Die explorative Visualisierung von Texten

Von den Herausforderungen der Darstellung geisteswissenschaftlicher Primär- und Annotationsdaten

Evelyn Gius und Marco Petris, Universität Hamburg

1. Zur Komplexität von Textdaten

Die Visualisierung von Textdaten ist innerhalb des Bereichs der Datenvisualisierung eine besondere Herausforderung, da es sich bei ihnen um unstrukturierte Daten handelt: Bevor man Textdaten visualisieren kann, muss aus ihnen eine Struktur abgeleitet werden. Hinzu kommt, dass Textdaten eine Vielzahl an Betrachtungsmöglichkeiten eröffnen, die durch die zahlreichen Bedeutungsdimensionen von Texten bedingt werden. Die einzelnen Dimensionen von Texten können durch Annotationen herausgearbeitet werden, wobei jede Annotationsschicht eine oder mehrere Dimensionen des Textes offenlegen kann. In diesem Sinne sind Textdaten also multidimensional. Insbesondere im Bereich der geisteswissenschaftlichen Textanalyse ist aufgrund des hermeneutischen Zugangs zu Texten auch in spezifischen Analysen nicht von vornherein klar, auf welche Weise Analyse und Interpretation zusammenhängen. Entsprechend muss eine sinnvolle Visualisierung von Textdaten im geisteswissenschaftlichen Kontext als exploratives Werkzeug zum Herausarbeiten möglicher Zusammenhänge fungieren können.[1]

Ein Blick in die einschlägige Literatur zur Datenvisualisierung zeigt, dass der Besonderheit von Textdaten häufig nicht Rechnung getragen wird. Zumindest scheint im traditionell mit Datenvisualisierung befassten informationswissenschaftlichen Bereich die Komplexität von annotierten Textdaten nicht immer im vollen Umfang wahrgenommen zu werden. So verweisen etwa Ward et al. (2010) in ihrer umfassenden Einführung zu Datenvisualisierung im Kapitel zu Textdaten auf drei mit Texten zusammenhängende Sucharten, die für die Anforderungen an die Visualisierung von Texten ausschlaggebend sind.[2] Die auf die Sucharten folgende Darstellung von Möglichkeiten der Textvisualisierung beschränkt sich allerdings auf die Darstellung von Texten und Korpora mit Metadaten (Erscheinungsjahr, Publikation o.ä.). Das Problem wird in der Zusammenfassung des Textvisualisierungskapitels offensichtlich: Die diskutierten Ansätze beträfen das "transforming unstructured text into structured data suitable for visualization and analysis" (Ward et al. 2010: 311). Die Option, dass der Text bereits mit Analysedaten in Form von

[1] Vorgehen, die darauf basieren, die Komplexität der Daten automatisiert zu reduzieren, erscheinen uns deshalb auch nicht geeignet für das beschriebene Problem (vgl. zu solchen Ansätzen z.B. Yang et al. 2003; Tatu et al. 2011).

[2] Typischerweise würden Zeichenketten in Form von Wörtern, Phrasen oder Themen gesucht, im Falle von partiell strukturierten Daten könnte außerdem nach Beziehungen zwischen Wörtern, Phrasen, Themen oder Dokumenten gesucht werden und schließlich ginge es in strukturierten Texten oder Textkorpora meistens um das Identifizieren von Mustern oder Auffälligkeiten innerhalb von Texten bzw. Dokumenten (vgl. Ward et al. 2010:291). Annotierte Textdaten fallen also potentiell unter die letzten beiden Fälle.

Annotationen angereichert sein könnte, wird nicht in Betracht gezogen. Das, obwohl in der Einleitung auf die drei Ebenen von Texten verwiesen – die lexikalische, die syntaktische und die semantische – und im Fall der syntaktischen Ebenen sogar explizit die Möglichkeit von Annotationen im Rahmen von *named entity recognition* (NER)-Prozessen erwähnt wurde (Ward et al. 2010:294).

2. Geisteswissenschaftliche Textdaten

In der stark geisteswissenschaftlich orientierten Position von Drucker (2014) werden hingegen die vielfältigen Interpretationsmöglichkeiten in den Fokus gerückt. Sie schreibt über die Visualisierung geisteswissenschaftlicher Interpretation: "The challenge is enormous, but essential, if the humanistic worldview, grounded in the recognition of the interpretive nature of knowledge, is to be part of the graphical expressions that come into play in the digital environment" (Drucker 2014: 136). Drucker geht es v.a. darum, die mit geisteswissenschaftlichen Analysen einhergehende Unsicherheit in der Darstellung des Wissens zu verdeutlichen, wobei sie sich nicht nur auf Texte beschränkt.

