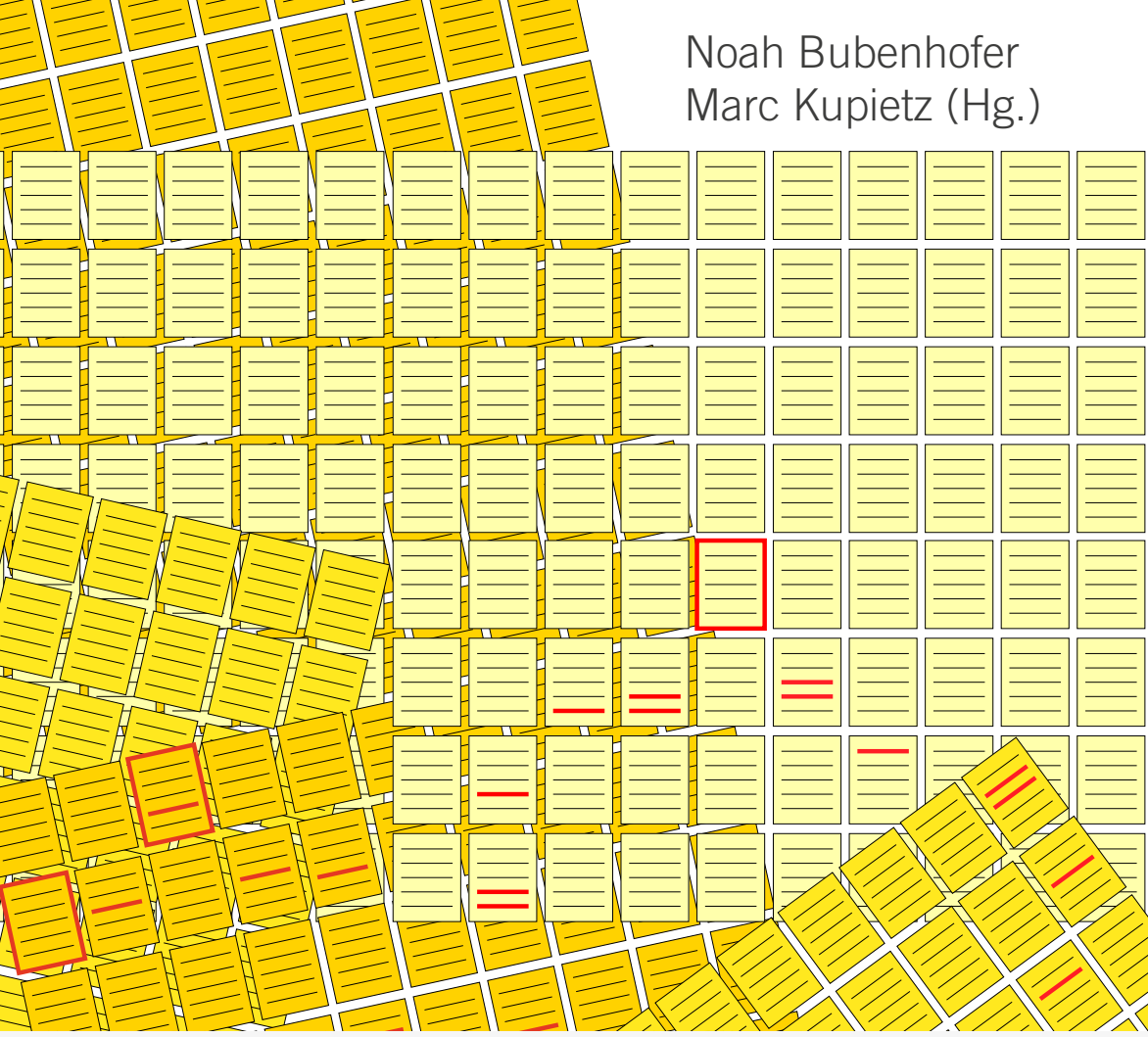
Noah Bubenhofer
Marc Kupietz (Hg.)



VISUALISIERUNG SPRACHLICHER DATEN

HEIDELBERG
UNIVERSITY PUBLISHING

Visualisierung sprachlicher Daten

Visualisierung sprachlicher Daten

Visual Linguistics – Praxis – Tools

Herausgegeben von
Noah Bubenhofer und Marc Kupietz

HEIDELBERG
UNIVERSITY PUBLISHING

Über die Herausgeber

Noah Bubenhofer leitet den Arbeitsschwerpunkt Digital Linguistics an der Zurich University of Applied Sciences, Winterthur. Zuvor war er Leiter des Projekts "Visual Linguistics" an der Universität Zürich. Seine Forschungsgebiete sind die Korpuslinguistik, Diskurslinguistik und Visualisierungen in der Wissenschaft.

Marc Kupietz leitet am Institut für Deutsche Sprache in Mannheim den Programmbereich Korpuslinguistik. Er forscht in den Bereichen Korpuslinguistik, empirisch fundierte Sprach- und Kognitionswissenschaft sowie Wissenschaftstheorie und Wissenschaftsmanagement.

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie. Detaillierte bibliografische Daten sind im Internet unter http://dnb.ddb.de abrufbar.

Die Online-Version dieser Publikation ist auf den Verlagswebseiten von HEIDELBERG UNIVERSITY PUBLISHING http://heiup.uni-heidelberg.de dauerhaft frei verfügbar (open access).
urn: urn:nbn:de:bsz:16-heiup-book-345-0
doi: https://doi.org/10.17885/heiup.345.474

ISBN 978-3-946054-77-1 (Hardcover)
ISBN 978-3-946054-75-7 (PDF)

Inhalt

Noah Bubenhofer / Marc Kupietz

Einleitung

1. Funktionen von Visualisierungen in den Wissenschaften

Visualisierungen von Daten spielen in den Wissenschaften eine wichtige Rolle im Forschungsprozess. Einerseits dienen sie der Illustration von gewonnener Erkenntnis, beispielsweise in der Form von Balken-, Streu- oder Liniendiagrammen, die Mess- oder Zählwerte repräsentieren. Solche Visualisierungen werden „Presentation Graphics" (Präsentationsgrafiken) genannt (Chen u. a. 2008, S. 4). Andererseits sind Visualisierungen aber auch eigenständige Mittel der Erkenntnisgewinnung, wenn andere Formen der Repräsentation von Wissen wie Listen, Tabellen oder Texte zu umfangreich oder zu komplex sind, um als Ganzes erfasst und gedeutet werden zu können. Visualisierungen dieser Art werden zur Gruppe der „Exploratory Graphics" (explorativen Visualisierungen) gezählt (Chen u. a. 2008, S. 5; Schumann und Müller 1999, S. 5).

Explorative Visualisierungsmethoden werden insbesondere im Bereich der Visual Analytics (Keim u. a. 2010; Chen u. a. 2008) eingesetzt. Visualisierungen transformieren, gewichten und filtern komplexe Daten und bringen sie dadurch in eine Form, die sie als Informationen erfassbar und interpretierbar machen. Visualisierungen sind damit keine Abbildungen der Wirklichkeit, sondern aufgrund von Relevanzkriterien geordnete und damit interpretative Reduktionen von Daten, die auf der Basis gestalterischer Vorgaben visuell repräsentiert werden. Visualisierungen könnten demnach auch als Scharnier zwischen quantitativ-maschinellen und qualitativ-interpretierenden Analysen angesehen werden, da durch einen iterativen Prozess der interpretativen Interaktion mit den Daten Modelle generiert werden (vgl. den Visual Analytics Process nach Keim u. a. 2010, S.10): „Visualisation becomes the medium of a semi-automated analytical process, where humans and machines cooperate using their respective, distinct capabilities for the most effective results" (Keim u. a. 2010, S.14). Insbesondere explorative Visualisierungen sind entsprechend nicht Endprodukt, sondern typischerweise Zwischenprodukt und Mittel, die (1) quantitativ-maschinelle mit der (2) qualitativ-interpretierenden Analyse in einem iterativen,

empirisch-hermeneutischen Erkenntnisprozess zu kombinieren. Dadurch wird etwa die Abduktion neuer, vielversprechender Hypothesen möglich (vgl. Jockers 2013; Kupietz/Keibel 2009, S. 48).

Der Wert visueller Methoden für die Analyse großer Datenmengen wird zwar in einigen Publikationen zur statistischen Theorie und Methode erkannt, trotzdem mahnen Chen et al. (2008, S. 4) die fehlende Reflexion über die gängigen Methoden der Visualisierung statistisch gewonnener Daten an: "Examples abound in almost every issue of every scientific journal concerned with quantitative analysis. There are occasionally articles published in a more theoretical vein about specific graphical forms, but little else" (Chen u. a. 2008, S. 4). Allerdings bleibt anzumerken, dass es eine Reihe von Arbeiten gibt, die die semiotischen und kognitiven Grundlagen der Visualisierung reflektieren und Regeln für die Erstellung von Grafiken formulieren (Bertin 1967; Tufte 1983; Tufte 1997; Schumann und Müller 1999; Unwin u. a. 2006), die Algorithmen zur Erstellung von Visualisierungen, etwa von Graphen, diskutieren (Tamassia 2013; Brandes u. a. 2013) oder den Wert von Visualisierungen aus Sicht der Benutzerinnen und Benutzer, z. B. im Web, darlegen (Hearst 2009; Hearst und Rosner 2008).

Insbesondere existiert eine Reihe von Arbeiten, die grundlegende semiotische und kulturelle Aspekte des Diagramms reflektieren: Welcher Art ist dieses Zeichen? Wie werden Diagramme eingesetzt, um Wissen zu ordnen, zu generieren oder zu kommunizieren? Welche historischen Grundfiguren des Diagramms gibt es und wie hängen diese mit ideengeschichtlichen Entwicklungen zusammen? Hierzu sind in den letzten Jahren einige Arbeiten entstanden, die eine Theorie der Diagrammatik skizzieren und elaborieren – oft im Rückgriff auf Charles Sanders Peirce (Bauer und Ernst 2010; Bender und Marrinan 2010; Bredekamp u. a. 2008; Krämer 2016; Liebsch und Mößner 2012; Reichert 2013; Siegel 2009; Stetter 2005; Stjernfelt 2007).

Die Wurzeln der explorativen Datenanalyse ("Exploratory Data Analysis", "EDA") gehen auf Tukey (1977) und Benzécri (1973b; 1973a) zurück, wobei die Geschichte der Visualisierung statistisch gewonnener Daten viel älter ist (Friendly 2005; Tufte 1983). So gilt Michael Florent van Langrens 1644 erstellte Grafik der Schätzungen verschiedener Astronomen zur Longitudinal-Differenz zwischen Toledo und Rom als erste visuelle Repräsentation statistischer Daten (Tufte 1997, S. 15). In der Folge wurden immer häufiger Visualisierungen verwendet, nicht nur, um komplexe statistische Zusammenhänge zu präsentieren, sondern auch, um sich einen Überblick über die Daten zu verschaffen und überhaupt die immer größer werdenden Datenbestände und darin auftretende Zusammenhänge analysieren zu können (Zhang 2008, S. IX). Die Grundlagen der visuellen Datenanalyse werden in einer Reihe von Werken erarbeitet und beispielhaft angewandt (Zhang 2008; Dill u. a. 2012; Chen u. a. 2008; Burkhart und Eppler 2004; Mazza 2009; Tufte 1983; Keim u. a. 2010; Arnold 2008).

2. Traditionelle Visualisierungen in der Sprachwissenschaft

In der Linguistik sind besonders in der Dialektologie Visualisierungen in Form von Karten schon lange gebräuchlich und sind sowohl „Dokumentations-“ als auch „Forschungsmittel“ (Naumann 1982) – dienen also sowohl der Präsentation von Ergebnissen als auch der Exploration von Daten. Es entwickelten sich verschiedene Typen von Themenkarten (Originalformkarten, Punktsymbolkarten, Flächenkarten, kombinierte Karten etc.), die sich zwischen Dokumentation und Interpretation bewegen. Auch in der Phonetik sind Visualisierungen für explorative Zwecke wichtig, z. B. Spektrogramme, um die Lage der Formanten zu erkennen und damit beispielsweise die Stimmqualität oder prosodische Eigenschaften zu messen (Reetz 2003).

In der strukturalen Syntax sind Baumgraphen und andere Visualisierungsmöglichkeiten von Strukturen (vgl. z. B. die syntagmatische Verkettung als Spirale bei Mikuš 1952; zit. nach Thümmel 1993a, S. 271) weit mehr als nur eine Darstellungshilfe, sie sind vielmehr Ausdruck strukturalistischer Theoriebildung (Thümmel 1993b; Heringer 1993). Ähnlich verhält es sich mit zahlreichen weiteren Visualisierungen von Modellen, die Sprachwirklichkeit beschreiben wollen. Exemplarisch sei auf die im germanistischen Raum relativ bekannte Visualisierung des soziolinguistischen Varietätenmodells von Löffler (1994) verwiesen. Es folgt dem Forschungsparadigma der Varietätenlinguistik, die Sprache als komplexe Menge von sprachlichen Varietäten und nicht als „unmittelbar gegebene[n] (homogene[n]) Gegenstand“ ansieht (Bußmann 2002, S. 729), was in der Visualisierung durch sich überlagernde Vektoren widerspiegelt wird. Ein weiterer wichtiger Anwendungsbereich von Visualisierungen sind Stammbäume in der vergleichenden Sprachwissenschaft. Schleicher (1860, S. 28) gilt als erster Sprachwissenschaftler, der eine Stammbaumdarstellung für den Sprachvergleich eingesetzt hat (Sutrop 2012, S. 299). Der Stammbaum reproduziert als gerichteter Graph bestimmte Ordnungsprinzipien, die sich nur bedingt mit modernen Auffassungen von Sprachfamilien vereinbaren lassen (Sutrop 2012, S. 320). Alternative Darstellungsformen ergeben sich dabei auch aus neuen methodischen, statistischen Zugängen (Fox 1995; Jäger 2014).

Weniger offensichtlich sind die Formen der Visualisierung in der Gesprächsanalyse: Dort müssen die flüchtigen Daten des Gesprächs dokumentiert werden, üblicherweise in Form einer Transkription (Redder 2001; Deppermann 2001). Erst die Transkription erlaubt anschließend die Exploration der Daten (Sager 2001). Die Art der Transkription richtet sich nach dem Erkenntnisinteresse, sodass verschiedene Transkriptionsstandards existieren, die als eine Form von Visualisierung je unterschiedliche Aspekte der komplexen Daten (Transliteration, Normalisierung, Intonation, Betonung etc.) hervorheben.

Interessant ist auch ein Blick über die engen disziplinären Grenzen zu den Literaturwissenschaften: Hier gilt Moretti (2000; 2009) als einer der Wegbereiter für eine neue, visuelle Sicht auf Literatur, die seinem Paradigma des Distant Reading folgt. Voraussetzung dafür ist die Computerphilologie, die beispielsweise die kritische Edition von Texten mit den Möglichkeiten der digitalen Aufbereitung (Annotation, dynamische Textdarstellung etc.) und Analyse verbindet (Jannidis 1999; Jannidis et al. 2017; Lauer 2011).

Einen erhellenden Überblick über Visualisierungen in der Linguistik bietet Harleman Stewart (1976). Sie reflektiert den Einfluss von grafischen Repräsentationen auf die Theoriebildung und analysiert Visualisierungen wie Baumgraphen in der vergleichenden Sprachwissenschaft, Phonetik, Syntax etc. Dabei wird die Komplexität des Visualisierungsprozesses deutlich, durch den nicht nur Daten interpretiert, sondern heuristisch Theorien modelliert werden. Diese ältere Publikation reflektiert jedoch noch nicht die neueren Entwicklungen im Bereich der Visualisierungen in der Sprachwissenschaft, die sich zudem seit den 1970er-Jahren in vielen Bereichen stark gewandelt hat.

3. Visuelle Textanalyse: Visualisierungen in der Korpuslinguistik und den datenintensiven Digital Humanities

In der Sprachwissenschaft und den Digital Humanities ist es insbesondere die Korpuslinguistik, bei der der Bedarf für neue Formen der visuellen Analyse stark ansteigt. Bei hypothesengeleiteten Ansätzen entstehen quantitative Analyseergebnisse, die visualisiert werden können („presentation graphics"). Doch die Verfügbarkeit großer Textmengen erlaubt es auch, datengeleitete Analyseverfahren anzuwenden, die der Hypothesengenerierung dienen. Im größeren Kontext der Digital Humanities zeigt sich zudem die Chance, nicht nur mit Textdaten zu arbeiten, sondern verschiedene Datentypen (Bilder, Daten historischer Ereignisse, geografische Informationen – GIS, soziodemografische Daten etc.) integrieren zu können. In der Korpuslinguistik werden deshalb vermehrt Methoden der analytischen Statistik und des Data Minings angewandt, um die verfügbaren Daten auswerten zu können (Manning und Schütze 2002; Baayen 2008; Gries 2009b).

Allerdings bringt die Analyse solcher Daten eine Reihe von Herausforderungen mit sich: 1) Textdaten gehören zu den unstrukturierten Datentypen und unterscheiden sich von anderen Daten, die dem Data Mining normalerweise zugrunde liegen. 2) Die Daten sind oft heterogen, da sie unterschiedliche Datentypen vereinen. 3) Die Daten sind oft komplex, da die einzelnen Datentypen wiederum eine Vielzahl von Ebenen umfassen können: Bei Textdaten sind das verschiedene Annotationsebenen, wie sie typischerweise in Korpora, die mit Methoden des Natural Language Processing aufbereitet worden sind, auftreten

(Wortartkategorien, Lemma, syntaktische Struktur), aber auch (halb-)manuell erzeugte Annotationen.

In anderen Disziplinen, die mit Big Data dieser Art arbeiten, erwiesen sich visuelle Analysemethoden, eben „exploratory graphics", als besonders fruchtbar (Tukey 1977; Thomas und Cook 2005; Unwin u. a. 2006; Chen u. a. 2008; Dill u. a. 2012). Als Teilgebiet der visuellen Analyse etablieren sich gegenwärtig die Visual Text Analytics (visuelle Textanalyse), die das Paradigma der Datenvisualisierung auf Textdaten anwendet (Risch u. a. 2008; Rohrdantz u. a. 2010). Erste Anwendungsbeispiele sind vielversprechend, doch fehlt noch weitgehend die theoretische und methodische Reflexion.

Visuelle Analysemethoden von Textdaten können dann gewinnbringend eingesetzt werden, wenn ein Analyseverständnis vorherrscht, bei dem nicht die Einzelbelege im Vordergrund stehen, sondern bei dem mit statistischen Mitteln Regularitäten im Sprachgebrauch aufgedeckt werden. Dazu stehen immer größere Korpora in Größenordnungen ab 1 Mia. Textwörter zur Verfügung, und es wird deutlich, dass ein Mehr an Daten auch ein Mehr an Analysemöglichkeiten bietet (Church und Mercer 1993). Die dafür nötigen statistischen Methoden werden gegenwärtig entwickelt und diskutiert, wie eine Vielzahl von methodologisch-statistischen Arbeiten (vgl. z. B. Kilgarriff 2005; Gries 2005; Gries 2008a; Gries 2009a; Gries 2010b; Hilpert & Gries 2009; Gries 2010a; Evert 2005; Rietveld & Hout 2005; Biber & Jones 2009) zeigt, wobei sich diese Forschungsrichtung auch bereits in Lehrbüchern niederschlägt (Gries 2008b; Baayen 2008). Fragen der Visualisierung generell, insbesondere auch der visuellen Analyse, scheinen dabei noch sekundär zu sein, obwohl z. B. Gries betont, dass die Visualisierung der Ergebnisse deskriptiver Analyse hilfreich ist (Gries 2008b, S.268) und die bereits verfügbaren statistischen Methoden, darunter auch Formen der Visualisierung, noch längst nicht ausgeschöpft sind (Gries 2010b, S. 24).

Korpus- und computerlinguistische Anwendungen visueller Textanalyse liegen z. B. für diskursive Daten vor (Luo u. a. 2012), um Sprachwandel oder semantische Variation zu analysieren (Hilpert 2011; Hao u. a. 2010; Rohrdantz, Hao u. a. 2012) oder die Lesbarkeit von Texten visuell auszudrücken (Oelke, Spretke u. a. 2012). Weiter gibt es Vorschläge, kontinuierlich entstehende Textdaten zu visualisieren (Rohrdantz u. a. 2011; Diakopoulos u. a. 2010) oder Ergebnisse von maschinellen semantischen Analysen, z. B. von Kundenrezensionen, zusammenfassend darzustellen (Alper u. a. 2011; Rohrdantz, Hao u. a. 2012; Shi u. a. 2010). Einige Arbeiten versuchen neue Visualisierungsformen für traditionelle, aber unbefriedigende Formate zu finden, wie z. B. Kollokationsgraphen oder Wortwolken (Gambette und Veronis 2009; Rockwell u. a. 1999; Wattenberg und Viegas 2008; Culy und Lyding 2010; Leblanc und Pérès 2010; Oelke, Eklund u. a. 2012; Collins u. a. 2009; Brandes u. a. 2006), diatopische Karten und andere Geotextdaten (Vriend u. a. 2011; Gregory und Hardie 2011), Netzwerke (Efer u. a.

2012), Syntax-Baumgraphen (Derrick & Archambault 2010) oder andere grammatische Muster (Elliott u. a. 2001; Säily u. a. 2011). Visualisierungen werden auch eingesetzt, um komplexe Korrelationen darzustellen, z. B. in Datensammlungen linguistischer Eigenschaften von Vornamen (Wattenberg 2005) oder bei der Berechnung von Ähnlichkeiten zwischen Sprachen oder Dialekten (Rohrdantz, Hund u. a. 2012; Zastrow 2011). Bei der maschinellen Textklassifikation dienen Visualisierungen auch dazu, die Auswahl der Variablen für die Modellierung zu unterstützen (May u. a. 2010; Oelke u. a. 2008; Chuang u. a. 2012).

Der Forschungsstand zeigt, dass in der Sprachwissenschaft sowohl für Darstellungszwecke als auch für die Datenexploration häufig Visualisierungstechniken eingesetzt werden, die Reflexion darüber jedoch oft fehlt. So werden beispielsweise Wortwolken zur Darstellung von häufigem Vokabular eingesetzt – insbesondere auch außerhalb der Wissenschaften scheint diese Darstellung, z. B. als Tag-Clouds, sehr attraktiv zu sein –, die Visualisierung weist aber den gravierenden Mangel auf, dass die Position des Wortes in der Wolke nicht semantisiert ist (die Wörter also zufällig angeordnet sind) oder die Semantisierung unwichtige Kriterien abbildet (bei alphabetischer oder gestalterisch optimierter Anordnung). Darüber hinaus sind die statistischen Berechnungsmethoden meist nicht transparent oder entsprechen nicht dem State of the Art, wenn z. B. die Größe des abgebildeten Wortes in ikonischem Verhältnis zur absoluten Auftretenshäufigkeit (unter Ausschluss von bestimmten Stoppwörtern) statt im Verhältnis zur statistischen Signifikanz der Frequenz im Vergleich zu einem Referenzkorpus steht.

4. Konzeption des Buches

Das vorliegende Buch ist aus dem im November 2014 veranstalteten Herrenhäuser Symposium *Visuelle Linguistik – Theorie und Anwendung von Visualisierungen in der Sprachwissenschaft*[1] entstanden. Ziel des Symposiums war es, vor dem oben dargelegten historischen und theoretischen Hintergrund die aktuellen und zukünftigen Herausforderungen und Chancen im Hinblick auf Visualisierungen in der Linguistik zu diskutieren. Ein wesentliches Anliegen des Symposiums war es dabei, ein sowohl horizontal breites Spektrum verschiedener Ansätze in der Linguistik und den angrenzenden Disziplinen in den Digital Humanities aufzugreifen als auch vertikal das Spektrum zwischen meta-theoretischen Überlegungen und konkreten Anwendungen abzubilden, um verschiedene Ansatzalternativen und ihre Integration in Methodologien und Arbeitsprozesse im Hinblick auf den möglichen Erkenntnisgewinn in einem interdisziplinären Teilnehmerfeld zu diskutieren.

1 Gefördert von der VolkswagenStiftung (Az. 88445).

Das Buch enthält schriftliche Ausarbeitungen ausgewählter Beiträge des Symposiums und gliedert sich in die drei großen Bereiche I Visual Linguistics, II Praxis und III Tools. Im ersten Teil sind Beiträge versammelt, die die theoretischen und methodologischen Grundlagen von Visualisierungen in der Linguistik diskutieren. Noah Bubenhofers Beitrag zur Eröffnung ist gleichzeitig als Einführung ins Thema und starkes Plädoyer für eine neue Sicht auf das Feld gedacht, er dient zudem auch dazu, die folgenden Beiträge zu kontextualisieren. Ebenfalls grundlegenden Charakter hat der Beitrag von Rainer Perkuhn und Marc Kupietz, in dem insbesondere die theoretisch-methodologischen Grundlagen für Visualisierungen im gegenwärtig wichtigsten Bereich der Linguistik, der Korpuslinguistik, dargelegt werden. Die weiteren Beiträge des ersten Teils adressieren weitere Bausteine einer Methode der visuellen Linguistik (Lauersdorf), wobei Jana Pflaegings Beitrag auch formal das Thema aufgreift, indem ihr Standpunkt zur Ästhetisierung linguistischer Wissensvermittlung auch als visuelles Statement gestaltet ist.

Im zweiten Teil folgen Beiträge, die einen Einblick vermitteln, wie in die Forschungspraxis mit visuellen Analysemöglichkeiten umgegangen wird. Einerseits werden darin linguistisch, literaturwissenschaftlich und psycholinguistisch interessante Erkenntnisse zu verschiedenen Themen – Referenzen in der Lyrik (Hoenen), Toponyme in literarischen Texten (Barbaresi) und Story Complexity (Flekova / Stoffel / Gurevych / Keim) – präsentiert, andererseits die Funktion und Bedeutung der dafür benutzten visuellen Analyseinstrumente kritisch reflektiert.

Der dritte Teil „Tools“ fokussiert stärker die Ebene des Analysewerkzeugs und zeigt anhand verschiedener Anwendungsgebiete Methoden der visuellen Analyse. Den Auftakt bildet ein Beitrag zur sehr universell einsetzbaren Programmiersprache R, die häufig für visuelle Analysen unter Einsatz verschiedenster Methoden eingesetzt wird (Wolfer, Hansen). Es folgen Beiträge über Werkzeuge, zunächst zur explorativen Analyse von Korpora mit einer dreistufigen Realisierung der Visualisierungsprozesse (Rüdiger), dann zur Darstellung und Interpretation von Topic Models (Hinneburg / Oberländer) und zur Signifikanzeinschätzung von Frequenzunterschieden bei N-Grammen (Todorova / Chinkina). Im letzten Beitrag wird der Einfluss von Visualisierungen auf die Usability von Annotationstools diskutiert (Burghardt), sodass ein breites Feld von visuellen Methoden und Anwendungsbereichen in der Linguistik skizziert werden kann.

Die Herausgeber bedanken sich bei der VolkswagenStiftung für die großzügige Finanzierung des Symposiums und der vorliegenden Publikation. Ebenfalls möchten wir uns beim Verlag für die umsichtige Betreuung und das sorgfältige Lektorat und die Produktion des Buches bedanken. Und schließlich geht der Dank an alle Autorinnen und Autoren dieses Bandes, ohne deren Engagement weder das Symposium erfolgreich durchgeführt noch das Buch hätte entstehen können.

Bibliografie

Alper, Basak, Huahai Yang, Eben Haber und Eser Kandogan. 2011. "OpinionBlocks: Visualizing Consumer Reviews." In *IEEE Workshop on Interactive Visual Text Analytics for Decision Making*. Providence, RI. http://vialab.science.uoit.ca/textvis2011/papers/textvis%202011-alper.pdf.

Arnold, Claus. 2008. *Visualisierung im Information Retrieval*. Saarbrücken: Dr. Müller.

Baayen, R. Harald. 2008. *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bauer, Matthias, Ernst, Christoph. 2010. *Diagrammatik: Einführung in ein kultur- und medienwissenschaftliches Forschungsfeld*. Bielefeld: transcript.

Bender, John und Michael Marrinan. 2010. *The Culture of Diagram*. Stanford, Calif: Stanford University Press.

Benzécri, Jean-Paul. 1976-1980. *L'analyse des données*. Paris: Dunod.

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

Benzécri, Jean-Paul. 1973b. *L'analyse des données : leçons sur l'analyse factorielle et la reconnaissance des formes et travaux du laboratoire de statistique de l'Université de Paris VI*. Paris: Dunod.

Bertin, Jacques. 1967. *Sémiologie graphique. Les diagrammes, les réseaux, les cartes*. Paris: Mouton.

Biber, Douglas und James K. Jones. 2009. „Quantitative methods in corpus linguistics." In *Corpus Linguistics*, herausgegeben von Anke Lüdeling und Merja Kytö. Berlin: Mouton de Gruyter, 1286–1304.

Brandes, Ulrik, Linton C. Freeman und Dorothea Wagner. 2013. Social Networks. In *Handbook of Graph Drawing and Visualization*, herausgegeben von Roberto Tamassia. London: Boca Raton.

Brandes, Ulrik, Martin Hoefer und Jürgen Lerner. 2006. *WordSpace: Visual Summary of Text Corpora*. https://dx.doi.org/10.1117/12.647867.

Bredekamp, Horst, Birgit Schneider, und Vera Dünkel, Hrsg. 2008. *Das Technische Bild: Kompendium zu einer Stilgeschichte wissenschaftlicher Bilder*. Berlin: Akademie-Verlag.

Burkhard, Remo A. und Martin J. Eppler. 2004. *Knowledge Visualization*. In: *Encyclopedia of Knowledge Management*. Hershey, PA: Idea Group Reference, 551–560.

Bußmann, Hadumod. 2002. *Lexikon der Sprachwissenschaft*. 3., aktual. und erw. Aufl.. Stuttgart: Kröner.

Chen, Chun-houh, Wolfgang Härdle und Antony Unwin, Hrsg. 2008. *Handbook of Data Visualization*. Berlin: Springer (Springer Handbooks of Computational Statistics).

Chuang, Jason, Christopher D. Manning und Jeffrey Heer. 2012. "Termite: Visualization Techniques for Assessing Textual Topic Models." In: *Advanced Visual Interfaces*. http://vis.stanford.edu/papers/termite (letzter Zugriff am 27. November 2017).

Church, KW und RL Mercer. 1993. "Introduction to the special issue on computational linguistics using large corpora." *Computational Linguistics*. 19 (1): 1–24.

Collins, Christopher, Fernanda B. Viegas und Martin Wattenberg. 2009. "Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora." In *2009 IEEE Symposium on Visual Analytics Science and Technology,* Atlantic City. 91 –98. https://doi.org/10.1109/VAST.2009.5333443.

Culy, Chris und Verena Lyding. 2010. "Double Tree: An Advanced KWIC Visualization for Expert Users." In *2010 14th International Conference Information Visualisation,* London, 98–103. https://doi.org/10.1109/IV.2010.24.

Deppermann, Arnulf. 2001. *Gespräche analysieren*. Opladen: Leske + Budrich.

Derrick, Donald und Daniel Archambault. 2010. „TreeForm: Explaining and Exploring Grammar Through Syntax Trees." *Literary and Linguistic Computing* 25 (1): 53–66. https://doi.org/10.1093/llc/fqp031.

Diakopoulos, Nicholas, Mor Naaman und Funda Kivran-Swaine. 2010. „Diamonds in the Rough: Social Media Visual Analytics for Journalistic Inquiry." In *2010 IEEE Symposium on Visual Analytics Science and Technology (VAST),* 115 –122. doi: https://doi.org/10.1109/VAST.2010.5652922.

Dill, John, Rae Earnshaw, David Kasik, John Vince und Pak Chung Wong, Hrsg. 2012. *Expanding the Frontiers of Visual Analytics and Visualization*. 2012. London: Springer.

Efer, Thomas, Jens Blecher und Gerhard Heyer. 2012. *Leipziger Rektoratsreden 1871–1933 Insights into Six Decades of Scientific Practice*. In: *International Conference on Historical Corpora*. http://asv.informatik.uni-leipzig.de/publication/file/239/HistCorp2012-EferBlecherHeyer-Rektoratsreden.pdf (letzter Zugriff am 18. Januar 2018).

Elliott, J., E. Atwell und B. Whyte. 2001. *Visualisation of long distance grammatical collocation patterns in language*. In: *Proceedings. Fifth International Conference on Information Visualisation*, 297–302. https://doi.org/10.1109/IV.2001.942073.

Evert, Stefan. 2005. *The Statistics of Word Cooccurrences. Word Pairs and Collocations*. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart. http://dx.doi.org/10.18419/opus-2556.

Finnegan, Ruth. 1992. *Oral Poetry*. Bloomington, Ind.: Indiana University Press

Fisher, Ronald A. 1950. *Statistical Methods for Research Workers*. 11. ed. London: Oliver and Boyd

Fox, Anthony. 1995. *Linguistic Reconstruction: An Introduction to Theory and Method.* Oxford: Oxford University Press.

Friendly, Michael. 2005. „Milestones in the History of Data Visualization: A Case Study in Statistical Historiography." In *Classification: The Ubiquitous Challenge,* herausgegeben von Claus Weihs und Wolfgang Gaul. New York: Springer, 34–52.

Gambette, Philippe und Jean Veronis. 2010. „Visualising a Text with a Tree Cloud." In: IFCS'09: International Federation of Classification Societies Conference, March 2009, Dresden, Germany. https://hal-lirmm.ccsd.cnrs.fr/lirmm-00373643v2 (letzter Zugriff am 28. November 2017).

