen Massendaten meist eine enorme Heterogenität auf. Diese Kombination aus großem Umfang, Produktions- und Analysegeschwindigkeit sowie Vielfalt ist mit traditionellen Methoden der Datenverarbeitung selbst für die Informatik nur schwer zu handhaben.[3]

Augenscheinlich treffen diese Merkmale auf Daten, die in den Geisteswissenschaften erzeugt und verarbeitet werden, nicht zu. Entsprechend findet der Begriff Big Data kaum Verwendung.[4] Möchte man den Begriff aber dennoch auf geisteswissenschaftliche Kontexte übertragen, wie z. B. Christoph Schöch[5] es tut, manifestiert sich vor allem das Merkmal Heterogenität nicht nur in Form heterogener Quellen und Formate, sondern vor allem durch die mangelnde Strukturierung von Texten, die keine weiteren Auszeichnungen haben. Diese mangelnde Strukturierung der Daten verweist bereits darauf, warum sie mit traditionellen Methoden schwer zu analysieren sind. Ein Beispiel für unstrukturierten Text sind die Ausgaben der *ZEIT*. Zwar existieren Metadaten zu den Texten, die sie z. B. als Zeitungsartikel eines bestimmten Datums und einer bestimmten Ausgabe kenntlich machen, eine Strukturierung innerhalb des Textes, die etwa Überschriften, Personen oder Orte erkennbar macht, fehlt jedoch. Big Data ist aufgrund ihrer mangelnden Strukturiertheit also vor allem ›messy data‹.[6]

2. Große Datenmengen strukturieren?

Von ›Big Data‹ lässt sich sogenannte ›Smart Data‹ unterscheiden, die (intellektuell oder halbautomatisch) strukturiert und angereichert ist und so bedeutend mehr und besser zugängliche Informationen bietet. Strukturierte Daten im typischen Sinne sind in Datenbanken organisiert, die Objekte (z. B. eine Person) und Relationen (z. B. die Herrschaftszeiträume dieser Person) deutlich erkennbar machen. Besonders Inschriftendatenbanken und in solcher Form strukturierte Werkverzeichnisse sind ›Smart Data‹, die seit langem etabliert und in den Geisteswissenschaften vorherrschend sind. Dazu kommen semi-strukturierte Daten, die durch XML-Auszeichnung mit TEI (*Text Encoding Initiative*) für Texte bzw. MEI (*Music Encoding Initiative*) für Musik zustande kommen. Diese semi-strukturierten Daten werden, wie besonders digitale Editionen zeigen, durch die Verknüpfung mit weiteren Datenbeständen angereichert. Dadurch werden Personen, Orte, Ereignisse, Zeiten und vieles mehr nicht nur maschinell erkenn- und auswertbar, sondern durch die Verknüpfung mit Normdaten (GND: *Gemeinsame Normdatei*, VIAF:

---

[2] https://de.wikipedia.org/wiki/Wikipedia:Technik/Datenbank/Download, 05.05.2017.

[3] Vgl. Chris Snijders, Uwe Matzat und Ulf-Dietrich Reips, »Big Data«: Big Gaps of Knowledge in the Field of Internet Science«, in: *International Journal of Internet Science* 7 (2012), Nr. 1, S. 1–5, http://www.ijis.net/ijis7_1/ijis7_1_editorial.pdf, 5.5.2017.

[4] Christoph Schöch, »Big? Smart? Clean? Messy? Data in the Humanities«, in: *Journal of Digital Humanities* 2 (2013), Nr. 3. http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/, 5.5.2017 sowie Anna Aurast u. a., »Big Data und Smart Data in den Geisteswissenschaften«, in: *Bibliothek Forschung und Praxis* 40 (2016), S. 200–206, doi:10.1515/bfp-2016-0033, 5.5.2017.

[5] Ebd.

[6] Ebd.

Virtual Authority File), weiteren kontrollierten Vokabularen (Thesauri, Ontologien) sowie mit Speziallexika auch eindeutig identifizierbar. Normdaten und kontrollierte Vokabulare können mit den Informationen im Text verknüpft werden. Kontrollierte Vokabulare bieten zudem Möglichkeiten zur differenzierten Erschließung von Inhalten. Diese Verbindung von Datenbeständen ermöglicht umfassende Analysen, die große Zusammenhänge aufdecken können, wie etwa die Erarbeitung von Korrespondenznetzwerken und Migrationsbewegungen zeigen. Besonders der Linked-Data-Ansatz bietet neue Möglichkeiten der Analyse, indem semantische Ansätze eingebracht werden.