Was bedeutet das im Falle von Texten? Betrachten wir die Problematik an mit CATMA[3] annotierten Texten, die durch die flexiblen Annotationsmöglichkeiten des Werkzeugs exemplarisch für die große Bandbreite und gleichzeitig eingeschränkte Vorhersagbarkeit geisteswissenschaftlicher Analysen sind.[4] Für die Visualisierung von in CATMA erzeugten Text- und Annotationsdaten ist die von Drucker angesprochene Unsicherheit geringer, da es um die Analyse von Texten geht: Sie beschränkt sich auf (Text-)Interpretationen und liegt zudem nur in Form von Annotationen vor, die diese Unsicherheit konzeptionell durch entsprechende Tags fassen. Die Tags selbst beinhalten aber keine Unsicherheit, die für die weitere Analyse berücksichtigt werden muss.[5] Trotzdem ist Druckers Beobachtung zur Besonderheit geisteswissenschaftlicher Aussagen auch für unseren Zweck gültig und muss für die Visualisierung der Text- und Annotationsdaten berücksichtigt werden: "[…] we need to conceive of every metric 'as a factor of X', where X is a point of view, agenda, assumption, presumption, or simply a convention. By qualifying any metric as a factor of some condition, the character of the 'information' shifts from self-evident 'fact' to constructed interpretation motivated by a human agenda." (Drucker 2014:131). Aufgrund des freien Annotationsschemas, das CATMA zur Verfügung stellt, ist die Art der "Information", die die Annotationen enthalten, nämlich nicht über die vorliegenden Daten zugänglich: Man kann in CATMA genauso gut strukturelle Textmerkmale wie inhaltliche Aspekte annotieren und dafür eine eigene Annotationshierarchie entwickeln, deren Struktur zwar von der Anlage her hierarchisch ist, die aber prinzipiell überlappendes und widersprüchliches Markup zulässt.

[3] CATMA = Computer Aided Text Markup and Analysis, vgl. www.catma.de (gesehen am 10.11.2014).
[4] In CATMA können Texte anhand von frei gewählten Tags annotiert werden, die zu so genannten Tagsets zusammengefasst werden. Die so entstehende Taxonomie oder Systematik kann wiederverwendet werden. Die Texte und die Annotationen können außerdem mit einer umfangreichen Suchfunktionalität durchsucht und analysiert – und ggf. weiter annotiert werden. Zum damit außerdem verbundenen Konzept des hermeneutischen Markups vgl. Bögel et al. (im Erscheinen).
[5] vgl. dazu Jacke & Meister (2014).

3. Anforderungen an Visualisierung als Exploration

Aufgrund der nicht a priori eingrenzbaren Zwecke der Annotation und der Analyse muss die Visualisierung von Textdaten so generisch wie möglich gestalten werden. Nur so kann sie ohne ein tieferes Verständnis über die jeweils vorliegenden Text- und Annotationsdaten eingesetzt werden und einen Mehrwert bei der Analyse der Daten erzeugen.[6] Grundsätzlich konzipieren wir Visualisierungen deshalb ausgehend von der Frage, wie viele und welche Dimensionen der Daten dargestellt werden sollen.[7]

Für die Auswahl der Dimensionen stellt CATMA über die Struktur der Ergebnismenge der Abfragen folgende Kategorien zur Verfügung:

- Metadaten der Dokumente (z.B. Titel, Autor, etc.),
- Tag bzw. Typ der Annotation,
- Properties der Annotation und die für den annotierten Text vergebenen Werte,
- annotierter Text,
- Position im Text (via Zeichen-Offset),
- Textkontext des annotierten Textes (variable Anzahl von Token),
- Vorkommenshäufigkeit des annotierten Textes,
- Vorkommenshäufigkeit der Annotation,
- weitere berechnete Kategorien, wie der z-Faktor oder der TF-IDF