Gregory, Ian N. und Andrew Hardie. 2011. „Visual GISting: bringing together corpus linguistics and Geographical Information Systems. " *Literary and Linguistic Computing* 26 (3): 297–314. https://doi.org/10.1093/llc/fqr022.

Gries, Stefan Thomas. 2010a. „Corpus linguistics and theoretical linguistics: A lovehate Relationship?" Not necessarily ... *International Journal of Corpus Linguistics* 15 (17): 327–343.

Gries, Stefan Thomas. 2008a. „Dispersions and adjusted frequencies in corpora." *International Journal of Corpus Linguistics* 13 (35): 403–437.

Gries, Stefan Thomas. 2009a. „Dispersions and adjusted frequencies in corpora: further explorations." *Language and Computers* 71 (1): 197–212.

Gries, Stefan Thomas. 2005. „Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff." *Corpus Linguistics and Linguistic Theory.* 1 (2): 277–294.

Gries, Stefan Thomas. 2009b. *Quantitative corpus linguistics with R: a practical introduction.* New York: Routledge.

Gries, Stefan Thomas. 2008b. *Statistik für Sprachwissenschaftler.* Göttingen: Vandenhoeck & Ruprecht (Studienbücher zur Linguistik).

Gries, Stefan Thomas. 2010b. „Useful statistics for corpus linguistics." In: *A mosaic of corpus linguistics Selected approaches,* herausgegeben von Aquilino Sánchez und Moisés Almela. Frankfurt am Main: Lang, 269–291.

Hao, M.C., M. Marwah, H. Janetzko u. a. 2010. „Visual analysis of frequent patterns in large time series". In *2010 IEEE Symposium on Visual Analytics Science and Technology (VAST),* 227 –228, https://doi.org/10.1109/VAST.2010.5650766.

Harleman Stewart, Ann. 1976. *Graphic representation of models in linguistic theory.* Bloomington: Indiana University Press.

Hearst, M.A., Rosner, D. 2008. Tag Clouds: „Data Analysis Tool or Social Signaller?" In: *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual,* 160–160, doi: https://doi.org/10.1109/HICSS.2008.422.

Hearst, Marti. 2009. *Search user interfaces.* Cambridge: Cambridge University Press.

Heringer, Hans Jürgen. 1993. „Basic Ideas and the Classical Model." In *Syntax*. Berlin: de Gruyter, 298–316 (Handbücher zur Sprach- und Kommunikationswissenschaft 9, 1).

Hilpert, Martin. 2011. „Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora." *International Journal of Corpus Linguistics*. 16 (4), 435–461, doi: https://doi.org/10.1075/ijcl.16.4.01hil.

Hilpert, Martin und Stefan Thomas Gries. 2009. „Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition." *Literary and Linguistic Computing*. 24 (4), 385–401, doi: https://doi.org/10.1093/llc/fqn012.

Jäger, Gerhard. 2014. „Lexikostatistik 2.0." In *Sprachverfall? Dynamik – Wandel – Variation*. Berlin: de Gruyter (Jahrbuch 2013).

Jannidis, Fotis. 1999. „Was ist Computerphilologie?" In *Jahrbuch für Computerphilologie* (1): 39–60.

Jannidis, Fotis, Hubertus Kohle, Malte Rehbein (Hrsg.). 2017. *Digital Humanities: Eine Einführung*. Stuttgart: Metzler.

Jockers, Mathew L. 2013. *Macroanalysis: Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.

Keim, Daniel A., Jörn Kohlhammer, Geoffrey Ellis und Florian Mansmann. 2010. *Mastering the Information Age - Solving Problems with Visual Analytics*. Goslar: Eurographics Association. http://www.vismaster.eu/wp-content/uploads/2010/11/VisMaster-book-lowres.pdf.

Kilgarriff, Adam. 2005. „Language is Never, Ever, Ever, Random." *Corpus Linguistics and Linguistic Theory*. 1 (2): 263–276.

Krämer, Sybille. 2016. *Figuration, Anschauung, Erkenntnis: Grundlinien einer Diagrammatologie*. Berlin: Suhrkamp.

Kupietz, Marc und Holger Keibel. 2009. „Gebrauchsbasierte Grammatik: Statistische Regelhaftigkeit." In: *Deutsche Grammatik – Regeln, Normen, Sprachgebrauch*, herausgegeben von Marek Konopka und Bruno Strecker. Berlin: de Gruyter, 33–50.

Lauer, Gerhard. 2011. „Bibliothek aus Daten." In *Die digitale Bibliothek*, herausgegeben von Christine Haug und Vincent Kaufmann. Wiesbaden: Harrassowitz, 79–86 (Kodex. Jahrbuch der Internationalen Buchwissenschaftlichen Gesellschaft 1).

Leblanc, Jean-Marc und Marie Pérès. 2010. „Visualiser les données textuelles : Propositions de fonctionnalités pour une modélisation tridimensionnelle du discours constructeur d'espaces." *Transeo Review* 2 (3): 16-25. https://halshs.archives-ouvertes.fr/halshs-01147433 (letzter Zugriff am 27. November 2017).

Liebsch, Dimitri, Mößner, Nikola. 2012. *Visualisierung und Erkenntnis. Bildverstehen und Bildverwenden in Natur- und Geisteswissenschaften*. Köln: Herbert von Halem.

Löffler, Heinrich. 1994. *Germanistische Soziolinguistik*. Berlin: E. Schmidt.

Lord, Albert Bates. 1960. *The Singer of Tales*. Cambridge, Mass.: Harvard University Press

Luo, Dongning, Jing Yang, Milos Krstajic, William Ribarsky und Daniel Keim. 2012. "EventRiver: Visually Exploring Text Collections with Temporal References." *IEEE Transactions on Visualization and Computer Graphics* 18 (1): 93–105, doi: https://doi.org/10.1109/TVCG.2010.225.

Manning, Christopher D. und Hinrich Schütze. 2002. *Foundations of Statistical Natural Language Processing*. 5. Aufl. Cambridge, Massachusetts: The MIT Press.

May, Thorsten, James Davey und Jörn Kohlhammer. 2010. "Combining statistical independence testing, visual attribute selection and automated analysis to find relevant attributes for classification." *2010 IEEE Symposium on Visual Analytics Science and Technology (VAST)*. 239–240, doi: https://doi.org/10.1109/VAST.2010.5654445.

Mazza, Riccardo. 2009. *Introduction to Information Visualization*. London: Springer.

Mikuš, Radivoj Francis. 1952. «Quelle est en fin de compte la structure-type du language." *Lingua* (3): 430–470.

Moretti, Franco. 2000. "Conjectures on World Literature." *New Left Review* (1): 54–68.

Moretti, Franco. 2009. *Kurven, Karten, Stammbäume. Abstrakte Modelle für die Literaturgeschichte*. Frankfurt am Main: Suhrkamp (edition suhrkamp).

Naumann, Carl Ludwig. 1982. „Kartographische Datendarstellung." In: *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung*. Berlin: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft), 667–692.

Oelke, Daniela, Peter Bak, Daniel A. Keim, Mark Last und Guy Danon 2008. „Visual evaluation of text features for document summarization and analysis." In: *2008 IEEE Symposium on Visual Analytics Science and Technology*. 75–82. https://doi.org/10.1109/VAST.2008.4677359.

Daniela Oelke, Ann-Marie Eklund, Svetoslav Marinov und Dimitrios Kokkinakis. 2012a. "Visual Analytics and the Language of Web Query Logs – A Terminology Perspective." In *Proceedings of the 15th EURALEX International Congress*, 541–548.

Daniela Oelke; David Spretke; Andreas Stoffel; Daniel A. Keim. 2012b. „Visual Readability Analysis: How to Make Your Writings Easier to Read." In *IEEE Transactions on Visualization and Computer Graphics* 18 (5): 662–674.

Parry, Milman. 1971. *The making of Homeric verse: the collected papers of Milman Parry*. Oxford: Clarendon Press.

Redder, Angelika. 2001. „Aufbau und Gestaltung von Transkriptionssystemen." In: *Text- und Gesprächslinguistik / Linguistics of Text and Conversation*. Berlin: de Gruyter, 1038–1059 (Handbücher zur Sprach- und Kommunikationswissenschaft 16, 2).

Reetz, Henning. 2003. *Artikulatorische und akustische Phonetik*. Trier: WVT, Wiss. Verlag.

Reichert, André. 2013. *Diagrammatik des Denkens Descartes und Deleuze*. Berlin: de Gruyter.

Rietveld, Toni und Roeland van Hout. 2005. *Statistics in Language Research: Analysis of Variance*. Berlin: de Gruyter.

Risch, John, Anne Kao, Stephen Poteet und Y.-J. Jason Wu. 2008. "Text Visualization for Visual Text Analytics." In *Visual Data Mining,* herausgegeben von Simeon J. Simoff, Michael H. Böhlen und Arturas Mazeika. Berlin: Springer (Lecture Notes in Computer Science), 154–171.

Rockwell, G., J. Bradley und P. Monger. 1999. "Seeing the text through the trees: visualization and interactivity in text applications." *Literary and Linguistic Computing*. 14 (1): 115–130. https://doi.org/10.1093/llc/14.1.115.

Rohrdantz, Christian, Ming C. Hao, Umeshwar Dayal, Lars-Erik Haug, Daniel A. Keim. 2012a. "Feature-based Visual Sentiment Analysis of Text Document Streams." In *ACM Transactions on Intelligent Systems and Technology, Special Issue on Intelligent Visual Interfaces for Text Analysis* 3, issue 2, article 26.

Rohrdantz, Christian, Michael Hund, Thomas Mayer, Bernhard Wälchli und Daniel A. Keim. 12012b. "The World's Languages Explorer: Visual Analysis of Language Features in Genealogical and Areal Contexts." *Computer Graphics Forum*, 31, issue 3pt1: 935–944, doi: 10.1111/j.1467-8659.2012.03086.x.

Rohrdantz, Christian, Steffen Koch, Charles Jochim, Gerhard Heyer, Gerik Scheuermann, Thomas Ertl, Hinrich Schütze und Daniel A. Keim. 2010. "Visuelle Textanalyse." *Informatik-Spektrum* 33 (6): 601–611, https://doi.org/10.1007/s00287-010-0483-x.

Rohrdantz, Christian, Daniela Oelka, Miloš Krstajić und Fabian Fischer. 2011. „Real-Time Visualization of Streaming Text Data: Tasks and Challenges." In *IEEE Workshop on Interactive Visual Text Analytics for Decision Making*. Providence, RI. http://vialab.science.uoit.ca/textvis2011/papers/textvis%202011-rohrdantz.pdf.

Sager, Sven F. 2001. „Formen und Probleme der technischen Dokumentation von Gesprächen." In *Text- und Gesprächslinguistik / Linguistics of Text and Conversation*. Berlin: de Gruyter, 1022–1033 (Handbücher zur Sprach- und Kommunikationswissenschaft, 16,2).

Säily, Tanja, Terttu Nevalainen, und Harri Siirtola. 2011. "Variation in noun and pronoun frequencies in a sociohistorical corpus of English." In *Literary and Linguistic Computing*. 26 (2): 167–188, doi: https://doi.org/10.1093/llc/fqr004.

Schleicher, August. 1860. *Die deutsche Sprache*. Stuttgart: Cotta.

Schumann, Heidrun und Wolfgang Müller. 1999. *Visualisierung: Grundlagen und allgemeine Methoden*. Berlin: Springer.

Shi, Lei, Furu Wei, Shixia Liu, Li Tan, Xiaoxiao Lian und Michelle X. Zhou. 2010. "Understanding text corpora with multiple facets." In *2010 IEEE Symposium on Visual Analytics Science and Technology (VAST),* 99 –106, doi: https://doi.org/10.1109/VAST.2010.5652931.

Siegel, Steffen. 2009. *Tabula: Figuren der Ordnung um 1600*. Berlin: Akademie-Verlag.

Stetter, Christian. 2005. *Bild, Diagramm, Schrift*. In *Schrift. Kulturtechnik zwischen Auge, Hand und Maschine*. München: Fink (Kulturtechnik).

Stjernfelt, Frederik. 2007. *Diagrammatology: an investigation on the borderlines of phenomenology, ontology, and semiotics*. Dordrecht: Springer.

Sutrop, Urmas. 2012. „Estonian Traces in the Tree of Life Concept and in the Language Family Tree Theory." *ESUKA – JEFUL*. 1 (3): 297–326.

Tamassia, Roberto, Hrsg. 2013. *Handbook of Graph Drawing and Visualization*. Boca Raton: CRC Press.

Thomas, James J. und Kristin A. Cook, Hrsg. 2005. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Los Alamitos, CA: IEEE Computer Society.

Thümmel, Wolfgang. 1993a. „Der europäische Strukturalismus." In *Syntax*. Berlin: de Gruyter, 257–280 (Handbücher zur Sprach- und Kommunikationswissenschaft 9,1).

Thümmel, Wolfgang. 1993b. „Geschichte der Syntaxforschung." In: *Syntax*. Berlin: de Gruyter, 130–199 (Handbücher zur Sprach- und Kommunikationswissenschaft 9,1).

Tufte, Edward R. 1983. *The visual display of quantitative information*. Cheshire, Conn.: Graphics Press.

Tufte, Edward R. 1997. *Visual explanations: images and quantities, evidence and narrative*. Cheshire, Conn.: Graphics Press.

Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley (Addison Wesley Series in Behavioral Science. Quantitative Methods).

Unwin, Antony, Martin Theus und Heike Hofmann. 2006. *Graphics of Large Datasets. Visualizing a Million*. Berlin: Springer (Statistics and Computing).

Vriend, Folkert de, Lou Boves und Roeland van Hout. 2011. "Visualization as a research tool for dialect geography using a geo-browser." *Literary and Linguistic Computing* 26 (1): 17–34, doi: https://doi.org/10.1093/llc/fqq027.

Wattenberg, M. und F. B. Viegas. 2008. „The Word Tree, an Interactive Visual Concordance." In: *IEEE Transactions on Visualization and Computer Graphics* 14 (6): 1221 –1228, https://doi.org/10.1109/TVCG.2008.172.

Wattenberg, Martin. 2005. *Baby Names, Visualization, and Social Data Analysis.* In: *INFOVIS '05 Proceedings of the 2005 IEEE Symposium on Information Visualization,* October 23–25, 2005. https://doi.org/10.1109/INFVIS.2005.1532122.

Zastrow, Thomas. 2011. „Neue Analyse- und Visualisierungsmethoden in der Dialektometrie." Dissertation, Universität Tübingen.

Zhang, Jin. 2008. *Visualization for Information Retrieval.* Berlin: Springer.

I. Visual Linguistics

Noah Bubenhofer

Visual Linguistics: Plädoyer für ein neues Forschungsfeld

Abstract Diagramme spielen auch in der Linguistik eine große Rolle. Ob der Verständlichkeit, mit der Diagramme erstellt und verwendet werden, geht die Reflexion über die diagrammatische Praxis manchmal verloren. Der folgende Beitrag ist ein Plädoyer, diese Praxis aus drei unterschiedlichen Perspektiven zu befragen: Aus diagrammatischer, algorithmischer und wissensgeschichtlicher Perspektive. Dieses Programm einer „Visual Linguistics" stellt Fragen nach dem Charakter von Diagrammen, dem Status von Diagrammen in Forschungsprozessen und insbesondere dazu, welchen Einfluss Digitalität auf die Visualisierung sprachlicher Phänomene ausübt. Schließlich kann mit Ludwik Fleck die diagrammatische Praxis in Beziehung zu wissenschaftlichen Denkstilen gesetzt werden. Vor dem Hintergrund dieser Überlegungen ergeben sich fünf diagrammatische Grundformen, die bei der Visualisierung von sprachlichen Daten eine wichtige Rolle spielen: Liste, Karte, Partitur, Vektoren, Graph/Netz. Listen und Partituren werden im vorliegenden Beitrag ausführlich diskutiert und es wird gezeigt, welche Rolle sie bei der Gegenstandskonstitution in der Linguistik haben.

1. Einleitung

Dialektkarten, Syntaxbäume, Kollokationsgraphen, Kommunikationsmodelle, Gesprächstranskripte, geclusterte Netzwerke: Die Linguistik ist geprägt von Diagrammen verschiedenster Art, um theoretische Modelle zu visualisieren, Daten zu explorieren oder Ergebnisse zu veranschaulichen. Damit nimmt die Disziplin keine Sonderstellung ein gegenüber anderen Disziplinen. Wissenschaftliche Visualisierungen sind überall ein wichtiges Mittel mit vielfältigen Funktionen.

Der Umgang mit Visualisierungen in der Linguistik ist aber, um es positiv zu formulieren, unbeschwert und vorwiegend kanonisch. Unbeschwert, weil weitgehend eine Reflexion darüber fehlt, welche semiotischen und hermeneutischen Eigenschaften und Effekte und welche wissenschaftsgeschichtlichen Implikationen Visualisierungen haben. Kanonisch, weil Visualisierungen in der Linguistik

Erschienen in: Noah Bubenhofer u. Marc Kupietz (Hrsg.): Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

mehrheitlich als Mittel zum Zweck angesehen werden und deren Gebrauch deswegen bewährten Praktiken folgt, das Experiment mit alternativen Visualisierungen aber gescheut wird.

Im Folgenden möchte ich für ein Programm plädieren, das zwar der Visualisierungspraxis in der Linguistik die Unbeschwertheit nimmt, dafür aber Erkenntnisse bietet, die für das Selbstverständnis des Fachs von Bedeutung sind, die Gründe für die Wirkmächtigkeit von Visualisierungskanons untersucht und zum Experiment anstiftet. Dieses Programm nenne ich „Visual Linguistics".

Die Eckpunkte des Programms ergeben sich aus drei Perspektiven, die mir fruchtbar für eine Analyse von Visualisierungspraktiken erscheinen und die ich im Folgenden näher beschreiben möchte:

- Die diagrammatische Perspektive: Was macht ein Diagramm zum Diagramm und welchen Status hat es in Forschungsprozessen?
- Die algorithmische Perspektive: Was ändert sich bei der Visualisierung von sprachlichen Phänomenen, wenn die Daten digital vorliegen und Visualisierungen computergeneriert sind?
- Die wissenschaftsgeschichtliche Perspektive: Warum ist das Diagramm als Drittes zwischen Sprache und wissenschaftlichem Arbeitsinstrument so wirkmächtig, dass es nicht nur Denkstile repräsentiert, sondern diese auch prägt?

Um es deutlich zu sagen: Für die jeweiligen Perspektiven gibt es einige Vorarbeiten aus der Philosophie und Semiotik (Diagrammatik), den Digital Humanities (Software Studies, Critical Code Studies), Medienwissenschaften (Computer als Medium) und der Wissenssoziologie und Wissenschaftsgeschichte. Die Verbindung dieser Erkenntnisse und die Anwendung auf die Linguistik ist jedoch nach wie vor ein Desiderat. Zudem ergaben sich innerhalb weniger Jahre und ergeben sich auch weiterhin durch die Masse digitaler Daten für die Geistes- und Kulturwissenschaften entscheidende Veränderungen, die mitbedacht werden müssen.

2. Die diagrammatische Perspektive

Wissenschaftliche Visualisierungen sind Diagramme, die eine dritte Position zwischen Bild und Text einnehmen, einer „Schriftbildlichkeit" (Krämer 2012a) angehören. Sie sind ein Ensemble von grafischen Ausdrucksmitteln, die im Verhältnis einer „entworfenen Ähnlichkeit" (Bauer und Ernst 2010: 18) zum Gemeinten stehen.

> „Diagramme sind, so könnte man vielleicht sagen, graphische Abkürzungsverfahren für komplexe Schematisierungen. Sie bewahren ein Minimum ästhetischer Anschauung, das wir benötigen, um zu verstehen, wovon die Rede ist, vor allen Dingen, um uns von abstrakten Sachverhalten in buchstäblichem Sinn ein Bild machen zu können." (Stetter 2005: 125)

Entscheidend ist bei einem Diagramm die Typenbildung (vgl. Abb. 1). Durch sie unterscheidet sich das Diagramm vom Bild (Stetter 2005: 125).

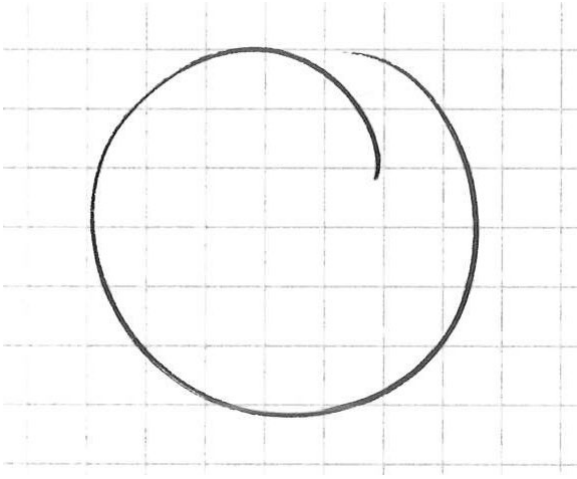


Abb 1: Ein Kreis als Diagramm.

Wenn diese Form als Realisierung eines abstrakten Typs, nämlich eines Kreises, wahrgenommen wird, ist die Form ein Diagramm. Wir abstrahieren von der tatsächlichen Form und sehen darin die idealisierte Form des Kreises. Die grafische Form steht in einem ikonischen Verhältnis zum Denotat: „Die ikonischen Zeichen geben einige Bedingungen der Wahrnehmung des Gegenstandes wieder" (Eco 2002: 205); die Visualisierung, das Diagramm beschränkt sich also auf bestimmte Aspekte des Denotats. Diese Beschränkung ist allerdings willkürlich und deswegen konventionalisiert; die Wiedergabe funktioniert also „erst, nachdem diese [Bedingungen] auf Grund von Erkennungscodes selektioniert und auf Grund von graphischen Konventionen erläutert worden sind" (Eco 2002: 205).

Diagramme werden mit Peirce (1994) gesprochen als „Sinzeichen" wahrgenommen, als „Verwirklichung eines abstrakten Modells" (Eco 1977: 58). Ihre grafischen Variationsmöglichkeiten wie Strichdicke, Färbung, leichte

Formabweichungen sind nicht bedeutungstragend,[1] im Unterschied beispielsweise zu Bildern, bei denen solche „Qualizeichenmerkmale" bedeutungstragend sind (Peirce 1994: 2.244; Eco 1977: 58–59). Sie entwerfen eine „proportionale Homologie", bei der sie „eine Logik darstellen, die vom selben Gesetz beherrscht wird wie die Diagramme" (Eco 1977: 143). Das wird deutlich, wenn beispielsweise ein Balkendiagramm betrachtet wird: Die Länge der Balken steht keineswegs in einem Ähnlichkeitsverhältnis zum Denotat, also einer Menge gezählter Entitäten, sondern die Balkenlängen stellen untereinander Proportionen dar, die homolog sind zu den denotierten abstrakten Zahlenverhältnissen. Der Zauber des Diagramms liegt darin, dass mit diesem grafischen Modus, mit dem die Balken gezeichnet werden können, beliebige Zahlenproportionen abgelesen werden können. Deswegen steht das Diagramm in einem Ähnlichkeitsverhältnis zum abstrakten Denotat der Zahlenverhältnisse.

Damit wird ein weiterer Aspekt deutlich, nämlich dass mit Diagrammen operiert werden kann (Krämer 2009): Besonders anschaulich dafür wird dies bei Karten. Auf ihnen wird mittels grafischer Elemente ein Schema eines bestimmten Ausschnittes des Realraums als Virtualraum dargestellt. Mit diesem Diagramm kann nun operiert werden, sobald die eigene Position und Ausrichtung darauf lokalisierbar ist. Die Karte macht eine Voraussage darüber, wo man einen Weg, ein Hindernis usw. antreffen wird, wenn man sich in eine bestimmte Richtung bewegt. Durch das Operieren mit der Karte ist es möglich, einen Weg zu finden (Krämer 2012b). Genauso ist das Sehen von Zahlenrelationen im Balkendiagramm eine solche Operation.

Besonders wichtig sind aber Operationen in Diagrammen bei Visualisierungen explorativer Datenanalysen (Chen u. a. 2008; Keim u. a. 2010): In solchen Visualisierungen wird bereits verfügbares Wissen diagrammatisch dargestellt, allerdings in einer für Operationen mit dem Diagramm optimierten Form. Durch die Arbeit mit dem Diagramm werden bestimmte Zusammenhänge erst sichtbar und dadurch neues Wissen aus dem Diagramm gezogen.

Obwohl solche explorativen Visualisierungen in neuerer Zeit etwa im Paradigma der Visual Analytics für Data-Mining-Aufgaben besonders wichtig sind, muss betont werden, dass dieses operative Element grundsätzlich in allen Diagrammen vorhanden ist (allerdings mit mehr oder weniger starken Ausprägung)

1 Selbstverständlich können in Diagrammen Strichdicken, Färbungen etc. sehr wohl bedeutungstragend sein. Es gibt aber immer ein Maß an Abweichung davon, das nicht mehr bedeutungstragend ist: Bei einem handgezeichnetem Diagramm, bei dem die Farben Rot und Blau zur Markierung von Klassen o. Ä. eingesetzt werden, werden alle ähnlichen Farbtönungen, die beispielsweise durch unterschiedlichen Druck des Stifts beim Zeichnen entstanden sind, unter der gleichen Farbe subsumiert.

und vor allem kein neues Phänomen ist. Ein Beispiel dafür ist eine Passage aus Platons *Menon*, wo die Erkenntnisfunktion des Diagramms deutlich wird (Krämer 2009): Die Aufgabe herauszufinden, welche Operationen nötig sind, um ein neues Quadrat mit der doppelten Fläche des ursprünglichen Quadrates zu erhalten, wird allein diagrammatisch durch Zeichnen gelöst. Durch die erste, fehlerhafte, Operation, nämlich das Verdoppeln der Seitenlängen, erwächst aus der so erweiterten Zeichnung die Erkenntnis, dass damit nicht die doppelte, sondern vierfache Fläche erzeugt worden ist. Der entscheidende Schritt ist nun zu sehen, dass die Hälfte dieser vervierfachten Fläche reichen würde und dass dies durch das Halbieren der vier entstandenen Quadrate mit einer Diagonale erreicht werden kann. Nur im Diagramm und nur durch die Operation damit kann diese Tatsache gesehen werden:

> „Wir sehen *in* der diagrammatischen Figur ein mathematisches Konzept; und wir sehen in den Transformationen der Figur den *verallgemeinerbaren Lösungsweg* eines generellen Problems, d. h. also eine mathematische Einsicht." (Krämer 2009)

Darin liegt die Hoffnung jeglicher Visual Analytics, also der explorativen, visuellen Datenanalysen, verborgen: Durch die geschickte Überführung von Wissen in ein Diagramm damit operieren und eine verallgemeinerbare Erkenntnis daraus ziehen zu können, weil das Diagramm eben in einem schematisierten, ikonischen Verhältnis zu den Daten steht. Dies bedeutet nicht, dass die diagrammatische Operation immer ausreicht, um Erkenntnisse zu generieren – oft werden aus solchen Operationen Hypothesen generiert, die danach wiederum mit anderen Methoden getestet werden müssen.

In den Beiträgen in diesem Buch finden sich einige Beispiele für visuelle Analysemethoden. Ein Beispiel ist der „Topic Explorer" von Hinneburg und Oberländer (S. 269), bei dem eine komplexe Darstellung von Topic Models gerechnet auf einem beliebigen Textkorpus die Interaktion mit ebendiesem Modell erlaubt. Deutlich wird bei diesem Beispiel aber auch, dass sich das Operieren mit solchen visuellen Analyseinstrumenten selten auf rein diagrammatische Aspekte beschränkt, sondern damit auch die statistische Modellierung davor beeinflusst wird. Hinneburg und Oberländer erhoffen sich somit nicht nur, durch das Operieren neue Erkenntnisse aus den Daten ziehen zu können, sondern damit auch eine „Plausibilitätsstruktur zur Interpretation der Themen" (S. 270) erschließen – also das statistische Modell des (in diesem Fall) Topic Models verstehen zu können „ohne detaillierte Kenntnisse der zugrunde liegenden mathematischen Theorie" (S. 270) besitzen zu müssen.

Ich werde weiter unten deshalb auch argumentieren, dass ein algorithmisch erstelltes Diagramm, um Operationalität zu ermöglichen, mehr sein muss als ein interaktives Diagramm.

Die meisten Diagramme erlauben zumindest eine minimale Form des Operierens. Besonders ausgeprägt ist dies bei den oben bereits erwähnten explorativen Visualisierungen, wie sie in der datenintensiven Geistes- und Sozialwissenschaften zur Anwendung kommen. Neben diesen explorativen Visualisierungen gibt es aber auch solche, die primär der Präsentation von bereits gewonnenen Erkenntnissen dienen. Solche Präsentationsgrafiken (Chen u. a. 2008: 5) sollen eine Erkenntnis klar und deutlich visualisieren. In der Varietätenlinguistik dienen Karten mit eingezeichneten Isoglossen diesem Zweck: Die zugrunde liegenden Daten, die zur Verortung der Isoglossen geführt haben, sind ggf. nicht mehr sichtbar. Eine Varietätenkarte aber, auf der viele erhobene Merkmalsausprägungen visualisiert werden, dient eher der Exploration; im besten Fall können durch die Arbeit mit der Karte Dialekträume bestimmt werden.

Schließlich dienen Diagramme aber oft auch dazu, mehr oder weniger komplexe Modelle zu skizzieren. Marc Richard Lauersdorf verweist in diesem Band auf eine Reihe von Diagrammen, die in der Linguistik dazu dienten, Theorien zu verdeutlichen und die zu „ikonischen" Visualisierungen geworden sind (S. 91). Der Beitrag von Jana Pflaeging in diesem Band (S. 123) zeigt Beispiele dafür, wie Bildmetaphern verwendet werden können, um theoretische Konzepte zu verdeutlichen – die, so Pflaeging, zu einem „lange verweilenden Blick" und „wilder Semiose" führen können. Die grundsätzliche Offenheit gegenüber unterschiedlichen Interpretationen von Diagrammen betont auch Lauersdorf und konzipiert Diagramme deswegen als *objets d'art* (S. 93). In der Linguistik haben Visualisierungen von Theorien eine lange Tradition, denkt man etwa an die verschiedenen Darstellungen des Zeichenkonzeptes bei Saussure („Ei"), Bühler, Morris etc., die als Abbreviationen der jeweiligen Konzepte zum Fachwissen gehören.

Fast alle Visualisierungen in diesem Band sind algorithmisch erstellt und dienen der Exploration von Daten. In der datenintensiven Linguistik und den Digital Humanities ist dies eine Visualisierungsform, die in den letzten Jahren besonders viel Aufmerksamkeit gewinnen konnte. Ich möchte mich deshalb im Folgenden auf diesen Typus fokussieren und als Nächstes diese algorithmische Perspektive einnehmen und ausführlicher diskutieren. Meine Argumentation, insbesondere anschließend bei der wissenschaftsgeschichtlichen Perspektive, beschränkt sich jedoch nicht grundsätzlich auf diesen Typus.

3. Die algorithmische Perspektive

Die Visualisierungen, wie sie in datenintensiven Geisteswissenschaften verwendet werden, sind Kinder einer „Welt der Maschinen" (Moles 1959; zit. nach Hörl 2008: 94), und das in zweifacher Weise: Einerseits modellieren sie auf algorithmische Weise ein Phänomen, simulieren dies also computertechnisch und lassen den Computer „zu einem wesentlichen epistemischen Tool" (Hörl 2008: 97) werden. Andererseits ist bereits das Phänomen selber digital. Auch wenn die digitalen Daten auf nichtdigitale Gegenstände der Welt referieren (Texte, Bilder, Personen), werden sie algorithmisch nur greifbar in der übersetzten, digitalen Form, sie werden in den „Bestand des Technischen" integriert (Rheinberger 1994: 409).

3.1 Computer als Metamedium

Die oben genannten methodischen Paradigmen – datengeleitete Analysen, Visual Analytics – sind möglich dank bestimmter technischer Voraussetzungen, nämlich Mitteln der Informatik. Zum einen benötigt man Rechner, also Hardware, um Daten digital verarbeiten zu können, zum anderen Software. Selbst wenn eine Visualisierung händisch skizziert wird, wird sie in fast allen Fällen für die Publikation digital weiterverarbeitet, nachdem sie gescannt worden ist. In den meisten Fällen entstehen Visualisierungen jedoch bereits im digitalen Raum – und meist sogar algorithmisch.

Für die Erstellung von wissenschaftlichen Visualisierungen hat sich eine Praxis entwickelt, in der bestimmte Programme dafür verwendet werden. Besonders weit verbreitet beispielsweise sind sog. Office-Programme wie Microsoft Excel oder OpenOffice Calc.[2] Neben Software, die leicht über eine grafische Benutzeroberfläche bedient kann, werden jedoch auch Programmiersprachen wie „R" oder „JavaScript" (und viele weitere) verwendet, insbesondere um algorithmische Visualisierungen zu erzeugen (siehe dazu den Beitrag von Wolfer / Hansen-Morath, S. 227). Eine Programmiersprache ist dabei ein Mittel, um sehr flexibel und genau Anweisungen an einen Rechner zu formulieren, bestimmte Operationen durchzuführen (vgl. dazu die gut lesbare Einführung von Ford 2015). Programmiersprachen gibt es auf unterschiedlichen Abstraktionsebenen; eine sog. höhere Programmiersprache erlaubt beispielsweise die Formulierung von

2 Interessanterweise entstehen Visualisierungen auch mit Software, die nicht dafür gedacht ist, etwa mit Microsoft PowerPoint, das für die Präsentation von Folien, nicht aber die Erstellung von druckfähigen Grafiken entwickelt wurde. Daran zeigt sich die Wirkmächtigkeit von Praktiken, die weit über die intendierten Verwendungsweisen eines Werkzeugs hinausgehen können.

