Preprint von: 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, S. 1-26.

Maschinelle Textanalyse im Zeichen von Big Data und Data-driven Turn – Überblick und Desiderate

Dr. Noah Bubenhofer, wissenschaftlicher Mitarbeiter, Professur für Angewandte Linguistik / Dresden Center for Digital Linguistics, Technische Universität Dresden, Helmholtzstr. 10, D-01062 Dresden, noah.bubenhofer@tu-dresden.de

Prof. Dr. Joachim Scharloth, Professur für Angewandte Linguistik, Technische Universität Dresden, Helmholtzstr. 10, D-01062 Dresden, joachim.scharloth@tu-dresden.de

Inhaltsverzeichnis

Die digitale Revolution verändert auch die Geistes-, Kultur- und Sozialwissenschaften und insbesondere ihren Umgang mit Texten. Drei Prozesse sind für den tiefgreifenden Wandel verantwortlich:

1. *Die Verdatung der Welt:* Immer mehr Informationen werden in ein digitales Format gebracht oder werden schon in digitaler Form produziert. "Digital" bedeutet "abzählbar sein", d.h. dass Informationen in eine numerische Form gebracht und mit mathematischen Methoden analysierbar werden. Parallel zur Entstehung von "Big Data" ermöglicht die Digitalisierung damit auch

2. *die Zusammenführung und damit kombinierte Analyse von Daten unterschiedlichster Provenienz:* die Repräsentation unterschiedlichster Informationstypen in einem numerischen Modell macht es möglich, unterschiedlichste Informationen durch Algorithmen miteinander zu verknüpfen und zu analysieren. Dies ist die Grundlage für

3. *die zunehmende Emanzipation der Daten von dem Zweck ihrer Produktion:* war bislang der Aufbau eines Datenarchivs eng mit einem Zweck verknüpft, der in der Struktur des Archivs und seiner Findemittel sichtbar wurde, erlaubt die Digitalisierung nun jede in einem mathematischen Modell mögliche Anfrage an die Daten und damit die Emanzipation des Nutzers von den Strukturen des Archivs. Damit verbunden ist freilich auch ein Kontrollverlust im Sinne einer Verfügungsmacht über die eigenen Daten.

In der Frühphase von „Big Data“ waren die großen Datenmengen, die im Zuge der Entwicklung des Web und erster Digitalisierungsinitiativen entstanden waren, ein „Retrievel“-Problem: Wie findet man effizient und treffsicher alle Dokumente zum Thema X? Web-Suchmaschinen, die diese Aufgabe gut beherrschten, waren wirtschaftlich erfolgreich. In der Sprachwissenschaft ermöglichten korpuslinguistische Suchmaschinen die Suche nach linguistischen Phänomenen in den entstehenden Korpora, die dank digitalisierter, aber auch originär digitaler Inhalte, schnell anwuchsen. Dank Fortschritten im Natural Language Processing in der Computerlinguistik konnten linguistisch annotierte Textkorpora nach komplexeren Phänomenen als nur Wortformen durchsucht werden: Findet sich Verbzweitstellung in Nebensätzen auch in Zeitungstexten? Wie verbreitete sich

das Schlagwort „soziale Marktwirtschaft“ in den politischen Diskursen Nachkriegdeutschlands? Solche Fragestellungen sind dank linguistisch annotierter Textkorpora leicht und schnell zu beantworten.

Bei der Bearbeitung solcher und ähnlicher Fragestellungen sind Korpora jedoch nicht mehr als – zugegeben sehr mächtige – elektronische Zettelkästen (Perkuhn/Belica 2006, 2). Die Ergebnisse mögen interessant sein, da sie auf großen empirischen Basen stehen, wie z.B. Schlagwortanalysen im Google n-Grams-Korpus, das „Millions of Digitized Books“ (Michel et al. 2011) umfasst. Als entscheidenden Paradigmenwechsel in den Sozial- und Geisteswissenschaften muss aber eine viel weitergehende Entwicklung bezeichnet werden, die wir als „data-driven turn“ (Scharloth et al. 2013) bezeichnen wollen. Statt von wohldefinierten und theoretisch begründeten Hypothesen auszugehen, die anhand von Daten überprüft werden, gesellt sich eine induktive Analyseperspektive zum Forschungskanon hinzu: Die algorithmische Vorstrukturierung der Daten erlaubt es, Zusammenhänge in den Daten zu finden, die vorher unbekannt waren und zu neuen Hypothesen führen. Dabei können in einem digitalen Forschungsdesign beliebige Daten miteinander kombiniert und in Beziehung gesetzt werden – da die Daten alle in einem digitalen Zustand sind, lassen sie sich ineinander verrechnen.

Eine ähnliche Entwicklung vollzieht sich auch in anderen wissenschaftlichen Disziplinen und außerhalb der Wissenschaft. Aufgrund vielfältiger Daten wie Einkaufsbelegen, Kreditkartenrechnungen, Mobilfunksignalen, Internetnutzungsprotokollen und dergleichen mehr können typische (oder auch untypische) Bewegungsmuster berechnet und Verhaltensweisen vorausgesagt werden. Dafür müssen Daten in einheitliche Formate überführt und statistisch ausgewertet werden. Dank den Methoden des Natural Language Processing wird es dabei auch immer besser möglich, sog. „unstrukturierte Daten“ wie Sprachdaten mit zu analysieren.

Wenn auch datengeleitete Methoden, die an allgemeinen Mustern (und ihren Abweichungen davon), aber nicht am Einzelfall interessiert sind, in vielen Bereichen genutzt werden, stellt der data-driven turn in der sozial- und kulturwissenschaftlichen Sprachwissenschaft (und generell in den Sozial- und Geisteswissenschaften) aber mehr dar. Denn die Zusammenhänge zwischen Sprache, sozialem Handeln, kultureller Praxis und Analysemethoden werden und müssen sorgfältig reflektiert werden:

- Der Zusammenhang zwischen Sprachgebrauch und sozialer Wirklichkeit wird in mehreren Disziplinen behauptet und bildet die Grundlage für korpuslinguistische Zugänge.
- Programmiercode, der für die algorithmische Aufbereitung der Daten verwendet wird, muss selber Gegenstand der Analyse sein, da er entscheidender Teil des Analyseprozesses ist.
- Algorithmisch analysierbare Daten erfordern neue Analyseinstrumente. Ergebnisse in Form von Belegen, Listen etc. schlagen bei der Untersuchung großer Datenmengen fehl. Stattdessen erlauben visuelle Analysemethoden eine Kombination von maschineller Rechenkraft und menschlicher Kunst der Mustererkennung. Obwohl also Visualisierungen ein attraktives Instrument sind, müssen ihre Probleme kritisch reflektiert werden.
- Zusammenhänge, Muster oder Phänomene, die wegen ihrer statistischen Auffälligkeit in den Fokus der Analyse gelangen, müssen interpretiert werden. Als verstehende, hermeneutische Wissenschaft muss der Zusammenhang von Analysemethode und Deutung auf eine neue Grundlage gestellt werden.

Diese Grundsatzfragen sind das Resultat einer längeren Entwicklung, die mit dem linguistic turn, der stärkeren Hinwendung auf Sprachgebrauch und den Anfängen der Korpuslinguistik begonnen hat. Im Folgenden zeichnen wir diesen Weg auf und skizzieren den Weg, der noch vor uns steht in Form von Desideraten.

1 Korpora und soziale Wirklichkeit

Mehrere Entwicklungen führten in den letzten Jahrzehnten zu einer „Rehabilitierung der

sprachlichen Oberfläche“ (Antos 1989): Die „pragmatische Wende“ schlägt sich ab den 1970er-Jahren verstärkt in der Linguistik nieder und führt zu einer Perspektive, die „sprachliche Tatbestände grundsätzlich vom Texthandeln“ und seiner Gelingensbedingungen her zu konzipieren und beschreiben versucht (Feilke 2000, 65). Dies führte zu einer Ausweitung der Untersuchungsgegenstände – gleichzeitig aber auch zu einem universalpragmatischen Erkenntnisinteresse, das erst in neuerer Zeit abgelöst wurde (Feilke 2003, 217): Abgelöst durch, vereinfacht ausgedrückt, ein Interesse für Medialität, Kulturalität, Musterhaftigkeit/Prägung, Kontextualisierung und Performanz.[1] Es sind Indikatoren auf der sprachlichen Oberfläche, die einen Text als Sprachhandlung situiert – etwa die flektierte Form „vereinzelt“ in Kombination mit „Durchzug von“, die sofort auf einen Wetterbericht schließen lässt (bei der Grundform „vereinzeln“ wäre diese Kontextualisierung nicht mehr eindeutig gegeben).

Insbesondere diese Sicht auf Sprachgebrauchsmuster, Formulierungen und Prägungen, die zum Sprachgebrauchswissen der Sprachteilnehmer/innen gehören, trifft sich produktiv mit korpuslinguistischen Sichtweisen. Bereits 1957 betont Firth (1957, 194) mit dem Begriff der „collocation“ den Aspekt des Sprachgebrauchs, der eigentlich arbiträre zu usuellen Wortverbindungen macht, die sich nicht über grammatische Regeln und semantisches Wissen über Einzellexeme herleiten lassen, sondern gelernt werden müssen. Das Interesse für Kollokationen entfaltet sich in der Folge einerseits im Kontext der Phraseologie (Burger 1998), andererseits in der Computerlinguistik und verwandten Gebieten (Carstensen et al. 2009; vgl. für eine ausführliche Darstellung Evert 2009). Die phraseologische Perspektive hat dabei oft Lernende von Fremdsprachen im Blick und macht dort auf die Bedeutung aufmerksam, Sprach*gebrauchs*wissen und nicht (nur) Wissen über Sprachsysteme zu vermitteln (Hausmann 1985).