Neben dem generischen Zugang über die Dimensionen der Daten muss auch ein Mechanismus zur Verfügung gestellt werden, mit dem der Zweck einer spezifischen Analyse in der Visualisierung der Daten herausgearbeitet werden kann – und der die erzeugten Visualisierungen als explorative Heuristik nutzbar macht. Für die damit zusammenhängenden spezifischen Erkenntnisinteressen werden deshalb zusätzliche Anpassungsmöglichkeiten in Form von wählbaren Parametern eingeführt. Diese sollen typische Varianten abfangen, wie etwa die Frage, ob die Häufigkeit einer Annotation oder aber der annotierte Textumfang dargestellt werden soll, wie mit überlappenden Annotationen verfahren werden soll (soll etwa eine zweifach annotierte Stelle zweimal oder nur einmal gezählt bzw. dargestellt werden?) oder ob die Struktur der Tagsets so ist, dass sich Tags auf derselben Hierarchieebene gegenseitig ausschließen oder ob sie sich ergänzen können.[8]

4. Beispiele

Die oben angestellten Überlegungen sollen an folgenden Beispielen demonstriert werden. Datengrundlage ist das in Gius (2013) beschriebene und analysierte umfangreich

[6] Dies gilt nicht für die Visualisierung zu Demonstrations- bzw. Kommunikationszwecken – also von Daten, die bereits analysiert und interpretiert wurden.

[7] Mit “Dimensionen” sind also nicht räumliche Dimensionen gemeint. Dies wird allerdings von einigen gängigen Ansätzen zur Visualisierung mehrdimensionaler Daten angenommen, die Textdaten nur als einen – eindimensionalen – Datentyp betrachten und allgemeine Modelle entwickeln (vgl. etwa Shneiderman 1996).

[8] Auch hier unterscheidet sich der vorgestellte Ansatz durch seinen Fokus auf die Spezifik von Texten wieder deutlich von Ward et al. (2010) oder Shneiderman (1996), die so genannte *tasks* als Basis für zusätzliche explorative Funktionalitäten betrachten.

annotierte Korpus.[9] Die hier nur kurz beschriebenen Visualisierungen werden ebenso wie eine Reihe weiterer Visualisierungen in CATMA zur Verfügung gestellt. Ihre Funktionen und der damit verbundene explorative Gewinn werden im Rahmen des Vortrags näher vorgestellt werden.