Anweisungen, die sich z. B. so paraphrasieren lassen: *Mache für jedes Wertepaar der Spalten 3 und 4 im Datensatz einen Punkt, wobei der Wert der Spalte 3 der Position auf der x- und der Spalte 4 der y-Achse entspricht.* Dabei kann die Art des Punktes von den Daten abhängig gemacht werden oder Interaktionen definiert werden, etwa: *Wenn der User mit der Maus über einen Punkt fährt, zeigte den Wert aus Spalte 1 der entsprechenden Zeile des Datensatzes an derselben Position an.*

Bei der Ausführung des Codes wird die hochsprachliche Anweisung letztlich in sog. Maschinencode überführt. Dies sind einfache Anweisungen an den Mikroprozessor, bestimmte Speicherbereiche mit bestimmten Werten zu belegen.

Nun lohnt es sich allerdings, genauer über die Funktionsweise eines Computers nachzudenken, und zwar aus linguistischer Perspektive. Erhellend ist dabei die Geschichte, die Heilmann (2012) erzählt: Die Geschichte des Computers als Schreibmaschine. Es gibt drei Typen des Schreibens im Zusammenhang mit Computern: „das Schreiben für Computer (aber nicht an ihnen), das Schreiben für und an Computern, und das Schreiben an Computern (aber nicht für sie)" (Heilmann 2012: 8). Diese drei Typen stehen für drei unterschiedliche Epochen der Computergeschichte und drei unterschiedliche Verhältnisse zwischen Menschen und Computern. Zunächst, in den 1940er-Jahren, dienten Computer als Automaten. Ihnen wurde ein Set von Anweisungen eingeschrieben, das sie für beliebige Eingabewerte abarbeiteten, z. B. um Flugbahnen von Geschossen zu berechnen. Das Regelset wurde dabei handschriftlich auf Formblättern formuliert und dann in Form von Lochkarten eingegeben, in denen die gewünschten Speicherbereiche und Operationen codiert waren. Um diesen beschwerlichen Vorgang zu vereinfachen, fasste man häufig verwendete Anweisungen zu „Subroutinen" zusammen, daraus entwickelten sich dann – verkürzt dargestellt – Programmiersprachen. In den 1960er-Jahren kam es dann zu einem entscheidenden Wandel: Der Anschluss von Schreibmaschinen oder Fernschreibern erlaubte „Interactive Computing": Die Anweisungen konnten sozusagen live während der Ausführung eines Programms abgesetzt und so das Programm beeinflussen. „Diese neue Art des Programmierens erforderte selbst passende Programme (sogenannte Editoren) und es stellte sich schnell heraus, dass Computer nicht nur zum Schreiben von Code taugten, sondern auch dem Verfassen von technischen Berichten, Artikeln und Dokumentationen dienlich sein konnten" (Heilmann 2012: 8).

Damit wurde Text offensichtlicher Gegenstand des Computers – textuell funktionierte er jedoch schon von Anfang an: *Codieren* bedeutet nämlich „anschreiben" von Werten in Form eines Systems diskreter Zeichen als Verweise auf Speicherorte. Und *kalkulieren* bedeutet schriftliches Operieren mit diesen Zeichen, genau so, wie wir beispielsweise durch schriftliche Operationen auf Papier eine Multiplikation durch geschicktes, regelhaftes Schreiben durchführen können (Heilmann 2012: 42 in Rückführung auf Kittler 1989; Turing 1936). Der Computer diente also bereits von Anbeginn der Verarbeitung von Text

(Programmcode, Quelltext), operierte schriftlich und gab Text aus. Die schriftliche Angabe der Zahl 452 in Form einer Lochkartenanweisung oder des in einer höheren Programmiersprache definierten Ausdrucks „x = 452“ ist für den Computer keine Zahl, sondern sind ein bestimmtes Set von Schaltzuständen von Transistoren. Der Ausdruck „x = 452“ wird semantisch entleert und rein *syntaktisch* übersetzt in Schaltzustände (Heilmann 2012: 65) und unterscheidet sich nicht grundsätzlich von der Anweisung „v = Haus“ – beides sind Schaltzustände, mit denen, ausgelöst durch Text, operiert wird.

Auffallend ist dabei die Ähnlichkeit zum Operieren mit einem Diagramm. Die manuelle Methode des „schriftlichen Rechnens“ zeigt bereits die Bedeutung des Diagrammatischen: Um mehrere Zahlen zu summieren, ordnet man sie als Liste untereinander stellenweise parallel rechtsbündig an, um dann Stelle für Stelle von rechts nach links summieren zu können. Die Positionierung der Zahlen im Raum ist dabei entscheidend und funktioniert, weil die Verschriftlichung der mathematisch abstrakten Zahlen in einer Systematik geschieht, die genau dies erlaubt. Ähnlich geschieht dies nun beim Computer, bei dem die abstrakten Zahlen nicht direkt verschriftlicht, sondern durch Schaltzustände als binäre Zahlen repräsentiert werden, die es aber dank ihrer Anordnung erlauben, genau so Operationen durchzuführen, wie wir das beim schriftlichen Rechnen im Dezimalsystem gewohnt sind. Insofern argumentiere ich, dass auch das „Rechnen“ des Computers einen diagrammatischen Charakter hat, da dieses Rechnen mit Schaltzuständen im Binärsystem eine proportionale Homologie entwirft: Das Setting der Schaltzustände (dessen Materialität bei den ersten Rechnern noch viel deutlicher hervortrat als heute) ist eine Art „Diagramm“, mit dem operiert wird, um im abstrakten mathematischen Raum eine Aufgabe zu lösen. In Anbetracht der Bedeutung des (schriftlichen) Codes, der diese Operation ermöglicht und der unzweifelhaft in seiner schriftlichen Form diagrammatische Mittel der Textformation nutzt (Listen, Gliederungen, Hierarchien, Einrückungen etc., vgl. Steinseifer 2013), könnte der Computer als eine Art „diagrammatische Maschine“ aufgefasst werden.

Hinzu kommt nun jedoch ein weiterer Aspekt: Mit der Wende hin zum Interactive Computing und damit den neuen Eingabe- und Ausgabemedien (Maus, Bildschirm) wird der Computer als eine Art Plattform wahrgenommen, um völlig unterschiedliche Dinge tun zu können. Kay und Goldberg (1977) charakterisieren ihn als „metamedium“:

> „Although digital computers were originally designed to do arithmetic computation, the ability to simulate the details of any descriptive model means that the computer, viewed as a medium itself, can be *all other media* if the embedding and viewing methods are sufficiently well provided.” (Kay und Goldberg 1977)“

Damit können mit diesem Metamedium auch neue Medien erfunden werden, die es bisher gar nicht gab. Manovich ist der Auffassung, dass dies weitreichende Konsequenzen hatte, wie heute Software als Medium funktioniert: „Once computers and programming were democratized enough, some of the most creative people of our time started to focus on creating these new structures and techniques rather than using the existing ones to make 'content'" (Manovich 2014: 81).

Entscheidend dafür ist, dass der Computer als „diagrammatische Maschine" mittels dreier Grundfunktionen – Codierung, Algorithmisierung und Formatierung (Heilmann 2012: 195) – Daten speichert, überträgt und verarbeitet. Die Codierung übersetzt Eingaben in Transistorzustände (ins Digitale), die Algorithmen operieren damit und die Formatierung macht die abstrakten Transistorzustände wieder sicht- und deutbar. Es gibt jedoch unendlich viele verschiedene Formatierungen, keine ist die „eigentliche" Repräsentation der Transistorzustände. Und wieder ist es Text, Programmcode, der die Formatierung der Daten bestimmt – im Falle von Textdaten „ist die Formatierung von Texten eine Sache der Beschreibung von Text *durch* Text. Zum ersten Mal in der Geschichte des Schreibens bestimmen damit Texte über die 'Materialität' der Schrift und nicht umgekehrt die Materialität der Schrift über die Gestalt von Texten" (Heilmann 2012: 239).

Welchen Unterschied macht es, wenn Text über die Materialität der Schrift bestimmt und nicht die Materialität der Schrift (auf Papier gedruckt, in Holz gekerbt, mit Pinsel und Tusche gemalt) über die Gestalt von Texten?

Zunächst ergibt sich ein ungeheures Potenzial an Transformationen von Text in unterschiedliche Formen (vgl. dazu Jägers „transkribierende Verfahrenslogik", Jäger 2007). Bedeutender ist jedoch, dass diese Transformationsprozesse live und rekursiv ablaufen können. Solange der Text den Computer nicht verlässt (z. B. ausgedruckt wird), so lange kann mit Text auf ihn Einfluss ausgeübt werden. Und der Text kann *sich selber* rekursiv beeinflussen.

Die „Visualisierung" von Text (z. B. einer Frequenztabelle) ist also eine textuell definierte Formatierung dieses Textes (in Form von Instruktionen), eine von vielen möglichen Transformationen von Text in eine bestimmte diagrammatische Form. Das Operieren mit diesem Diagramm beschränkt sich deshalb nicht nur darauf, mit dem formatierten Diagramm zu interagieren (gemeinhin *Interaktivität* genannt), sondern umfasst alle Transformationsprozesse, die im Computer möglich sind. Dies im Unterschied etwa zur Skizze auf Papier: Dort kann durch Zeichnen beispielsweise ein geometrischer Beweis geführt werden (vgl. oben Platons Menon-Beispiel), die materiellen Vorbedingungen, Papier, Stift, können jedoch nicht verändert werden. Bei der algorithmischen Visualisierungen jedoch wird selbst die Materialität operationabel – das Diagramm kann verschiedene Zustände der Materialität einnehmen: angezeigt auf einem Bildschirm, zweidimensional, [simuliert] dreidimensional etc. –, denn alles, was mit den digitalen

Daten im Computer passiert, sind diagrammatische Operationen. Der Computer ist eine diagrammatische Maschine.

Wenn aus diagrammatischer Perspektive also deutlich gemacht wird, dass mit Diagrammen operiert werden kann, dann wird aus algorithmischer Perspektive klar, dass diese Operationalität umfassend ist und das Diagramm nicht bloß eines der „Bilder" ist, die durch Transformationen entstanden sind, sondern alle Prozesse davor bereits diagrammatisch sind. Eine interaktive Visualisierung ist dann umfassend interaktiv, also operationabel im diagrammatischen Sinn, wenn sie alle Interaktionen mit den Daten zulässt, also wenn die digitale Aufbereitung und algorithmische Beeinflussung der Daten genauso dazugehören wie die Beeinflussung der visuellen Darstellung. Es leuchtet natürlich ein, dass die Operationalität sinnvollerweise nicht bis auf die Ebene des Maschinencodes reichen soll: Wir möchten uns nicht damit herumschlagen, wie die Zeichen im Computer digital repräsentiert und gegenseitig verrechnet werden. Die Operationen von Interesse finden auf einer viel höheren Ebene statt – es geht um die Frage, welcher Art die Indizes sind, mit denen die Daten gespeichert werden, welche Aspekte der Daten überhaupt gespeichert, welche Transformationen in der Folge möglich sind und in welchen Formen die Daten materialisiert werden. Wenn mit Daten diagrammatisch operiert werden soll, dann sollten diese Aspekte potenziell kontrolliert werden können.

Diese umfassende Kontrolle über das Digitale ist natürlich genau der Witz des Interactive Computings, also des Computers, wie wir ihn heute kennen. Dieser Grundgedanke, mit dem Computer ein Metamedium bedienen zu können, wird allerdings durch Betriebssysteme, die möglichst viele potenzielle Interaktionsmöglichkeiten hinter einer schönen Metapher verstecken, oder durch Software, die besonders benutzerfreundlich gestrickt ist und Interaktionen nur auf der Oberfläche zulässt, sabotiert.

Muss man sich demnach als Geisteswissenschaftlerin oder -wissenschaftler bei der Nutzung von visuellen Analysemethoden mit Algorithmen auseinandersetzen? Unbedingt – und man sollte sich sogar mit Programmierkulturen auseinandersetzen, wie ich im Folgenden zeigen möchte.

3.2 Topoi in der Programmierpraxis

Es ist das Verdienst der Software Studies, auf die kulturelle Verfasstheit von Software aufmerksam gemacht zu haben (Fuller 2003; Mackenzie 2006). Und es ist einsichtig, dass die Gestaltung von Software eine kulturelle Praxis widerspiegelt und gleichzeitig reproduziert: Die grafischen Benutzeroberflächen, die „eine ganze Maschine ihren Benutzern entziehen" (Kittler 1993: 233), ermöglichen (und verunmöglichen) gewisse Praktiken, die Art, wie Wissen als verarbeitbare

Daten repräsentiert wird (Datenbanktypen), ist nicht unabhängig von kulturellen Übereinkünften zu Wissenrepräsentationen (Manovich 2002; Dourish 2014).

Auch auf der Ebene des Programmiercodes begegnen einem Hinweise über die Einbettung in kulturelle Praktiken auf Schritt und Tritt. Die Wahl einer Programmiersprache „is the most important signaling behavior that a technology company can engage in" (Ford 2015) und Ford nennt in seinem Text die überspitzten Klischees, die in der Gemeinde der Programmierer/innen den unterschiedlichen Programmiersprachen zugeschrieben werden:

> „Tell me that you program in Java, and I believe you to be either serious or boring. In Ruby, and you are interested in building things quickly. In Clojure, and I think you are smart but wonder if you ship. In Python, and I trust you implicitly. In PHP, and we sigh together. In C++ or C, and I nod humbly. In C#, and I smile and assume we have nothing in common. In Fortran, and I ask to see your security clearance. These languages *contain entire civilizations.*" (Ford 2015, Hervorhebungen NB)

Larry Wall, bezeichnenderweise Linguist, entwickelte eine der wichtigsten Programmiersprachen: Perl. 1999 hielt er an der Konferenz „LinuxWorld" eine Rede mit dem Titel „Perl, the first postmodern computer language" (Wall 1999). Auf sehr unterhaltsame Weise argumentiert Wall darin, dass Perl – und bis damals nur Perl – eine postmoderne Programmiersprache sei, im Gegensatz zu allen anderen, die in der Moderne stecken geblieben seien. Sie richte sich gegen vier die Informatik beherrschende Kulte: „spareness", „originality", „seriousness" und „objectivity". Perl ist eine quelloffene Sprache, die auf vielen Bruchstücken der Betriebssysteme Unix und Linux beruht und versucht, Lösungsdogmen zu vermeiden und dafür Diversität zuzulassen: „Perl programming is unabashedly genre programming. It has conventions. It has culture. Perl was the first computer language whose culture was designed for diversity right along with the language" (Wall 1999).[3] Walls Vortrag endet mit der These, dass es gerade die

3 Um keine falschen Vorstellungen zu wecken: Trotz der Diversität und dem Perl-Mantra „There's More Than One Way To Do It" handelt es sich um eine rigide Programmiersprache, die auf eine korrekte Verwendung ihrer Syntax angewiesen ist, um zu funktionieren. Im Vergleich zu anderen Programmiersprachen ist sie jedoch tatsächlich in vielen Aspekten wenig rigid; zu nennen ist beispielsweise, dass in Perl bei der Definition von Variablen nicht festgelegt werden muss, welche Inhaltstypen (Ganzzahlen, Kommazahlen, Buchstaben) darin vorkommen dürfen, sondern dies im entsprechenden Kontext automatisch geschieht.

Open-Source-Bewegung sei, die die oben genannten Kulte umstoße und so für einen kulturellen Wandel stehe.

Walls „Kulte", von Fuller (2003: 15) als „idealist tendencies in computing" bezeichnet, drehen sich um einen starken Topos der Informatik: Die Gleichsetzung von Purismus von Zahlen und Schönheit. Die ganze Welt kann in Verhältnisse von Zahlen aufgelöst werden und wird dann zu „purer Mathematik". Je stärker sich diese Repräsentation durch Zahlen der puristischen Form annähern,

> „the more beautiful they become. There is an endpoint to this passage to beauty which is absolute beauty. Access to and understanding of this beauty is allowed only to those souls that are themselves beautiful." (Fuller 2003: 15)

Fuller kritisiert in der Folge Tendenzen, „ästhetisches Programmieren" zu einem Dogma zu erklären, und warnt, dass solche Ästhetiken zu sozialer Kontrolle führen. Die „märchenhaften" [„fabulatory"] Programmieransätze seien weit interessanter, denn:

> „Numbers do not provide big answers, but rather opportunities to explore further manifold and synthetic possibilities–that is to say, they provide access to more figures." (Fuller 2003: 16)

Eng verwandt mit dem Purismus-Topos ist der Utilitarismus-Topos: „Computer programming seems to carry to an extreme an understanding of technology as a utilitarian tool predicated on a reframing of the world as a set of calculable quantities" (Goffey 2014: 21). Dieser Topos ist nicht nur für Computertechnik wirkmächtig, sondern für Technologie generell, wie Böhme (2006) zeigt. Er nennt Karl Marx als prägend für diesen Topos, der wohl als einer der ersten eine „*theory* of technology" entwickelte, „to have a conception of technology as a social enterprise" (Böhme 2006: 55). Die Auffassung, dass Technologie dem „Reich der Notwendigkeit, nicht der Freiheit" angehört und dazu dient, die Lebensbedingungen zu verbessern und somit letztlich der Reproduktion der Menschheit dient, verankerte sich in der Folge fest in unserem Verständnis (Böhme 2006: 55–56). Allerdings entspringt diese Auffassung von Technologie einer vorherrschenden Theorie von Technologie, nicht ihrer tatsächlichen Geschichte (Böhme 2006: 56).

Die Geschichte nämlich brachte immer wieder Beispiele hervor für Technologien, die der Unterhaltung, dem Genuss, der Zerstreuung dienten, unabhängig jeglicher Nützlichkeitsziele. Böhme zeigt dies anhand der Technologien an den königlichen Höfen, Goffey nennt – neben den naheliegenden Computerspielen

– Beispiele aus der Computergeschichte: „The fascination with automata evinced within Department of Defense-funded research in artificial intelligence (AI) could perhaps be adduced as evidence that technologies of enjoyment were alive" (Goffey 2014: 27).

Entscheidend in unserem Kontext ist aber, dass durch das Metamedium Software eine Möglichkeit entstanden ist, mit Programmcode nicht nur Modelle der *Welt*, also „Aneignungen von Natur" (Böhme 2006: 55) als zweckrationale Tools zu entwickeln, sondern Modelle *virtueller* Welten, die selber Produkte des Computers sind: „Strictly speaking, many of the things that software models are not real-world processes at all, they are things that are brought into being by computational technologies themselves" (Goffey 2014: 34). Software weist also eine „demiurgische" Qualität auf, sie kann etwas erzeugen, was vorher noch gar nicht existierte (Goffey 2014: 35). Dabei geht eine Praxis der Programmierung in ein Stück Software auf, der die Praxis der Codeerzeugung nicht mehr anzusehen ist:

> „clever tricks with assembly language, manipulating the side effects of an algorithm, devising workarounds, the smoke and mirrors of user interface design, and so on. The virtuoso smarts of programming practice that get software working can of course eventually be explained logically if the code compiles, the app works and users accept it. But in doing so, the process disappears into the product and the experimenting, the 'bricolage' gets forgotten." (Goffey 2014: 35)

Die semiotische Qualität von Programmiersprache wird damit deutlich: Programmiersprachen sind Sets von Codes, die im Rahmen einer Praxis des Programmierens in einem kulturellen Kontext neue Codes erzeugen, die wiederum als Zeichen gelesen werden. Software wird zu einer Äußerung: „software is a semiotic artefact, a set of operations in and on codes that implies the ongoing, repeated fact of enunciation" (Goffey 2014: 36). Äußerungen sind aber nicht formallogisch vollständig erfassbar; „Enunciation brings the messiness of the world back and forces us to connect coding or programming with practice in more rigorous ways" (Goffey 2014: 37).

3.3 Coding Cultures

Ein anschauliches Beispiel für die Einbettung von Programmierpraktiken in Kulturen bieten Hacker-Szenen. *Hacking* wird in diesem Zusammenhang verstanden als Tätigkeit, mit der nicht nur ein bestimmtes konstruktives Ziel erfüllt

wird, sondern einen Selbstzweck dient – im Fall des Computers beispielsweise das Programmieren des Programmierens willen; aus purer Freude (Levy 2010: 10). Dabei spielen Ideale wie freier Zugang zu Wissen, Redefreiheit, Transparenz, Chancengleichheit, Öffentlichkeit und ein meritokratisches Verständnis eine wichtige Rolle. In der Geschichte des Computerhackings sind diese Werte stark mit einem Engagement für liberale Softwarelizenzen verbunden: Erzeugter Code gehört allen, jede und jeder darf ihn modifizieren (Coleman 2012: 3). Diese Ansicht stellt ein deutliches Gegengewicht zur kommerziellen Softwareindustrie dar; ein Kampf gegen die oben bereits von Wall erwähnten „Kulte".

Im Selbstverständnis der Hacker-Szenen spielen diese Werte, zunächst implizit, manchmal auch explizit ausformuliert in einer „Hacker-Ethik", eine wichtige Rolle:

> „It was a philosophy of sharing, openness, decentralization, and getting your hands on machines at any cost to improve the machines and to improve the world. This Hacker Ethic is their gift to us: something with value even to those of us with no interest at all in computers." (Levy 2010: IX)

Diskussionen über diese ethischen Prinzipien und Formulierungen von Manifesten begleiten die Geschichte des Hackens bis heute und werden teilweise heftiger diskutiert als technische Fragen (Coleman 2012: 18). Im Zeitalter der Start-ups des Silicon Valleys spricht Scott (2015) etwa von der gehackten, nämlich „gentrifizierten" Hacker-Kultur: Der kommerziellen Vereinnahmung dieser Kultur, die damit des rebellischen, kritischen Moments beraubt wird.

Es liegt nahe zu untersuchen, welchen Einfluss Hacker-ethische Überlegungen auf die tatsächliche Praxis des Hackens – angefangen beim Auseinandernehmen und Modifizieren von Hardware über das Verändern bestehender Programme und dem Programmieren neuer Programme bis zum Hacken von Sicherheitssystemen – haben. Levy und Coleman (Coleman 2012; Levy 2010) zeigen dies über unterschiedliche Ansätze; Coleman argumentiert beispielsweise, wie sich in unterschiedlichen Varianten, die gleiche Funktion in der gleichen Programmiersprache zu programmieren, Witz und Können gleichzeitig manifestiert, was natürlich nur für die Eingeweihten überhaupt ersichtlich ist (Coleman 2012: 93ff.). Anders als bei proprietärer Software bleibt bei quelloffenen, geteilten Programmen die gefundene Lösung, die „Bricolage" (vgl. oben; Goffey 2014: 35), sichtbar.

Die Computerwissenschaften waren quelloffenen und freien Software-Prinzipien grundsätzlich immer aufgeschlossen, zumal die ersten Hacker-Bewegungen an Universitäten entstanden sind (Levy 2010). In den Geistes- und

Sozialwissenschaften spielte und spielt jedoch proprietäre Software eine wichtige Rolle, seien es Office-Programme zur Textproduktion, Datenverwaltung und Präsentation oder Statistikprogramme wie SPSS.[4] In jüngerer Zeit und vor allem im Bereich der Digital Humanities, der Korpus- und Computerlinguistik und der elektronischen Datenverarbeitung werden mit Programmiersprachen wie R, Perl, Python, Javascript etc. Techniken angewandt, die zu quelloffener und freier Software führen. Diese Programmiersprachen tragen ihre kulturelle Genese und Bedeutung in sich, auch wenn sie uninformiert und naiv eingesetzt werden – ihre Verwendung enthält eine Botschaft.

Es finden sich im Netz viele Texte, die sich an programmiertechnische Laien richten und argumentieren, warum eine bestimmte Programmiersprache besser ist als eine andere. Mit R lassen sich beispielsweise statistische Berechnungen und Visualisierungen erstellen, was teilweise natürlich auch mit dem viel bekannteren Office-Paket „Excel" von Microsoft geht.[5] Paradigmatisch für die Nennung der Vorteile von R gegenüber Excel ist beispielsweise der Text von Isaac Petersen mit dem Titel „Why R is Better Than Excel for Fantasy Football (and most other) Data Analysis".[6] Neben einigen technischen Argumenten werden auch Aspekte genannt, die eher ideologischer Natur sind, beispielsweise Quelloffenheit, Lauffähigkeit auf vielen Plattformen und die Möglichkeit, selber als Teil einer großen Community sich aktiv an der Weiterentwicklung zu beteiligen. Ähnlich argumentiert ein Blogeintrag von Michael Milton, der zudem ein schönes Beispiel für den oben genannten Purismus-Topos darstellt:

> „The visualizations you can create in R are much more sophisticated and much more nuanced. And, philosophically, you can tell that the visualization tools in R were created by people more interested in good thinking about data than about beautiful presentation. (The result, ironically, is a much more beautiful presentation, IMHO.)"[7]

4 Wobei es immer auch Gegenströmungen gab wie für die Textproduktion zum Beispiel die Verwendung des quelloffenen Satzsystemes LaTeX.

5 Die Möglichkeiten im Bereich statistischer Analyse und Visualisierung, aber auch der Datenaufbereitung und insbesondere der Einbettung in andere Programmierroutinen sind bei R deutlich vielfältiger als bei einer Software wie Microsoft Excel. Trotzdem begnügen sich viele Wissenschaftler/innen mit Excel, ggf. auch in Arbeitsroutinen, die R-Nutzer/innen als sehr unelegant bezeichnen würden.

6 http://fantasyfootballanalytics.net/2014/01/why-r-is-better-than-excel.html (25. August 2015).

7 http://www.michaelmilton.net/2010/01/26/when-to-use-excel-when-to-use-r/ (24. August 2015).

Wenn man sich als Novize oder Novizin auf R einlässt, wird man automatisch seine Praxis ändern, sobald das erste Problem auftaucht und man nach einer Lösung sucht: Die Lösung findet sich nicht in einem singulären Handbuch, sondern in zig Diskussionsschnipseln im Netz. Dabei wird es nicht nur eine Lösung geben, sondern viele verschiedene. Bald wird auch klar werden, wodurch sich die Lösungen unterscheiden: Unterschiedliche Codevarianten mit gleichem Ausgang oder Abhängigkeit von bestimmten Zusatzpaketen mit Variationen der Ausgabe, wobei Argumente wie „Eleganz", „Ästhetik" oder „Sauberkeit" für die eine oder andere Lösung auftauchen werden.

Obwohl der Umgang mit Computern von einem starken Utilitarismus-Topos durchdrungen ist, zeigen die Ausführungen oben, dass Praktiken der Programmierung und Softwarenutzung zutiefst kulturelle Praktiken sind. Die Verwendung von bestimmten Programmiersprachen und Programmen kommt einer kulturell bedeutsamen Botschaft gleich. Es macht einen Unterschied, ob die Visualisierung mit Excel oder R erstellt worden ist; die Entscheidung über das Programm oder die Sprache beeinflusst den Erkenntnisprozess genauso wie die Wahl der Daten, die Formulierung der Forschungsfrage oder die theoretische Grundierung der Analyse.

4. Die wissenschaftsgeschichtliche Perspektive

Eine „visuelle Linguistik" muss sich zwingend mit wissenschaftsgeschichtlichen Theorien auseinandersetzen, um die Rolle der Visualisierungen in der Disziplin synchron als auch diachron zu verstehen. Neben Thomas S. Kuhns Buch *Struktur wissenschaftlicher Revolutionen* von 1962 (Kuhn 1996), das zweifellos eine bedeutende Stellung in der Wissenschaftsgeschichte einnimmt, sind die Arbeiten von Ludwik Fleck besonders gut anschlussfähig an die Desiderate einer visuellen Linguistik. Fleck hat in mehreren Arbeiten, ausgehend vom 1935 erschienenen Text *Entstehung und Entwicklung einer wissenschaftlichen Tatsache* (Fleck 1980), seine Konzepte Denkstil und Denkkollektiv entwickelt und plausibilisiert (Fleck 1983, 2011): Die Angehörigen einer wissenschaftlichen (Teil-)Disziplin eint ein gemeinsamer „Denkstil", der sie zu einem „Denkkollektiv" macht (Fleck 2011: 87). Das erschwert nicht nur das gegenseitige Verständnis über Denkkollektive hinweg, sondern es entsteht auch „eine spezifische Bereitschaft, dem Stil entsprechende Gestalten wahrzunehmen" und es „verschwindet dagegen parallel das Vermögen, nicht stilgemäße Phänomene wahrzunehmen" (Fleck 1983: 107). Ein Denkkollektiv ist also blind gegenüber Evidenzen, die nicht zum eigenen Denkstil passen.

Auf die Kommunikation innerhalb eines Denkkollektivs richtet Fleck ein besonderes Augenmerk. Erkenntnisse und Wissen werden innerhalb des

Kollektivs gemäß dem herrschenden Denkstil kommuniziert und verändern sich dadurch laufend, festigen aber gleichzeitig den Denkstil innerhalb des Kollektivs.

Flecks Beobachtungen zur Wirkmächtigkeit von Denkstilen in der Wissenschaft sind auch in der Linguistik – mit einigen Jahrzehnten Verzögerung – auf fruchtbaren Boden gefallen. Dabei stößt der Stilbegriff auf reges Interesse: Denn „das, was wir mit Denkstil meinen und am Denkstil beobachten, [muss sich] ja immer sprachlich materialisiert haben […], um wahrnehmbar, beobachtbar zu sein" (Fix 2011: 1; Möller 2007). Fleck nennt Mittel, die dazu dienen, einen Denkstil zu pflegen: „Ich möchte nur noch zwei Mittel erwähnen, über die der wissenschaftliche Denkstil verfügt, um seinen Produkten den Charakter einer Sache zu verleihen. Eines von ihnen sind *technische Termini* […]. Das zweite Mittel ist *das wissenschaftliche Gerät* […]. Wer es versteht, in ein Fernrohr zu schauen und an den Saturn zu denken, benutzt damit allein bereits einen bestimmten abgegrenzten Denkstil" (Fleck 1983: 121f.). Ulla Fix hat plausibel gezeigt, wie anschlussfähig Flecks Überlegungen zur Rolle des Sprachgebrauchs zur Ausbildung von Denkstilen an Konzepte der Linguistik sind (Fix 2011). Sie nennt Kategorien wie Stil, Text, Wort, Varietät und Metapher als sprachliche Elemente des Wahrnehmbaren eines Denkstils.

Doch gehören Diagramme nicht ebenfalls zu den Mitteln, die wissenschaftliche Denkstile prägen? Und gehören sie zu den „technischen Termini" oder zum „wissenschaftlichen Gerät"? Ich möchte argumentieren, dass Diagramme eine eigenartige Doppelfunktion zwischen Sprache und Gerät einnehmen – und gerade deshalb besonders wirkmächtig sind.

4.1 Visualisierungen als Zeichen

Zweifellos sind Visualisierungen komplexe Zeichen, die in einem ikonischen Verhältnis zum Referenten stehen. Sie weisen damit semiotische Qualitäten auf und funktionieren in der wissenschaftlichen Praxis genauso als kommunikatives Mittel wie sprachliche Zeichen, können also, z. B. Bühler folgend, darstellen, ausdrücken und appellieren (Bühler 1934). Mit Fleck gedacht können Visualisierungen dank dieser Funktionen innerhalb eines Denkkollektivs „den Produkten den Charakter einer Sache" verleihen. Dies wird sofort evident, wenn man sich vergegenwärtigt, in welchen wissenschaftlichen Disziplinen gewisse Typen von Visualisierungen üblich, zwingend notwendig oder geradezu verpönt sind. Jedes Denkkollektiv pflegt eine Praxis des Visualisierens oder Nicht-Visualisierens, angefangen bei leichten Formen der Visualisierung wie textstrukturierenden Merkmalen (Spiegelpunkte, Listen, Tabellen) über übliche und kanonisierte

Formen (Balkendiagramme, Streudiagramme, Karten etc.) bis hin zu komplexen Formen (interaktive Visualisierungen).[8]

In der Linguistik entwickelten sich in bestimmten Teildisziplinen mehr oder weniger stabile Formen, etwa in der Dialektologie und Varietätenlinguistik, wo Karten eine wichtige Rolle einnehmen und sich bestimmte Praktiken durchgesetzt haben oder zumindest das Ziel der Standardisierung definiert wird. So formuliert etwa Naumann (1982) in der Dialektkartografie das folgende Desiderat:

> „Die *Methodik* wird […] immer deutlicher einen Rahmen für große Teile der Arbeitsschritte geben und damit dem Erfindungsreichtum Grenzen setzen, soweit das die Kommunizierbarkeit von Kartierung und Karte verbessert." (Naumann 1982: 687)

Es werde zwar weiterhin ein „sozusagen künstlerischer Anteil" notwendig sein, wobei Naumann darunter alle Entscheidungen versteht, die nicht einer Systematik folgen (Naumann 1982: 668). Auf das „Künstlerische", offensichtlich ein eigentlich in der Wissenschaft unerwünschter Aspekt, müsste noch zurückgekommen werden (vgl. aber die Beiträge von Pflaeging und Lauersdorf in diesem Band); der Wunsch einer Systematisierung von Visualisierungspraktiken, der in vielen wissenschaftlichen Disziplinen zu beobachten ist, zeugt natürlich davon, dass Visualisierung dort nicht als Ausschmückung, sondern als Arbeitsinstrument angesehen wird. Gleichzeitig wird daran sichtbar, wie wichtig definierte Visualisierungspraktiken für die Durchsetzung von Denkstilen in einem Denkkollektiv sind: Wer den kanonisierten Praktiken nicht folgt, muss eine überzeugende Innovation anbieten oder aber riskiert seine disziplinäre Glaubwürdigkeit.