Für eine sozial- und kulturwissenschaftlich interessierte Sprachwissenschaft war als drittes Element jedoch der Einfluss der Diskursanalyse nach Foucault entscheidend. Die linguistischen Antworten darauf in Form der Postulate der Begriffsgeschichte (Busse et al. 1994; Hermanns 1995) und der diskurslinguistischen Forschungsprogramme (Bluhm et al. 2000; Busse/Teubert 2013; Spitzmüller/Warnke 2011; Warnke 2007) schärften das Bewusstsein für Merkmale der Textoberfläche, die jedoch als Ergebnis von sozialem Handeln und damit Indikatoren für Diskurse sind.

Als Melange dieser drei Paradigmen, schlagwortartig gefasst mit „pragmatische Wende“, „Korpuslinguistik“ und „Diskurslinguistik“, entstanden in der Folge eine Reihe von Arbeiten, die mit korpuslinguistischen Methoden Diskurse untersuchten und dafür das notwendige Methodenrepertoire erarbeiteten. Wir möchten solche Arbeiten im Folgenden „korpuspragmatisch“ nennen, da sie alle „rekurrente sprachliche Muster mit kulturellen oder sozialen Phänomenen in Zusammenhang“ bringen und die Muster „entweder als deren Symptom oder als diese (mit-)konstituierend“ deuten (Scharloth/Bubenhofer 2011, 196). Da also sprachliche Muster im Vordergrund des Interesses stehen, sind insbesondere die Analysekategoren Kollokationen, n-Gramme und Keywords, die im Folgenden kurz skizziert werden sollen, zentral.

1.1 Kollokationen und n-Gramme

Die Ausführungen oben haben bereits deutlich gemacht, dass typische Formulierungen oder Sprachgebrauchsmuster für eine korpuspragmatische Analyse von besonderem Interesse sind. Das Ziel des Konzeptes der Kollokation ist es, empirisch typische Wortverbindungen in den Daten zu finden und zu deuten.

Ein kurzer Blick in das DWDS-Korpus[2] (Geyken 2007) nennt als signifikanteste Kollokatoren zu „Islam“ – und in Abgrenzung zu „Christentum“:

> *Christentum, Prägung, Sufis, sunnitischen, als Ideologie, Terrorgruppe, gleichgesetzt, Islamismus, vereinbar, gelehrt, (als) Kultur, beleidigt, gleichzusetzen, Ideologie, Kultur* etc.

Für „Christentum“ – in Abgrenzung zu „Islam“ werden folgende Kollokatoren genannt:

1 Vgl. für eine vertiefte Diskussion dazu Feilke (1996, 2000, 2003) sowie Feilke/Linke (2008).
2 Vgl. www.dwds.de (letzter Zugriff: 27. 8. 2014).

als Kulturgrenze, Sekte, Antike, Marxismus, Spätantike, Humanismus, Heidentum, Wesen, Kriminalgeschichte, Aufklärung, Stifter, Ursprung, Apologie

Daneben gibt es eine Reihe von Kollokatoren, die sowohl für „Islam" als auch „Christentum" gleichenfalls typisch sind:

(als) Religion, (als) Staatsreligion, Judentum, übergetreten zum, Religionen wie, konvertiert zum, Hindusimus, Westen, als Weltreligion, Stätten etc.

Die Analyse der Kollokatoren – hier im Vergleich zwischen zwei Begriffen – erlaubt also, Indizien für typische Sprachgebrauchsmuster zu finden und Diskurse zu beschreiben.

Die Terminologie um die verschiedenen Konzepte von Mehrworteinheiten ist in der Literatur nicht ganz einheitlich. Mit „Kollokationen" sind üblicherweise binäre Verbindungen von Wortformen oder Lemmata gemeint, die innerhalb eines „Fenster" von *x* Wörtern typischerweise miteinander vorkommen (Evert 2009). Dieses „typische Kovorkommen" wird als „Assoziation" gemessen und statistisch mit einem Signifikanztest modelliert: Wenn die beiden Einheiten in den Daten signifikant häufiger zusammen vorkommen als es bei einer zufälligen Verteilung der Wörter im Korpus anzunehmen wäre, handelt es sich um eine Kollokation. Besonders in der Phraseologie wird jedoch ein engerer Kollokationsbegriff verwendet, bei der die Verbindung zusätzlich in einem syntaktischen Bezug stehen muss (Bartsch 2004, 76; Evert 2009, 1213). In der Korpuslinguistik wird normalerweise jedoch der weiten Definition von Kollokationen, die Evert als „empirische" von den „theoretischen" Kollokationen oder „Multiword Expressions" unterscheidet (Evert 2009, 1213), gefolgt.

Als Kollokationen können prinzipiell auch Einheiten von mehr als zwei Worteinheiten aufgefasst werden: „wie sollen Sie eigentlich". Dafür wird dann aber wiederum eher der Begriff „n-Gramm" verwendet wird, wobei *n* für eine beliebige Anzahl Worteinheiten steht (Manning/Schütze 2002, 192ff.). Eine Erweiterung des n-Gramm-Konzeptes insbesondere für korpuspragmatische Zwecke verwenden unter dem Terminus „komplexe n-Gramme" Scharloth/Bubenhofer (2011; und Bubenhofer/Scharloth 2011), bei dem ein n-Gramm nicht nur aus einer Folge von n Wortformen, sondern aus Kombinationen von Wortformen und Wortartklassen bestehen können. Dadurch können eine Reihe von ähnlichen Wortformen-n-Grammen zu einem Muster zusammengefasst werden (Bubenhofer 2015):

VMFIN Sie ADV ADV

Hinter der Kombination von einem finiten Modalverb mit der Wortform „Sie", gefolgt von zwei Adverbien[3], stehen Realisierungen wie:

Können Sie bitte einmal
Können Sie hier noch
können Sie schon jetzt
könnten Sie hier sofort
müssen Sie endlich einmal
müssen Sie schon selbst
wollen Sie dann weiterhin
wollen Sie denn da
wollen Sie denn eigentlich
wollen Sie doch sicherlich
wollen Sie sogar noch

Im Vergleich zu (binären) Kollokationen erfassen n-Gramme, und insbesondere komplexe n-Gramme, umfangreichere Sprachgebrauchsmuster, die z.B. Hinweise für typische Sprachhandlungen (z.B. KRITISIEREN) geben können (Bubenhofer et al. 2009; vgl. für Analysen dieser Art Bubenhofer/Scharloth 2011; Scharloth/Bubenhofer 2011).

3 Die Wortartklassen folgen dem Stuttgart-Tübingen-Tagset (Schiller et al. 1995); vgl. zu maschineller Part-of-Speech-Annotation auch den Beitrag von Zinsmeister in diesem Band.

1.2 Keywords

Durch den Vergleich von mindestens zwei Korpora können für einzelne sprachliche Einheiten, normalerweise Wort- oder Grundformen (hier „Wörter"), Frequenzen verglichen werden. Wenn sich die Frequenz eines Wortes in einem Korpus A im Vergleich zu einem oder mehreren anderen Korpora (B, C, …) erheblich voneinander unterscheidet, ist dieses Wort für Korpus A ein Keyword, bzw. der Signifikanzwert ist die „Keyness", also die „Typizität" des Wortes für das jeweilige Korpus (Bondi/Scott 2010; Bubenhofer 2015; Scott/Tribble 2006). Es wird also auch hier oft ein statistischer Signifikanztest vorgenommen, um abzuschätzen, ob der Frequenzunterschied in Relation zu den jeweiligen Korpusgrößen genug groß ist, um einen zufälligen Unterschied ausschließen zu können.

Der korpuslinguistische Begriff von „Keyword" oder „Schlagwort" muss also vom nicht-korpuslinguistischen, nicht-empirischen, klar abgegrenzt werden, da ersterer zunächst nur auf einem statistischen Vergleich von Frequenzen beruht. Trotzdem zielt das Keywords-Konzept natürlich darauf hin, einen nicht-empirischen Schlagwortbegriff zu operationalisieren, also die Wörter zu finden, die ein Korpus im Vergleich zu einem Referenzkorpus besonders gut charakterisieren.

Ein typisches Setting für einen Keywords-Vergleich ist die Berechnung der für eine bestimmte politische Partei typischen Schlagwörter. So zeigen sich z.B. in einer Keywords-Berechnung in einem Korpus aller Wortprotokolle des Deutschen Bundestags der Wahlperiode 17 (2009-2013) die folgenden typischen Nomen für die Fraktionen B90/Die Grünen und CDU/CSU – immer im Vergleich zum Gesamtkorpus (Bubenhofer 2015):

> B90/Die Grünen: *Bundesregierung, Staatssekretär, Atomkraftwerk, Schwarz-Gelb, Atomkraft, Klimaschutz, Leiharbeitskräfte, Einwanderer, Laufzeitverlängerungen, Subventionen, Ministerin, Frage* etc.
>
> CDU/CSU: *Land, Weg, Opposition, Zukunft, Bereich, Rahmen, Erfolg, Kernenergie, Sicherheit, Wettbewerbsfähigkeit, Entwicklung, Maßnahme* etc.

Unter den berechneten Keywords befinden sich politische Schlagwörter, Fahnen- und Stigmawörter etc., deren Bedeutung in einem bestimmten Diskurs dadurch beschrieben werden kann.

Neben Signifikanzmaßen werden für Frequenzvergleiche von Wörtern in Korpora auch andere Maße verwendet. Im einfachsten Fall die „relative Frequenz" (absolute Frequenz in Relation zur Korpusgröße) oder die Häufigkeitsklassen (Perkuhn et al. 2012, 80), die bei stark unterschiedlich großen Korpora ein stabileres Maß für den Vergleich darstellt.