Interaktive TreeMap

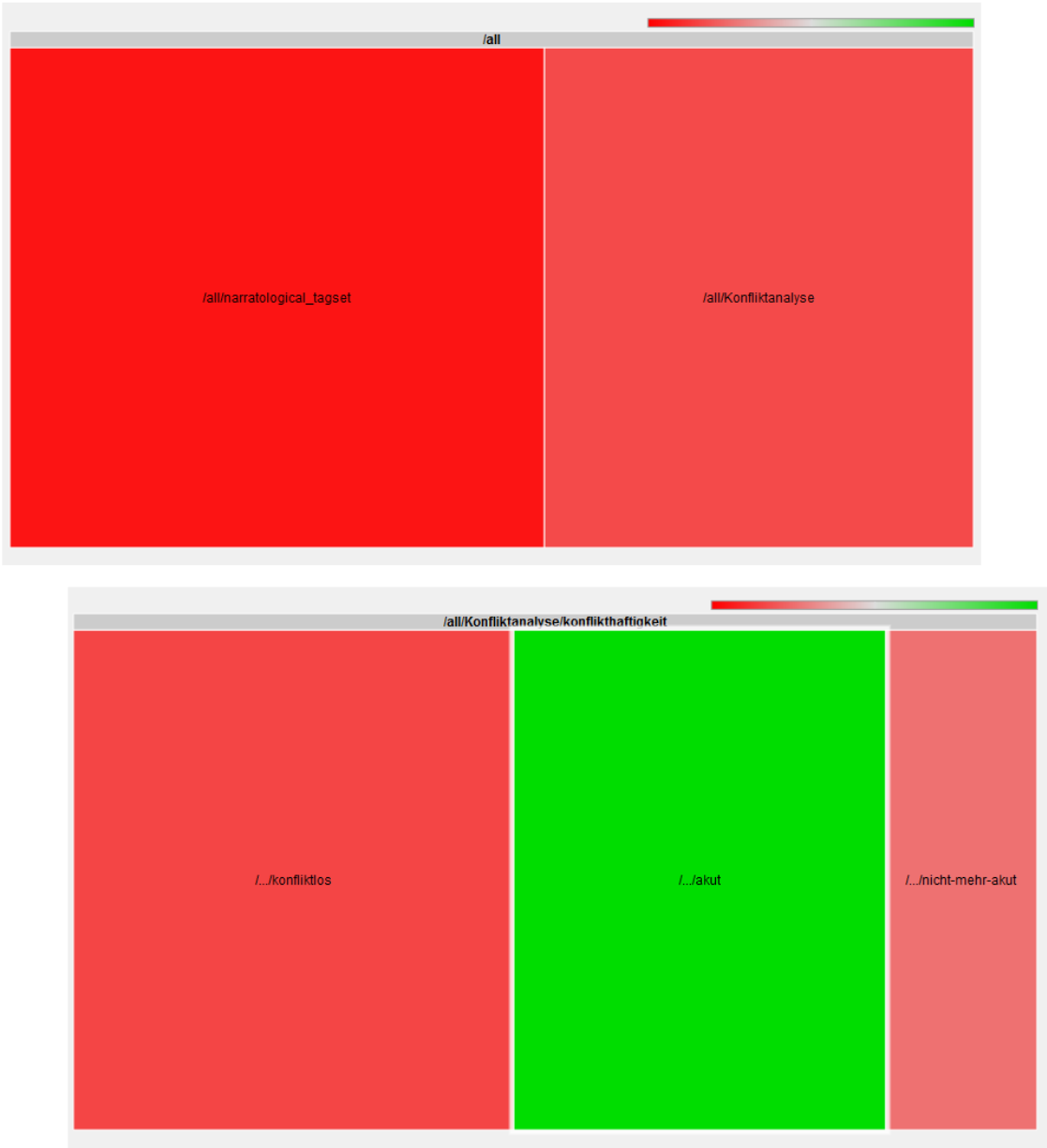

Abbildung 1: Interaktive TreeMap[10]

[9] Das Korpus besteht aus 24 Texten mit insgesamt 86.246 Wörter, die auf etwa 150 als Tags eingeführte narratologische Konzepte untersucht und mit insgesamt 24.347 Annotationen versehen wurden.

[10] Erstellt auf Basis von Google Charts: https://developers.google.com/chart/interactive/docs/gallery/treemap?hl=de (gesehen am 10.11.2014).

Das erste Beispiel ist eine interaktive TreeMap. Jede einzelne Sicht zeigt zwei Dimensionen: (1) Die Vorkommenshäufigkeit der Abfrageergebnisse (Tags, Wörter, o.ä.) als Größe des zugehörigen Rechtecks und (2) die durchschnittliche annotierte Textmenge als Farbintensität auf einer Skala von rot (weniger) bis grün (mehr). Die interaktive Komponente ermöglicht das Ergründen einer dritten Dimension: Durch Klicken der einzelnen Rechtecke kann man durch (3) die Hierarchie des Tagsets navigieren.
Abbildung 1 zeigt zwei Ebenen: Rechts die höchste Ebene mit den beiden Top-Level Tags "narratological_tagset" und "Konfliktanalyse" und links die Ebene 1 den Zweig entlang dem Tag "Konfliktanalyse" mit den Tags der darunter liegenden Ebene. Gezeigt wird also die Verteilung von Vorkommenshäufigkeit und annotierter Textmenge für die Hierarchieebene. Für die dargestellte Datenbasis ist das insofern interessant, als hier die Konflikthaftigkeit von Erzählungen bzw. die als konflikthaft oder konfliktlos annotierten Passagen dargestellt werden. Für die Analyse des Korpus ist sowohl die Frage nach der Häufigkeit, in der konflikthafte Passagen auftauchen (sie unterbrechen nämlich von den Erzählerinnen eigentlich als konfliktlos deklarierte Erzählabschnitte und deshalb ist ihre Anzahl relevant), als auch die reine Textmenge, die sie umfassen (wird ausgiebiger über konfliktlose oder über konflikthafte Situationen erzählt?), interessant. Die Visualisierung als dreidimensionale TreeMap ermöglicht es, die beiden Betrachtungsweisen – Anzahl vs. Textmenge – überblickshaft in Beziehung zu setzen und dabei durch die hierarchisch angeordneten Tags zu navigieren, also zusammengefasste und detailliertere Perspektiven zu wählen.