Bemerkenswert beim Stellenwert von Visualisierungen in unterschiedlichen Disziplinen (und auch in historischer Perspektive) sind die Differenzen der Einschätzung darüber, wie akkurat Visualisierungen einen Gegenstand wiedergeben können und welchen Zwecken sie dienen. In stark empirisch ausgerichteten Disziplinen, z. B. der Korpuslinguistik, spielen Visualisierungen eine wichtige Rolle, die die Eigenschaften von Datenmengen wiedergeben sollen. Eine der einfachsten Formen ist dabei das Balkendiagramm, eine professionellere Praxis verwendet aber eine breite Palette von Streudiagrammen mit zusätzlichen

8 Karin Knorr Cetina (2001) etwa zeigt am Beispiel der Hochenergiephysik die Bedeutung von Visualisierungen für die Disziplin, da sie die Grundlage der gesamten wissenschaftlichen Kommunikation überhaupt bilden. Sie prägte dafür den Begriff der „Viskurse".

Indikatoren, die die Verteilung von Datenpunkten möglichst exakt beschreiben sollen. Das Vertrauen in die Visualisierung ist dabei so groß, dass sie oft als erstes Mittel der Datenanalyse empfohlen wird und sich daran dann statistische Tests anschließen, um den über die Visualisierung gewonnenen Eindruck zu prüfen (Gries 2008). Dies ist akzeptiert, da die Visualisierung auf wohldefinierten und in der Statistik akzeptierten Mess- und Testverfahren beruht und der Transfer der daraus entstandenen Werte in grafische Elemente (sog. „Mapping") ebenfalls systematisiert und standardisiert ist.

Ebenso stark von Denkstilen durchsetzt sind die Rollen, die in den unterschiedlichen Disziplinen Visualisierungen zugedacht werden. Verfahren des Data Minings, die visuelle Analysemethoden nutzen, sind beispielsweise stark von einem Utilitarismus-Topos durchdrungen, bei dem davon ausgegangen wird, dass die Daten eine Wahrheit (eine „ground truth", einen „Schatz") enthalten, der gefunden werden kann (Bubenhofer 2016; Bubenhofer u. a. 2018; vgl. auch den Beitrag von Barbaresi in diesem Band, S. 167). Die Rolle des Tools ist damit klar definiert: diesen Schatz zu heben. Damit kann ein visuelles Analyseinstrument anhand eines „Goldstandards" evaluiert und seine Güte ausgedrückt werden. In geisteswissenschaftlichen Kontexten herrscht allerdings oft Skepsis gegenüber einer solchen Sicht, da eher von einem Gewebe von Theorie, Methodologie und deren Operationalisierung und Implementierung, bei dem die/der Forscher/in bedeutender Bestandteil ist, ausgegangen wird. Aus dieser Sicht sind deshalb visuelle Analysemethoden eher nicht effektive Analysetools, die den „Informationsüberfluss" zähmen sollen, sondern eher eine Möglichkeit, reiche Nahrung für Deutungen zu bieten, etwa im Sinne einer „dichten Beschreibung" (Geertz 1987).

4.2 Visualisierungen als wissenschaftliches Gerät

Genau so, wie Visualisierungen als Zeichen rhetorische Funktionen einnehmen, dienen sie in der Forschung oft als Instrument. Dies gilt insbesondere für explorative Visualisierungen: Mit den Daten kann erst gearbeitet werden, wenn diejenigen Aspekte davon, die von Interesse sind, in grafische Formen überführt wurden. Die Operationen im visuellen Ensemble der grafischen Formen führen dann im besten Fall zu neuen Erkenntnissen, analog der Arbeit mit dem Mikroskop.

Im Fall von algorithmisch erstellen Visualisierungen sind nach den Ausführungen oben zu den diagrammatischen und algorithmischen Perspektiven zwei Aspekte wichtig: 1) Zum wissenschaftlichen Gerät gehört nicht nur die Visualisierung selber, sondern der Computer insgesamt als diagrammatische Maschine. 2) Programmcode, Algorithmen und Software als Bestandteile des wissenschaftlichen Geräts sind kulturell geprägt und machen das Gerät dadurch

zu einem besonders wirkmächtigen Mittel wissenschaftlicher Denkstile. Denn das Metamedium Computer ist fluider und unbestimmter als etwa ein Mikroskop und damit weit stärker kulturellen Praktiken unterworfen, die allerdings häufig den Anwenderinnen und Anwendern von Visualisierungssoftware nicht bewusst sind.

4.3 Kanons und Kulturen

Visualisierungen sind also Ausdruck von wissenschaftlichen Denkstilen. Die Visualisierungspraxis eines Denkkollektivs zeigt sich daran, ob Visualisierungen in der jeweiligen Disziplin akzeptiert sind, welche Typen und welche Ausprägungen üblich sind und zu welchen Zwecken sie verwendet werden. Für jede Disziplin wird in bestimmten Zeiträumen ein bestimmter Bildstil erkennbar sein: „Aus dem Abstand von hundert Jahren wird man die Darstellungen der Nanoforschung auf einen Blick auf zwei oder drei Jahre genau datieren können; in Bezug auf die fraktale Mathematik gilt das gleiche." (Bredekamp u. a. 2008: 41f.) Ebenso existieren in jeder Disziplin „ikonisierte Diagramme" (vgl. Lauersdorf in diesem Band S. 91), die alternative Visualisierungen behindern.

Insbesondere bei algorithmischen Visualisierungen müssen dabei auch die programmiertechnischen und algorithmischen Grundlagen mitbedacht werden. Visualisierungen, die mit der Javascript-Bibliothek D3 (Bostock u. a. 2011) erstellt wurden, ähnelt sich vom Typus her, obwohl diese Bibliothek eine enorme Vielfalt an neuen Visualisierungsformen von Daten auslöste. Die Designprinzipien der Bibliothek sind in eine bestimmte Softwarekultur eingebettet (Open Source, browserbasiert, leicht nutzbar, modular etc.), die u. a. vom Erfinder selber auch explizit verbalisiert wird, etwa im Rahmen der neuen Version 4 der Bibliothek: „Programming interfaces are user interfaces. Or, to put it another way: Programmers are people, too" (Bostock 2016).

Um aber Innovation im Bereich wissenschaftlicher Visualisierung zu ermöglichen, muss dieser Kanon disziplinärer Visualisierungspraktiken immer wieder hintergangen werden. Die *andere, nicht-kanonisierte* Visualisierung ist notwendig, um Innovation zu ermöglichen. Pflaeging (in diesem Band S. 123) plädiert beispielsweise dafür, gezielt visualisierte Metaphern hinzuzuziehen, um linguistische Theorien zu erklären und so durch eine „Ästhetisierung von Linguistikvermittlung […] rezipientenseitig zu einer verlängerten und bewussteren Wahrnehmung der musterbrechenden Elemente" zu gelangen. Relativ pragmatisch machen Keim et al. (2006) auf dieses Problem aufmerksam: „User acceptability is a further challenge; many novel visualization techniques have been presented, yet their widespread deployment has not taken place, primarily due to the users' refusal to change their working routines." Als Rezept zur Akzeptanzförderung

wird schlicht und einfach vorgeschlagen, den Benutzerinnen und Benutzern die Vorteile der Visualisierung zu erklären – durchaus ein gangbarer Weg, doch dürfte das alleine nicht reichen, solange der jeweilige Denkstil diese Form noch nicht akzeptiert.

4.4 Listen und Partituren

Im letzten Teil dieses Beitrags möchte ich den Blick nochmals auf die Visualisierungspraxis in der Linguistik lenken. Neben Karten, Graphen/Netzen (etwa zur Darstellung von Kollokationen) oder Bäumen (für syntaktische Strukturen, Sprachfamilien etc.) und Vektoren (visualisiert als Werte in einem Koordinatensystem) gibt es in den Sprachwissenschaften zwei unauffälligere, aber deswegen nicht weniger wirkmächtige Methoden, die ich unter den Schlagworten „Listen“ und „Partituren“ fassen möchte (vgl. zu typischen Visualisierungsformen in der Linguistik auch den Beitrag von Perkuhn und Kupietz in diesem Band, S. 63). An beiden folgenden Beispielen zu Listen und Partituren kann gleichzeitig diskutiert werden, wie sich durch die Transformation oder Formatierung der Sprachdaten in diese Formen der Analysegegenstand verändert.

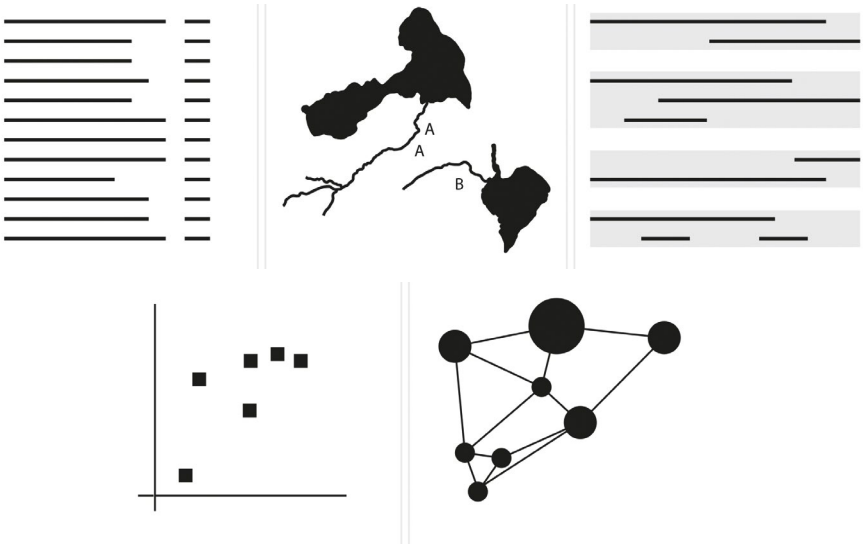


Abb. 2: Fünf in der Linguistik häufig anzutreffende Visualisierungstypen: Liste, Karte, Partitur, Vektoren, Graph/Netz (gerichtet: Baum; ungerichtet: Netz).

Aus diagrammatischer Perspektive ist klar, dass jegliche Strukturierung von Text, beispielsweise in Form von Spiegelpunkten, Aufzählungen o. Ä. bereits diagrammatischen Charakter hat (Steinseifer 2013). Im linguistischen Kontext sind zwei Beispiele naheliegend, bei denen Sprache umstrukturiert wird, um sie analysierbar zu machen: 1) In der Korpuslinguistik ist vor allem die „Key Word in Context"-Darstellung – KWiC – bekannt. Sie wird verwendet, um eine Ergebnismenge von Treffern übersichtlich darzustellen. Diese sehr alte und prinzipiell simple Form eines Index widerspiegelt eine völlig neue Lektüretechnik. 2) In der Gesprächsanalyse dienen verschiedene Notationssysteme, darunter insbesondere die Transkription in die sog. Partiturschreibweise dazu, gesprochene Sprache in ihrer Flüchtigkeit und Gleichzeitigkeit analysierbar zu machen. Beide Beispiele lohnen eine Analyse aus diagrammatischer Sicht.

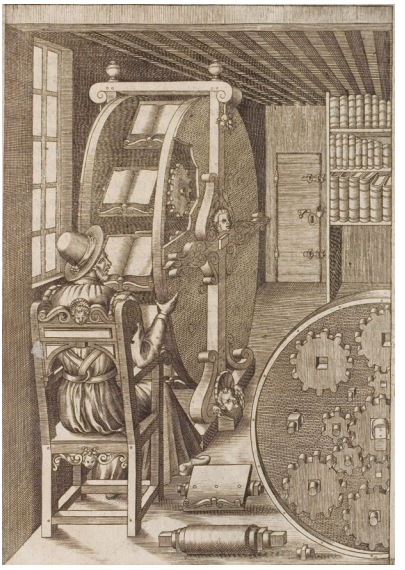


Abb. 3: Das Bücherrad von Agostino Ramelli (Ramelli 1588: 317).

1588 erfindet Agostino Ramelli ein Bücherrad (Ramelli 1588: 317): Eine hölzerne Konstruktion, auf der gleichzeitig mehrere Bücher aufgeschlagen Platz finden und vor der die Leserin/der Leser Platz findet und bequem am Rad drehen kann, um jederzeit zum nächsten Buch wechseln zu können (Siegel 2009: 28). Zwar blieb die Erfindung ein Plan auf Papier, sie widerspiegelt jedoch einen Paradigmenwechsel humanistischer Lektüretechniken in der Frühen Neuzeit: „Anstelle kontemplativer Versenkung in einen einzigen, durch die Buchdeckel begrenzten Text, bedeutet Lektüre hier einen stetigen Akt des aufeinander Beziehens von ursprünglich distinkten Sinneinheiten." (Siegel 2009: 33). Dafür werden neue Methoden notwendig, etwa die Glossierung am Seitenrand oder das Exzerpieren

und die Verwaltung dieser Exzerpte, z. B. in Form des „Liber excerptorum" von Placcius, bei dem Zettel flexibel in einem Buch eingesetzt werden können (Placcius 1689: 68ff.), oder dem „Scrinium litteratum", dem Karteikasten, der noch größere Freiheit beim Reorganisieren der Daten erlaubt (Placcius 1689: 121ff.). In der gleichen Zeit entstehen die großen Enzyklopädien zur Systematisierung des Wissens und erweitern damit Lektüre zu einem komplexen Akt, „der im Ganzen die Struktur des gedruckten Fließtextes zu Formularen erweitert, deren hoher Anspruch eine systematische Ordnung der gelehrten Überlieferung ist" (Siegel 2009: 47).

Die Nähe dieser Lektüretechniken zur Korpuslinguistik und dem modernen Data Mining ist unübersehbar: „Die […] praktisch geübte und methodisch reflektierte Isolation einzelner Textpartien zum *locus [communis]* bedeutet, die Vielfalt tradierter Texte als einen riesenhaften Speicher zu behandeln, der in sich die Summe des verfügbaren Wissens inkorporiert" (Siegel 2009: 37). Die Konzeptualisierung der Texte als Speicher geht mit einem neuen Zugriff darauf einher. Genauso bricht die Suche in einem Textkorpus nach sprachlichen Einheiten und die Repräsentation der Ergebnisse als Konkordanz – also die Isolation der Textstellen aus dem jeweiligen Text – mit dem traditionellen hermeneutischen Lektüreerlebnis. Gleichzeitig stellt die Konkordanz aber einen eigenen „locus" dar, ist also von eigenem Aussagewert und ist die Voraussetzung dafür, daraus eine emergente Struktur ableiten zu können.

Natürlich sind auch Unterschiede zwischen Korpuslinguistik / Data Mining und den neuen Lektüreformen in der Frühen Neuzeit sichtbar, was in der digitalen Form der Daten begründet liegt. Das betrifft sowohl die Möglichkeiten der Erstellung der Indizes als auch deren weitere Verarbeitung, bei der in der digitalen Welt die menschliche Leserin / der menschliche Leser an verschiedenen Stellen durch Algorithmen ersetzt werden kann.

Die KWiC-Liste wird in der Linguistik (und benachbarten Disziplinen) zuweilen durchaus als Provokation wahrgenommen. Dies ist beispielsweise im Kontext der Diskurslinguistik sichtbar, wo keine Einigkeit über den Nutzen des korpuslinguistischen Zugangs herrscht, gerade auch, weil die bekannteste korpuslinguistische Präsentationsform von Ergebnissen, die Liste von Textstellen, die Texteinheit aufbricht. So wird argumentiert, die Diskurslinguistik arbeite mit dem Begriff des einheitlichen Textes, die Korpuslinguistik dagegen mit Textfragmenten (Leech 2000: 678; Spitzmüller und Warnke 2011: 32). Andererseits kann der korpuslinguistische Zugang auch gerade als passend angesehen werden, um textübergreifende Aussagensysteme zu untersuchen (Bubenhofer und Scharloth 2013: 247). Es handelt sich um eine Art „maschinengeleitete Lesetechnik" (Scholz und Mattissek 2014: 87):

> „Die dekontextualisierte Darstellung erlaubt es den Forschenden, frei vom ‚hermeneutischen Reflex', der die Lektüre von Texten und Textpassagen bestimmt, kreativ Ideen zu möglichen diskursiven Zusammenhängen einzelner Korpusteile zu entwickeln, die bei einer subjektiven Lektüre möglicherweise verdeckt blieben." (Scholz und Mattissek 2014: 87)

Diese dekontextualisierte Darstellung ist in der Korpuslinguistik, wie erwähnt, zwingendes Ergebnis eines quantitativen Zugangs, der eine Lektüre des Einzeltextes je nach Ansicht ersetzen, ergänzen und/oder aber gerade nicht ersetzen kann. Die KWiC-Darstellungen (und viele weitere, avanciertere synthetisierende Darstellungen von quantitativen Ergebnissen) werden damit zum Symbol einer am Sprachgebrauch orientierten Sicht auf Diskurse und so im größeren Kontext zum Gegenstand lebhafter Diskussionen über die verschiedenen Spielarten von Diskursanalyse (Angermüller 2014; Meier u. a. 2014; Niehr 2015).

Ich meine damit zeigen zu können, wie stark eine diagrammatische Form – die Liste als Konkordanz – die Lektüre von Text beeinflusst und zu einem neuen Verständnis von Sprache führt: Ein Verständnis, das geprägt ist vom Interesse an der sprachlichen Form, an repetitiven Segmenten, Mustern und ihrer soziokulturellen Deutung – an Oberfläche und Performanz (Feilke und Linke 2008).

Eine ähnliche Verschränkung von diagrammatischer Form und neuer Sicht auf Sprache ist in der Analyse gesprochener Sprache zu beobachten. Im Rahmen der Dialektologie und der sprachtypologischen Forschung war die Transkription von Äußerungen gesprochener Sprache schon länger ein Thema (Redder 2001), doch erst mit Arbeiten wie jenen von Sacks et al. (1974) entsteht eine neue Perspektive, die auch radikal neue Darstellungsformen bedingt: Die Keimzelle dieser Perspektive liegt wahrscheinlich in der Fokussierung auf den „turn": Eine Einheit, die Gespräche strukturiert und organisiert und sich nicht mit den klassischen grammatischen Kategorien geschriebener Sprache fassen lässt. Damit einher gehen Mechanismen des Sprecherwechsels, bei denen sich „turns" nacheinander ablösen, aber oft sich eben auch überlappen. Die Sequenzialität sprachlicher Äußerungen, die bei der geschriebenen Sprache die Normalität ist, wird in der gesprochenen Sprache um die Dimension der Gleichzeitigkeit ergänzt.

Dies bedingt neue diagrammatische Formen der Transkription, mit der die Gleichzeitigkeit mehrerer „turns" dargestellt werden kann. Es etablierten sich verschiedene Lösungen dafür, besonders erfolgreich ist jedoch die Partiturschreibweise (vgl. Abb. 4). Dabei werden die verschiedenen Sprecher/innen als „Stimmen" eines Orchesters aufgefasst und die in der Musik übliche Notation der Partitur übernommen, bei der jede Stimme eine eigene Zeile einnimmt und sich übereinander angeordnet die Gleich- und Ungleichzeitigkeit der verschiedenen Stimmen

[142]
f2 jahre de fakto;=und da drüber brauchen wir jetzt gar ned lang diskutieren-
m2 ((klopft auf pult, senkt blick,
? ( )

[143]
f2 (.) weil wenn die komission in ihrer WUNderbaren vielfalt nach 6 jahren beschliesst, dass
m2 bläst luft aus, blickt sie wieder an))

[144]
f2 sie alle noch VI:EL lieber und noch VI:EL billiger bei uns durchfahren
m1 <<all>kann sie gar nicht>
m2 hehehe
? (aso)

[145]
f2 möchten, (.) ä:h dann sind WIR der billige jakob? (.)
m1 ((grinst))
m2 frau lichtenberger ich schätze sie.

[146]
f2 <<wehrt äusserungen gestisch ab>na bitte do (redens) mer da jetzt nit rein?> dann (-) sein
m2 machen wir (kei innerpolitisches parkett da); ((grinst, senkt

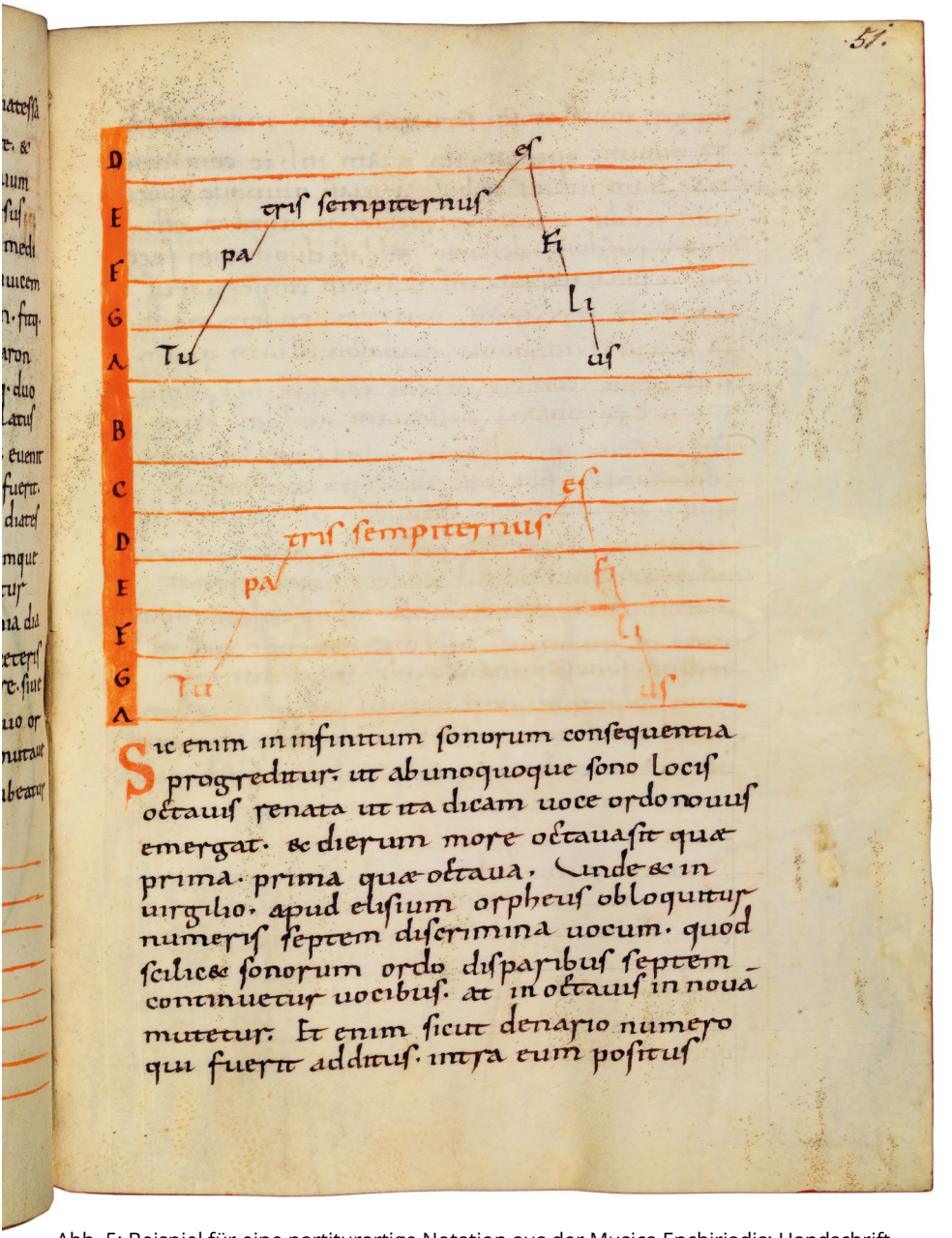
[147]
f2 mir der billige jakob von ÖSCHterreich? (-) wir haben nicht einmal eine geringfügige
m2 blick, lehnt sich zurück))

Abb. 4: Beispiel für ein Transkript in Partiturschreibweise (aus: Stocker u. a. 2004).

sofort überblicken lässt. Transkriptionsstandards wie GAT (Selting u. a. 1998) oder HIAT (Ehlich und Rehbein 1976, 1979; Rehbein u. a. 2004) definieren die Details der Umsetzung.

Es lohnt sich, einen Blick in die Geschichte der Notation von Musik und dort insbesondere der Partitur zu werfen, um die diagrammatische Innovation dieser Schreibweise, gleichzeitig aber auch die funktionalen Differenzen zwischen der Musikpartitur und der Partiturschreibweise von Gesprächstranskripten herauszuarbeiten. Bei der Notation polyphoner Musik muss unterschieden werden zwischen Ensemble-Musik, bei der mehrere Personen gleichzeitig spielen, und Solo-Musik, bei der eine Person auf dem gleichen Instrument gleichzeitig mehrere Stimmen spielt, beispielsweise die linke und rechte Hand beim Klavierspiel (Apel 1961: XXV; Sachs und Röder 1989). Bereits aus dem 9. Jahrhundert sind Notationen bekannt, die als Vorläufer von Partituren angesehen werden, so z. B. in der *Musica Enchiriadis*, einer Art Handbuch, das dazu dienen sollte, das Singen von Gregorianischen Chorälen zu unterrichten (vgl. Abb. 5[9]). Dort ergab sich die Notwendigkeit, die sich hauptsächlich in Quart- oder Quintabständen parallel bewegenden Stimmen untereinander, aber im gleichen Notensystem, abzubilden.

9 Die gesamte Handschrift ist bei der Staatsbibliothek Bamberg digital einsehbar: http://bsbsbb.bsb.lrz.de/~db/0000/sbb00000078/images/ (letzter Zugriff: 20. Januar 2018).



Sic enim in infinitum ſonorum conſequentia
progreditur: ut ab unoquoque ſono locis
octauis renata ut ita dicam uoce ordo nouus
emergat. ac dierum more octauaſit quæ
prima. prima quæ octaua. Unde ac in
uirgilio. apud eliſium orpheus obloquitur
numeris ſeptem diſcrimina uocum. quod
ſcilicet ſonorum ordo diſpariſ ſeptem
continuetur uocibuſ. at in octauis in noua
mutetur. Et enim ſicut denario numero
qui fuerit additus. intra eum poſitus

Abb. 5: Beispiel für eine partiturartige Notation aus der Musica Enchiriadis; Handschrift, Kopie Msc.Var.1, fol. 51r, Scolica enchiriadis de arte musica – u. a. musiktheoretische Texte. Werden (?), um 1000. Staatsbibliothek Bamberg.Foto: Gerald Raab.

Damit weist das Notationssystem zwei Achsen auf: Die vertikale Achse gibt das Tonsystem in der sog. Dasia-Notation (später Tonbuchstaben) wieder, während in der horizontalen der Verlauf der Melodien notiert war (Sachs und Röder 1989). Mit der Notation war ein didaktischer Nutzen verbunden, um durch die „Verbindung von Hören und Sehen" (Sachs und Röder 1989) die Intervall-Bezüge zwischen den Stimmen deutlich zu machen.

Allen Partitur-ähnlichen Notationen bis ins 15. Jahrhundert ist gemein, dass eine exakte vertikale Anordnung nicht angestrebt wurde, da sich erst eine systematische Notation von Tonlängen, Taktstrichen und konventionalisierter vertikaler Anordnung entwickeln musste. Auch danach schienen die ersten Partituren, die nun für jede Stimme ein eigenes Notensystem mit entsprechendem Schlüssel nennen (etwa 1537 bei Lampadius, vgl. Sachs und Röder 1989: 1431), eine nebengeordnete Rolle zu spielen: Sie dienten als Vorlage, um daraus die Einzelstimmen zu extrahieren, wobei bei der Aufführung eine dieser Einzelstimmen auch als Dirigiergrundlage verwendet wurde. Im Verlauf des 16. Jahrhunderts etabliert sich die Partitur allmählich, da sie als Aufzeichnungssystem diente, um „ein (fremdes) Werk genauer prüfen zu können", und dazu, aus „aufführungspraktischen Zwecken […] ein Tasteninstrument, in der Regel die Orgel, als Stütze heranzuziehen" und die dafür nötige Notengrundlage als (Teil-)Partitur zu erstellen (Sachs und Röder 1989: 1429).

Deutlich ist also auch bei musikalischen Partituren, dass sie zunächst dazu dienten, die Polyphonie von Musik sichtbar zu machen, weniger als Vorlage für die musikalische Wiedergabe. Die Erfindung der Partiturschreibweise ermöglicht einen neuen Blick auf den bereits bekannten Gegenstand (das Lied, das Musikstück, das Gespräch), indem die zeitliche Interaktion der Stimmen bzw. Sprecher/innen hervortritt. Die Partitur visualisiert die Gleichzeitigkeit von Gleichwertigem – im Unterschied zu Glossen oder Marginalien, die zwar auch ein diagrammatisches Ausdrucksmittel von Gleichzeitigkeit sind, doch deutlich priorisieren, eben: marginalisieren.

Trotzdem sind auch Differenzen zwischen Musik- und Gesprächspartitur ersichtlich: Die Musikpartitur hat zwar eine ähnliche ordnende Funktion von Stimmen wie Gesprächspartituren, nutzt aber zusätzlich mit dem Notationssystem von Tonhöhen, -längen und Rhythmen ein System, das sich in die Partiturschreibweise problemlos integriert: Die Viertelnote der ersten Stimme wird, ergänzt mit Freiraum, denselben Platz einnehmen wie die gleichzeitig spielenden zwei Achtelnoten der zweiten Stimme – die Taktmarkierungen sind eine zusätzliche Hilfe, um diesen Effekt zu erreichen. Bei Gesprächstranskripten hingegen entspricht die orthografische und typografische Umsetzung des gesprochenen Lautes nicht dessen Länge; visuell kann deshalb ein „turn" der einen Sprecherin denjenigen des zweiten Sprechers überdauern, ohne es real ebenso gemacht zu haben (Redder 2001: 1048). Moderne Transkriptionsprogramme begegnen

diesem Problem durch die Integration einer grafischen Umsetzung des Frequenzspektrums in Form eines Spektrogramms.

Den beiden Anwendungsbereichen gleich ist wiederum die Einsicht, dass die Notationssysteme nicht reichen, den Gegenstand, das Gespräch bzw. die Musik, komplett zu erfassen (vgl. für die Musik z. B. Schneider 1987: 317), sie aber trotzdem unverzichtbar sind, um die Flüchtigkeit des Gegenstands zu fixieren und die Komplexität handhabbar zu machen.

Ähnlich wie die Liste erlaubte die Übernahme der Partiturschreibweise aus der Musik in die Linguistik einen neuen Blick auf Sprache – oder schuf damit einen neuen Analysegegenstand: Das Gespräch. Doch nicht nur in den engen Grenzen der Gesprächslinguistik wirkte diese Darstellungsweise innovativ: Das Prinzip der Annotation von Textdaten beruht in der Darstellung ebenfalls auf dem Partiturprinzip, da die Gleichzeitigkeit von Text und Annotation(en) in der ganzen Komplexität dargestellt werden muss (siehe zu den Schwierigkeiten solcher Darstellungen den Beitrag von Burghardt in diesem Band, S. 315). Bereits die in der Korpuslinguistik übliche Darstellung eines annotierten Textes in einer Mischung aus XML-Auszeichnung und vertikalisierter Spaltendarstellung folgt einer (vertikalen) Partiturschreibweise (vgl. Abb. 6).

5. Fazit

Diagrammatische Formen haben das Potenzial, Denkstile zu verändern und wissenschaftliche Tatsachen zu schaffen. In der Linguistik verändern sie den Gegenstand „Sprache“ auf ihre je eigene Art, wie ich am Beispiel von Listen und Partituren gezeigt habe. Karten, Bäume und Netze wären weitere in der Linguistik gängige Formen, deren Auswirkungen untersucht werden müssten.