Es ist gerade diese Form der Strukturierung und Anreicherung, die Smart Data auszeichnet und damit zu ›Qualitätsdaten‹ macht. Semi-strukturierte Daten sind gut in strukturierte Formate überführbar, indem Informationen aus TEI- und MEI-annotierten Daten automatisiert in Datenbanken übertragen werden können. Strukturierende Eingriffe bedeuten jedoch – das zeigen vor allem digitale Editionen und Nachweisdatenbanken – extrem hohen Arbeitsaufwand und erlauben daher nur relativ kleine Datenmengen zu bearbeiten, wie etwa eine digitale Mozart-Edition. Echte Massendaten im eingangs definierten Sinne sind also mit den bislang verwendeten Methoden nicht in Qualitätsdaten umzuwandeln und aufgrund ihrer mangelnden Strukturiertheit einer Analyse zugänglich zu machen.

3. Zugänge zu Massendaten?

Wie kann dennoch ein Zugang zu textuellen oder auditiven Massendaten geschaffen werden? Hier bieten die eingangs erwähnten quantitativen, probabilistischen Methoden, die latente Strukturen in großen Text- und Musik(signal)massen aufdecken können, gute Zugriffsmöglichkeiten.

In Bezug auf unstrukturierte Textdaten (Sprachkorpora, digitale Volltexte von Romanen oder Libretti) haben sich in der Linguistik und darüber hinaus Methoden des Textmining eingebürgert, die sich zwischen Statistik, natürlicher Sprachverarbeitung (natural language processing, kurz: NLP) und Machine learning verorten lassen. Am Anfang solcher Textanalysen steht häufig eine explorative Datenvisualisierung auf der Basis ganzer Wörter oder Wortbestandteile bzw. Graphemfolgen (sogenannte ›n-grams‹), wie hier exemplarisch anhand des Vorkommens des Begriffs ›Gutmensch‹ in Google Books von 1800 bis 2016 erhoben (vgl. Abbildung 1).[7]

Abb. 1: Explorative Datenvisualisierung zum Vorkommen von ›Gutmensch‹ in Google Books zwischen 1800 und 2016 (https://tinyurl.com/ngrams-gutmensch).

[7] N-Gramme können etwa zur Autorenattribution verwendet werden, vgl. Alexis Antonia, Hugh Craig und Jack Elliott, »Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution«, in: *Literary and Linguistic Computing* 29 (2014), S. 147–163, doi:10.1093/llc/fqt028, 5.5.2017 sowie Alexander Koplenig, »The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets – Reconstructing the composition of the german corpus in times of WWII«, in: *Digital Scholarship in the Humanities* 32 (2017), doi:10.1093/llc/fqv037, http://dsh.oxfordjournals.org/content/early/2015/09/02/llc.fqv037, 5.5.2017.

Ziel ist es, mittels solcher Visualisierungen erste makroperspektivische Zugänge zu den Daten zu gewinnen, die dann Schritt für Schritt mittels weiterer Verfahren verfeinert werden können.[8] Beispielsweise können in der nächsten Analysestufe clusteranalytische Verfahren zum Einsatz kommen wie etwa bei Franco Moretti[9] in der Literaturwissenschaft oder Oliver Čulo[10] in Bezug auf die linguistische Translationswissenschaft, wenn es darum geht, anhand großer Textkorpora Literaturgeschichte zu untersuchen oder Schlüsselwörter zur Bestimmung von Bedeutungskontexten zu extrahieren. Mit solchen statischen Verfahren können beispielsweise durch Sentiment-Analysen auch die emotionale Ausrichtung von Texten zugänglich gemacht und diese Texte (halb-)automatisch zu Gruppen zusammengefasst oder vorgegebenen Gruppen zugeordnet werden.[11] Diese und weitere Verfahren der computerbasierten Analyse von Texten setzen jedoch in der Regel eine zumindest grundlegende vorherige Strukturierung der Daten voraus. Hierbei kann bis zu einem bestimmten Grad die automatische Erkennung von Wortarten (sogenanntes ›part-of-speech-Tagging‹) und Entitäten (Namen, Orte, Relationen) aus der natürlichen Sprachverarbeitung helfen.