Keyness-Maße können mit der Berechnung von Kollokationen und n-Grammen kombiniert werden: Wenn für ein Untersuchungs- und ein Referenzkorpus alle kombinatorisch möglichen Kollokationen oder n-Gramme berechnet werden, können anschließend die jeweiligen Frequenzen in den beiden Korpora verglichen und so die Mehrworteinheiten ausgewählt werden, die besonders typisch (oder untypisch) für das Untersuchungskorpus sind (Bubenhofer 2009, 2015).

1.3 Corpus-Based und Corpus-Driven

Mit der Verfügbarkeit von großen Korpora, die für kultur- und sozialwissenschaftliche Fragestellungen interessant sind, wurden diese zunächst als moderne Zettelkästen verwendet, um Belege für bestimmte Wortverwendungen zu finden. Dieser Zugang, der als „corpus-illustrated" (Atkins et al. 1992, 14) bezeichnet werden kann, wurde bald ergänzt um sog. korpusbasierte („corpus-based") Zugänge, bei denen möglichst systematisch Korpora nach bestimmten, gegebenen Phänomenen durchsucht und die Ergebnisse analysiert wurden. Kollokationsanalysen zu Lexemen, die der Forscher/die Forscherin hypothesengeleitet als untersuchenswert definiert hat, sind ein typisches Beispiel dafür.

Schon bald forderten Korpuslinguist/innen aber, die Empirie ernster zu nehmen und kritisieren, bei einem korpusbasierten Verfahren bestehende linguistische Kategorien nicht ernsthaft in Frage stellen zu können (Sinclair 1991; Teubert 2005). So formuliert etwa Tognini-Bonelli:

> [. . .] corpus-based linguists adopt a „confident“ stand with respect to the relationship between theory and data in that they bring with them models of language and descriptions which they believe to be fundamentally adequate, they perceive and analyse the corpus through these categories and sieve the data accordingly. (Tognini-Bonelli 2001, 66)

Eine korpusbasierte Perspektive sei zwar in der Lage, bestehende Kategorien zu verfeinern, sie jedoch nicht grundsätzlich in Frage zu stellen. Deshalb fordert die corpus-driven-Perspektive, die Empirie als Ausgangsbasis für jegliche Analysen zu setzen – Teubert formuliert dies explizit für diskurslinguistische Interessen so:

> While corpus linguistics may make use of the categories of traditional linguistics, it does not take them for granted. It is the discourse itself, and not a language-external taxonomy of linguistic entities, which will have to provide the categories and classifications that are needed to answer a given research question. This is the corpus-driven approach. (Teubert 2005, 4)

Die Forderung, Korpora nicht einfach nur als Beispielspender zu verwenden, sondern die Daten mit allen Widersprüchen ernst zu nehmen, fand auch im deutschsprachigen Raum in vielen Teildisziplinen der Linguistik ihren Niederschlag, so beispielsweise für semantische/phraseologische (Belica/Steyer 2008; Steyer 2000, 2013; Steyer/Brunner 2009; Steyer/Lauer 2007), grammatische (Kupietz/Keibel 2008; Dürscheid et al. 2011; Bubenhofer et al. 2014) aber auch diskurslinguistische Fragestellungen (Bubenhofer 2009; Vogel 2010; Scharloth/Bubenhofer 2011), um nur einige wenige Vertreter zu nennen.

Methodisch beschränkte sich die corpus-driven-Perspektive in der Korpuspragmatik auf die oben genannte Konzepte Kollokationen, n-Gramme und Keywords – auch in Kombination – und Analysen zur Verteilung von bestimmten Phänomenen in den Daten. Durch die Berechnung der typischen n-Gramme auf der Basis aller kombinatorisch möglichen n-Gramme in Untersuchungs- und Referenzkorpus etwa beschränkt sich die Analyse zunächst nicht auf bestimmte Lexeme, sondern führt zu auffälligen Sprachgebrauchsmustern, die anschließend gedeutet und zu neuen Kategorien gefasst werden können. Um der großen Datenmengen Herr zu werden, können zudem Clustering-Methoden eingesetzt werden, um ähnliche n-Gramme zu gruppieren (Bubenhofer 2015). In anderen Teildisziplinen und auch in der Computerlinguistik wurden zwar bereits früh auch anspruchsvollere statistische Verfahren eingesetzt, etwa zur datengeleiteten Stilbeschreibung mittels multivariater Analysen (Biber/Jones 2005), diese fanden zunächst aber noch nicht Eingang in die korpuspragmatische Forschung.

2 Maschinelle Textanalyse im Zeichen von Big Data und Data-driven Turn

„Big Data“ wird in der Informatik in den 1990er-Jahren zu einem Thema: In großen Unternehmen fallen riesige Datenmengen an, die verwaltet und genutzt werden sollen („Data Warehouse“). Diese Entwicklung fällt mit technischen Fortschritten der Datenverarbeitung (Prozessorleistung, günstigere und umfangreichere Speichermöglichkeiten) zusammen; in der Kombination werden neue Formen der Datenanalyse möglich, etwa um Korrelationen zwischen Kennzahlen datengeleitet zu erkennen.

In den Sprachwissenschaften entstehen in der gleichen Zeit die ersten großen Textkorpora wie das British National Corpus oder das Deutsche Referenzkorpus (DeReKo) am Institut für Deutsche Sprache (Perkuhn et al. 2012, 47) – mit dem „Brown Corpus“ als Vorläufer bereits Mitte der 60er-Jahre (Francis/Kučera 1964). In den letzten Jahren stiegen die Korpusgrößen enorm an: Das DeReKo (Kupietz et al. 2010) umfasste 1992 28 Mio. laufende Wortformen, 2012 etwa 7 Mia. und 2014 24 Mia. laufende Wortformen.[4] Dabei bietet sich natürlich auch das Web an, um große Korpora aufzubauen, wie z.B. das „Corpora from the Web (COW)“-Projekt zeigt (Schäfer/Bildhauer 2012).

In den Digital Humanities werden naturgemäß nicht nur Sprachdaten verwendet, sondern

4 Vgl. http://www1.ids-mannheim.de/kl/projekte/korpora/archiv.html (4. September 2014).

Datensätze unterschiedlicher Art, die digitalen Analysemethoden unterzogen werden. Diese sollen in den Daten „patterns, dynamics, and relationships“ (Rieder/Röhle 2012, 70) aufdecken. Unter Verwendung der Google n-Grams-Datenbank, Frequenzlisten von Wörtern und Wortketten in den von Google gescannten Büchern, postulieren Michel et al. (2011) eine neue Wissenschaft: „Culturomics“. Das Studium von Zeitreihenanalysen von n-Gram-Frequenzen erlaube es, Forschungsfragen aus Bereichen wie „lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology“ (Michel et al. 2011, 1) zu beantworten.

Für Wissenschaftsdisziplinen, die es gewohnt sind, klare Hypothesen zu formulieren, diese sorgfältig zu operationalisieren und dann anhand empirischer Daten zu testen, ergeben sich plötzlich neue Möglichkeiten: „The End of Theory“, wie Anderson (2008) im Wired-Magazin ausruft und mit den Methoden von Google argumentiert, mit denen beispielsweise algorithmisch durch die statistische Analyse von Webnutzungsverhalten Kunden die passende Werbung angezeigt wird, ohne „to know anything about the culture and conventions of advertising — it just assumed that better data, with better analytical tools, would win the day“. Was in der Wirtschaft üblich ist, soll auch in die Wissenschaft Eingang finden: Ausgangspunkt für Analysen sind nicht von Theorien abgeleitete Hypothesen, Ausgangspunkt sind die Daten und Tools, die Muster in den Daten entdecken sollen.

Sowohl im Bereich der Digital Humanities als auch in der Sprachwissenschaft (und im Teilbereich der Korpuspragmatik) sind eine Reihe von datengeleiteten Studien auf der Basis großer Datenbasen oder Textkorpora entstanden – viele auch unter Verwendung visueller Analysemethoden (vgl. dazu weiter unten).

Mit dem Wunsch, datengeleitete Verfahren anzuwenden, steigt in den Geisteswissenschaften das Interesse für Methoden des Data-Minings. Im Bereich der Textanalyse kommt dabei eine breite Palette von Methoden des textbasierten Informationsmanagements (Carstensen et al. 2010, 577) in Frage. Dabei geht es in den überwiegenden Fällen darum, Texteinheiten maschinell zu klassifizieren. Bei sog. „überwachten“ Verfahren, wird auf der Basis einer manuell klassifizierten Trainingsmenge ein statistisches Modell erzeugt, das neue Fälle klassifizieren kann. Bei „unüberwachten“ Lernverfahren, sog. Clustering-Methoden, werden die Texte aufgrund ihrer textuellen Eigenschaften maschinell in möglichst homogene Gruppen aufgeteilt (Carstensen et al. 2010, 591ff.).

Solche Clustering-Methoden können natürlich für beliebige Daten verwendet werden. In der maschinellen Textanalyse und insbesondere auch in geisteswissenschaftlichen Zusammenhängen, werden insbesondere verschiedene Topic Modelling-Verfahren (Anthes 2010) immer häufiger angewandt (vgl. die an programmierende Historiker gerichtete Anleitung von Graham et al. 2012). Sie dienen dazu, aufgrund der Wortverteilung, gemessen an einer bestimmten Wahrscheinlichkeitsverteilung, Texte in Klassen aufzuteilen und die dafür charakterisierenden Wörter zu nennen. In den Digital Humanities werden damit beispielsweise zeitliche Veränderungen in der Themenzusammensetzung in Korpora untersucht. Rohrdantz et al. (2012) verwenden die Methode aber auch, um semantische Veränderungen darzustellen, indem nicht Texte, sondern Belege von bestimmten Lexemen klassifiziert werden. Ein Teilergebnis einer ähnlichen Analyse zeigt Tabelle 1: In einem Korpus aller Spiegel-Artikel von 1947 bis 2010 wurden 50-Wort-Belege von „Terror(ismus)“ einer LDA-Klassifikation (Blei et al. 2003) unterzogen. Zwei der 20 gerechneten Themen – oder besser: Lesarten – wurden vom Algorithmus mit den Schlagwörtern „al, Laden, Qaida, Pakistan, Bin, Afghanistan“ etc. und mit „RAF, 1977, SPD, Terrorist, CDU, Hans, Bonn“ etc. umschrieben. Mit Blick auf die zeitliche Distribution der beiden Klassen werden die beiden Lesarten von „Terror“, nämlich „islamistischer Terror“ und „RAF-Terror“ deutlich sichtbar.