Doch welcher Art sind diese Transformationen von „Sprache“? Noch von einer erschöpfenden Systematik weit entfernt, nenne ich im Folgenden sechs grundlegende Transformationstypen, die mir für die Analyse von Sprache besonders relevant zu sein scheinen (aber natürlich auch bei anderen Daten auftauchen):

— Rekontextualisierung: Die Listendarstellung, z. B. von Belegen in einem Korpus als KWiC-Liste, aber auch die Platzierung von sprachlichen Daten auf einer Karte, erzeugen für diese sprachlichen Daten einen neuen Kontext. Weiter oben habe ich im Zusammenhang der Korpuslinguistik von „Dekontextualisierung“ gesprochen – ich meine, es ist angemessener von einer Rekontextualisierung zu sprechen, da durch die neuen Darstellungsweisen eine neue Einheit gebildet wird, die durchaus in einem Kontext steht. Bei der KWiC-Liste beispielsweise formt das Paradigma der einzelnen

```
<s lang="de" n="a2-s87">
Unter   APPR    unter   a2-s87-w1
den ART d   a2-s87-w2
zahllosen   ADJA    zahllos a2-s87-w3
Bergen  NN  Berg    a2-s87-w4
vor APPR    vor a2-s87-w5
unsern  PPOSAT  unser   a2-s87-w6
Augen   NN  Auge    a2-s87-w7
ragte   VVFIN   empor+ragen a2-s87-w8
der PDS d   a2-s87-w9
ferne   ADV ferne   a2-s87-w10
<mountain id="g_38" stid="g23" level="geo">
Mont    NE  Mont    a2-s87-w11
Blanc   NE  Blanc   a2-s87-w12
</mountain>
über    APPR    über    a2-s87-w13
die ART d   a2-s87-w14
andern  PIS ander   a2-s87-w15
empor   PTKVZ   empor   a2-s87-w16
;   $.  ;   a2-s87-w17
</s>
<s lang="de" n="a2-s88">
näher   ADJD    nah a2-s88-w1
bei APPR    bei a2-s88-w2
uns PRF wir a2-s88-w3
thronten    VVFIN   thronen a2-s88-w4
<mountain id="g_39" stid="s7302510" level="geo">
Schreckhorn NE  Schreckhorn a2-s88-w5
</mountain>
,   $,  ,   a2-s88-w6
<mountain id="g_40" stid="s7296734" level="geo">
Wetterhorn  NE  Wetterhorn  a2-s88-w7
</mountain>
und KON und a2-s88-w8
<mountain id="g_41" stid="s7308060" level="geo">
Jungfrau    NE  Jungfrau    a2-s88-w9
</mountain>
,   $,  ,   a2-s88-w10
dem PRELS   d   a2-s88-w11
Scheine NN  Schein  a2-s88-w12
nach    APPO    nach    a2-s88-w13
weniger ADV weniger a2-s88-w14
hoch    ADJD    hoch    a2-s88-w15
als KOKOM   als a2-s88-w16
unser   PPOSAT  unser   a2-s88-w17
Standort    NN  Standort    a2-s88-w18
.   $.  .   a2-s88-w19
</s>
```

Abb. 6: Vertikalisierter Text mit XML-Auszeichnungen, Beispiel für die Partiturschreibweise in der Korpuslinguistik.

Syntagmen einen neuen Kontext. Deutlicher noch bei einem Kollokationsprofil: Diese Liste ist Ausdruck eines statistischen Distributionsverhaltens; die Distribution ist der Kontext, in dem die einzelnen Einträge der Liste gelesen werden müssen.

- Desequenzialisierung: Sie ist manchmal ein Nebeneffekt der Rekontextualisierung. Der springende Punkt vieler Diagramme in der Linguistik ist die Auflösung der der Sprache innewohnenden Sequenzialisierung. Die Wortfrequenzliste rekombiniert den Text zu einer geordneten Liste von Wörtern und ignoriert deren ursprüngliche Sequenz. Nicht jede diagrammatische Transformation von Sprache muss zwingend zu kompletter Desequenzialisierung führen. Die Visualisierung von typischen Narrativen in seriellen Geschichten (Bubenhofer u. a. 2013) führt zu einer Rekontextualisierung (Typizität von Narrativen), ohne die ursprüngliche Sequenz zu zerstören. Allgemeiner könnte Desequenzialisierung als Typus einer Dimensionsreduktion angesehen werden.
- Dimensionsanreicherung: Alle Formen von Partituren, aber auch Netze, Bäume oder Karten reichern sprachliche Daten um weitere Dimensionen an. Partituren ermöglichen die Darstellung beliebig vieler weiterer Ebenen der Gleichzeitigkeit, Netze stellen Bezüge zu anderen Entitäten her, gerichtete Graphen wie Bäume fügen die Dimension der hierarchischen Gliederung hinzu und Karten bieten eine geografische Anreicherung.
- Rematerialisierung: Die diagrammatische Transformation überführt sprachliche Daten in eine neue Materialität. Das Kollokationsprofil stellt die statistische Zusammenfassung eines Distributionsverhaltens des entsprechenden Lexems dar und ist damit ein neuer Gegenstand, der z. B. als semantisches Lesartenspektrum des Lexems behandelt werden kann. Mit einer Karte assoziierte sprachliche Einheiten ergeben einen Gegenstand von Sprache „in situ". Der Kollokationsgraph ergibt den Gegenstand des Bedeutungsgewebes.

Die ersten vier Typen formen unmittelbar sprachliche Daten in eine Form um, die die neue Perspektive auf sie ermöglicht. Daraus ergeben sich zwei für Diagramme allgemein typische Funktionen, Operationalität und das Potenzial für Emergenz:

- Ermöglichung von Operationalität: Erst wenn durch die oben genannten Transformationen das Diagramm entsteht, kann damit operiert werden: Aus einer Karte, auf der Aussprachevarianten markiert sind, können Dialekträume herausgelesen werden; im Kollokationsgraph können Knoten hervorgehoben werden, um deren Position im Netz genauer zu untersuchen. Bei algorithmisch erstellten und digital verfügbaren Visualisierungen

dienen oft Mittel der interaktiven Beeinflussung dazu, mit dem Diagramm operieren zu können. Allerdings ist digitale Interaktivität keinesfalls notwendig, um Operationalität zu erzeugen, denn diese kann auch mit Stift und Papier oder nur in Gedanken vollführt werden. Und andererseits geht bei digitalen Daten diagrammatische Operationalität über Interaktivität hinaus, da sie auch die zugrunde liegenden Algorithmen berührt.

— Emergenz-Erzeugung: Die Visualisierung der Daten ermöglicht im besten Fall, darin ein emergentes Phänomen sichtbar zu machen (vgl. dazu auch Barbaresi Seite 167 in diesem Band). Bei Netzgraphen gibt die generelle Ausprägung des Netzes – etwa ob viele oder wenige Cluster von Knoten sichtbar sind, wie dicht das Netz ist etc. – eine Information über die Art des Netzes. Bei der Betrachtung eines Kollokationsprofils erscheinen auf einer emergenten Ebene „Bedeutungen" des Lexems, auf einer Dialektkarte sind es „Dialekte", bei anderen Darstellungen vielleicht „Diskursräume" oder „sprachliche Muster", also alles reichlich abstrakte Phänomene, die aus der Einzelbetrachtung von Belegen oder Datenpunkten nicht abgeleitet werden könnten und erst mit der Visualisierung als generelle Form sichtbar werden.

Generell handelt es sich bei diesen diagrammatischen Operationen um Verfahren „‚rekursive[r] Transkriptivität' der Sprache" (Jäger 2010: 315):

> „Symbolische Systeme tendieren dazu, als Gewinn aus der für sie charakteristischen Verfahrensform der *rekursiven Selbstverarbeitung* Eigensinn zu generieren. Dies gilt insbesondere für Sprache, die in paradigmatischer Weise über die Eigenschaft verfügt, sich rekursiv auf sich selbst zurückzubiegen und so die eigene Zeichenverwendung fortlaufend zum Gegenstand weiterer thematisierender, kommentierender, explizierender oder zitierender Zeichenverwendungen zu machen, zum Objekt also selbstbezüglicher semiologischer Operationen, in denen sich das zeigt, was man die ‚rekursive Transkriptivität' der Sprache nennen könnte." (Jäger 2010: 315)

Grundlegende Fragestellung einer Visual Linguistics ist also, genauer zu verstehen, welche diagrammatischen Operationen in welchen wissenschafts- und technikkulturellen Umgebungen zu welchen Deutungsmöglichkeiten (Semantiken) von Sprache führen.

Die Beiträge in diesem Band sind Zeugnis einer lebendigen Visualisierungspraxis in der Sprachwissenschaft. Die Digitalisierung der Daten und das neu erwachte Interesse im Fach für die Analyse großer Textdatenmengen ermöglicht

Visualisierungsexperimente, die den Kanon gegenwärtig erheblich erweitern und Einfluss auf die herrschenden Denkstile haben. Es scheint mir dabei zwingend, aber auch lohnenswert, gerade aus geisteswissenschaftlicher Sicht den Algorithmus und das Programmieren nicht Spezialist/innen komplett zu überlassen, sondern zumindest die kulturellen Prägungen dieser Praktiken zu verstehen. Ebenso wichtig ist, die Visualisierungspraktiken im Fach nicht als Mittel zum Zweck, sondern als gegenstandskonstituierend zu verstehen, als Praktiken, die sowohl Ausdruck von Denkstilen sind als auch diese mitkonstituieren.

Danksagung: Der vorliegende Beitrag entstand im Rahmen des vom Schweizer Nationalfonds (SNF) geförderten Projektes „Visual Linguistics".

6. Bibliografie

Angermüller, Johannes. 2014. „‚Der' oder ‚das' Korpus? Perspektiven aus der Sozialforschung." In *Diskursforschung. Ein interdisziplinäres Handbuch*. Berlin, Boston: de Gruyter, 604–611.

Apel, Willi. 1961. *The notation of polyphonic music, 900-1600*. Cambridge, Mass.: Mediaeval Academy of America.

Bauer, Matthias und Christoph Ernst. 2010. *Diagrammatik: Einführung in ein kultur- und medienwissenschaftliches Forschungsfeld*. Bielefeld: transcript.

Böhme, Gernot. 2006. „Technical Gadgetry: Technological Development in the Aesthetic Economy." In *Thesis Eleven* 86 (1): 54–66, doi: 10.1177/0725513606066240.

Bostock, Michael, Vadim Ogievetsky und Jeffrey Heer. 2011. „D3: Data-Driven Documents". In *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis),* 2301-2309.

Bostock, Mike. 2016. What Makes Software Good? https://medium.com/@mbostock/what-makes-software-good-943557f8a488 (letzter Zugriff am 24. August 2016).

Bredekamp, Horst, Birgit Schneider und Vera Dünkel, Hrsg. 2008. *Das Technische Bild: Kompendium zu einer Stilgeschichte wissenschaftlicher Bilder*. Berlin: Akademie Verlag.

Bubenhofer, Noah. 2016. „Drei Thesen zu Visualisierungspraktiken in den Digital Humanities". *Rechtsgeschichte Legal History – Journal of the Max Planck Institute for European Legal History* (24): 351–355.

Bubenhofer, Noah, Nicole Müller und Joachim Scharloth. 2013. „Narrative Muster und Diskursanalyse: Ein datengeleiteter Ansatz". *Zeitschrift für Semiotik, Methoden der Diskursanalyse* 35 (3–4): 419–444.

Bubenhofer, Noah, Klaus Rothenhäusler, Katrin Affolter und Danica Pajovic. 2018. „The Linguistic Construction of World – an Example of Visual Analysis and Methodological Challenges". In *Quantifying Approaches to Discourse for*

Social Scientists, herausgegeben von Ronny Scholz. Basingstoke: Palgrave Macmillan.

Bubenhofer, Noah, Joachim Scharloth. 2013. „Korpuslinguistische Diskursanalyse: Der Nutzen empirisch-quantitativer Verfahren". In *Diskurslinguistik im Spannungsfeld von Deskription und Kritik,* herausgegeben von Ingo Warnke, Ulrike Meinhof und Martin Reisigl. Berlin: Akademie-Verlag, 147–168 (Diskursmuster – Discourse Patterns).

Bühler, Karl. 1934. *Sprachtheorie*. Stuttgart: G. Fischer.

Chen, Chun-houh, Wolfgang Härdle und Antony Unwin, Hrsg. 2008. *Handbook of data visualization*. Berlin: Springer (Springer Handbooks of Computational statistics).

Coleman, E. Gabriella. 2012. *Coding Freedom: The Ethics and Aesthetics of Hacking*. Princeton: Princeton University Press.

Dourish, Paul. 2014. „No SQL: The Shifting Materialities of Database Technology : Computational Culture". *Computational Culture. A Journal of Software Studies* (4). http://computationalculture.net/no-sql-the-shifting-materialities-of-database-technology/ (letzter Zugriff am 08. Dezember 2017).

Eco, Umberto. 2002. *Einführung in die Semiotik*. 9., unveränd. Aufl.. München: Fink.

Eco, Umberto. 1977. *Zeichen. Einführung in einen Begriff und seine Geschichte*. Frankfurt am Main: Suhrkamp (es).

Ehlich, Konrad und Jochen Rehbein. 1979. „Erweiterte halbinterpretative Arbeitstranskriptionen (HIAT2): Intonation." *Linguistische Berichte* 59: 51–75.

Ehlich, Konrad und Jochen Rehbein, Jochen. 1976. „Halbinterpretative Arbeitstranskriptionen (HIAT)." *Linguistische Berichte* 45: 21–41.

Feilke, Helmuth und Angelika Linke. 2008. „Oberfläche und Performanz – Zur Einleitung." In *Oberfläche und Performanz,* herausgegeben von Helmuth Feilke und Angelika Linke. Berlin: de Gruyter, 3–18.

Fix, Ulla. 2011. *Denkstile und Sprache. Die Funktion von „Sinn-Sehen" und „Sinn-Bildern" für die „Entwicklung einer wissenschaftlichen Tatsache". home.uni-leipzig.de/fix/Fleck.pdf.*

Fleck, Ludwik, Sylwia Werner und Claus Zittel, Hrsg. 2011. *Denkstile und Tatsachen: Gesammelte Schriften und Zeugnisse*. Originalausgabe. Berlin: Suhrkamp.

Fleck, Ludwik, Lothar Schäfer und Thomas Schnelle, Hrsg. 1980. *Entstehung und Entwicklung einer wissenschaftlichen Tatsache: Einführung in die Lehre vom Denkstil und Denkkollektiv*. 10. Aufl. Frankfurt am Main: Suhrkamp.

Fleck, Ludwik, Lothar Schäfer und Thomas Schnelle, Hrsg. 1983. *Erfahrung und Tatsache: gesammelte Aufsätze*. Frankfurt am Main: Suhrkamp.

Ford, Paul. 2015. *What Is Code? If You Don't Know, You Need to Read This*. In *Businessweek*. (June 11, 2015), https://www.bloomberg.com/graphics/2015-paul-ford-what-is-code/ (letzter Zugriff am 18. Januar 2018).

Fuller, Matthew. 2003. *Behind the blip: essays on the culture of software*. Brooklyn: Autonomedia.

Geertz, Clifford. 1987. „Dichte Beschreibung. Bemerkungen zu einer deutenden Theorie von Kultur". In *Dichte Beschreibung. Beiträge zum Verstehen kultureller Systeme*. Frankfurt am Main: Suhrkamp, 7–43 (stw).

Goffey, Andrew. 2014. „Technology, Logistics and Logic: Rethinking the Problem of Fun in Software". In *Fun and software: exploring pleasure, paradox, and pain in computing*, herausgegeben von Olga Goriunova. New York: Bloomsbury Academic, 21–40.

Gries, Stefan Thomas. 2008. *Statistik für Sprachwissenschaftler*. Göttingen: Vandenhoeck & Ruprecht (Studienbücher zur Linguistik).

Heilmann, Till A. 2012. *Textverarbeitung: Eine Mediengeschichte des Computers als Schreibmaschine*. Bielefeld: Transcript (MedienAnalysen).

Hörl, Erich. 2008. „Wissen im Zeitalter der Simulation. Metatechnische Reflexionen". In *Simulation: Präsentationstechnik und Erkenntnisinstrument*, herausgegeben von Andrea Gleiniger und Georg Vrachliotis. Berlin: de Gruyter, 93–106.

Jäger, Ludwig. 2010. „Intermedialität – Intramedialität – Transkriptivität. Überlegungen zu einigen Prinzipien der kulturellen Semiosis." In *Sprache intermedial: Stimme und Schrift, Bild und Ton*, herausgegeben von Arnulf Deppermann und Angelika Linke. Berlin: de Gruyter, 301–324.

Jäger, Ludwig. 2008. „Transkriptive Verhältnisse. Zur Logik intra- und intermedialer Bezugnahmen in ästhetischen Diskursen." In *Transkription und Fassung in der Musik des 20. Jahrhunderts: Beiträge des Kolloquiums in der Akademie der Wissenschaften und der Literatur, Mainz, vom 5. bis 6. März 2004*, herausgegeben von Gabriele Buschmeier, Ulrich Konrad und Albrecht Riethmüller. Stuttgart: Steiner, 103–134.

Kay, A. und A. Goldberg. 1977. „Personal Dynamic Media." *Computer* 10 (3): 31–41. doi: 10.1109/C-M.1977.217672.

Keim, Daniel A., Jörn Kohlhammer, Geoffrey Ellis und Florian Mansmann. 2010. *Mastering the Information Age – Solving Problems with Visual Analytics*. Goslar: Eurographics Association. http://diglib.eg.org/handle/10.2312/14803 (letzter Zugriff am 8. Dezember 2017).

Keim, Daniel A., Florian Mansmann, Jörn Schneidewind und Hartmut Ziegler. 2006. „Challenges in Visual Data Analysis." In *Tenth International Conference on Information Visualization (IV 2006)*, 9–16. https://doi.org/10.1109/IV.2006.31.

Kittler, Friedrich. 1989. „Die künstliche Intelligenz des Weltkriegs: Alan Turing." In *Arsenale der Seele. Literatur- und Medienanalyse seit 1870*, herausgegeben von Friedrich Kittler und Georg Christoph Tholen. München: Fink, 187–202.

Kittler, Friedrich. 1993. *Draculas Vermächtnis. Technische Schriften.* Leipzig: Reclam.

Knorr Cetina, Karin. 2001. „‚Viskurse' der Physik. Konsensbildung und visuelle Darstellung". In Heintz, Bettina; Huber, Jörg (Hrsg.) *Mit dem Auge denken: Strategien der Sichtbarmachung in wissenschaftlichen und virtuellen Wellen.* Zürich: Wien; New York: Voldemeer; Springer, 305–320 (Theorie:Gestaltung).

Krämer, Sybille. 2009. „Operative Bildlichkeit. Von der ‚Grammatologie' zu einer ‚Diagrammatologie'?" In *Logik des Bildlichen. Zur Kritik der ikonischen Vernunft,* herausgegeben von Martina Heßler und Dieter Mersch. Bielefeld: Transcript, 94–123 (Metabasis).

Krämer, Sybille. 2012a. „Punkt, Strich, Fläche. Von der Schriftbildlichkeit zur Diagrammatik." In *Schriftbildlichkeit. Wahrnehmbarkeit, Materialität und Operativität von Notationen,* herausgegeben von Sybille Krämer, Eva Cancik-Kirschbaum und Rainer Totzke. Berlin: Akademie, 79-100.

Krämer, Sybille. 2012b. „Was ist eigentlich eine Karte? Wie Karten Räume darstellen und warum Ptolemaios zur Gründerfigur wissenschaftlicher Kartografie wird." In *Politische Räume in vormodernen Gesellschaften. Gestaltung – Wahrnehmung – Funktion,* herausgegeben von Friederike Fless, Rudolf Haensch, Felix Pirson, Susanne Sievers, Ortwin Dally, Rahden/Westf.: Leidorf, 47–53.

Kuhn, Thomas S. 1996. *Die Struktur wissenschaftlicher Revolutionen.* 13. Aufl. Frankfurt am Main: Suhrkamp.

Leech, Geoffrey 2000. „Grammars of Spoken English: New Outcomes of Corpus-Oriented Research." *Language Learning* 50 (4): 675–724, doi: 10.1111/0023-8333.00143.

Levy, Steven. 2010. *Hackers: Heroes of the Computer Revolution – 25th Anniversary Edition.* Sebastopol, CA: O'Reilly and Associates.

Mackenzie, Adrian. 2006. *Cutting Code: Software And Sociality* . New York: Lang (Digital Formations).

Manovich, Lev. 2014. „Software is the Message." *Journal of Visual Culture.* 13 (1): 79–81, doi: 10.1177/1470412913509459.

Manovich, Lev. 2002. *The Language of New Media.* Reprint. Cambridge, Mass.: The MIT Press.

Meier, Stefan, Martin Reisigl und Alexander Ziem. 2014. „Vom (Kon-)Text zum Korpus. Ein diskursanalytisches Kamingespräch." In *Diskursforschung: Ein interdisziplinäres Handbuch.* Berlin: de Gruyter, 436-464.

Moles, Abraham. 1959. „Kybernetik, eine Revolution in der Stille." In *Epoche Atom und Automation: Enzyklopädie des technischen Jahrhunderts in zehn Bänden; VII: Kybernetik, Elektronik, Automation.* Genf: Limpert, 7.

Möller, Torger. 2007. „Kritische Anmerkungen zu den Begriffen Denkkollektiv, Denkstil und Denkverkehr – Probleme der heutigen Anschlussfähigkeit an Ludwik Fleck." In *Von der wissenschaftlichen Tatsache zur Wissensproduktion: Ludwik Fleck und seine Bedeutung für die Wissenschaft und Praxis,* herausgegeben von Bożena Chołuj und Jan C. Joerden. Frankfurt am Main : Lang, 397–413.

Naumann, Carl Ludwig. 1982. „Kartographische Datendarstellung". In *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung.* Berlin: de Gruyter, 667–692 (Handbücher zur Sprach- und Kommunikationswissenschaft).

Niehr, Thomas. 2015. „Die Universität im öffentlichen Sprachgebrauch. Ein Plädoyer für das Zusammenwirken von quantitativen und qualitativen Methoden der Diskursforschung." In *Universität, Öffentlichkeit: Festschrift für Jürgen Schiewe,* herausgegeben von Kersten Sven Roth, Jürgen Spitzmüller, Birte Arendt und Jana Kiesendahl. Bremen: Hempen, 134–146.

Peirce, Charles S., Charles Hartshorne, Paul Weiss und Arthur W. Burks, Hrsg. 1994. *The collected papers of Charles Sanders Peirce.* Charlottesville: InteLex Corp.

Placcius, Vincentius. 1689. *De arte excerpendi. Vom Gelahrten Buchhalten.* Hamburg: Liebezeit.

Ramelli, Agostino. 1588. *Le diverse et artificiose machine del capitano Agostino Ramelli : nellequali si contengono varij et industriosi movimenti, degni digrandissima speculatione, per cavarne beneficio infinito in ogni sorte d'operatione : composte in lingua Italiana et Francese.* doi: 10.3931/e-rara-8944.

Redder, Angelika. 2001. „Aufbau und Gestaltung von Transkriptionssystemen". In *Text- und Gesprächslinguistik / Linguistics of Text and Conversation.* Berlin: de Gruyter, 1038–1059 (Handbücher zur Sprach- und Kommunikationswissenschaft).

Rehbein, Jochen, Thomas Schmidt, Bernd Meyer, Franziska Wazke und Annette Herkenrath. 2004. *Handbuch für das computergestützte Transkribieren nach HIAT.* Hamburg: Universität Hamburg (Arbeiten zur Mehrsprachigkeit, Folge B. Working Papers in Multilingualism, Series B).

Rheinberger, Hans-Jörg. 1994. „Experimentalsysteme, Epistemische Dinge, Experimentalkulturen: Zu einer Epistemologie des Experiments." *Deutsche Zeitschrift für Philosophie* 42 (3): 405–418.

Sachs, Klaus-Jürgen und Thomas Röder. 1989. „Partitur." In *Die Musik in Geschichte und Gegenwart,* herausgegeben von Ludwig Finscher (Die Musik in Geschichte und Gegenwart).

Sacks, Harvey; Emanuel A. Schegloff und Jefferson, Gail. 1974. „A Simplest Systematics for the Organization of Turn-Taking for Conversation." *Language* 50 (4): 696–735, doi: 10.2307/412243.

Schneider, Albrecht. 1987. „Musik, Sound, Sprache, Schrift: Transkription und Notation in der Vergleichenden Musikwissenschaft und Musikethnologie." *Zeitschrift für Semiotik* 9 (3–4): 317–343.

Scholz, Ronny und Annika Mattissek. 2014. „Zwischen Exzellenz und Bildungsstreik Lexikometrie als Methodik zur Ermittlung semantischer Makrostrukturen des Hochschulreformdiskurses." In *Diskursforschung: Ein interdisziplinäres Handbuch*. Berlin: de Gruyter, 86-112.

Scott, Brett. 2015. *How yuppies hacked the hacker ethos. Aeon Magazine*. https://aeon.co/essays/how-yuppies-hacked-the-original-hacker-ethos (Letzter Zugriff am 10. August 2015).

Selting, Margret, Peter Auer, Birgit Barden, Jörg Bergmann, Elizabeth Couper-Kuhlen, Susanne. Günthner, Christoph Meier, Uta Quasthoff, Peter Schlobinski, Susanne Uhmann. 1998. „Gesprächsanalytisches Transkriptionssystem (GAT)." *Linguistische Berichte* 173: 91–122.

Siegel, Steffen. 2009. *Tabula: Figuren der Ordnung um 1600*. Berlin: Akademie-Verlag.

Spitzmüller, Jürgen und Ingo H. Warnke. 2011. *Diskurslinguistik: eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse*. Berlin: de Gruyter.

Steinseifer, Martin. 2013. „Texte sehen – Diagrammatologische Impulse für die Textlinguistik". *Zeitschrift für germanistische Linguistik* 41 (1): 8–39.

Stetter, Christian. 2005. „Bild, Diagramm, Schrift". *Schrift. Kulturtechnik zwischen Auge, Hand und Maschine*. München: Fink, 115–136 (Kulturtechnik).

Stocker, Christa, Daniela Macher; Rebekka Studler, Noah Bubenhofer, Daniel Crvelin, Reto Liniger, Martin Volk. 2004. *Studien-CD Linguistik: multimediale Einführungen und interaktive Übungen zur germanistischen Sprachwissenschaft*. Tübingen: Niemeyer.

Turing, Alan M. 1936. „On Computable Numbers". *Proceedings of the London Mathematical Society, 2nd series* 42 (3-4): 230–265.

Wall, Larry. 1999. *Perl, the first postmodern computer language*. http://www.wall.org/~larry/pm.html (letzter Zugriff am 19. August 2015).

Rainer Perkuhn / Marc Kupietz

Visualisierung als aufmerksamkeitsleitendes Instrument bei der Analyse sehr großer Korpora

Abstract Sehr große Korpora – wie das Deutsche Referenzkorpus DeReKo – bieten eine breite Basis für die empirische Forschung. Sie bringen aber auch Herausforderungen mit sich, da sich weder Eigenschaften ihrer Zusammensetzung noch derer von Recherche- und Analyseergebnissen mit einfachen Mitteln erschließen lassen. Dafür bedarf es Verfahren geschickter Sortierung, Gruppierung oder des Clusterings, kurzum: strukturentdeckender Methoden. In Kombination mit Visualisierungstechniken kann so die Wahrnehmung bestimmter Eigenschaften und Zusammenhänge unterstützt und die Aufmerksamkeit auf bestimmte Phänomene, ggf. in Anlehnung an präferenzrelationale Befunde, gelenkt werden. Neben der illustrativen Funktion geht es in diesem Beitrag vor allem um das erkenntnisleitende Potenzial derartiger Verfahren in Kombination. Aus verschiedenen Bereichen werden Beispiele gezeigt, die am IDS oder in Kooperationen zum Einsatz kommen, sowohl zur dokumentarischen und reflexiven Kontrolle von Eigenschaften der Korpuszusammensetzung als auch hinsichtlich korpusanalytischer Methodik, um die qualitative Interpretation von Analysebefunden und die Abduktion von Hypothesen stimulierend zu unterstützen.

1. Einleitung

Visualisierung, insbesondere die Forschung zur Visualisierung, ist nicht das primäre Forschungsfeld der Projekte, aus deren Arbeit im Folgenden berichtet wird. Die Schwerpunkte der Projekte sind vielmehr angesiedelt im Umfeld des Deutschen Referenzkorpus (DeReKo) des Instituts für Deutsche Sprache (IDS, Institut für Deutsche Sprache 2016a). Der Bericht soll daraus überblicks- und querschnittsartig Themenbereiche kurz vorstellen, in denen Visualisierungstechniken zum Einsatz kommen. Neben vereinzelten Anwendungen zu illustrativen Zwecken wird – und perspektivisch in noch stärker zunehmendem Maße – ein Aspekt in den Mittelpunkt gerückt, den Schumann/Müller (2000) als Informationsvisualisierung

Erschienen in: Noah Bubenhofer u. Marc Kupietz (Hrsg.): Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

dem Bereich Data Mining bzw. Knowledge Discovery in Databases (KDD) zuordnen. Als Fortführung unseres Forschungsparadigmas, das das Aufspüren präferenz-relationaler Zusammenhänge im Sprachgebrauch zum Ziel hat, setzt diese Art der Visualisierung auf „die Idee, das menschliche visuelle System mit seinen unnachahmlichen Fähigkeiten zum Auffinden von Strukturen und Korrelationen zur Analyse der Informationen zu nutzen" (ebda, S. 342). Ihre Aufgabe in diesem Kontext ist z. T. weniger die einer adäquaten und objektiven Ergebnisdarstellung als vielmehr die eines Hilfsmittels zur Abduktion vielversprechender Hypothesen. Die Güte des gesamten Vorgehens hängt aber nicht nur von einer geeigneten Visualisierung ab, sondern insgesamt von einem harmonischen Zusammenspiel einer angemessenen Datengrundlage, geeigneter Analyseverfahren und der Fähigkeit, die durch die Visualisierung hervorgehobenen Aspekte zu interpretieren. Die Rolle des Interpretierenden übernehmen wir dabei teilweise vollständig selbst. Insbesondere bei der Erschließung der Zusammensetzung des Korpus fließen die Erkenntnisse in die Dokumentation und rückgekoppelt in die weitere Akquisitionsstrategie ein. Für die Entwicklung oder Verfeinerung von Analysemethoden beziehen wir darüber hinaus als Interpretierende aber auch die Methodenanwender für ihre diversen linguistischen Fragestellungen mit ein.

Der stärker rückgekoppelte Einsatz der Visualisierungstechniken ergibt sich in unserem Umfeld aus der extremen Größe des Archivs und einer zum Teil „opportunistischen" Akquisitionsstrategie. Im Vergleich zu anderen Korpora, die gezielt nach vorgegebenen Kriterien aufgebaut werden, wird das Archiv des IDS dynamisch weiterentwickelt und kontinuierlich ausgebaut. Der Erfolg der Bestrebungen hängt dabei von konzeptuell Gewünschtem ab, aber natürlich auch von dem, was rechtlich, finanziell und technisch – grundsätzlich von den gegebenen Kapazitäten her – machbar ist (vgl. Kupietz/Schmidt 2015). Da das Archiv als Ur-Stichprobe dienen soll, ist eine ungleichmäßige Zusammensetzung der Daten weniger relevant. Jeder Nutzende kann eine Arbeitsversion aus dieser Datensammlung als ein sogenanntes virtuelles Korpus zusammenstellen (vgl. Kupietz et al. 2010). Trotzdem ist es natürlich hilfreich, möglichst viele Informationen über die Zusammensetzung der Daten zu sammeln, einerseits, um den Nutzenden bei der Definition des Arbeitskorpus zu unterstützen, andererseits, um Bereiche aufzuspüren, die noch besser ausgebaut werden könnten.

Zu Anfang des nächsten Abschnitts wollen wir aber zunächst am Beispiel eines als ausgewogen geplanten Korpus zeigen, wie Visualisierung im primär illustrierenden Sinne eingesetzt werden kann. Anstelle von hierarchischen Strukturen, die für didaktische und/oder dokumentarische Zwecke syntaktische oder morpho-syntaktische Zusammenhänge veranschaulichen oder auch die Genese und Verwandtschaft von Sprachfamilien (vgl. Bubenhofer in diesem Band, S. 63), deuten wir nur einige Diagrammtechniken für Verteilungen an, wie sie in den meisten Tabellenkalkulationsprogrammen integriert angeboten werden.

2. Eigenschaften/Zusammensetzung des Archivs

Korpora werden vielfach nach bestimmten Vorgaben geplant, etwa dass (zumindest abschnittsweise) ein bestimmter Umfang angestrebt wird oder dass die Zusammensetzung nach bestimmten Kriterien (wie Textsorte o. Ä.) einer bestimmten Verteilung folgt. Für das DWDS-Kernkorpus ist laut Webseite (http://www.dwds.de/ressourcen/kernkorpus/, letzter Zugriff am 22. August 2016) ein Umfang von 10.000.000 Token je Dekade des 20. Jahrhunderts und eine Zusammensetzung von Belletristik : Gebrauchsliteratur : Wissenschaft : Zeitung im Verhältnis 28,42% : 21,05% : 23,15% : 27,36% vorgegeben (verschriftlichte gesprochene Sprache ist bei diesen Zusammenstellungen herausgenommen). Ohne die Plausibilität dieser Vorgaben diskutieren zu wollen, zeigen wir hier eine selbst erzeugte Grafik für die geplante Verteilung nach Textsorten, ein sogenanntes Tortendiagramm (s. Abb. 1).

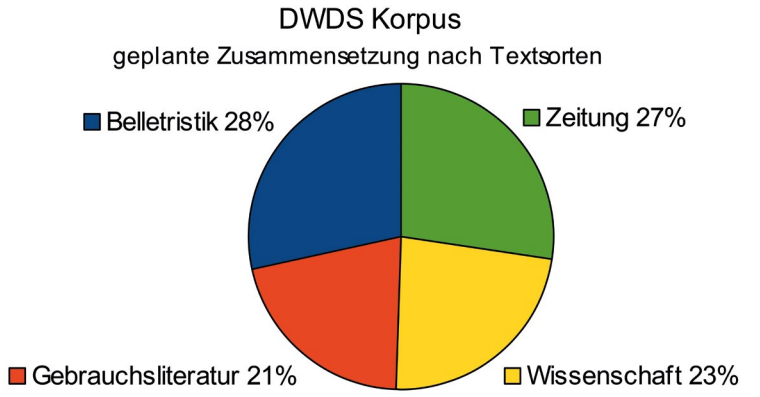


Abb. 1: Geplante Verteilung nach Textsorten (DWDS-Kernkorpus) (Copyright IDS).

Die reine Darstellung der SOLL-Werte liefert nur wenig zusätzliche Information – am wenigsten bei einer Darstellung der Umfänge pro Jahrzehnt. Interessant wird es aber auch in diesem Umfeld schon, wenn die geplanten und die tatsächlich erreichten Werte nebeneinandergestellt präsentiert werden. Eine Darstellung der beiden Dimensionen getrennt voneinander lässt sich noch gut umsetzen (vgl. Abb. 2).