In Bezug auf Musik bzw. Audiosignale gibt es bereits eine bis in die 1990er Jahre zurückreichende Tradition, Methoden der maschinellen Mustererkennung für Vergleich und Klassifikation von Musikstücken zu verwenden. Mustererkennungsalgorithmen werden dabei zur Erkennung der Komplexität von Rhythmen[12], Musikstilen[13] und Soundscapes[14] verwendet. Besonders die Verbindung der computerunterstützten Erkennung von Audiosignalen mit Methoden des Textmining bietet Raum für neue Forschungsfragen. Beispielsweise können in Gesangbüchern sprachliche Inhalte auf ihre Verknüpfung mit bestimmten musikalischen Formen auf der Basis mittlerer und großer Mengen hin untersucht werden, um Makrostrukturen aufzudecken, die qualitativen Analysen, die nur wenige Beispiele untersuchen können, verborgen bleiben.

---

[8] Zu makroperspektivischen Zugängen in der Literaturwissenschaft siehe Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History*, Urbana (Illinois) u. a. 2013.

[9] Franco Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History*, London und New York 2005. Zum Aufdecken struktureller Ähnlichkeiten über Texte hinweg mittels Ansätzen aus der natürlichen Sprachverarbeitung (NLP) siehe Nils Reiter, Anette Frank und Oliver Hellwig, »An NLP-based cross-document approach to narrative structure discovery«, in: *Literary and Linguistic Computing* 29 (2014), S. 583–605, doi:10.1093/llc/fqu055, 5.5.2017.

[10] Oliver Čulo, »Die digitale Perspektive auf das UeLex: Ein Zwischenruf«, in: *Das Germersheimer Übersetzerlexikon*, hrsg. von Andreas Kelletat und Aleksey Tashinskiy, Berlin 2016, S. 281–296.

[11] Zur Verwendung von Sentiment-Analysen zur Bestimmung der emotionalen Ausrichtung von Forenbeiträgen siehe William N. Robinson, Tianjie Deng und Zirun Qi, »Developer Behavior and Sentiment from Data Mining Open Source Repositories«, in: *49th Hawaii International Conference on System Sciences (HICSS)*, IEEE (2016), S. 3729–3738, http://dx.doi.org/10.1109/HICSS.2016.465, 5.5.2017. Einen vergleichenden Überblick über Verfahren zur Textklassifikation bietet Bei Yu, »An evaluation of text classification methods for literary study«, in: *Literary and Linguistic Computing* 23 (2008), S. 327–343, doi:10.1093/llc/fqn015, 5.5.2017.

[12] Ilya Shmulevich und Dirk-Jan Povel, »Rhythm complexity measures for music pattern recognition«, in: *1998 IEEE Second Workshop on Multimedia Signal Processing (Cat. No.98EX175)*, IEEE (1998), S. 167–172, http://dx.doi.org/10.1109/MMSP.1998.738930, 5.5.2017 sowie Ilya Shmulevich u. a., »Perceptual Issues in Music Pattern Recognition: Complexity of Rhythm and Key Finding«, in: *Computers and the Humanities* 35 (2001), S. 23–35, http://dx.doi.org/10.1023/A:1002629217152, 5.5.2017.

[13] Pedro J. Ponce de León und José M. Iñesta, »Pattern Recognition Approach for Music Style Identification Using Shallow Statistical Descriptors«, in: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37 (2007), S. 248–257, doi:10.1109/TSMCC.2006.876045, 5.5.2017.

[14] Jean-Julien Aucouturier, Boris Defreville und François Pachet, »The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music«, in: *The Journal of the Acoustical Society of America* 122 (2007), Nr. 2, S. 881–891, doi:10.1121/1.2750160, 5.5.2017.

4. Potenziale von Makroanalysen

Quantitative Verfahren wie die angerissenen bieten eine Makroperspektive auf unstrukturierte und strukturierte Daten, die große Zusammenhänge aufdecken und mögliche Parallelen aufzeigen können, wo vorher nur exemplarische Analysen und darauf basierend begrenzte Thesenformulierungen möglich waren. Die dadurch eröffnete Vogelperspektive auf Daten bietet eine Form des »digitalen Querlesens«[15], die den Blick auf große Strukturen eröffnen und auf partikuläre Phänomene erweitern kann.

Dadurch werden Forschungsfragen umfassender oder bisweilen überhaupt erst bearbeitbar, neue Fragen werden aufgeworfen: Wortkarrieren und Begriffsgeschichten können zum Beispiel aus großen diachron ausgerichteten Textkorpora wie dem Deutschen Textarchiv[16], das Volltexte vom 17.–19. Jahrhundert vorhält, und weiteren Textkollektionen aus Zeitungen und Magazinen abgeleitet werden, narrative Techniken sind aus hunderten von Romanen oder Libretti ableitbar, statt nur aus wenigen – sofern ausreichend Daten vorhanden sind. Dadurch wird empirische Theoriebildung in einem vorher nicht vorstellbaren Maße möglich: In Massendaten sind Makrostrukturen zu entdecken, die zu neuen Hypothesen und Theorien führen. Umgekehrt ist auch die Validierung bestehender Theorien, die aus exemplarischen, qualitativen Analysen entwickelt wurden, an großen Datensamples möglich. So können etwa bestehende Kategoriekonstruktionen überprüft, hinterfragt und überarbeitet werden.