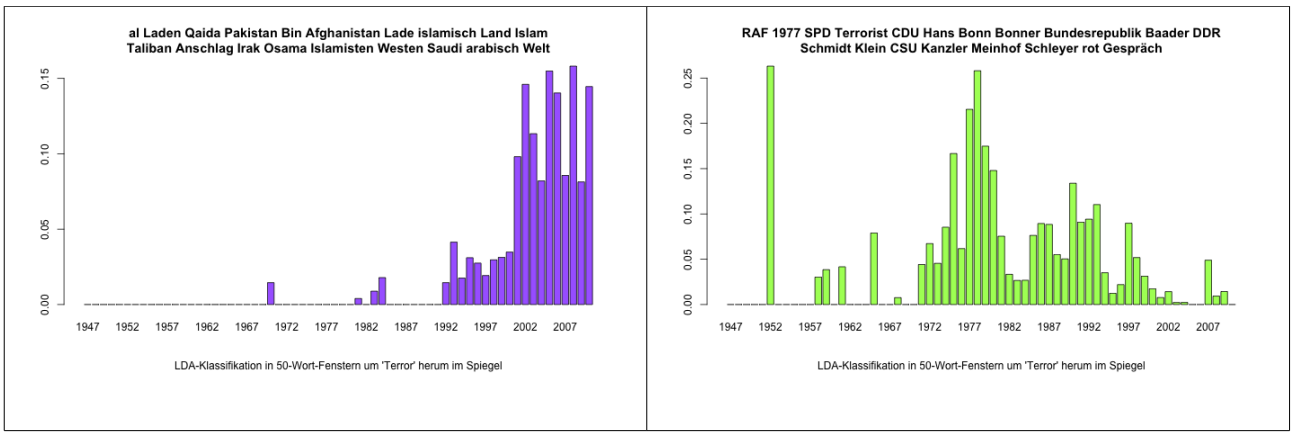

Tabelle 1: LDA-Klassifikation von "Terror" im Spiegel-Korpus – Beispiel von zwei Lesarten

Die gegenwärtige Phase der automatischen Textanalyse in den Geistes- und Sozialwissenschaften zeichnet sich durch große Experimentierfreudigkeit im Umgang mit verschiedenen Methoden und Datensätzen aus. Es ist wichtig, dass diese Disziplinen Anschluss an den State of the Art des Text Minings, der Computerlinguistik und der Informatik finden (vgl. dazu Scharloth et al. 2013). Trotzdem könne solche Methoden in diesen Disziplinen nicht einfach mit dem Entscheidungskriterium „funktioniert“ oder „funktioniert nicht“ angewendet werden, wie das etwa in der Computerlinguistik üblich und sinnvoll ist. Dort bewährt sich eine Methode dann, wenn sie ein maschinelles System, z.B. ein maschinelles Übersetzungssystem, messbar verbessert, also gemessen an einem Goldstandard erfolgreicher ist. Für geistes- und sozialwissenschaftliche Anwendungen reicht das jedoch nicht aus – der letzte Abschnitt skizziert, weshalb.

3 Desiderate

3.1 Die maschinelle Textanalyse braucht einen integrierten Textbegriff

Ein zentrales Problem bei der Anwendung maschineller Methoden der Textanalyse ist ihr unterkomplexer Textbegriff. Im Text Mining waren Texte lange "Bags of Words". Die grundlegende Idee hinter diesem schon in den 1960er Jahren entwickelten Ansatz ist, dass sich die Bedeutung eines Textes mit Hilfe des Gewichts der im Dokument vorkommenden Terme operationalisieren lässt. Ein Text wird dann als Vektor repräsentiert, dessen Elemente die dokumentenspezifischen Werte jedes einzelnen Terms enthält. Ein Korpus aus vielen Texten entspricht dann einer Matrix. (Vgl. Heyer / Quasthoff / Wittig 2005; Weiss et al. 2005: 15-46; Berry / Kogan 2010: 22ff. Für einen differenzierteren Überblick über die Entwicklung von Text Mining-Modellen und ihre Anwendung in den Sozial- und Geisteswissenschaften vgl. Wiedemann 2013.) Trotz der Ausdifferenzierung des Spektrums berücksichtigter Textmerkmale (Vgl. hierzu exemplarisch die Entwicklung der Feature-Vektoren bei der Autorschaftsattribution, Koppel / Schler / Argamon 2009) und dem Einsatz von Latent Semantic Analysis (LSA) und semantischen Wissensbasen bleibt der Textbegriff des Text Mining doch von weitgehend atomistischen Vorstellungen von den Konstituenten eines Textes geprägt und versäumt es, Texte als komplexes Gewebe zu operationalisieren.

Da verwundert es nicht, dass schon früheste textlinguistische Textbegriffe, die die Einheit eines Textes aus transphrastischen semantischen und syntaktischen Beziehungen herleiten ("durch ununterbrochene pronominale Verkettung konstituiertes Nacheinander sprachlicher Einheiten", Harweg 1968: 148) über den Textbegriff des Textmining hinausgehen. Kognitive und textpragmatische Dimensionen des Textbegriffs, etwa die der Kohärenz als durch Wissen konstituierbarer Sinnzusammenhang (Beaugrande/Dressler 1981: 8) oder die Auffassung von Text als Realisierung eines "erkennbaren Illokutionspotenzials" und damit die Signalisierung einer kommunikativen Funktion (Schmidt 1976: 150), spielen bei den in der Informatik wurzelnden Ansätzen der maschinellen Textverarbeitung keine Rolle. Dagegen gibt es im Kontext der

eHumanities bereits Ansätze, pronominale Koreferenz, Thema-Rhema-Verkettungen, diskursive Strukturen und Argumentationsmuster maschinell zu operationalisieren. (Vgl. etwa Bunescu / Mooney 2006 für Analyse der Abhängigkeit von Phrasen, Bamman et al. 2014 für die Analyse literarischer Charaktere mit Hilfe von pronominaler Koreferenz oder das VisArgue-Projekt (http://www.visargue.uni-konstanz.de, Bögel et al. 2014), das kausale Begründungsmuster analysiert.) Andere bedeutungskonstituierende Aspekte von Texten wie ihre Gestaltganzheit (Einheitlichkeit von Stilprinzipien) und Kulturalität (Wissen um Textsorten/Gattungen in ihrer kulturellen Geprägtheit) sowie ihre Materialität (formale Sichtbarmachung und Gestaltung der Zeichen), Medialität (technische Mittel für die Übertragung und Speicherung) und Lokalität (der institutionalisierte Ort der Publikation mit seiner kulturell verfestigten Bedeutung) (Fix 2008: 345) sind bislang nicht im Blick der maschinellen Textanalyse.

Um für kultur- und sozialwissenschaftliche Fragestellungen aber als Methode attraktiv zu sein, muss sich die maschinelle Textanalyse stärker um eine gegenstandsadäquate Modellierung des Textbegriffs bemühen und sich hierbei an Theorien der Textlinguistik orientieren; sie braucht einen integrierten Textbegriff.

3.2 Die maschinelle Textanalyse braucht valide Modelle

Problematisch an der Verwendung maschinelle Methoden der Textanalyse für sozial- und kulturwissenschaftliche Fragestellungen ist, dass maschinelle Methoden und Algorithmen für Problemstellungen entwickelt wurden, die nur bedingt zur Forschungslogik der Kultur- und Sozialwissenschaften, die am Verstehen und Erklären von Sachverhalten und Phänomenen orientiert ist, passt.

Ein Beispiel mag dies illustrieren: Die informatische Forschung im Bereich Stilometrie und Autorschaftsattribution scheint zu belegen, dass Buchstaben-n-Gramme die beste Merkmalskategorie bei der Klassifizierung von anonymen Dokumenten sind (vgl. Stamatatos 2009: 24). Um die Autorschaft eines anonymen oder pseudonymen Schreibens aus einem Set von möglichen Kandidaten zu ermitteln, werden für jedes Autorenkorpus alle Buchstaben-n-Gramme berechnet, die über einer bestimmten Mindestfrequenz liegen. Daraus werden eine größere Zahl von Buchstaben-n-Grammen ausgewählt, deren Distribution für das Sample die höchste Informativität haben. Das Korpus der autorspezifischen Texte wird dann als Matrix repräsentiert, die die relative Frequenz der Buchstaben-n-Gramme als Werte enthält. Mit Methoden maschinellen Lernens wird nun ein Klassifikator berechnet, der die Texte aufgrund ihrer Merkmale optimal in Klassen einteilt. Dieser Klassifikator wird dann dazu benutzt, das Anonymschreiben einem Autor zuzuordnen. Er ist aus stilometrischer Sicht ein Modell über die stilistischen Eigenschaften der Texte im Sample. Aus linguistischer Perspektive hat dieses Modell freilich eine nur geringe Erklärungskraft, denn es ist unklar, ob überhaupt und wenn ja welche Dimensionen von Stil Buchstaben-n-Gramme messen. Und selbst wenn das Ergebnis stimmen würde, so würde sich eine literaturwissenschaftliche oder linguistische Stilanalyse nicht nur für die Attribution interessieren, sondern auch dafür, wie diese qualifiziert ist, d.h. welche stilistischen Eigenschaften eines Textes für einen Autor typisch sind. Und hier helfen Buchstaben-n-Gramme kaum weiter, weil sie keine Aspekte linguistischer oder literaturwissenschaftlicher Stilbegriffe operationalisieren. Die Forschungslogik der Informatik freilich verlangt nach dem Modell, das mit der höchsten Trefferquote den Autor bzw. die Autorin eines anonymen Dokuments ermittelt.