Der Versuch, beide Dimensionen zusammenzuführen, zeigt ansatzweise die Grenzen des Machbaren auf (vgl. Abb. 3). Um die Bereiche lokalisieren zu können, die auf der Webseite als unterbesetzt erwähnt werden, braucht es ein glückliches Händchen für die Wahl der Perspektive bzw. der Anordnung in der Tiefe zwischen Vorder- und Hintergrund.

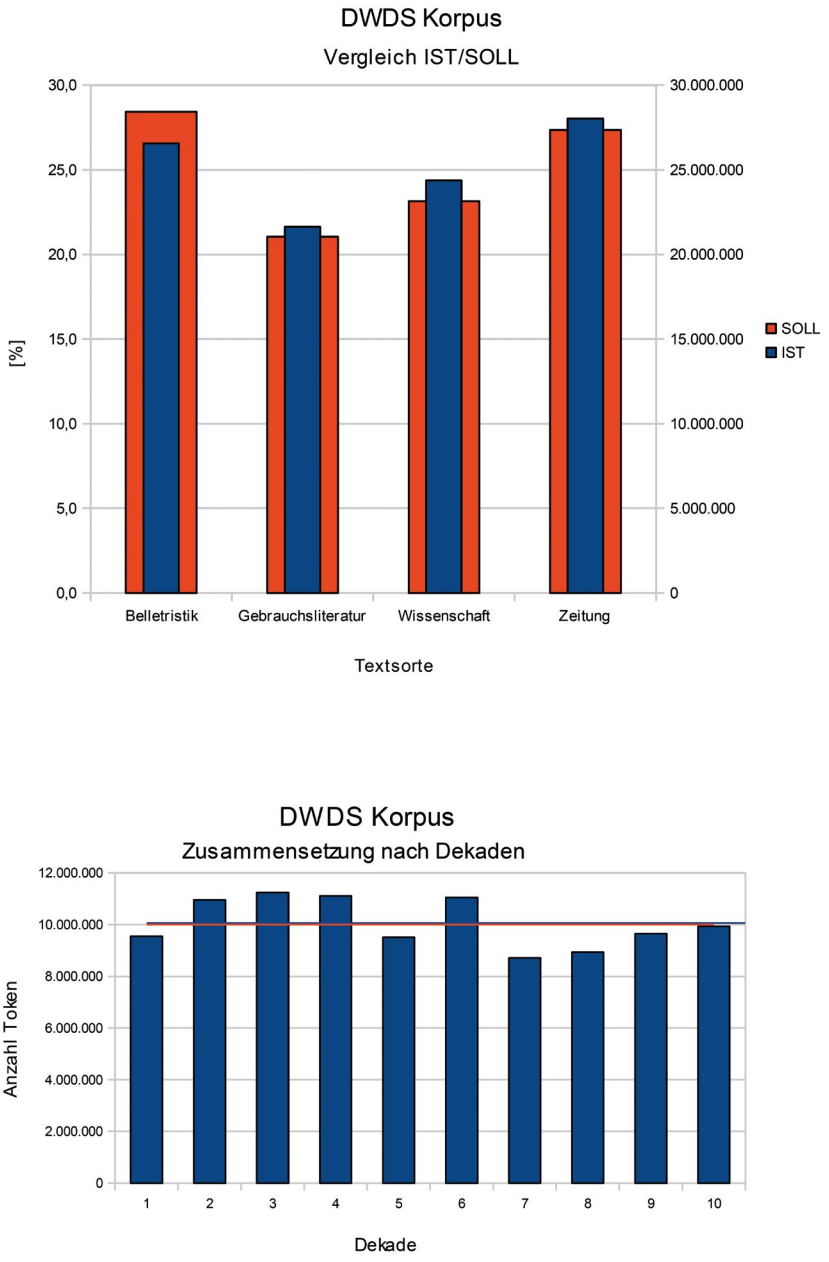


Abb. 2: Vergleich IST/SOLL getrennt nach Textsorte bzw. Umfang (DWDS-Kernkorpus) (Copyright IDS).

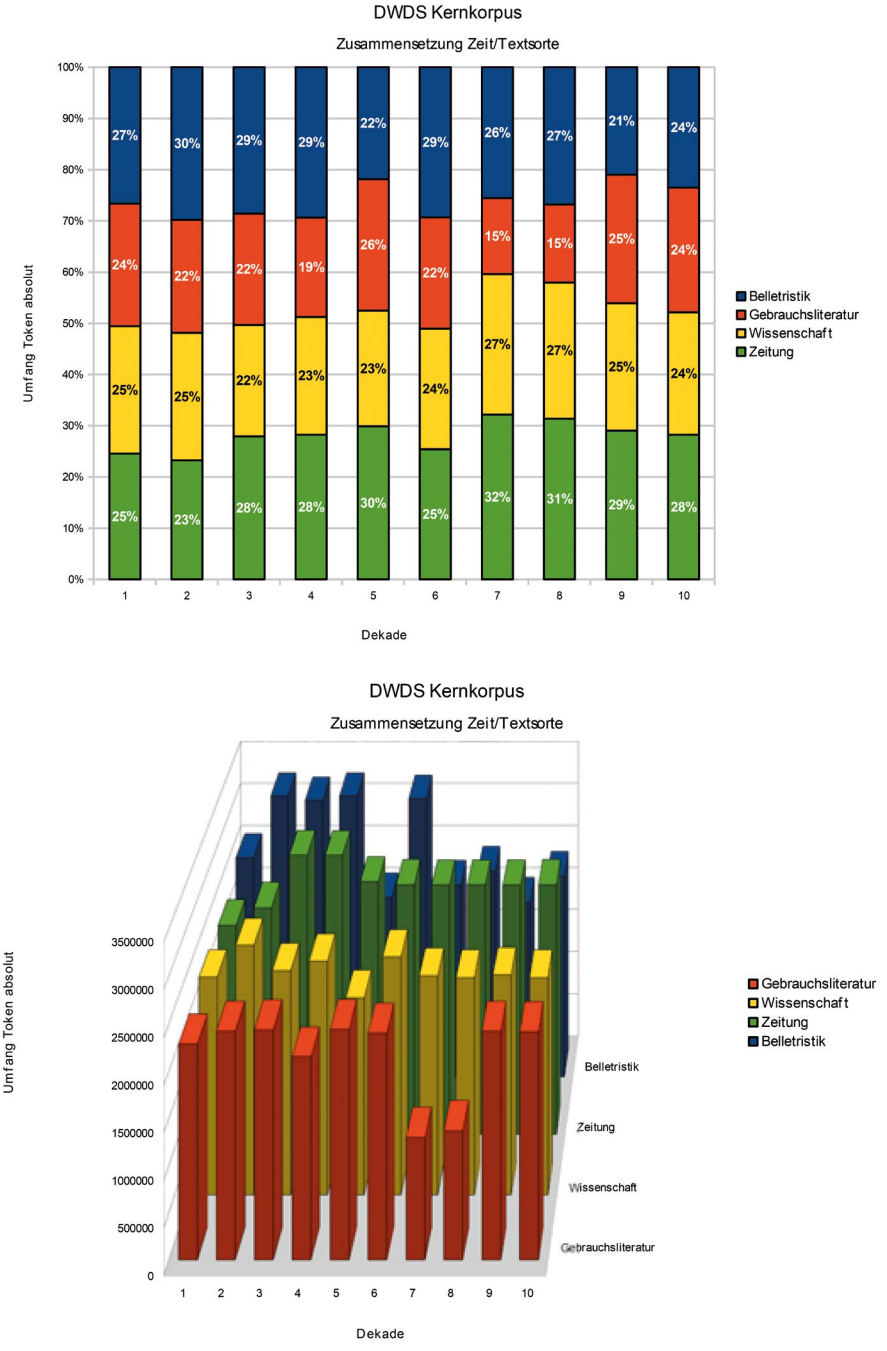


Abb. 3: Differenzierung der Dekaden-IST-Werte nach Textsorte und Umfang, gestapelt vs. dreidimensional (DWDS-Kernkorpus) (Copyright IDS).

Den Umfang der Korpora kumuliert nach Dekaden vor- und anzugeben, eröffnet einen schwer einzuschätzenden Spielraum für die Verteilung auf die einzelnen Jahre von absoluter Gleichverteilung bis hin zu extremen Schieflagen. Um die Zusammensetzung des IDS-Archivs zu dokumentieren, beziehen wir uns auf die Einheit „pro Publikationsjahr“ (vgl. Abb. 4).

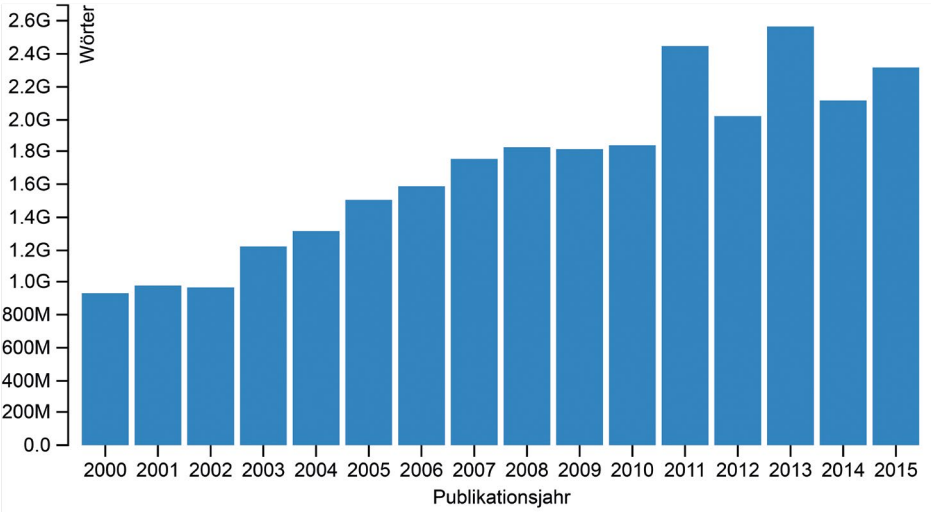


Abb. 4: DeReKo-Umfang pro Publikationsjahr ab 2000 (Copyright IDS).

Die Botschaft dieses Diagramms ist nicht die Dokumentation des Vergleichs mit einem geplanten Ziel. Auch wenn Recherchen im gesamten Archiv durchaus sinnvoll sein können, soll dieser ansteigende Verlauf vor allem dafür sensibel machen, dass Trefferzahlen pro Jahr nur in engen Bereichen absolut miteinander verglichen werden können. Auch die Relativierung auf den jeweiligen Jahresumfang kann bei extremen Abweichungen zu Überraschungen führen. Für Untersuchungen, die auf die Dimension Zeit schauen, ist es in vielen Fällen ratsam, ein virtuelles Korpus zu definieren, dessen Umfang je nach zugrunde gelegter Zeiteinheit nicht allzu sehr schwankt (vgl. Abschnitt zu Zeitverläufen). Interessant hierfür sind vor allem auch Zeitungsdaten, da sie sich – im Vergleich zu internetbasierter Kommunikation – gut datieren lassen. Zudem erscheinen Zeitungstexte – im Kontrast zu Belletristik – in einem kurzen Takt, und Textproduktion und -publikation liegen zeitlich nah beieinander. Sie bilden verhältnismäßig gut und vor allem eben zeitnah den allgemeinen Sprachgebrauch ab, zugegebenermaßen stark vom Zeitgeschehen, teilweise mit einem gewissen Lokalkolorit, beeinflusst.

2.1 Geografische Verteilung der Archiv-Quellen

Zeitungsdaten, auf die wir uns in diesem Abschnitt konzentrieren wollen, weisen regionale Unterschiede auf. In geringem Maße ergibt sich dies sicher auch aus dem umgebungsbedingten Substandard der allgemeinen Sprache (wobei vermutlich die wenigsten Redakteure „unverfälschte" gebürtige Sprecher der jeweiligen Region sind). Vielfach sind aber gerade die Themen, über die geschrieben wird, und somit Wortwahl und das Vokabular stark durch die Region geprägt. Für Untersuchungen, die gerade gezielt darauf eingehen wollen oder die Effekte in einem größeren Zusammenhang gedämpft sehen möchten, wird häufig der Wunsch an uns herangetragen, mehr Daten aus möglichst vielen verschiedenen Regionen zur Verfügung zu stellen. In einem benachbarten IDS-Projekt „Deutsch heute", das sich zum Ziel gesetzt hat, flächendeckend Audio-Aufnahmen für ein Korpus gesprochener Sprache zu erheben, wurde dieses Bestreben Teil der Projektplanung (vgl. Abb. 5).

Demhingegen wollen wir für die schriftsprachlichen Korpora im Allgemeinen nicht zur Textproduktion auffordern, sondern uns bei Quellen bedienen, die in einem „natürlichen" Prozess Texte gestalten. Eine entsprechende Karte wäre deutlich dünner besetzt als die in Abb. 5 gezeigte, liegt uns aber nicht vor. Anstelle der Produktionsorte haben wir die Verlagsorte unserer Textspender auf eine Karte geplottet (vgl. Abb. 6). Sie stellt einen Ausschnitt aus Mitteleuropa dar, im Kern bestehend aus Deutschland und den Nachbarländern, in denen Deutsch zumindest als Minderheitensprache gesprochen wird. Dazu haben wir auf die Orte Kreissignaturen platziert, die in ihrer Größe dem Umfang der Texte entsprechen, die aus den Quellen des gleichen Ortes in unser Archiv eingespeist sind.

Um den Erfolg einer der letzten Akquisitionsbestrebungen zu dokumentieren, sind die Kreise in unterschiedlichen Farben jeweils für die Zeit vor und nach der Aktion markiert. Wie die Karte zeigt, sind sehr viele Lücken geschlossen worden. Sie zeigt aber auch, dass auch vor der Akquisition der Schwerpunkt gar nicht so sehr durch die Großregion Mannheim geprägt war (was dem Archiv gelegentlich unterstellt wird) – und dass noch einige Lücken geblieben sind. Neben vielen anderen Aspekten bleibt bei dieser Darstellung auch unberücksichtigt, wie sich die Bevölkerung auf die Regionen verteilt und wie hoch etwa die Auflagenstärke der Printmedien ist, um womöglich einen rezeptiven Wirkungsgrad abschätzen zu können.

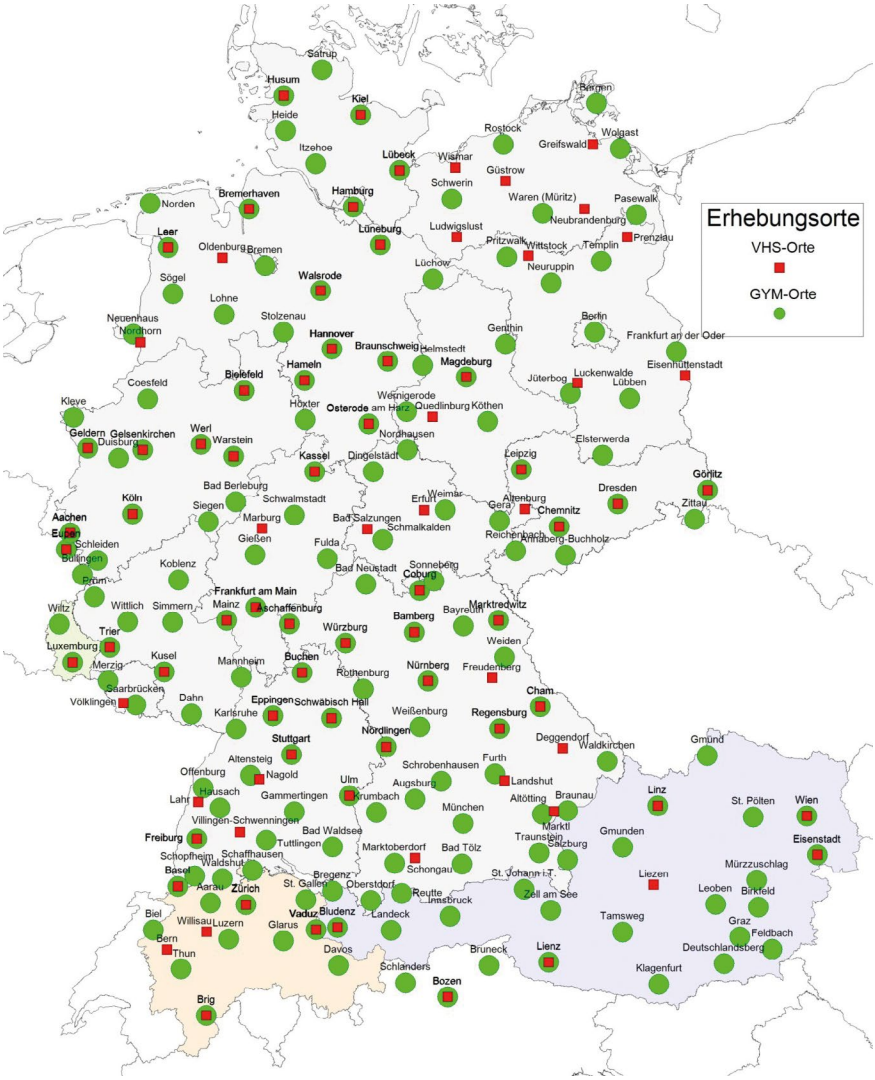


Abb. 5: Geplante Orte der Datenerhebung (Deutsch heute) (Copyright IDS, http://www1.ids-mannheim.de/fileadmin/prag/AusVar/Deutsch_heute/Erhebungsorte_DH_70.jpg).

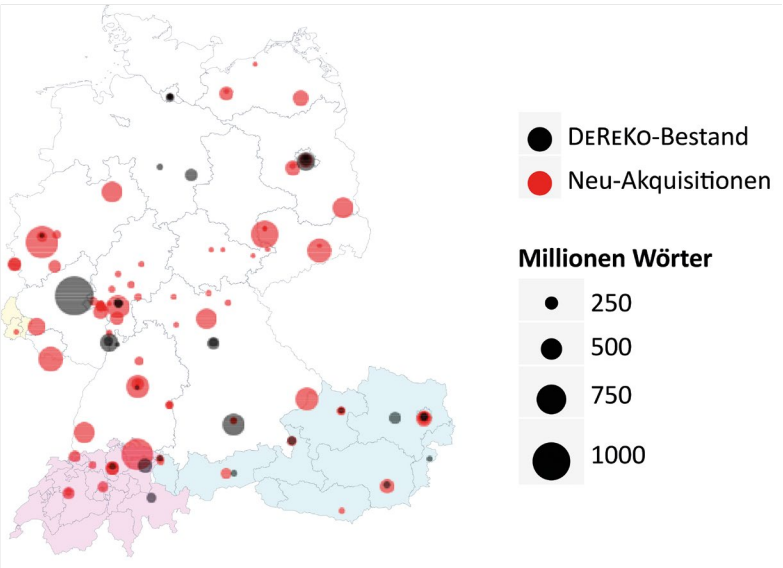


Abb. 6: Regionale Verteilung der bestehenden DeReKo-Zeitungsquellen und der in 2013 neu akquirierten Quellen. (Copyright IDS)

2.2 Textdublettenerkennung

Aus verschiedenen Gründen kann es vorkommen, dass nahezu identische Texte zur Aufnahme in das Archiv bereitgestellt werden. Dies kann versehentlich zustande kommen aufgrund von Überschneidungen des redaktionellen Prozesses mit der Übernahme der Daten in unsere Arbeitsabläufe. In anderen Fällen kann es sich um schematisch angelegte Texte handeln, wie etwa Wetterberichte, Kinoankündigungen oder -rezensionen, oder auch um Variationen von Agenturmeldungen. Während im ersten Fall die Einschätzung relativ eindeutig ausfallen sollte, dass es sich dabei um (echte) unerwünschte Textdubletten handelt, ist dies bei den anderen Fällen weniger klar. Je nach Fragestellung (z. B. Untersuchung der Produktion vs. Untersuchung der Rezeption) können gerade die Variationen von Serientexten zum Gegenstand der Betrachtung werden. Die Ähnlichkeiten zwischen den Texten unseres Archivs werden mithilfe eines Dublettenerkennungsverfahrens (Kupietz 2005) ermittelt, dem zum Zweck einer einfacheren Qualitätsprüfung eine Visualisierung nachgelagert ist. Diese basiert auf einer einfachen Alignierung und farblichen Hervorhebung der Gemeinsamkeiten und Unterschiede, setzt damit aber genau die erforderliche Expressivität um (vgl. Abb. 7).

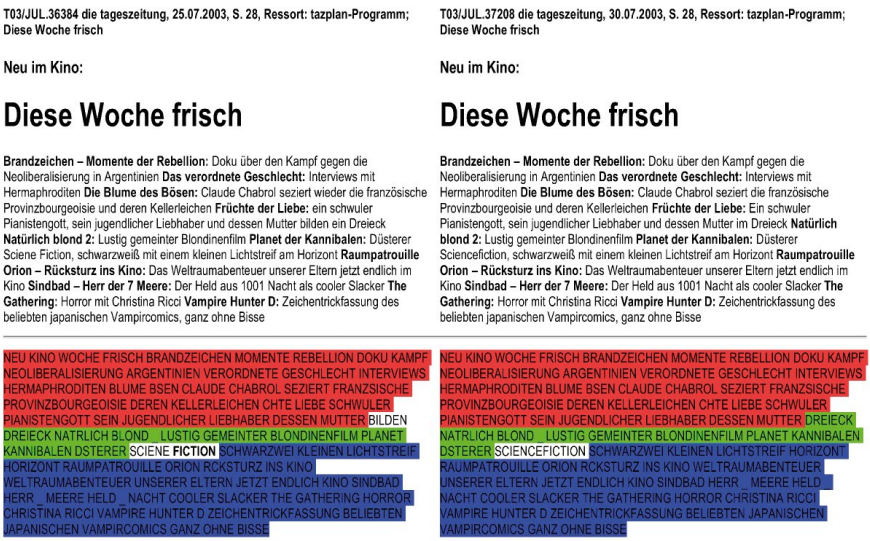


Abb. 7: Visualisierung von Textdubletten (DeReKo) (Copyright IDS).

Die Dublettenerkennung wird in erste Linie dazu eingesetzt, um die Beziehungen zwischen den Texten zu dokumentieren und als Metadaten festzuhalten. Nur in eindeutigen Fällen werden die Texte tatsächlich aussortiert. Perspektivisch soll es den Nutzenden virtuell ermöglicht werden, diese Entscheidung für die Gesamtmenge der markierten Texte über einen einstellbaren Schwellwert selbst zu treffen.

2.3 Inhaltliche Eigenschaften der Archiv-Quellen

Neben der stärksten Ausprägung der Ähnlichkeit von Texten als nahezu vollständige Übereinstimmung gibt es weitere Beweggründe, um Texte oder ganze Korpora bezüglich weicherer Aspekte zu vergleichen. Zum Beispiel: Entstammen Texte etwa ähnlichen Registern oder handeln sie von ähnlichen Themen? Bereits beim Vergleich unseres Archivs mit dem zum Zeitpunkt der Untersuchung größenmäßig vergleichbaren webbasierten Korpus deWaC (Baroni et al. 2009) hat sich gezeigt, wie aussagekräftig Vergleiche schon allein auf der Ebene des lexikalischen Inventars, also des Vokabulars, sind. In Fortführung dieser Gedanken haben wir alle Teilkorpora unseres Archivs mithilfe eines Maßes von Kilgarriff (2001) verglichen, das auf den vorderen Ausschnitten der frequenzsortierten Vokabulare basiert (Kupietz et al. 2012). Hintergrund der

Studie war der Versuch, wiederum aus einer webbasierten Ressource durch optimierende wiederholte Stichprobenziehungen eine Sammlung von Texten („litDeWaC“) zusammenzustellen, die einem echten Literaturkorpus („lit“: zusammengesetzt aus allen Belletristik-Teilkorpora aus DeReKo) möglichst nahekommt. In Abb. 8 sind die Abstände zwischen den Korpora mithilfe nichtmetrischer multidimensionaler Skalierung (NMDS) zweidimensional projiziert (vgl. Cox & Cox 2001). Sie zeigt durchaus plausible räumliche Anordnungen, wobei sich in diesem Fall die Semantik des Hintergrunds erst aus der Interpretation der Gruppen ergibt.

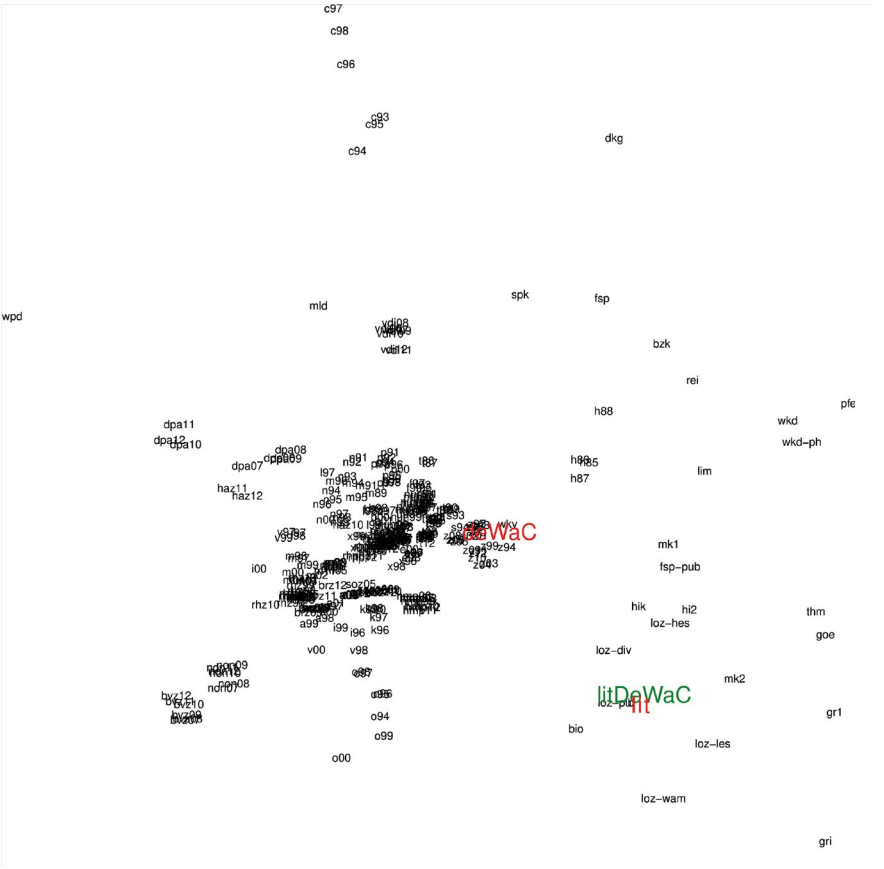


Abb. 8: Ergebnisdarstellung der Extraktion eines Pseudo-Literaturkorpus litDeWaC nach dem Vorbild des Korpus lit aus deWaC vor dem Hintergrund aller DeReKo-Teilkorpora (weitere Abkürzungen siehe http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html (interaktiv)). Die Ähnlichkeitsmatrix der Teilkorpora wurde wie in Kilgarriff (2001) mit Spearmans Rangkorrelationskoeffizienten berechnet. Zur Projektion der Ähnlichkeitsmatrix auf eine Karte wurde NMDS verwendet. (Copyright IDS)

Dabei ist weniger die Positionierung des neuen Korpus litDeWaC überraschend, da dessen Zusammenstellung ja gerade auf die Nähe zu bekannten Literaturdaten (lit) ausgerichtet war. Beeindruckend ist aber schon das Gesamtbild, das sich aus der Anordnung der verschiedenen Teilkorpora abzeichnet, beispielsweise im unteren rechten Bereich die Nähe der LOZ-Korpora (**L**iteratur, **O**riginalsprache: Deutsch, des **Z**wanzigsten Jahrhunderts) untereinander und zu den Korpora einzelner Schriftsteller (thm – Thomas Mann, goe – Goethe), aber auch zu den Märchen der Gebrüder Grimm (gri). Insgesamt deutet sich eine Topographie anhand verschiedener Texteigenschaften wie Register, Medium, Textsorte, Thematik u. Ä. an: Im mittleren rechten Bereich schimmern etwa Aspekte mündlicher Interaktion durch, durch das gemeinsame Arrangement der Korpora „Reden und Interviews“ (rei), dem Wendekorpus (wkd) und dem Pfeffer-Korpus (pfe).

3. Ergebnisübersichten der Treffermengen

Neben den bisher diskutierten, stärker kumulativen Betrachtungen von Texten oder Korpora finden Visualisierungstechniken auch bei eher wortbezogenen Auswertungen Anwendung. Bereits seit den Anfängen der Korpuslinguistik werden die auf die Suchobjekte passenden Texteinheiten bei der Beleganzeige durch Textattribute (fett, Farbe) hervorgehoben. Bei der auf einen unmittelbaren Vergleich ausgerichteten kompakten Darstellung einer Konkordanz wird dies zusätzlich durch die positionelle Anordnung unterstützt: Mehrere Treffer werden zeilenweise untereinander so angeordnet, dass die gefundenen Objekte mittig aligniert untereinander platziert werden. Nach links und rechts werden dann – je nach Platz aufgrund des genutzten Mediums – gleichermaßen so viele Zeichen aufgefüllt, wie eine Zeile aufnehmen kann. Während man in den frühen Phasen aufgrund der gegebenen Rahmenbedingungen dafür auf nichtproportionale Schriften zurückgegriffen hat, lässt sich dies mit heutigen Mitteln auch mit proportionalen Schriften quasi tabellenartig darstellen.

Vor allem für große Treffermengen ist es oft hilfreich, sich zunächst anhand eines Ergebnisüberblicks einen ersten Eindruck zu verschaffen. Wie viele Treffer gibt es etwa pro Jahr, wie viele pro Quelle oder je Region? So wie oben aber bereits angedeutet, sind absolute Häufigkeiten nur bei annähernd gleich großen Schnitten vergleichbar. Aber auch relative Frequenzen verlieren dann ihre Aussagekraft, wenn die Schnitte um mehrere Größenordnungen auseinanderliegen. Das Recherchesystem des IDS Cosmas II (Bodmer Mory 2014, Institut für Deutsche Sprache 2016b) bietet Ergebnisübersichten nur in tabellarischer Form an. Diagramme wie die hier gezeigten können mit anderen Softwaretools auf der Grundlage dieser Angaben in einem weiteren

nachgelagerten Bearbeitungsschritt erzeugt werden, wie das Balkendiagramm für die Häufigkeiten pro Jahr oder das Tortendiagramm für die Verteilung nach Quellen, Themen oder Region (vgl. Abb. 9 und 10).

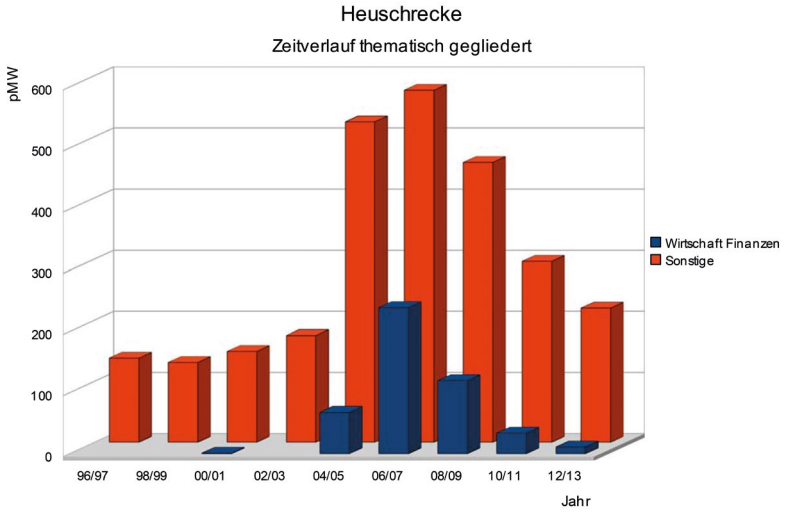


Abb. 9: Thematische Verteilung der Ergebnismenge zu dem Lemma „Heuschrecke“ (DeReKo/Cosmas II) (Copyright IDS).

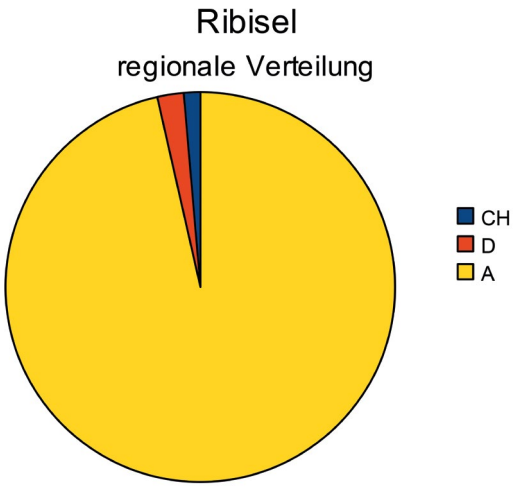


Abb. 10: Regionale Verteilung der Ergebnismenge zu dem Lemma „Ribisel“ (DeReKo/Cosmas II) (Copyright IDS).

Die Darstellung der regionalen Verteilung für das Wort *Ribisel* zeigt dabei einen eindeutigen Befund, der unabhängig von absoluten oder relativen Häufigkeiten ist: Dieses Wort wird fast ausschließlich im österreichischen Deutsch verwendet.