Bei aller ›Schöne-neue-Datenwelt-Musik‹, die in solchen Potenzialen anklingt, deuten sich auch Grenzen der Erkenntnismöglichkeiten aus Massendaten an: Auch die maschinelle Analyse erfordert mitunter extensive intellektuelle Bearbeitung, um Daten zu bereinigen, und bleibt weiterhin interpretationsbedürftig. Quantitative Ansätze wie die hier angerissenen können Strukturen zwar offenlegen und mittels geeigneter Methoden Zusammenhänge zwischen Phänomenen erklären; Sinnverstehen kann jedoch häufig nur rekonstruktiv auf Basis qualitativer Methoden erreicht werden.[17]

Hinzu kommt, dass es nach wie vor neben der notwendigen Interpretation statistischer Maße und visueller Outputs quantitativer Daten der Reflexion der gewählten Daten an sich und ihres Zustandekommens bedarf. Hier kommt vor allem dem Zusammenstellen der Daten, z. B. in Form von Text- oder Datenkorpora, also für spezifische Forschungsfragen generierte Datenkollektionen, hohe Relevanz zu.

Will man die Potenziale quantitativer Zugänge voll ausnutzen und sie für das Verstehen von sozio-kultureller Bedeutung ausbauen, sind Forschungsdesigns, die quantitative und qualitative Methoden vereinen, enorm gewinnbringend.

Statistische Aussagen über prozentuale Anteile von Vorkommnissen an einem Gesamtphänomen sind interessant, das Aufdecken latenter Zusammenhänge durch multivariate Datenanalysen noch interessanter für die Forschungsarbeit. Will man sich jedoch nicht allein mit solchen erklärenden Ansätzen im Weber'schen Sinne begnügen, sondern Bedeutungen verstehen, bietet es sich an, die Analyse durch qualitative Methoden zu ergänzen.

Dabei wird die Makroperspektive des digitalen Querlesens mit der Mikroperspektive exemplarischer Analysen verbunden. Dies ermöglicht nicht nur tiefe Einblicke in latente Strukturen und Phänomene, sondern auch Serendipitätseffekte, die auf die weitere Analyse zurückwirken können.[18]

---

[15] Čulo, »Die digitale Perspektive auf das UeLex«, S. 281.
[16] http://www.deutschestextarchiv.de/, 29.4.2017.
[17] Hier klingt bewusst Max Webers Unterscheidung zwischen Erklären und Verstehen für die Sozialwissenschaften an. Vgl. Max Weber, *Wirtschaft und Gesellschaft*, Tübingen 1922, dort besonders S. 20f.
[18] Unter Serendipität wird in den Sozialwissenschaften in der Folge von Robert K. Merton (vgl. Robert K. Merton, *Auf den Schultern von Riesen*, Frankfurt a. M. 2004 sowie Robert K. Merton und Elinor Barber, *The Travels and Adventures of Serendipity:*

5. Fazit

Die Analyse von Massendaten eröffnet eine Makroperspektive, die es ermöglicht, latente Strukturen zu erkennen und aus qualitativen Analysen entwickelte Hypothesen zu überprüfen. Jedoch stellen große, unstrukturierte Datenmengen maschinelle Analyseverfahren weiterhin vor große Herausforderungen. Computerlinguistische Verfahren ermöglichen Zugänge zu unstrukturierten Daten, indem sie schwer handhabbare Massendaten (vor)strukturieren und damit den besser analysierbaren Qualitätsdaten im Sinne von ›Smart Data‹ annähern. Dabei gilt nach wie vor, dass Analyseergebnisse und Erkenntnisgewinn in höchstem Maße von der Zusammenstellung der richtigen Daten und möglichst hoher Qualität abhängig sind. Ein Ansatz, zu relevanten und interessanten Ergebnissen zu gelangen, liegt einerseits in der Verbindung von Massendaten (Big Data) und Qualitätsdaten (Smart Data) und andererseits in der Verknüpfung von qualitativer Mikroperspektive und quantitativer Makroperspektive.

A Study in Sociological Semantics and the Sociology of Science, Princeton 2004) das zufällige Auffinden ursprünglich nicht gesuchter, aber fragestellungsrelevanter Phänomene verstanden.