Small Multiples

Das zweite Beispiel (vgl. Abbildung 2) zeigt die Vorkommenshäufigkeit von zwei Annotationen (Wiedergabe von mentalen Prozessen und Wiedergabe von Rede) im Textverlauf bei neun Texten des Korpus. Die Vorkommenshäufigkeit wird auf der y-Achse und der Textverlauf in 10%-Schritten auf der x-Achse dargestellt. Für jeden ausgewählten Text wird jeweils ein Koordinatensystem als dritte Dimension erstellt, in dem die Annotationen als farbige Linien abgebildet werden.[11]

Diese Darstellung ermöglicht eine explorative Betrachtung der Verteilung der beiden annotierten Phänomene in den Einzeltexten und einen ersten Überblick über mögliche Muster im gesamten Korpus. Für eine weitere Analyse können auffällige Stellen – wie etwa besondere Häufigkeiten in einem Textabschnitt oder der Wechsel von dominierender Redewiedergabe zu dominierender Wiedergabe von mentalen Prozessen – genauer betrachtet werden: Das Anklicken der entsprechenden Punkte im Graphen erzeugt eine KWIC(=*KeyWord In Context*)-Anzeige der Annotationen im betreffenden Textabschnitt, von denen aus wiederum durch Klicken in den Volltext gesprungen werden kann.

[11] Die Darstellung als Linie wurde aus Gründen der Übersichtlichkeit gewählt, mathematisch gesehen handelt es sich natürlich um diskrete Werte.