3.1 Ergebnisübersicht der Treffermengen nach Zeit (Zeitverlaufsgrafiken)

Die relativen Vorkommen einer sprachlichen Einheit pro Zeitabschnitt genießen für verschiedene Fragestellungen eine besondere Aufmerksamkeit. Eine geeignete Datengrundlage vorausgesetzt, lassen sich Häufigkeitsverschiebungen als Veränderungen im Sprachgebrauch deuten: als Indizien für die Lebendigkeit bestimmter Diskurse oder, im einfachsten Fall, für das Aufkommen neuer (oder das Aussterben alter) Wörter. Für Wörter, die in ihrer Form neu in einem bestimmten Zeitabschnitt zu beobachten sind, sogenannte Neulexeme (z. B. der Jahre 2000 bis 2010), generieren wir auf der Grundlage eines speziell dafür zusammengestellten virtuellen Korpus Zeitverlaufsgrafiken (Lüngen/Keibel 2013, vgl. Abb. 11), die über das Online-Informationssystem OWID, Rubrik Neologismenwörterbuch, öffentlich angeboten werden.

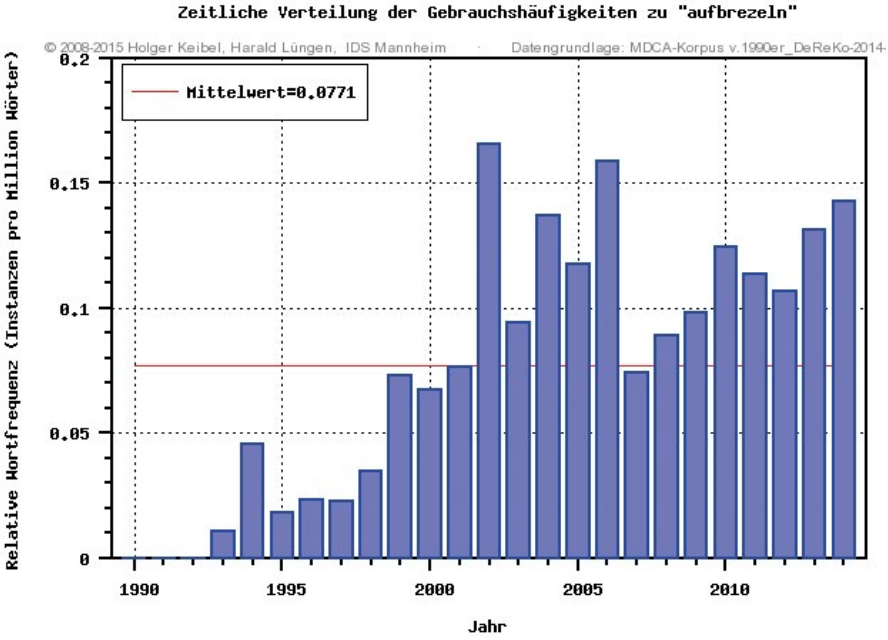


Abb. 11: Zeitverlaufsgrafik eines Neulexems (DeReKo/OWID) (Copyright IDS, http://www.owid.de/artikel/298252, http://www.ids-mannheim.de/kl/neoplots/owid/298252.html).

Neben dieser Visualisierung als Bestätigung eines Befundes setzen wir auf den Gesamtbestand des Korpus auch Verfahren auf, um Neulexem-Kandidaten aufzuspüren (Keibel et al. 2010). Visualisierung spielt in diesem Zusammenhang aber nur eine untergeordnete Rolle, um didaktisch zu illustrieren, dass Neulexem-Kandidaten als notwendige Bedingung einen Zeitverlauf aufweisen müssten, der auf eine schematisch angedeutete Zeitverlaufsschablone passt. Wörter, die eine zusätzliche, neue Bedeutung angenommen haben (Neubedeutungen), lassen sich mit diesen einfachen quantitativen Verfahren nicht aufspüren und somit darauf aufbauend auch nicht visualisieren.

3.2 Ergebnisübersicht der Treffermenge nach Kontext (Kookkurrenz)

Eine andere Form, die Treffermenge eines Suchausdrucks zu sortieren, basiert auf der Idee, dafür die Wörter in der unmittelbaren textuellen Umgebung heranzuziehen, die besonders systematisch in der Nähe des Suchausdrucks vorkommen. Für die Berechnung dieser typischen Wortverbindungen bietet das IDS das parametrisierbare Verfahren der Kookkurrenzanalyse (Belica 1995) an, das auch in dem System Cosmas II integriert ist. Die Präsentation des Analyseergebnisses ähnelt in diesem System, wenn auch mit höherer Komplexität, den bereits bekannten tabellarischen Darstellungen. Gerade für diesen Ergebnisüberblick sind aber speziellere Zugangsformen wünschenswert, da die Skala der Sortierung nicht vorgegeben ist, sondern sich erst aus der Analyse heraus ergibt. Ein Vorgehensmodell für die Erschließung dieser Ergebnisstrukturen wurde für ausgewählte Aspekte als Prototyp operationalisiert (Perkuhn 2007a/b). Für dieses Werkzeug sind die zwei zentralen Aspekte die Visualisierung der Gesamtstruktur und die Möglichkeit, in Form von Annotationen gewonnene Erkenntnisse festhalten zu können, wobei wir auf den zuletzt genannten Aspekt hier nicht weiter eingehen werden. Die Darstellung der Gesamtstruktur steht vor der Herausforderung, dass bei den üblicherweise verwendeten Medien nur begrenzt Raum zur Verfügung steht. Der Ausweg, nur einen kleinen ausgewählten Ausschnitt anzubieten – z. B. über scrollbare Fensterausschnitte wie bei der Cosmas II-Präsentation –, steht im Widerspruch zur Forderung einer Gesamtsicht. Um alle Elemente der Gesamtstruktur anzeigen zu können, müssten diese jedoch so weit verkleinert werden, dass sie quasi nicht mehr zu erkennen sind. Ein Ausweg aus diesem Dilemma bietet ein einfacher Ansatz, der diese Übersicht von „Miniaturen“ um interaktiv angebotene Möglichkeiten ergänzt. So können beispielsweise einzelne Elemente oder Ausschnitte ausgewählt werden, die vergrößert und dadurch erkennbar dargestellt werden (analog Abb. 15). Die Interaktionsmöglichkeiten sind zurzeit über Maus-Bewegungen und -Aktionen umgesetzt. Die sogenannten Fokus&Kontext-Techniken (vgl. Lamping et al. 1995) könnte

man als Variante dieses Ansatzes verstehen, bei denen ein ausgewählter Ausschnitt zu Anfang bereits gesetzt ist. Das, was im Fokus steht, wird gut erkennbar dargestellt eingebettet in dessen Kontext, d. h. vor dem Hintergrund der ggf. zur Unkenntlichkeit verkleinerten Gesamtstruktur. Unsere Präsentation bedient sich hierzu eines hyperbolischen Modells (vgl. Abb. 12). Dabei wird eine hierarchische Struktur sozusagen auf eine Halbkugel, die von oben betrachtet wird, projiziert. Hätten wir die hierarchische Struktur plan in der Ebene radial um die Wurzel herum aufgezeichnet, wären die Verbindungslinien zwischen den verschiedenen Hierarchiestufen gleich lang. Dadurch, dass dieses netzartige Gebilde quasi über die Halbkugel gelegt wird, wirken die Linien in der Nähe der Wurzel fast so lang wie in der Ebene, während die weiter außen liegenden quasi perspektivisch in der dritten Dimension „nach hinten“ nahezu verschwinden. Die Größen der verbundenen Objekte werden entsprechend angepasst, so dass die oben liegenden (der oberste und sein enger Kontext) gut zu erkennen sind, die weiter außen liegenden im Normalfall aber unkenntlich klein sind.

Die Projektion auf die Halbkugel ist der Trick, um auf dem begrenzt zur Verfügung stehenden Raum eine beliebig komplexe hierarchische Struktur abbilden zu können. Der äußere Rand sammelt die Objekte der untersten Ebene der Hierarchie mit der maximalen Verzweigung der Verästelung. Jedes Element der Struktur kann aber zum obersten Punkt der Halbkugel verschoben, somit

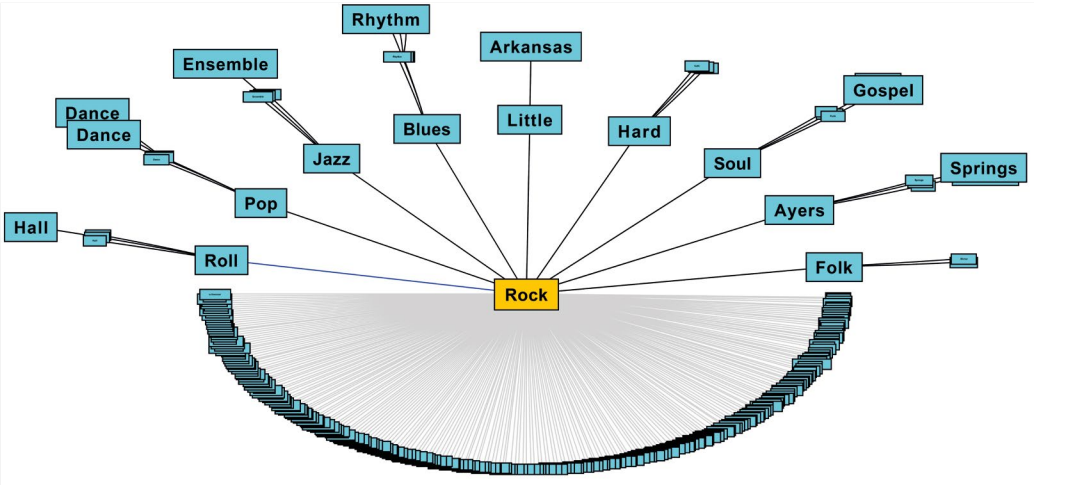


Abb. 12: Hyperbolische Repräsentation des Kookkurrenzprofils des Lemmas „Rock“ (Vicomte) (Copyright IDS).

in den Fokus gerückt und gut erkennbar dargestellt werden. Auf diesem Weg lässt sich nach und nach die gesamte Struktur erschließen.

Bei unseren Analyseergebnissen handelt es sich meist um sehr flache, aber sehr schnell stark verzweigende Strukturen. Die herkömmliche Art der Fokusverschiebung ermöglicht eine nur sehr kleinschrittige Navigation durch die Gesamtstruktur. Neben der interaktiven Vergrößerung von Elementen über die Maus-Bewegung bietet unser Ansatz deshalb eine überlagernde Fokussierung auf einen Ausschnitt der ersten Ebene der Hierarchie an. In der oberen Hälfte der Halbkugel wird ein sehr kleiner Ausschnitt auf den zur Verfügung stehenden Platz gestreckt, während sich der Rest der Gesamtstruktur mit der unteren Hälfte begnügen muss. Zusätzlich zu den anderen Fokussierungsmöglichkeiten lässt sich die Gesamtstruktur drehen, sodass nach und nach jeder Abschnitt der ersten Hierarchieebene gut erkennbar dargestellt wird. Einzelne Elemente können hervorgehoben werden, was auch fixiert für die Darstellung im unteren Bereich beibehalten wird.

Untersuchungen (z. B. in Storjohann 2007a/b, Schnörch 2015) haben gezeigt, dass diese Visualisierungsform wie auch die hier nicht näher ausgeführte Möglichkeit der Annotation die Erschließung der Gesamtstruktur für verschiedene Fragestellungen gut unterstützt und sie dadurch nachvollziehbar dokumentiert wird.

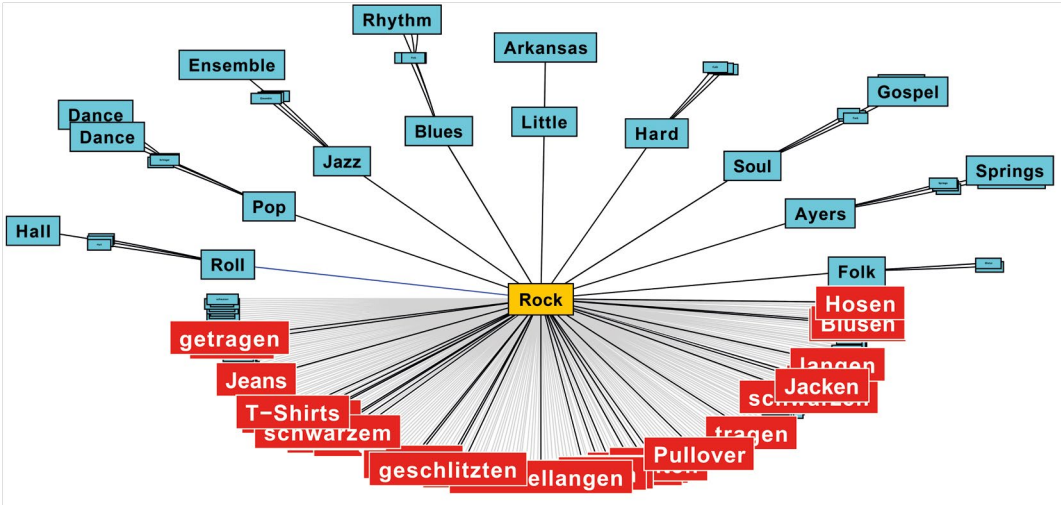


Abb. 13: Hervorgehobene Übereinstimmungen des Profils des Lemmas „Rock“ im Vergleich zum Lemma „Kleid“ (Vicomte) (Copyright IDS).

3.3 Kookkurrenz und Zeit

Während sich Neulexemkandidaten über einen typischen Zeitverlauf charakterisieren lassen, ist dieser für viele Neubedeutungen eher unscheinbar. Die Ausschläge der Amplituden sind nicht auffälliger als bei bedeutungsstabilen Wörtern, bei denen diese durch Veränderungen in der Zusammensetzung der Daten, durch das Weltgeschehen oder einfach durch „Modeerscheinungen" bedingt sind. Vor allem fehlt aber das Gegenstück für eine einfache quantitative Messung: Anders als bei einem Neulexem, dessen Häufigkeit vor dessen Initiation vernachlässigbar klein gewesen sein muss, kann das Aufkommen einer neuen Bedeutung nicht ohne Weiteres zu einem bestimmten Zeitpunkt zahlenmäßig erfasst werden. Einen Anhaltspunkt gibt es aber dennoch: Wenn sich Bedeutungsaspekte im Kookkurrenzverhalten eines Wortes niederschlagen, so sollten sich auch Bedeutungsveränderungen in Änderungen des Kookkurrenzverhaltens abzeichnen. Um die oben angedeuteten alternativen Gründe für Veränderungen ein wenig kontrollieren zu können, achten wir für entsprechende Untersuchungen verstärkt auf eine durchgängig homogene Zusammensetzung des zugrunde gelegten virtuellen Korpus. Der Preis, den wir dafür zahlen, besteht in kleineren Treffermengen. Wenn diese so gering ausfallen, dass Kookkurrenzanalysen für einzelne Jahrgänge zu unergiebig sind, sind die Definitionen der Zeitscheiben auf mehrere Jahrgänge zu erweitern. Ein weiterer Nachteil des kritischen Mindestdatenumfangs besteht darin, dass lediglich Untersuchungen für ca. die letzten 25 Jahre umsetzbar sind. Aufgrund noch vieler anderer Unwägbarkeiten können wir auch noch keine Methode präsentieren, die auffälligen Bedeutungswandel aufdeckt. Wir benutzen in diesem Zusammenhang allerdings Visualisierungen, die die Plausibilität des Ansatzes unterstützen. Die Ergebnisse explorativer Untersuchungen zeigen, dass die Rangverläufe im Kookkurrenzverhalten bestimmter Partnerwörter, die die Neubedeutungen indizieren, Parallelen zu den Zeitverläufen von Neulexemen aufweisen (vgl. Abb. 14).

Für andere Fragestellungen lässt sich derselbe Ansatz verwenden, zum Beispiel, um die Entwicklung auffälliger Diskurse nachzuzeichnen (vgl. das Beispiel Konflikt in Perkuhn/Belica 2016). Das automatische Erkennen derartiger Indikatoren gestaltet sich zurzeit noch schwierig, da einerseits das gesamte Kookkurrenzprofil „ständig in Bewegung" ist und andererseits für eine genauere Isolierung tatsächlich markanter Veränderungen mehr Messpunkte, somit breitere homogene Datengrundlagen erforderlich wären.

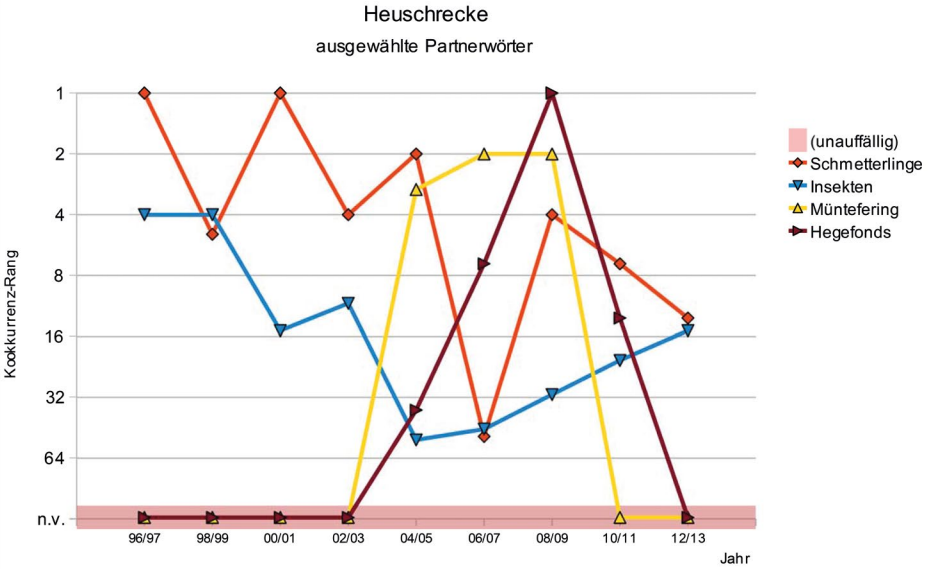


Abb. 14: Rangverlauf auffälliger Partnerwörter des Lemmas „Heuschrecke“ (DeReKo) (Copyright IDS).

3.4 Kookkurrenzverhalten im Vergleich

Wir haben im vorherigen Abschnitt beschrieben, wie das Koookkurrenzverhalten eines Wortes im Laufe der Zeit anhand zeitscheibenbezogener Kookkurrenzprofile analysiert werden kann, um den Bedeutungswandel eines Wortes zu dokumentieren. Der elementare Schritt war hierbei der Vergleich zweier Kookkurrenzprofile. Ausgangspunkt des Vergleichs war in diesem Fall dasselbe Wort, die Profile wurden auf der Grundlage unterschiedlicher virtueller Korpora ermittelt. Übertragen wir jetzt die Vorgehensweise auf die Profile zweier verschiedener Wörter, die auf der Grundlage desselben Korpus erstellt wurden, so können wir hoffen, etwas über die Beziehung zwischen den Wörtern zu lernen: Wörter, die eine enge semantische Beziehung vermuten lassen, sollten viele Gemeinsamkeiten im Kookkurrenzverhalten aufweisen. Eine Erweiterung zu dem oben beschriebenen Vorgehensmodell für die Erschließung von Kookkurrenzanalysen visualisiert die Verteilung der Partnerwörter bei einem Vergleich von bis zu drei Bezugswörtern, wobei topographisch zwischen den dedizierten und den paarweise (ggf. auch allen drei) gemeinsamen Partnern unterschieden wird. Erstere werden vertikal nach Rang, letztere nach gemitteltem Rang und horizontal näher bei dem Bezugswort mit dem höheren Rang angeordnet (vgl. Abb. 15).

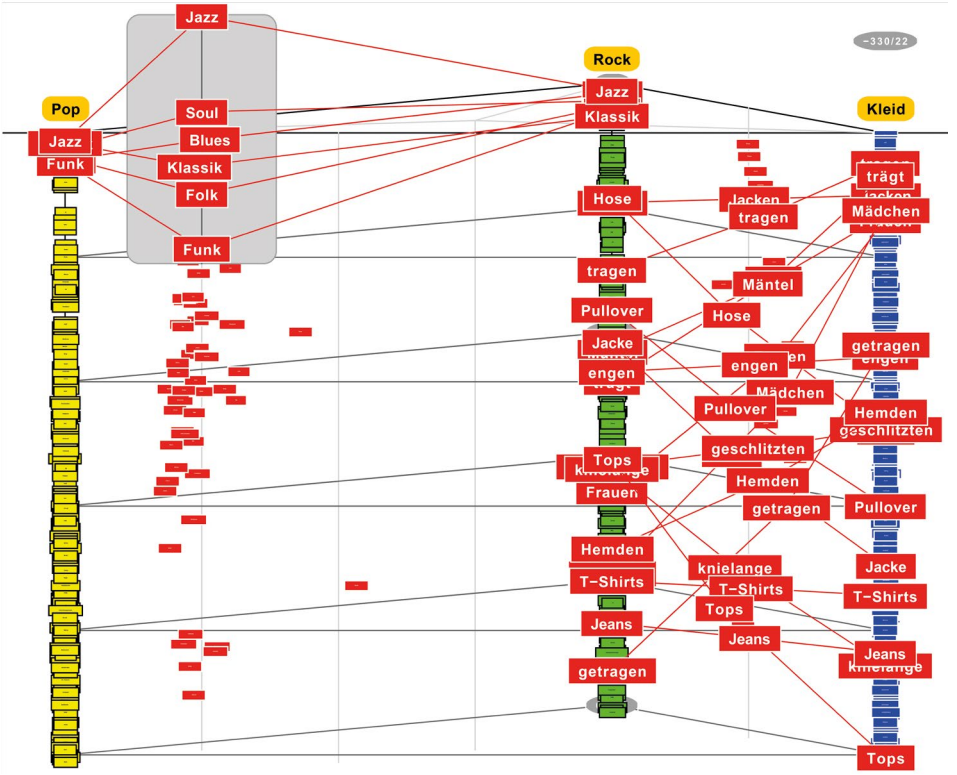


Abb. 15: Hervorgehobene Übereinstimmungen des Profils des Lemmas „Rock“ im Vergleich zu den Lemmata „Pop“ (Ausschnittsvergrößerung) und „Kleid“ (Markierungen) (Vicomte) (Copyright IDS).

Die Darstellung dieser Struktur leidet unter demselben Dilemma, das komprimierte Ansichten wie die Fokus-und-Kontext-Technik erforderlich gemacht hat. Auch hier wird nur überblicksartig die gesamte Information verkleinert dargestellt; eine Vergrößerung von Bestandteilen ist interaktiv über einen Ausschnitt oder durch explizite Markierung möglich. Die Gesamtansicht liefert einen ersten Eindruck von der Verteilung der Partnerwörter etwa bei Synonymen oder Paronymen. Im Detail zeigen sich dann die charakteristischen Partnerwörter. Bei den gemeinsamen Partnern wird durch die gewichtete Anordnung bei der Visualisierung der Eindruck unterstützt, wie relevant das jeweilige Partnerwort bei den jeweiligen Bezugswörtern ist.

Diese eher anschauliche Erklärung und Visualisierung einer gewichteten Ähnlichkeitsbeziehung zwischen Kookkurrenzprofilen ist bereits seit längerer Zeit durch ein formales Vergleichsmaß modelliert, das in der Kookkurrenzdatenbank CCDB Anwendung findet.

3.5 Kartierung von Gebrauchsaspekten

Die Kookkurrenzdatenbank CCDB (Belica 2007) wurde als Sammlung von Kookkurrenzprofilen zu über 220.000 Einträgen aufgebaut und dient als Denk- und Experimentierplattform für die weitere methodische Auswertung der Kookkurrenz, u. a. für die systematische Anwendung des Ähnlichkeitsvergleichs zwischen allen vorhandenen Einträgen. Das Ergebnis dieses Abgleichs wird für jedes Wort als Liste der verwandten Profile absteigend nach dem ermittelten Maß (Related Collocation Profiles) in der CCDB angeboten. So überzeugend und plausibel fast alle Einträge in diesen Listen für sich alleine stehend wirken, so sehr deutet sich eine Vielfalt unterschiedlicher Begründungen der Ähnlichkeit an – im Extremfall bis hin zu disjunkten Aufteilungen des Kookkurrenzverhaltens aufgrund mehrerer Lesarten eines Homonyms. Ein weiteres Verfahren versucht, das Verwendungsspektrum eines Wortes, das diese Vielfalt begründet, in ein grobes Raster einzuordnen. Dazu wird angenommen, dass sich die Wörter, die einem vorgegebenen Wort ähnlich sind, je in Gruppen einordnen lassen, innerhalb derer alle Elemente dem Bezugswort auf eine vergleichbare, aber auch von anderen Gruppen abgrenzende Art ähneln. Als Ausgangsmaß für diese Einordnung wird die Ähnlichkeit nach dem beschriebenen Maß für alle Paare von Einträgen in der Liste der ähnlichen Profile herangezogen. Nach Vorgabe eines Rasters arrangiert dann ein selbst-organisierendes Verfahren eine Anordnung aller ähnlichen Profile auf einer zweidimensionalen Karte (SOM, Kohonen 1990). Die hochdimensionale Vielfalt der Ähnlichkeitsbeziehungen wird so auf eine planare Topologie reduziert, bei der die geometrische Distanz den bestmöglichen Kompromiss aus Ähnlichkeit bzw. Unähnlichkeit aller Einträge widerzuspiegeln versucht. Oberhalb eines Schwellwerts wird allerdings nicht weiter differenziert. Die Gruppen von Wörtern, die die höchste Ähnlichkeit untereinander aufweisen, werden gemeinsam in einem Feld des Rasters abgebildet, da zu vermuten ist, dass sie tendenziell denselben Gebrauchsaspekt des Bezugswortes zum Ausdruck bringen.

Als Raster hat sich in vielen Anwendungen eine 5×5-Matrix bewährt, deren Felder in fließenden Farbtönen hinterlegt sind (vgl. Abb. 16). Damit soll visuell unterstützt werden, dass die Aspekte, die in den einzelnen Feldern ausgedrückt werden, weich ineinander übergehen. In den Fällen, in denen aufgrund (Un-)Ähnlichkeitsbeziehungen kein weicher Übergang möglich ist (etwa wenn Aspekte unterschiedlicher, scharf trennbarer Lesarten aneinanderstoßen sollen), wird dieser Abstand durch unbesetzte und ungefärbte Felder umgesetzt, die sich häufig (z. T. in Kombination mit weiteren, dünn besetzten Feldern) wie ein Trenngraben durch das gesamte Diagramm ziehen.

Bei der (meist lexikographisch motivierten) Interpretation der Karten hat sich eine semiotische Ausrichtung herauskristallisiert: Ausgehend von einzelnen Feldern werden auch die umgebenden Felder miteinbezogen, um zu sichten,

Rock

Musikstil Klassik Stilrichtung Mixtur Salsa Calypso Folklore Melange	Reggae Techno Funk Musikrichtung Rap Punkrock Mix Weltmusik	Pop Folk Hop Jazz Ska Punk Grunge Rockabilly	Soul Blues Hip Hardcore Dancefloor Groove Disco Rave	Acid Jam Sampler Independent
Rockmusik Popmusik Schlager Dixieland Chanson Tanzmusik Volksmusik Sprechgesang	tanzbar fetzig funkig eingängig melodiös jazzig Rocksong Siebziger	Gospel Swing grooven rockig Ragtime Mambo Eigenkomposition soulig	Country Boogie Sound Revival rappen	Jump Attack Chart Rocker Underground Floor Pistol Voodoo
knallig	poppig	Dixie Ballade Evergreen Gassenhauer Popsong Spiritual Schnulze Band	Song Twist rocken Ohrwurm Musical Medley Abba Feeling	covern Rocks Coverversion Soundcheck Cover Dancer Hot Nirvana
Jeans Hemd Blouson Pullover Sweatshirt Latzhose beigen hellblau	pinkfarben gestickt		Hit Welthit Album Count Creole Alben Maid King	Diskografie Trouble Dust Sweet Live Boy Skin Straight
Hose Jacke Shirt Mantel Shorts Pulli Blazer Strickjacke	ärmellos knielang tailliert Bluse geschlitzt hauteng Hosenanzug Oberteil	Slip Petticoat Smoking	Let Titelsong Look Springfield	My Girl Go Rain Out Want On Gon

Abb. 16: Self-organized Map des Lemmas „Rock" (CCDB) (Copyright IDS).

inwieweit diese gemeinsame Aspekte zum Ausdruck bringen. Durch die Kartierung der Gebrauchsaspekte und ihre Interpretation konnten aufschlussreiche Aspekte über die Verwendungsspektren der betrachteten Wörter gewonnen werden (Vachkova/Belica 2009).

Eine Erweiterung des Verfahrens ist für die Anwendung auf Wortpaare konzipiert und operiert auf der Vereinigungsmenge aller Profile, die zu einem Wort (oder beiden) als ähnlich eingestuft wurden. Abweichend von der SOM-Farbgestaltung wird hierbei dann durch die Einfärbung eine weitere Information kodiert: Je nach Verhältnis der Ähnlichkeiten der Felder zu einem der beiden vorgegebenen Wörter wird die Feldfarbe aus Anteilen von Rot und Gelb zusammengemischt. Ein klares Votum für das eine oder andere Wort spiegelt

Abb. 17: Self-organized Map der Kontrastierung der Lemmata „Rock“ und „Kleid“ (CCDB) (Copyright IDS).

sich in den ihnen zugeordneten Primärfarben wider; unklare Zuordnungen zeigen sich durch entsprechende Abstufungen von Orangetönen.

Angewandt auf Paare, die als Synonyme oder Paronyme gelten (vgl. Abb. 17), zeigt das Verfahren und diese Einfärbung auf, in welchen Domänen oder Diskursbereichen die Wörter sich nahestehen oder sich scharf trennen lassen. Studien in diesen Bereichen (Marková 2012, Storjohann/Schnörch 2014) zeigen, dass der reflektierte Einsatz dieser Methoden gerade auch durch die Interpretation der Visualisierungen einen erheblichen Mehrwert empirischer Analysen darstellt.

Eine Besonderheit der Kartierung mithilfe von SOMs ist, dass sie kleinere Unterschiede vollständig ausblendet, indem sie verschiedene Wörter einer Zelle

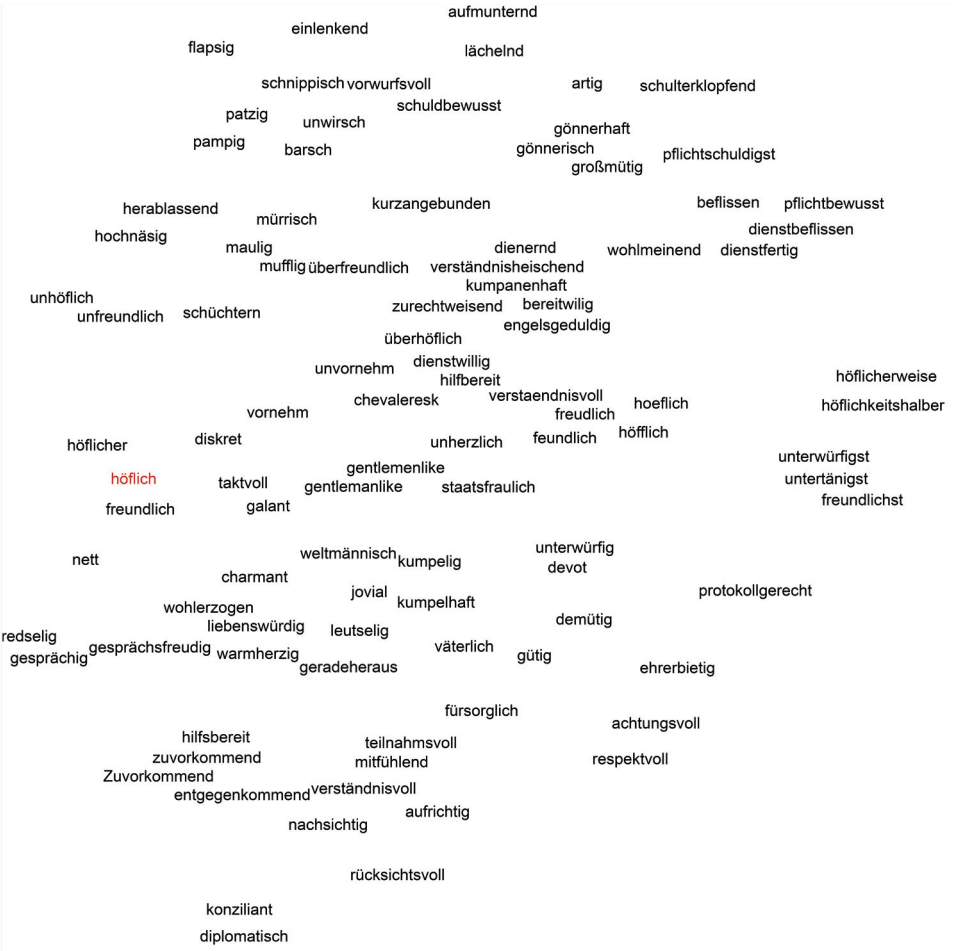


Abb. 18: „höflich“ und seine 99 distributionell ähnlichsten Nachbarn in DeReKo dargestellt mithilfe von t-SNE (Copyright IDS).

auf der Karte zuordnet, ohne deren Ähnlichkeitsbeziehungen und Abstände innerhalb dieser Gruppe im Diagramm darzustellen. Oft ist das Weglassen solcher potenziell ablenkender Detailinformation aber ein gewünschter Effekt, um den Blick auf das Wesentliche nicht zu verstellen. So kann eine solche Quantisierung bzw. ein solches Clustering erfahrungsgemäß die Abduktion neuer Hypothesen erleichtern. Dabei muss natürlich beachtet werden, dass man es nicht mit Ergebnissen zu tun hat, sondern nur mit Hypothesen, die erst noch überprüft

werden müssen (z. B. mithilfe neuer Analysen oder einer manuellen Untersuchung von Belegstellen).