Das Problem, das im Beispiel sichtbar wird, ist jedoch unseres Erachtens kein kategorisches. Die Forschungslogiken der Informatik einerseits und der Kultur- und Sozialwissenschaften schließen sich nicht gegenseitig aus, vielmehr ist die Frage zu klären, unter welchen Umständen ein mathematisches Modell als Erklärung im sozialwissenschaftlichen Sinn bzw. als Grundlage für das Verstehen eines Sachverhalts oder eines Phänomens gelten kann. Die Antwort auf diese Frage muss in der Modelltheorie gesucht werden.

Modelle sind nach Stachowiak (1973: 131-134) durch ihren Abbildcharakter ("Modelle sind stets Modelle von etwas, nämlich Abbildungen, Repräsentationen natürlicher oder künstlicher Originale"), die Notwendigkeit der Verkürzung ("Modelle erfassen [...] nicht alle Attribute des

durch sie repräsentierten Originals, sondern nur solche, die den jeweiligen Modellerschaffern und/oder Modellbenutzern relevant scheinen") und dem ihnen innewohnenden Pragmatismus (Modelle erfüllen Ersetzungsfunktion für bestimmte Subjekte, in einer bestimmten Zeit und eingeschränkt auf bestimmte Operationen) charakterisiert. Mathematische Modelle für die sozial- und kulturwissenschaftliche Textanalyse müssen die Besonderheiten der Gegenstände in allen drei Bereichen berücksichtigen. Dies bedeutet, der Tatsache Rechnung zu tragen, dass der Abbildcharakter des Modells sich in den Sozial- und noch mehr in den Kulturwissenschaften sich auf Interpretationen bezieht, Modelle also Interpretationen von Interpretationen (im Sinne von Geertz) sind, mithin ein konstruktorientierter Modellbegriff zugrundegelegt werden muss. Hinsichtlich der Verkürzung bedeutet dies, dass Validität ein wesentliches Kriterium für Modelle sein muss und auf allen Ebenen ihrer Konstruktion als Kriterium berücksichtigt werden muss. Um beim Beispiel zu bleiben: Auch wenn Buchstaben-n-Gramm-Modelle bessere Ergebnisse liefern, sind sie wegen mangelnder Konstruktvalidität für die sozial- und kulturwissenschaftliche Forschung fragwürdig. Davon betroffen sind auch die zum Einsatz kommenden Algorithmen: Statt auf sog. Black-box-Data-Mining-Algorithmen wie etwa Support Vector Machines zu setzen, sollte die sozial- und kulturwissenschaftliche interessierte maschinelle Textanalyse White-box-Algorithmen benutzen, um die Konstruktivität des Algorithmus transparent und nachvollziehbar zu machen (vgl. Rieder/Röhle 2012). Im Hinblick auf die Pragmatizität der Modelle bedeutet dies, dass Out-of-the-box-Lösungen nur in wenigen Fällen anwendbar sind. Dies hat auch Konsequenzen die die Forschungsförderung, die lange auf langfristige eHumanities-Infrastrukturprojekte gesetzt hat. Es bedeutet hinwiederum auch, dass Kultur- und Sozialwissenschaftlerinnen und -wissenschaftler sich mehr um das Verständnis mathematischer Modelle bemühen müssen, um das Potenzial maschineller Ansätz ausschöpfen zu können.

3.3 Die maschinelle Textanalyse braucht neue Methoden der Visualisierung

Zur Analyse von großen Datenmengen haben sich in vielen wissenschaftlichen Diziplinen bereits die Methoden der „Visual Analytics“ etabliert. Diese helfen weiter, wenn traditionelle Formen der Repräsentation von Wissen wie Listen, Tabellen oder Texte zu komplex sind, um als Ganzes erfasst und gedeutet zu werden (Chen et al. 2008; Keim et al. 2010). Für visuelle Analysen von Sprachdaten werden gegenwärtig eine Reihe von Methoden und Werkzeugen entwickelt und erprobt (Risch et al. 2008; Rohrdantz et al. 2010). Damit können große Textkorpora datengeleitet auf auffällige Muster hin untersucht werden.

An verschiedenen Stellen wird aber auch auf die Gefahr datengeleiteter Zugänge verwiesen. Berry (2012) fordert z.B. eine vertiefte Auseinandersetzung mit „computationality“, der Wissenstransformation durch Software – ein grundlegendes Problem visueller Analysetools. Rieder/Röhle (2012) nennen unter anderen Herausforderungen für die Digital Humanities die „Macht der visuellen Evidenz“ als grundlegendes Problem: Wie kann der rhetorischen Qualität der mitunter hoch-ästhetischen Visualisierungen begegnet und die angebliche visuelle Evidenz hinterfragt werden?

Es fehlt also eine tiefgehende Reflexion über Visualisierungen im Forschungsprozess: Wie können und sollen visuelle Analysemethoden in den Analyseprozess integriert werden? Welche Möglichkeiten und Beschränkungen gibt es? Wie können Visualisierungen interpretiert werden, ohne vorschnell der visuellen Evidenz zu erliegen? Wie verändern sich die Praktiken der Visualisierung (vgl. für die forschungsethischen Aspekte dazu den folgenden Abschnitt) und welche Auswirkungen hat das auf den Forschungsprozess? Welche Auswirkungen hat der Kanon von Visualisierungsformen (Linien-, Balken-, Punktdiagramme etc.) und dessen Ausweitung (unterschiedliche Graphentypen, Interaktivität) auf die Praxis der visuellen Analyse?

Mit der Diagrammatik (Bauer/Ernst 2010; Stjernfelt 2007) und dem Interesse für „operative Bildlichkeit“ (Krämer 2009) wächst von theoretischer Seite die Diskussion um die Funktionsweise und den zeichentheoretischen Status von Visualisierungen an und stellt eine Ergänzung zur Arbeit der Pioniere der explorativen Visualisierung dar (Benzécri 1973; Tukey 1977; Tufte 1983; vgl. für

eine Überblicksdarstellung Friendly 2005). In Verbindung mit der Praxis von visueller Textanalysemethoden werden die Grundlagen für eine „New Visual Hermeneutics“ (Kath/Schaal/Dumm in diesem Band) geschaffen.

3.4 Die maschinelle Textanalyse braucht eine Forschungsethik

Die Digitalisierung hat in den Geistes- und Sozialwissenschaften neue Datentypen und Datenmengen, neue Forschungsfragen und neue Methoden initiiert, deren praktische Konsequenzen für Subjekte, Objekte und Adressaten der Forschung bislang nur selten reflektiert werden. Denn Digitalisierung ist mehr als Zahlen, Algorithmen, Daten und Informationen. Digitalisierung und der forschende Umgang mit digitalen Daten ist auch Praxis, ganz gleich ob an Universitäten oder in Unternehmen. Dort wo Akteure digitale Handlungsräume zum Erforschen, Rechtfertigen, Erproben, Teilen und Kommunizieren neuen Wissens und neuer Kompetenzen betreten, bedarf es einer Ethik der Digital Humanities, die die impliziten normativen Orientierungen dieser Akteure und ihrer Handlungsformen reflektiert.

Im Praxisfeld Digital Humanities werden dabei insbesondere solche Fragen virulent, die sich aus der Digitalisierung der Relation der Forschenden zu den Beforschten ergeben. Sie lassen sich als Fragen der *informationellen Selbstbestimmung* und der Forschungsethik im engeren Sinne beschreiben. Die häufig als datengeleitet und theorielos daherkommende Forschung auf großen Datenmengen (Anderson 2007) stellt nämlich einige Grundprinzipien des Datenschutzes in Frage: am offensichtlichsten das Prinzip Datensparsamkeit, denn „more data is better data“. Datenschutzrechtliche Standards wie die Freiwilligkeit und informierte Einwilligung der Beforschten sowie die enge Zweckbindung der erhobenen Daten werden in Zeiten von im Internet erhobenen Big (Social) Data, Open Data Repositories und der Möglichkeit zur Zusammenführung von Metadaten ausgehöhlt. Datengeleitete Methoden ermöglichen schließlich auch bei anonymisierten Daten eine Individuierung von Fällen.

Auch *forschungsethische Grundprinzipien* geraten durch Forschung im Medium des Digitalen zunehmend ins Wanken, etwa die fünf aus dem Belmont Report, dem 2002er-Standard der American Psychological Association (APA) und dem 1999 verabschiedeten Code der American Statistical Association (ASA) amalgamierten ethischen Standards (Beneficence, Nonmaleficence, Justice, Integrity und Respect). Der Beitrag von Forschung zum Gemeinwohl (Beneficence) steht grundsätzlich dann in Frage, wenn privatwirtschaftliche Unternehmen große Mengen personenbezogener Daten für Marktanalysen, Targeting oder Verhaltensvorhersagen analysieren. Die Vermeidung negativer Folgen für Einzelne oder Gruppen (Nonmaleficience) ist als Prinzip etwa dann bedroht, wenn aufgrund von Forschungsergebnisse Zuschreibungen an Personen möglich werden, die bei einer Veröffentlichung geeignet sind, diesen zu schaden, aber auch wenn Sicherheitsbehörden digitale Kommunikation mit innovativen Methoden massenhaft auswerten und dadurch den Eindruck umfassender Überwachung erzeugen. Die Verpflichtung, die beforschten Individuen als autonome Subjekte zu behandeln (Respect), sind durch die Arbeit mit Big Data, in denen der Einzelne zunächst gänzlich unsichtbar wird, ohnehin ausgehöhlt. Gleichbehandlung (Justice) steht dann in Frage, wenn der Zugang zu Daten und Analysetools beschränkt ist oder Personen von der Datenerfassung ausgeschlossen sind (eine Trendanalyse auf Twitterdaten schließt alle Nicht-Twitter-Nutzer aus).