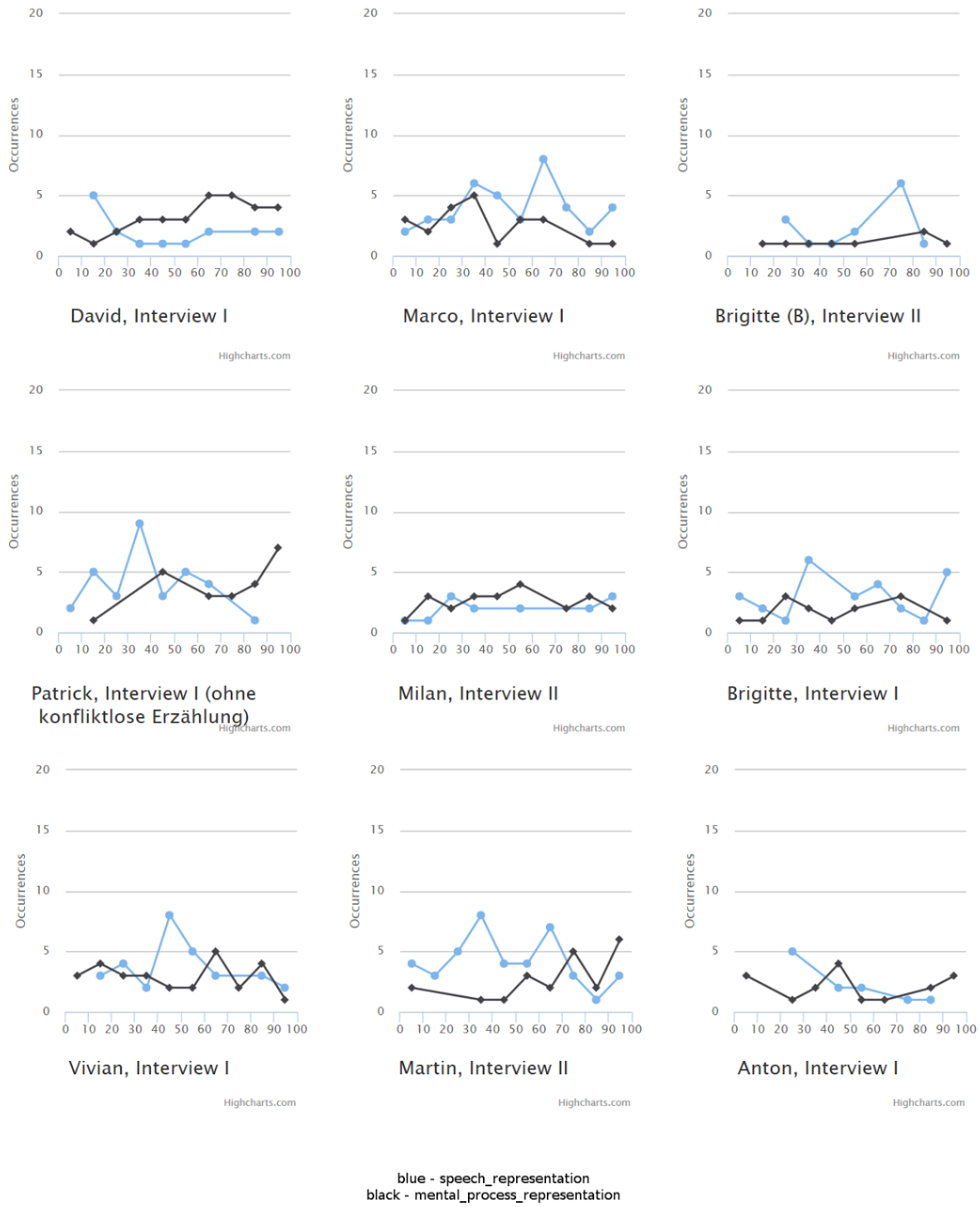

Abbildung 2: Small Multiples: Distributionsgraphen[12]

[12] Erstellt auf Basis von Highcharts: http://www.highcharts.com/ (gesehen am 10.11.2014).

5. Ausblick

Für das dargestellte Korpus sind oben beschriebene Visualisierungen von großem Gewinn. Sie ermöglichen einen Überblick über die Daten, für die nicht bereits bei der Annotation festgelegt wurde, welche Dimensionen genauer betrachtet und in Zusammenhang gebracht werden müssen. Dadurch sind die für die Analyse der Daten von großem Nutzen. Inwiefern sich nach in diesem Beitrag vorgestellten Überlegungen entwickelte Visualisierungen auch systematisch als heuristisches Werkzeug eignen und ob sie sich als Alternative gegen generelle Datenvisualisierungen durchsetzen können, ist zum momentanen Zeitpunkt allerdings noch nicht abschätzbar. Neben der Vieldimensionalität der Daten ist dafür insbesondere die Frage der explorativen Funktion der Visualisierungen zentral: Reichen (1) die dargelegte Aufschlüsselung der Analysen nach ihrer Dimensionalität und (2) die Explorationsmöglichkeit durch einstellbare Parameter aus, um Visualisierungen zu erzeugen, die systematische Rückschlüsse auf die dargestellten Daten und Strukturen zulassen – und nicht nur assoziative Denkanstöße zu liefern?

Dies wird in breit angelegten Nutzerstudien zu ergründen sein, ebenso wie untersucht werden muss, ob und auf welche Weise die in den Visualisierungen zum Einsatz kommenden visuellen Metaphern den Verstehensprozess beeinflussen.

Referenzen

Bögel, Thomas, Michael Gertz, Evelyn Gius, Janina Jacke, Jan Christoph Meister, Marco Petris, and Jannik Strötgen. “Collaborative Text Annotation Meets Machine Learning: heureCLÉA, a Digital Heuristics of Narrative.” *DHCommons Journal*, im Erscheinen.

Drucker, Johanna. *Graphesis: Visual Forms of Knowledge Production*. MetaLABprojects. Cambridge, Massachusetts: Harvard University Press, 2014.

Gius, Evelyn. “Erzählen Über Konflikte. Eine Computergestützte Narratologische Untersuchung von Narrativen Interviews Zu Arbeitskonflikten.” Dissertation, Universität Hamburg, 2013.

Jacke, Janina, und Jan Christoph Meister. „Pushing Back the Boundary of Interpretation: Concept, Practice and Relevance of a Digital Heuristic“. In *Digital Humanities 2014 – Book of Abstracts*, 264–66. Lausanne, 2014.

Shneiderman, Ben. “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations.” In *In IEEE Symposium on Visual Languages*, 336–43, 1996.

Tatu, Andrada, Georgia Albuquerque, Martin Eisemann, Peter Bak, Holger Theisel, Marcus Magnor, and Daniel Keim. “Automated Analytical Methods to Support Visual Exploration of High-Dimensional Data.” *IEEE Transactions on Visualization and Computer Graphics* 17, no. 5 (May 2011): 584–97.

Ward, Matthew, Georges G. Grinstein, and Daniel Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications*. Natick, Mass: A K Peters, 2010.

Yang, J., M. O. Ward, E. A. Rundensteiner, und S. Huang. „Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets“. In *Proceedings of the*

Symposium on Data Visualisation 2003, 19–28. VISSYM ’03. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2003.