Gleichwohl bergen die neuen Datentypen, Methoden und Forschungsfragen ein großes Potenzial für ein tieferes, empirisch fundiertes Verständnis von Kulturen und die Lösung gesellschaftlicher Probleme. Die Grenzen dessen, was als legitime Forschungspraxis aufzufassen ist, müssen daher neu verhandelt werden.

Daneben muss für die Bereitstellung von und den Zugang zu Rohdaten und Analyseergebnissen gleichermaßen ein neuer Ausgleich zwischen Urheber- und Verwertungsrechten einerseits und informationsethischen Grundprinzipien gefunden werden, der sich stärker in Richtung der Informationsethik orientiert. Floridi (2008) unterscheidet drei Dimensionen der Informationsethik: (1) Informationsehtik als eine Ethik informationeller Ressourcen (availability, accessibility, and accuracy of informational resources), in der etwa Fragen der Digital Divide, der Zuverlässigkeit und

Glaubwürdigkeit informationeller Ressourcen verhandelt werden (Floridi 2008: 6), (2) Informationsethik als eine Ethik informationeller Erzeugnisse (Informationen im Kontext von moralisch zu bewertenden Handlungen wie Werbung, Desinformation, Plagiat, Propaganda etc.), (3) Informationsethik als eine Ethik der informationellen und informationstechnischen Umwelt, in der die Folgen des Handelns in digitalen Umgebungen für andere Akteure reflektiert werden (Hacken, Überwachung, Zensur, Open Source, Piraterie etc.). Eine ganzheitliche Annäherung an informationsethische Fragen muss letztlich alle drei Dimensionen und ihren Einfluss auf die Informationssphäre als Ganze im Sinne einer Makroethik berücksichtigen (Floridi 2008: 11). Eine Orientierung an diesen Prinzipien würde auch die maschinelle Textanalyse zu einer Wissenschaft mit größerer gesellschaftlicher Relevanz und höherer kritischer Potenz machen.

4 Überblick über die Artikel

Roxana Kath, Gary S. Schaal und Sebastian Dumm legen in ihrem Text einen Grundstein für eine „New Visual Hermeneutics“ und fordern dringend eine grundlagentheoretische Fundierung der Digital Humanities. Die New Visual Hermeneutics zielt darauf, die epistemischen und methodischen Grundlagen zu erarbeiten, um „aus großen unstrukturierten Datenkorpora durch explorative Analysen von Visualisierungen“ neue „insights“, also „relevantes neues Wissen“, zu generieren. Die Autoren fordern im Richtungsstreit der DH zwischen „doing research“ und „theorizing“ auf die letztere Seite, denn durch den „computational turn“ (Berry) wurden die epistemischen Grundlagen der DH entscheidend verändert: Code, der verwendet wird, um unstrukturierte Daten zu analysieren, verhält sich nicht neutral, Algorithmen sind Theorie. Ebenso wurden bislang die in der DH praktizierten Forschungsprozesse zu wenig reflektiert und es existieren kaum best practice-Empehlungen.

Die Autoren schlagen einen zirkulären Forschungsprozess mit den Schritten „Sampling und Aufbereitung der Daten“, „Algorithmen basierte Analyse“, „Visualisierung der Daten“ und (in phänomenologischer Perspektive) die „hermeneutische Interpretation der Visualisierung“ vor. Es wird deutlich, dass bereits bei der Datengewinnung und -aufbereitung weitreichende Entscheidungen und Interpretationen getroffen werden, die den Analyseprozess beeinflussen. Eingesetzte Verfahren und Algorithmen müssen „konstruktvalide“ sein und deren Anwendung selber theoretischen Kriterien unterliegen. Die Visualisierungen zielen darauf, „high quality insights“ anzuregen und folgen deshalb nicht bestimmten Visualisierungsstandards, sondern sollen ihre Polylatenz offenbaren. Auf Analyseebene plädieren die Autoren dafür, den phänomenologischen mit einem hermeneutischen Zugang zu kombinieren, um einerseits mit mentaler Offenheit das Neue zu sehen und andererseits dieses in bekannte Kontexte einzubetten – unter Kenntnis der „methodischen, theoretischen und epistemischen Implikationen in allen Phasen des Forschungsprozesses“.

Matthias Lemke und Alexander Stulpe zeigen in ihrem Beitrag „Text und soziale Wirklichkeit“ auf, wie Hermeneutik und Wissenssoziologie erst in Folge des linguistic turns konvergieren und einen „erheblichen textanalytischen Empiriebedarf entwickeln“. Vorher sind die Differenzen deutlich: Während die Hermeneutik davon ausgeht, auktoriale Intentionen und Selbstverständnisse verstehen zu können, negiert die Wissenssoziologie eine autonome Sphäre des Geistigen als idealistisch und thematisiert stattdessen die „Seinsverbundenheit des Denkens“ (Mannheim). Infolge des linguistic turns kommt es aber zu einer Annäherung der beiden Sichtweisen, nämlich einer „konzeptuelle[n] Sozialphänomenologisierung von Texten und Versprachlichung der sozialen Wirklichkeit“. Texte sind nicht nur ideeller Ausdruck, sondern soziale Phänomene, die sprachlich soziale Wirklichkeit konstituieren. Dabei wird klar: Je mehr Texte analysiert werden können, desto weniger beliebig sind die daraus ableitbaren Aussagen.

Die wissensoziologischen Diskurs- (Foucault) und Semantik-Begriffe (Luhmann), sowie die hermeneutischen Diskurs- (Skinner/Pocock) und Semantik-Begriffe (Koselleck) rechtfertigen alle „prinzipiell textanalytische Verfahren sozialwissenschaftlicher Wirklichkeitserfassung“, sind aber unterschiedlichen Denkstilen verpflichtet und stellen unterschiedliche Analysekategorien zur Verfügung. Alle Verfahren präferieren eine transtextuelle Textanalyse über möglichst große

Textmengen hinweg, um Regelmäßigkeiten und Wiederholungen in den Blick zu nehmen.

Im Analysebeispiel zeigen die Autoren mit Frequenz- und Kookkurrenzanalysen zwei Verfahren des Text-Minings, die Operationalisierungen der genannten Diskurs- und Semantik-Konzeptionen sind. Erstere dienen als Zeitreihenanalysen der Periodisierung der Daten und machen auf mögliche Einstiegspunkte für close-reading aufmerksam. Mit den Kookkurrenzanalysen auf der Basis der vorher erarbeiteten Periodisierung werden Veränderungen des Sprachgebrauchs aufgedeckt.

Die Autoren sehen in den Methoden des Text-Minings eine Navigationshilfe zur Orientierung in der Masse von Texten (Strukturierung), die als Indikatoren von nicht-sprachlicher sozialer Wirklichkeit verstanden werden (Repräsentation). Über eine Plausibilitätsprüfung, bestenfalls eine Falsifizierung theoriegeleiteter Hypothesen, wird Wissen aus der Beobachtung der Daten hinaus generiert.

Die Computerlinguistik bietet eine Reihe von Tools, um Sprachdaten mit linguistischen Informationen automatisiert anzureichen. **Heike Zinsmeister** zeigt in ihrem Beitrag die „Chancen und Grenzen von automatischer Annotation“. Es wird deutlich, dass mit Wortarten annotierte Daten eine große Erleichterung bei der Recherche darstellen, es aber wichtig ist, über die Funktionsweise, Möglichkeiten und Probleme von den dafür verwendeten Methoden Bescheid zu wissen. So werden Annotationstools normalerweise anhand von manuell annotierten Daten trainiert, wobei dafür Entscheidungen für ein bestimmtes Annotationsschema getroffen worden sind. Die anschließenden Analysen in entsprechend annotierten Korpora sind deshalb bis zu einem gewissen Grad an die dem Annotationsschema zugrundeliegenden linguistischen Kategorien gebunden.

Ebenso beeinflusst die Architektur des Annotationstools die spätere Analyse. Es wird zwischen regelbasierten und statistischen Systemen unterschieden, wobei erstere in ihren Annotationen dem implementierten Regelsystem, letztere den darauf trainierten Daten unterworfen sind. Zudem arbeiten die Tools nie fehlerfrei. Deshalb ist es wichtig, den Einfluss der Fehlerrate auf die gewünschten Analysen abschätzen zu können.

Die Anwendung computerlinguistischer Tools zur maschinellen Annotation von Korpora ist demnach sinnvoll, solange der Forscher und die Forscherin im Forschungsprozess die den Tools zugrundeliegenden Algorithmen versteht und kritisch reflektieren kann.

„Big Data“ sehen **Claudia Fraas und Christian Pentzold** für diskursanalytische Forschungsvorhaben einerseits als Chance, aber auch als Herausforderung. Die Verfügbarkeit von großen (digitalen) Datenmengen aus verschiedenen Domänen erlaubt es, eine „breite kulturelle Vielfalt und multimodale Varianz an diskursiv zum Ausdruck gebrachten Bedeutungen zu berücksichtigen“. Digitale Daten weisen zudem reiche Metadaten auf (Zeitstempel, Angaben zur Identifikation von Sendern, Marker zur Lokalisierung etc.), die ausgewertet werden können. Gleichwohl ergeben sich verschiedene Probleme bei der Zusammenstellung der empirischen Basis. Darunter sicher besonders einschneidend: Als Referenz für das zu gewinnende Sample kann nicht verlässlich eine Grundgesamtheit definiert und auf die Daten kann nur über z.B. technisch und rechtlich restringierte Zugänge (Suchmaschinen, Programmierschnittstellen) zugegriffen werden.

Aber auch die Auswertung der Daten ist komplex, da diese unterschiedlicher Kommunikationsformen angehören und „die empirische Vielfalt an multimodalen Zeichenensembles“, die sich zudem durch technologische Neuerungen rasch verändert, in Blick genommen werden soll.

Für die „softwareunterstützte Analyse transmedialer multimodaler Diskurse“ schlagen die Autoren die Verwendung des Framekonzeptes als Analysekategorie vor und favorisieren den Forschungsstil der Grounded Theory. Letztere ermöglicht eine systematisches, aber reflexiv-zirkuläres, interpretatives Vorgehen zur Datenerhebung und Auswertung. Die Autoren zeigen an einem Beispiel unter Einsatz der QDA-Software Atlas.ti diesen Forschungsprozess und heben die Vorteile (darunter: dynamische, variable Kodiermöglichkeiten, quantitative Auswertungen der qualitativ vergebenen Kodierungen) aber auch die Beschränkungen hervor. Die Autoren plädieren abschließend für einen „blended approach“, bei dem manuell-verstehende mit computergestützten Methoden kombiniert werden.

Die Beiträge dieser Sonderausgabe mögen eine lebendige Diskussion zu den theoretischen Implikationen und methodischen Möglichkeiten und Problemen der automatischen Textanalyse in den Geistes- und Sozialwissenschaften initiieren.

Literatur Teil Joachim:

Anderson, Chris (2007): The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. In: Wired Magazine 16/07. Published online: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory (19.5.2014)

Bamman, David / Ted Underwood / Noah Smith, "A Bayesian Mixed Effects Model of Literary Character," ACL 2014. 370-379.

Berry, Michael W. / Jacob Kogan (2010): Text Mining Applications and Theory. Hoboken, NJ: John Wiley & Sons.

Beaugrande, Robert-Alain de / Wolfgang Ulrich Dressler (1981): Einführung in die Textlinguistik (= Konzepte der Sprach- und Literaturwissenschaft 28). Tübingen: Niemeyer.

Bögel, Tina / Annette Hautli-Janisz / Sebastian Sulger / Miriam Butt. 2014. Automatic Detection of Causal Relations in German Multilogs. In Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL), Association for Computational Linguistics, pp. 20-27. Gothenburg, Sweden.

Brinker, Klaus (Hrsg.) (1991): Aspekte der Textlinguistik (= Germanistische Linguistik 106/107). Hildesheim/Zürich/New York: Olms.

Bunescu, Razvan C. / Raymond J. Mooney (2006): Extracting Relations from Text: From Word Sequences to Dependency Paths. In: Anne Kao / Stephen R. Poteet (Hrsg.): Natural Language Processing and Text Mining. Springer: London. S. 29-44.

Fix, Ulla (2008): Nichtsprachliches als Textfaktor: Medialität, Materialität, Lokalität. In: Zeitschrift für Germanistische Linguistik 36/3, S. 343-354.

Floridi, Luciano (2008): Foundations of Information Ethics. In: The Handbook of Information and Computer Ethics. Hg v. Kenneth Einar Himma und Herman T. Tavani. Hoboken, New Jersey: Wiley. S. 3-23.

Harweg, Roland (1968): Pronomina und Textkonstitution. München.

Heyer, Gerhard / Uwe Quasthoff / Thomas Wittig (2005): Wissensrohstoff Text. Text Mining: Konzepte, Algorithmen, Ergebnisse. Bochum: W3L.

Koppel, Mosche / Jonathan Schler / Shlomo Argamon (2009): "Computational Methods in Authorship Attribution", JASIST 60 (1): 9–26, doi:10.1002/asi.20961

Mehler, Alexander / Christian Wolff (2005): Einleitung: Perspektiven und Positionen des Text Mining. In: LDV-Forum. 2005;20(1), S. 1-18.

Rieder, Bernhard/Röhle, Theo (2012): Digital Methods: Five Challenges. In: Berry, David M. (Hg.): Understanding Digital Humanities. Basingstoke, 67–84.

Schmidt, Siegfried J. (1976): Texttheorie. Probleme einer Linguistik der sprachlichen Kommunikation (= UTB 202). 2. Auflage. München: Fink.

Stachowiak, Herbert (1973): Allgemeine Modelltheorie. Wien, New York: Springer-Verlag.

Stamatatos, Efstathios (2009): A Survey of Modern Authorship Attribution Methods Journal of the American Society for Information Science and Technology, 60(3), S. 538-556.

Weiss, Sholom M. / Nitin Indurkhya / Tong Zhang / Fred J. Damerau (2005): Text Mining: Predictive Methods for Analyzing unstructured Information. New York: Springer.

Wiedemann, Gregor (2013). Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences [54 paragraphs]. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 14(2), Art. 13,

http://nbn-resolving.de/urn:nbn:de:0114-fqs1302231.

5 Bibliographie

Anderson, Chris (2008): The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. In: Wired Magazine 16 (07), Abgerufen am 04.09.2014 von http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory.

Anthes, Gary (2010): Topic Models Vs. Unstructured Data. In: Communications of the ACM (53), 16–18.

Antos, Gerd (1989): Textproduktion. Ein einführender Überblick. In: Antos, Gerd/Krings, Hans P. (Hg.): Textproduktion: ein interdisziplinarer Forschungsuberblick. Tübingen, 5–57.

Atkins, Sue/Clear, Jeremy/Ostler, Nicholas (1992): Corpus Design Criteria. In: Literary and Linguistic Computing 7 (1), 1–16.

Bartsch, Sabine (2004): Structural and functional properties of collocations in english: a corpus study of lexical and pragmatic constraints on lexical co-occurrence. Tübingen.

Bauer, Matthias/Ernst, Christoph (2010): Diagrammatik / Einführung in ein kultur- und medienwissenschaftliches Forschungsfeld. Bielefeld.

Belica, Cyril/Steyer, Kathrin (2008): Korpusanalytische Zugänge zu sprachlichem Usus. In: Vachková, Marie (Hg.): Beiträge zur bilingualen Lexikographie. Prag, 7–24.

Benzécri, Jean-Paul (1973): L'Analyse des correspondants: introduction, théorie, applications diverses notamment à l'analyse des questionnaires, programmes de calcul. [S.l.].

Berry, David M. (Hg.) (2012): Understanding Digital Humanities. Abgerufen am 4.3.2014 von http://slub.eblib.com/patron/FullRecord.aspx?p=868344.

Biber, Douglas/Jones, James K (2005): Merging corpus linguistic and discourse analytic research goals: Discourse units in biology research articles. In: Corpus Linguistics and Linguistic Theory 1 (2), 151–182.

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

Bluhm, Claudia/Deissler, Dirk/Scharloth, Joachim et al. (2000): Linguistische Diskursanalyse: Überblick, Probleme, Perspektiven. In: Sprache und Literatur in Wissenschaft und Unterricht 88 , 3–19.

Bondi, Marina/Scott, Mike (2010): Keyness in texts. Amsterdam/Philadelphia.

Bubenhofer, Noah (2015): Kollokationen, n-Gramme, Mehrworteinheiten. In: Roth, Kersten/Wengeler, Martin/Ziem, Alexander (Hg.): Handbuch Sprache in Politik und Gesellschaft. Berlin / New York (Sprachwissen).

Bubenhofer, Noah (2009): Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. Berlin, New York (Sprache und Wissen, 4).

Bubenhofer, Noah/Dussa, Tobias/Ebling, Sarah et al. (2009): „So etwas wie eine Botschaft." Korpuslinguistische Analysen der Bundestagswahl 2009. In: Sprachreport 4 , 2–10.

Bubenhofer, Noah/Konopka, Marek/Schneider, Roman (2014): Präliminarien einer Korpusgrammatik. Tübingen (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache CLIP, 4).

Bubenhofer, Noah/Scharloth, Joachim (2011): Korpuspragmatische Analysen alpinistischer Literatur. In: Elmiger, Daniel/Kamber, Alain (Hg.): La linguistique de corpus – de l'analyse quantitative à l'interpretation qualitative / Korpuslinguistik – von der quantitativen Analyse zur qualitativen Interpretatino. Neuchâtel (Travaux neuchâtelois de linguistique, 55), 241–259.

Burger, Harald (1998): Phraseologie. Eine Einführung am Beispiel des Deutschen. Berlin (Grundlagen der Germanistik, 36).

Busse, Dietrich/Hermanns, Fritz/Teubert, Wolfgang (Hg.) (1994): Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik. Opladen.

Busse, Dietrich/Teubert, Wolfgang (2013): Linguistische Diskursanalyse: neue Perspektiven. Auflage: 2013. Wiesbaden.

Carstensen, Kai-Uwe/Ebert, Christian/Ebert, Cornelia et al. (2010): Computerlinguistik und Sprachtechnologie. 3. Aufl. Heidelberg, Berlin Abgerufen am 14.01.2013 von http://www.springer.com/spektrum+akademischer+verlag/informatik/informatik+und+it+%C3%BCbergreifend/book/978-3-8274-2023-7.

Carstensen, Kai-Uwe/Ebert, Christian/Ebert, Cornelia et al. (2009): Computerlinguistik und Sprachtechnologie: Eine Einführung. 3. überarb. u. erw. Aufl.

Chen, Chun-houh/Härdle, Wolfgang/Unwin, Antony (Hg.) (2008): Handbook of data visualization. (Springer handbooks of computational statistics), Abgerufen am 4.3.2014 von http://sfx.ethz.ch/sfx_locater?sid=ALEPH:EBI01&genre=book&isbn=9783540330370&id=doi:10.1007/978-3-540-33037-0 Online via SFX.

Dürscheid, Christa/Elspaß, Stephan/Ziegler, Arne (2011): Grammatische Variabilität im Gebrauchsstandard – das Projekt „Variantengrammatik des Standarddeutschen". In: Konopka, Marek/Kubczak, Jacqueline/Waßner, Ulrich H. (Hg.): Grammatik und Korpora 2009. Tübingen, 123–140.

Evert, Stefan (2009): 58. corpora and collocations. In: Lüdeling, Anke/Kytö, Merja (Hg.): Corpus Linguistics. Berlin, New York (Handbücher zur Sprach- und Kommunikationswissenschaft, 29), 1212–1248.

Feilke, Helmuth (2000): Die pragmatische Wende in der Textlinguistik. In: Brinker, Klaus (Hg.): Text- und Gesprächslinguistik/Linguistics of Text and Conversation. Berlin/New York (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science, 16), 64–82.

Feilke, Helmuth (1996): Sprache als soziale Gestalt. Ausdruck, Prägung und die Ordnung der sprachlichen Typik. Frankfurt am Main.

Feilke, Helmuth (2003): Textroutine, Textsemantik und sprachliches Wissen. In: Linke, Angelika/Ortner, Hanspeter/Portmann-Tselikas, Paul R (Hg.): Sprache und mehr. Ansichten einer Linguistik der sprachlichen Praxis. Tübingen (Reihe Germanistische Linguistik), 209–230.

Feilke, Helmuth/Linke, Angelika (2008): Oberfläche und Performanz – Zur Einleitung. In: Feilke, Helmuth/Linke, Angelika (Hg.): Oberfläche und Performanz. Berlin / New York, 3–18.

Firth, John Rupert (1957): Modes of meaning. In: Papers in Linguistics 1934–1951. London, 190–215.

Francis, Nelson W./Kučera, Henry (1964): Brown Corpus Manual. Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English for Use with Digital Computers. Providence, Rhode Island Abgerufen am 4.3.2014 von http://www.hit.uib.no/icame/brown/bcm.html.

Friendly, Michael (2005): Milestones in the History of Data Visualization: A Case Study in Statistical Historiography. In: Weihs, Claus/Gaul, Wolfgang (Hg.): Classification: The Ubiquitous Challenge. New York, 34–52.

Geyken, Alexander (2007): The dwds corpus: a reference corpus for the german language of the 20th century. In: Fellbaum, Christiane (Hg.): Collocations and Idioms: Linguistic, lexicographic, and computational aspects. London, 23–42.

Graham, Shawn/Weingart, Scott/Milligan, Ian (2012): Getting Started with Topic Modeling and MALLET. Abgerufen am 4.3.2014 von http://programminghistorian.org/lessons/topic-modeling-and-mallet.

Hausmann, Franz Josef (1985): Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In: Bergenholtz, H./Mugdan, J. (Hg.): Lexikographie und

Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 1984. Tübingen (Lexicographica Series Maior), 118–129.

Hermanns, Fritz (1995): Sprachgeschichte als Mentalitätsgeschichte. Überlegungen zu Sinn und Form und Gegenstand historischer Semantik. In: Gardt, Andreas/Mattheier, Klaus/Reichmann, Oskar (Hg.): Sprachgeschichte des Neuhochdeutschen. Gegenstände, Methoden, Theorien. Tübingen, 69–101.

Keim, Daniel A./Kohlhammer, Jörn/Ellis, Geoffrey et al. (2010): Mastering the information age - solving problems with visual analytics. Goslar Abgerufen am 4.3.2014 von http://www.vismaster.eu/book/.

Krämer, Sybille (2009): Operative Bildlichkeit. Von der ‚Grammatologie' zu einer ‚Diagrammatologie'? In: Heßler, Martina/Mersch, Dieter (Hg.): Logik des Bildlichen. Zur Kritik der ikonischen Vernunft. Bielefeld (Metabasis, 2), 94–123.

Kupietz, Marc/Belica, Cyril/Keibel, Holger et al. (2010): The german reference corpus dereko: a primordial sample for linguistic research. In: Proceedings of the 7th conference on International Language Resources and Evaluation. Valletta, Malta, 1848–1854.

Kupietz, Marc/Keibel, Holger (2008): Gebrauchsbasierte Grammatik: Statistische Regelhaftigkeit. In: Konopka, Marek/Strecker, Bruno (Hg.): Deutsche Grammatik. Regeln, Normen, Sprachgebrauch. Berlin/New York, 33–50.

Manning, Christopher D/Schütze, Hinrich (2002): Foundations of statistical natural language processing. 5. Aufl. Cambridge, Massachusetts.

Michel, Jean-Baptiste/Shen, Yuan Kui/Aiden, Aviva Presser et al. (2011): Quantitative Analysis of Culture Using Millions of Digitized Books. In: Science 331 (6014), 176–182.

Perkuhn, Rainer/Belica, Cyril (2006): Korpuslinguistik – Das unbekannte Wesen. Oder Mythen über Korpora und Korpuslinguistik. In: Sprachreport 22 (1), 2–8.

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): Korpuslinguistik. Stuttgart.

Rieder, Bernhard/Röhle, Theo (2012): Digital Methods: Five Challenges. In: Berry, David M. (Hg.): Understanding Digital Humanities. Basingstoke, 67–84.

Risch, John/Kao, Anne/Poteet, Stephen et al. (2008): Text visualization for visual text analytics. In: Simoff, Simeon/Böhlen, Michael/Mazeika, Arturas (Hg.): Visual Data Mining. Berlin, Heidelberg (Lecture Notes in Computer Science), 154–171.

Rohrdantz, Christian/Hautli, Annette/Mayer, Thomas et al. (2012): Towards tracking semantic change by visual analytics. Abgerufen am 04.03.2013 von http://kops.ub.uni-konstanz.de/handle/urn:nbn:de:bsz:352-186381.

Rohrdantz, Christian/Koch, Steffen/Jochim, Charles et al. (2010): Visuelle Textanalyse. In: Informatik-Spektrum 33 (6), 601–611, doi: 10.1007/s00287-010-0483-x.

Schäfer, Roland/Bildhauer, Felix (2012): Building Large Corpora from the Web Using a New Efficiant Tool Chain. In: Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry et al. (Hg.): Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, 486–493.

Scharloth, Joachim/Bubenhofer, Noah (2011): Datengeleitete Korpuspragmatik: Korpusvergleich als Methode der Stilanalyse. In: Felder, Ekkehard/Müller, Marcus/Vogel, Friedemann (Hg.):

Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen von Texten und Gesprächen. Berlin, New York, 195–230.

Scharloth, Joachim/Eugster, David/Bubenhofer, Noah (2013): Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn. In: Busse, Dietrich/Teubert, Wolfgang (Hg.): Linguistische Diskursanalyse. Neue Perspektiven. Wiesbaden, 345–380.

Schiller, Anne/Teufel, Simone/Thielen, Christine (1995): Guidelines für das Tagging deutscher Textcorpora mit STTS. Stuttgart.

Scott, Mike/Tribble, Chris (2006): Textual patterns: key words and corpus analysis in language education.

Sinclair, John (1991): Corpus, Concordance, Collocation. Oxford.

Spitzmüller, Jürgen/Warnke, Ingo H. (2011): Diskurslinguistik: eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse.

Steyer, Kathrin (2000): Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikografische Möglichkeiten. In: Deutsche Sprache 28 , 101–125.

Steyer, Kathrin (2013): Usuelle Wortverbindungen: Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht. Tübingen.

Steyer, Kathrin/Brunner, Annelen (2009): Das UWV-Analysemodell. Eine korpusgesteuerte Methode zur linguistischen Systematisierung von Wortverbindungen. (Online publizierte Arbeiten zur Linguistik OPAL, 1).

Steyer, Kathrin/Lauer, Meike (2007): „Corpus-Driven“: Linguistische Interpretation von Kookkurrenzbeziehungen. In: Eichinger, Ludwig M/Kämper, Heidrun (Hg.): Sprach-Perspektiven. Germanistische Linguistik und das Institut für Deutsche Sprache. Tübingen (Studien zur deutschen Sprache), 493–509.

Stjernfelt, Frederik (2007): Diagrammatology: an investigation on the borderlines of phenomenology, ontology, and semiotics. Dordrecht; London.

Teubert, Wolfgang (2005): My version of corpus linguistics. In: International Journal of Corpus Linguistics 10 (1), 1–13.

Tognini-Bonelli, Elena (2001): Corpus linguistics at work. Amsterdam (Studies in Corpus linguistics, 6).

Tufte, Edward R (1983): The visual display of quantitative information. Cheshire, Conn.

Tukey, John W (1977): Exploratory Data Analysis. Reading, Masachusetts [etc.] (Addison Wesley Series in Behavioral Science. Quantitative Methods).

Vogel, Friedemann (2010): Linguistische Imageanalyse (LIma). Grundlegende Überlegungen und exemplifi zierende Studie zum öffentlichen Image von Türken und Türkei in deutschsprachigen Medien. In: Deutsche Sprache (4), 345–377.

Warnke, Ingo H (2007): Diskurslinguistik nach Foucault – Dimensionen einer Sprachwissenschaft jenseits textueller Grenzen. In: Warnke, Ingo H (Hg.): Diskursanalyse nach Foucault. Theorie und Gegenstände. Berlin, New York (Linguistik – Impulse & Tendenzen), 3–24.