Wenn gerade nicht das Ausblenden von Detailinformationen zur Abduktion allgemeinerer Hypothesen, sondern eine detaillierte, topographieerhaltende Darstellung engerer Ähnlichkeitsbeziehungen das Ziel ist, hat sich in den letzten Jahren t-SNE (van der Maaten & Hinton 2008) als eine in der Regel gut geeignete Methode zur Dimensionsreduktion etabliert. Abb. 18 zeigt „höflich" und seine 99 bezüglich ihrer distributionellen Eigenschaften ähnlichsten Nachbarn auf einer mithilfe von t-SNE erzeugten Karte. Zur Vektorrepräsentation der Wörter wurden in diesem Fall nicht Kookkurrenzprofile verwendet, sondern sogenannte „word embeddings" (Mikolov et al. 2013), die mittels einer Erweiterung der Programme word2vec bzw. wang2vec (Ling et al. 2015) auf der Basis von DeReKo-2016-I (Institut für Deutsche Sprache 2016a) berechnet wurden. Auch solche Streudiagramme können natürlich so angereichert werden, dass sie bestimmte Zusammengehörigkeitshypothesen nahelegen, indem etwa die Ergebnisse von Clusteranalysen über gemeinsame Farben oder Rahmen um zusammengehörige Knoten kodiert werden.

4. Fazit und Ausblick

In vielen Bereichen unserer Arbeitsfelder hat es sich als sinnvoll und hilfreich erwiesen, eine Mischung von quantitativ-qualitativen Vorgehensweisen mit Visualisierungstechniken zu kombinieren. Dabei können es auch durchaus schlichte Ansätze sein, die ausreichen, um die Interpretation auf Interessantes zu lenken. Aufwändigere Techniken und insbesondere Interaktionsmöglichkeiten bergen neben dem Einarbeitungsaufwand bisweilen die Gefahr, verstärkt Aufmerksamkeit zu binden und den Status der angebotenen Signale überzubewerten.

Man sollte stets im Hinterkopf behalten, dass nicht unbedingt neue Fakten, sondern nur zu interpretierende Hinweise angeboten werden. Diese können in verzerrender Weise zu Unrecht überspitzt sein, sie können auch weniger relevant sein als andere, die nicht in den Vordergrund gerückt wurden.

Auch Visualisierungstechniken müssen reflektiert und mit der nötigen Distanz eingesetzt werden. Sie sollten, wenn möglich, in verschiedenen Variationen verglichen werden können, um die ersten, schnellen Hypothesen auf den Prüfstand zu stellen. Ein in jeglicher Hinsicht kritischer Punkt der Verfahren ist dabei, die Menge an Information auf das Wesentliche zu reduzieren und auf das Relevante zu fokussieren. Denn gerade dafür erhofft sich der/die NutzerIn Unterstützung bei der Bewältigung des Reichtums an Hinweisen, die in einem sehr großen Korpus stecken.

Es ist absehbar, dass der Bereich der Visualisierung zukünftig an Bedeutung gewinnt und etwa durch Andockmöglichkeiten diverser Visualisierungsverfahren an unser Recherchesystem für ein breites Fachpublikum zu einer selbstverständlichen Ergänzung des üblichen Arbeitens werden wird (s. a. Kupietz et al. 2015.).

5. Bibliographie

Baroni, Marco, Silvia Bernardini, Adriano Ferrares und Eros Zanchetta. 2009. "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora." *Language Resources and Evaluation* 43 (3): 209–226. wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wacky_2008.pdf.

Belica, Cyril. 1995. „Statistische Kollokationsanalyse und -clustering. Korpuslinguistische Analysemethode." http://corpora.ids-mannheim.de/ (letzter Zugriff am 12. Oktober 2016).

Belica, Cyril. 2007. „Kookkurrenzdatenbank CCDB – V3: Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs." http://corpora.ids-mannheim.de/ccdb/ (letzter Zugriff am 12. Oktober 2016).

Bodmer Mory, Franck. 2014. „Mit COSMAS II ‚in den Weiten der IDS-Korpora unterwegs'." In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache,* herausgegeben vom Institut für Deutsche Sprache. Mannheim: Institut für Deutsche Sprache, 376–385.

Cox, Trevor F. und Michael A. A. Cox. 2001. *Multidimensional Scaling.* Boca Raton, Fla.: Chapman and Hall.

Institut für Deutsche Sprache. 2016a. „Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2016-I (Release vom 31.03.2016)." Mannheim: Institut für Deutsche Sprache. www.ids-mannheim.de/DeReKo (letzter Zugriff am 12. Oktober 2016).

Institut für Deutsche Sprache. 2016b. „*Cosmas II*-Recherchesystem." https://cosmas2.ids-mannheim.de/cosmas2-web/ (letzter Zugriff am 12.Oktober 2016).

Keibel, Holger, Sophie Hennig und Rainer Perkuhn. 2010. *Effiziente halbautomatische Detektion von Neologismuskandidaten.* Mannheim: Institut für Deutsche Sprache (Technical Report IDS – KL-2010-01). www.ids-mannheim.de/kl/dokumente/ids-kl-2010-01.pdf

Kilgarriff, Adam. 2001. "Comparing Corpora." *International Journal of Corpus Linguistics,* 6 (1): 97–133.

Kohonen, Teuvo. 1990. "The Self-Organizing Map. New Concepts in Computer Science." In Informatique: nouveaux concepts scientifiques. Colloque en l'honneur de Jean-Claude Simon, Paris. AFCET, 181–190.

Kupietz, Marc. 2005. *Near-Duplicate Detection in the IDS Corpora of Written German.* Mannheim: Institut für Deutsche Sprache (Tech. Rep. KT-2006-01*). www1.ids-mannheim.de/fileadmin/kl/misc/ids-kt-2006-01.pdf.*

Kupietz, Marc, Cyril Belica, Holger Keibel und Andreas Witt. 2010." The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research." In *Proceedings of the seventh conference on International Language Resources and Evaluation* (LREC 2010), herausgegeben von Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner und Daniel Tapias. Malta : ELRA, 1848–1854. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf.

Marc Kupietz, Harald Lüngen, Cyril Belica, Cyril und Rainer Perkuhn. 2012. "Webkorpora als qualitätsgesicherte Forschungsdaten." Unveröffentlichter Vortrag im Rahmen des GSCL-Workshops Webkorpora in Computerlinguistik und Sprachforschung am 27. September 2012.

Kupietz, Marc, Nils Diewald, Michael Hanl und Eliza Margaretha. 2016. „Möglichkeiten der Erforschung grammatischer Variation mithilfe von KorAP." In *Grammatische Variation – empirische Zugänge und theoretische Modellierung*, herausgegeben von Marek Konopka und Angelika Wöllstein. Berlin: de Gruyter, 319–330.

Marc Kupietz und Thomas Schmidt. 2015. „Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung." In *Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven,* herausgegeben von Ludwig M. Eichinger. Berlin: de Gruyter, 297–322(Jahrbuch des Instituts für Deutsche Sprache 2014).

Lamping, John, Ramana Rao und Peter Pirolli. 1995. "A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies." In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems.* New York: ACM, 401–408.

Ling, Wang, Chris Dyer, Alan Black und Isabel Trancoso. 2015. "Two/Too Simple Adaptations of word2vec for Syntax Problems." In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics.* Denver, CO: Human Language Technologies. www.cs.cmu.edu/~lingwang/papers/naacl2015.pdf.

Lüngen, Harald und Holger Keibel. 2013. „Zur Erstellung und Interpretation der Zeitverlaufsgrafiken." In *Neuer Wortschatz: Neologismen im Deutschen 2001–2010,* herausgegeben von Doris Steffens und Doris al-Wadi. Mannheim: Institut für Deutsche Sprache, 561–567.

Marková, Věra. 2012. Synonyme unter dem Mikroskop: Eine korpuslinguistische Studie. Tübingen: Narr (CLIP 2).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado und Jeffrey Dean. 2013. "Distributed representations of words and phrases and their compositionality."

In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality (letzter Zugriff am 27. November 2017).

Perkuhn, Rainer. 2007a. "Systematic Exploration of Collocation Profiles." In: *Proceedings of the 4th Corpus Linguistics Conference (CL 2007)*. Birmingham: University of Birmingham. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/4715 (letzter Zugriff am 27. November 2017).

Perkuhn, Rainer. 2007b. „'Corpus-driven': Systematische Auswertung automatisch ermittelter sprachlicher Muster." In *Sprach-Perspektiven: Germanistische Linguistik und das Institut für Deutsche Sprache,* herausgegeben von Heidrun Kämper und Ludwig M. Eichinger. Tübingen: Narr, 465–491 (Studien zur Deutschen Sprache 40).

Perkuhn, Rainer. 2012. *Diachrone Kookkurrenzanalyse.* Mannheim: Institut für Deutsche Sprache (Technical Report IDS-KL-2012-02).

Perkuhn, Rainer und Cyril Belica. 2016. „Konflikt, Sprache, korpuslinguistische Methodik." In *Linguistische Zugänge zu Konflikten in europäischen Sprachräumen. Korpus – Pragmatik – kontrovers,* herausgegeben von Friedemann Vogel, Stefaniya Ptashnyk und Janine Luth. Heidelberg: Winter 4), 339–364 (Schriften des Europäischen Zentrums für Sprachwissenschaften (EZS).

Schnörch, Ulrich. 2015. „Wortschatz." In *Handbuch „Wort und Wortschatz"*, herausgegeben von Ulrike Haß und Petra Storjohann. Berlin/Boston: de Gruyter, 3–26 (Handbücher Sprachwissen 3).

Schumann, Heidrun und Wolfgang Müller. 2000. *Visualisierung–- Grundlagen und allgemeine Methoden.* Berlin: Springer.

Storjohann, Petra. 2007a. „Wie viel Diskurs braucht ein Wörterbuch?" *German Life and Letters* 60, Issue 4: 569–592.

Storjohann, Petra. 2007b. „Der Diskurs ‚Globalisierung' in der öffentlichen Sprache. Eine korpusgestützte Analyse kontextueller Thematisierungen." *Aptum: Zeitschrift für Sprachkritik und Sprachkultur* 2007, Heft 2: 139–155.

Storjohann, Petra und Ulrich Schnörch. 2014. "Empirical Approaches to Paronyms." In: *Proceedings of the XVI EURALEX International Congress*, herausgegeben von Andrea Abel, Chiara Vettori und Natascia S. Ralli, 463–476. Bozen: Institute for Specialised Communication and Multilingualism.

Vachková, Marie und Cyril Belica. 2009. "Self-Organizing Lexical Feature Maps. Semiotic Interpretation and Possible Application in Lexicography." *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis* 13 (2): 223–260.

van der Maaten, Laurens und Geoffrey Hinton. 2008. "Visualizing High-Dimensional Data Using t-SNE." *Journal of Machine Learning Research* 9: 2579–2605.

Mark Richard Lauersdorf

Linguistic Visualizations as *objets d'art*?

Abstract This article undertakes a broad-ranging examination of the practice of data visualization in linguistic research, whether for elucidation and elaboration of theoretical models, analysis and interpretation of datasets, or summarization and presentation of research outcomes. Roman Jakobson's cube model for Russian case theory, and the concept of *objet d'art* (Chvany 1987) as a notional frame, are deployed to draw attention to a range of issues that must be considered in the use of linguistic visualizations. Following a survey of traditional and newer visualization techniques in linguistic research, the specific example of data analysis in historical sociolinguistics is used to make an argument for linguistic visualization practices that "use all the data", "view all the data", "view all the combinations", "view all the angles", and "use all the techniques".

1. Introduction

Jakobson's cube

The inspiration for the title of this article is a study by Catherine Chvany entitled "Jakobson's Cube as *Objet d'Art* and as Scientific Model" (Chvany 1987).[1] In her article Chvany discusses Roman Jakobson's "famous cube model for the interrelated meanings of the eight ($= 2^3$) cases encoded in the Russian language" (199) and the "three binary (+ or −) features ($= 2^3$): [± MARGINAL] (represented as the vertical dimension), [± QUANTIFYING] (the depth dimension), [± DIRECTIONAL] (the width dimension)" expressed by the eight Russian cases (200); and she reproduces a graphic representation of the Jakobson cube model, seen here in Figure 1.

1 Another version of the work appeared as Chvany 1984.

Published in: Noah Bubenhofer, Marc Kupietz: Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

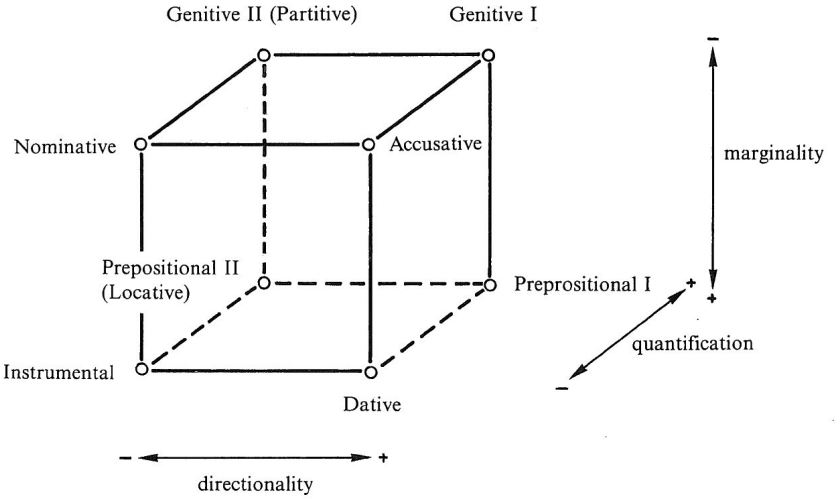


Figure 1: Mel'čuk's 1983 illustration of the Jakobson cube model, adapted from Jakobson 1958 (image from Chvany 1987, 199).

The purpose of Jakobson's cube is, of course, to provide a visual representation of a complex set of information.[2] Specifically it seeks to provide a maximally informative, and at the same time clearly interpretable, visualization of a theoretical model describing a set of features and the relationships between them in a morpho-syntactic system. The interpretive value of the cube visualization of this theoretical model becomes clearer when we consider it alongside an example (Figure 2) of one alternative way that these features and their interrelationships can be presented.

The intent here is not to say that the grid presentation in Figure 2 provides no meaningful access to the theoretical model and is of no assistance in the analysis of the information contained in the model, but when compared with the cube visualization in Figure 1, it is clear that we perceive the theoretical model and its information in a different way in each of the two visual representations.

This specific example of linguistic visualization – Roman Jakobson's Russian case theory represented as a cube – serves in the following discussion as our point of entry into a more broad-ranging examination of linguistic visualization under the general notion of *objet d'art*.

2 The practice of providing visual representations of complex information is certainly not unique to academic presentation of scientific research; it is also found quite readily in everyday use, with many types of visual illustrations employed to explain difficult concepts or rich information. The value of this everyday use of visualizations is clearly reflected in the common expression "a picture is worth a thousand words".

	Marginal	Quantifying	Ascriptive
Nominative	–	–	–
Accusative	–	–	+
Genitive$_1$	–	+	+
Genitive$_2$	–	+	–
Locative$_2$	+	+	–
Locative$_1$	+	+	+
Dative	+	–	+
Instrumental	+	–	–

Figure 2: Neidle's 1982 grid for the eight-case system in Russian. Note that Neidle uses "ascriptive" for "directional" (image from Chvany 1987, 218).

Jakobson's cube as *objet d'art*

In her discussion of graphic representations of linguistic systems, Chvany states: "[…] each geometric figure has its own semantics; it can be ambiguous (have homonyms), and it can have approximate synonyms, just as words do. The meanings of figures, governed by principles of visual perception, necessarily combine with the meanings assigned by the linguist" (1987, 208). Put another way, graphic representations themselves carry meaning – meaning that is correlated with other graphic representations, and that is sometimes ambiguous or interpretable in multiple ways, i.e., visual representations are embedded in systems of meaning, governed in part by "principles of visual perception". So, it can be argued that the cube itself, as a geometric figure, carries meaning and has the potential for variation and ambiguity in its meaning, as interpreted by individual observers – some might see one thing and others might see something else. Consider the illustrations in Figure 3 and Figure 4 by Joseph Jastrow (1899) of variant perceptions of the cube.[3]

3 For Figure 3, a version of the "Necker cube" (Necker 1832), Jastrow (1899) describes the multiple interpretations as follows: "Figs. 13a and 13b are added to make clearer the two methods of viewing Fig. 13. The heavier lines seem to represent the nearer surface. Fig. 13a more naturally suggests the nearer surface of the box in a position downward and to the left, and Fig. 13b makes the nearer side seem to be upward and to the right. But in spite of the heavier outlines of the one surface, it may be made to shift positions from foreground to background, although not so readily as in Fig. 13" (308). "The presence of the diagonal line makes the change more striking: in one position it runs from the left-hand rear upper corner to the right-hand front lower corner; while in the other it connects the left-hand front upper corner with the right-hand rear lower corner." (309). For the possible interpretations of the stacked cube illustration in Figure 4 he provides the following description: "If viewed in one way – the

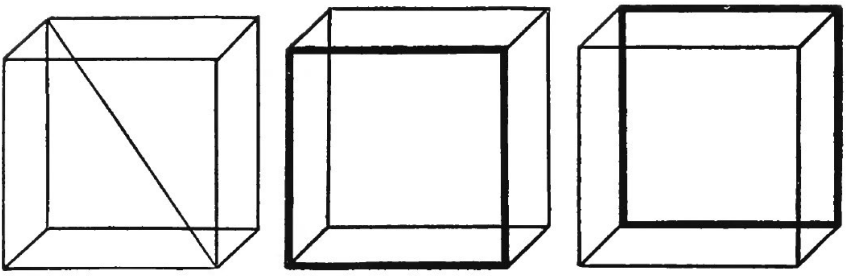


Figure 3: Jastrow's illustration of variation in perception of a cube (1899, 308).

This is not unlike our experiences when observing artistic designs and works of art. Each of us potentially sees something different first, or has a different interpretation of what is seen; and sometimes the longer you look at it, the more you see one thing instead of another or favor one interpretation over another; or the more often you look at it, the more you see different aspects of it that allow for different interpretations, or perhaps the less sure you are of what you actually see. As illustration of this effect, consider whether the artistic sketch in Figure 5 depicts the image of a rabbit or a duck.[4]

Jastrow, in summarizing his observations concerning the phenomenon of variant perceptions/interpretations of geometric shapes and artistic designs, including those illustrated in Figures 3, 4, and 5, notes:

> All these diagrams serve to illustrate the principle that when the objective features are ambiguous we see one thing or another according to the impression that is in the mind's eye; what the objective factors lack in definiteness the subjective ones supply, while *familiarity, prepossession, as well as other circumstances influence the result.* These illustrations show conclusively that seeing is not wholly an objective matter depending upon what there is to be seen, but is very considerably a

black surface forming the tops of the blocks – there seem to be six … ; but when the transformation has taken place and the black surfaces have become the overhanging bottoms of the boxes, there are seven …" (310).

4 This classic image was first published on page 147 of the 23 October 1892 issue (issue no. 2465) of the German magazine *Fliegende Blätter* (I. Schneider, ed. München: Braun & Schneider) with the wording "Welche Thiere gleichen einander am meisten? Kaninchen und Ente." (Which animals resemble each other the most? Rabbit and duck.) See http://digi.ub.uni-heidelberg.de/diglit/fb97/0147 for a digital facsimile edition of the issue containing the original image.

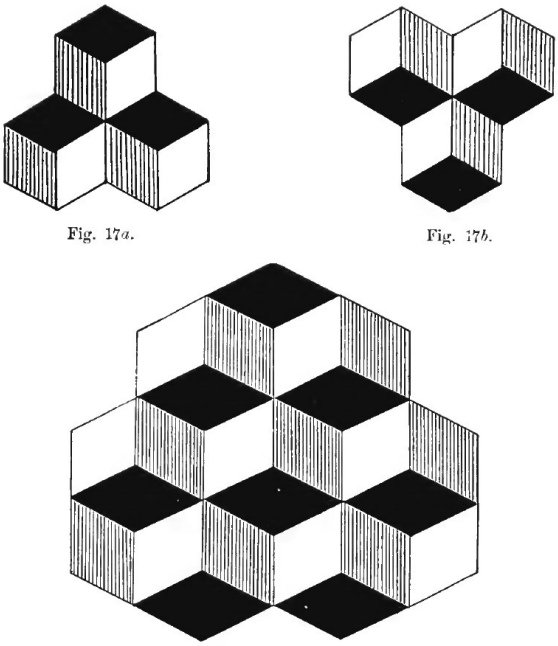


Figure 4: Jastrow's illustration of variation in perception of a stack of cubes (1899, 311).

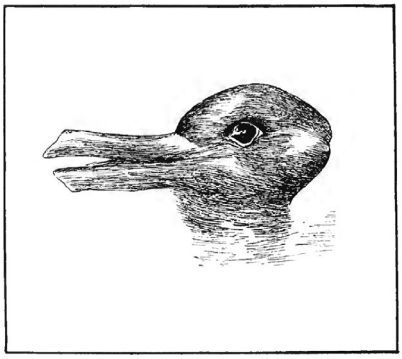


Figure 5: Jastrow's adaptation of the rabbit ~ duck illusion (1899, 312).

> subjective matter depending upon the eye that sees. *To the same observer a given arrangement of lines now appears as the representation of one object and now of another; and from the same objective experience, especially in instances that demand a somewhat complicated exercise of the senses, different observers derive very different impressions* [emphasis added, MRL] (1899, 310–311).

From the discussion above, the conclusion can be drawn that the cube itself (or any other visual illustration), like a work of art (an *objet d'art*), is open to potential variant perceptions that may *broaden* the possible range of observers' interpretations of the visualization (and the data it represents) well beyond any specific meaning assigned to it by the person using the cube (or other graphic illustration) as a visualization of their data and information.

Chvany in her continued discussion of visual representations of linguistic systems also points us in the opposite direction to the possible *constraining* influence of visualizations, describing the potential for a visualization to activate only a specific interpretation or range of interpretations of the data and thus ultimately influence the direction of the linguistic theory derived from the interpretation. "Moreover, the graphic representation, be it matrix, tree, box diagram or polyhedron, may, through its own semantics, influence the perception of the modeled system. As Stewart (1976) points out in her Introduction, 'the relationship of analogy between figure and datum, between design and meaning, is what enables graphic representation to influence linguistic theory'" (Chvany 1987, 208).

Thus, once again, as with *objets d'art*, where a specific artwork may become for many observers the standard interpretation (a sort of iconic representation) of the subject it is depicting, the specific form of a data visualization has the potential to narrow our perception and interpretation of the data. In our example of Jakobson's cube, the fact that he visualized his case theory with a cube could lead us to favor certain interpretations of the data, and it could ultimately become the primary, or even the only, way in which we see the data and conceptualize it theoretically, causing us to overlook, or even exclude, other possible theoretical interpretations.

Chvany highlights this potential constraining factor of a chosen visualization in the specific case of Roman Jakobson's cube representation of the Russian case system: "The 1958 model … expands the prism, closes the unfinished cube, answers its questions, *removes choice … the cube is a nonnegotiable model*" (1987, 215 – emphasis added MRL), adding further:

> Even those of us who disagree with one or another aspect of the cube model have used some of its component claims, whether as supporting argument, stipulation or axiom. For there are occasions where it is possible to use one or another part of the model, without regard – or need – for internal consistency of the whole. […] The cube's unfalsifiable claims, while undesirable in a theory, do not interfere with these limited but useful applications, so *there is little motivation to change the model*. […] In the area of applications, *it's 'love it or leave it'*. The system is so tight, it hangs together so well, that adjusting one opposition would entail changes in the rest, destroying the parts that one cannot disagree with (Chvany 1987, 216–217 – emphasis added MRL).[5]

In the end, a specific visualization could constrain the possible interpretations of a dataset or model to the point that we focus on the visualization rather than on the data or model that it represents, and we objectify the visualization thereby fixing (locking in) the form or type of that visualization, considering it immutable/unchangeable; and we then interpret all new data and arguments through that fixed form – the visualization becomes an *iconic* representation of the underlying information. This should cause us to wonder, with Chvany, "But the question remains, is the cube [or any other given visualization] the best possible icon of the system?" (1987, 218). We will return to this question of "best possible icon" in the discussion further below.

2. Tasks of linguistic visualizations

Elucidation and elaboration of theoretical models

The preceding discussion, of Roman Jakobson's cube visualization for his theory of the Russian case system, provides a good example of one of the common tasks for which linguistic visualizations are deployed – the elucidation and elaboration of theoretical models. A long-standing example of the use of visualizations to represent theoretical frameworks is the use of tree diagrams to illustrate the concept of genetic relatedness among languages (see Figure 6).

5 Chvany later adds to these thoughts a direct *objet d'art* reference, "The cube's take-it-or-leave-it, love-it-or-leave-it fate resembles the history of an art object more than the normal development of a scientific model" (1987: 222).

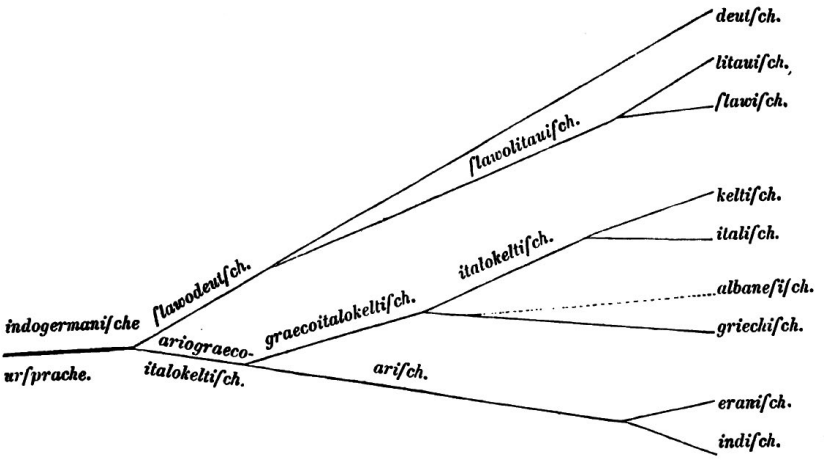


Figure 6: Schleicher's *Stammbaum* visualization of the genetic relatedness of the Indo-European languages (1861, 7).

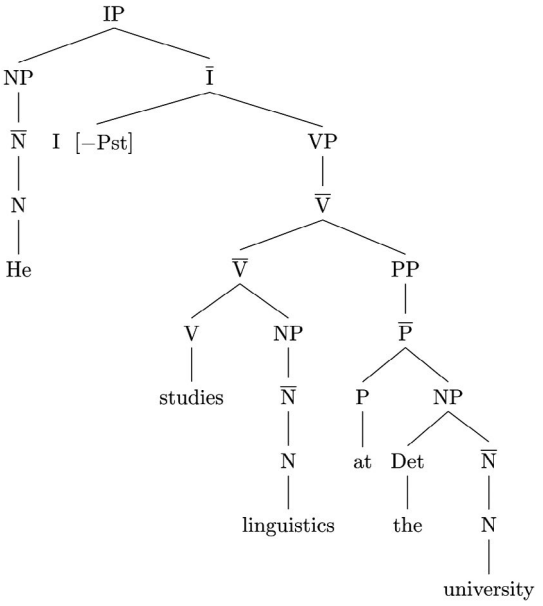


Figure 7: Syntax tree illustrating X-bar theory (https://commons.wikimedia.org/wiki/File:Xbarst1.svg – accessed 09 October 2016).

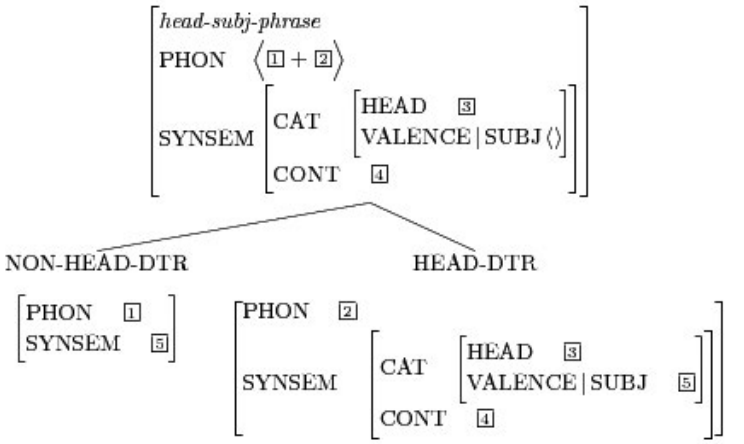


Figure 8: Tree visualizing HPSG theory (https://commons.wikimedia.org/wiki/File:Head-subj-tree.png – accessed 09 October 2016)

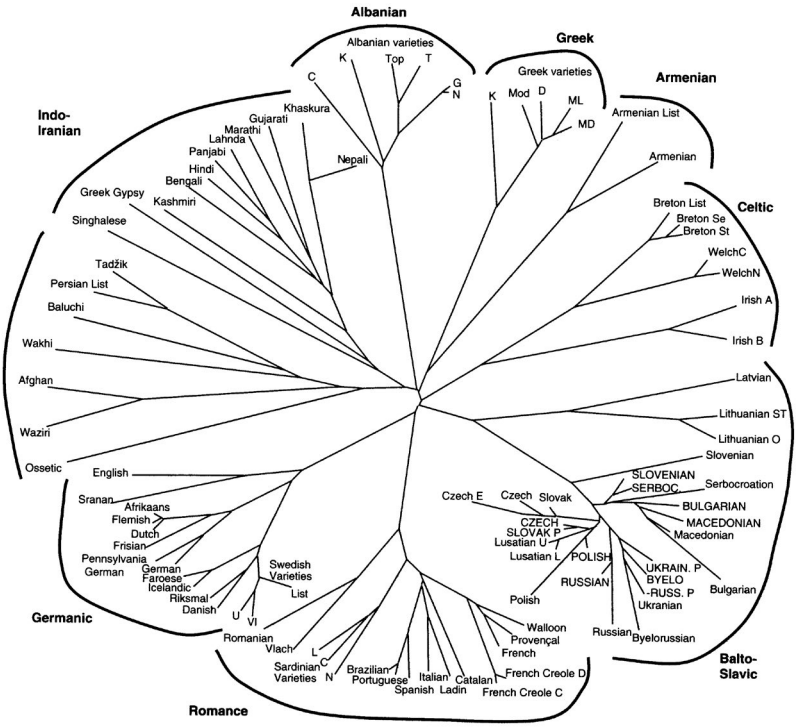


Figure 9: McMahon and McMahon's unrooted Indo-European tree generated on the basis of quantitative statistical analysis of relatedness data (2005, 101).

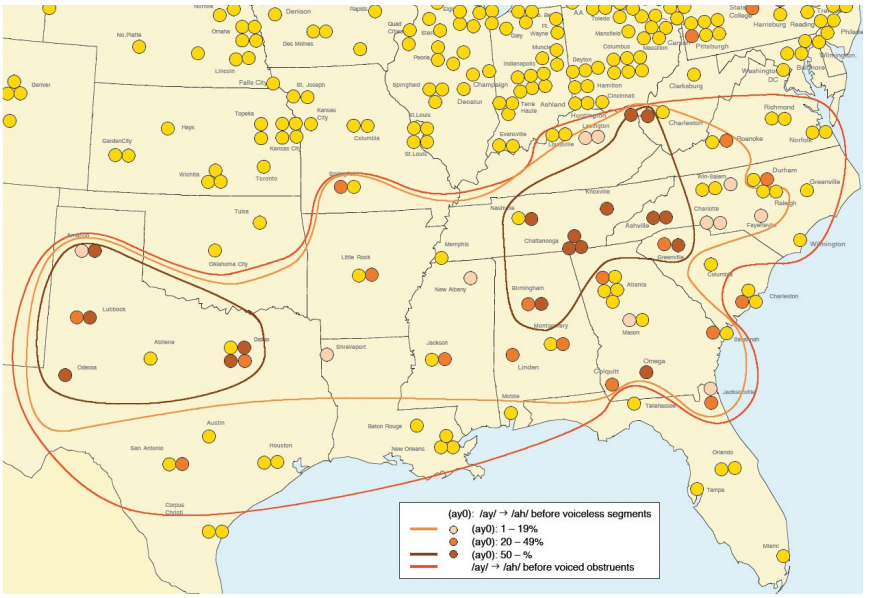


Figure 10: Labov, Ash, Boberg's map of monophthongization before voiceless consonants in North American English (2006, 129).

Of course, tree structures have also become a highly common visualization tool in other areas of linguistics as graphic representations of theoretical constructs as seen in Figure 7 and Figure 8.

Technological advances (most recently digital) in the tools and instruments available to us have allowed us to apply more sophisticated visualization techniques to existing theoretical models, checking those models and also refining and elaborating them in ways not before possible (or accomplished only with difficulty). Figure 9 provides an example of statistical and computational advances in tree diagrams for visualizing linguistic relatedness.

Analysis and interpretation of datasets

In addition to their use in elucidating and elaborating theoretical models, visualizations are commonly deployed on linguistic datasets with the hope of aiding the analysis of the data and the interpretation of the results of that analysis. One common example of this use of visualizations for data analysis and interpretation is the geospatial plotting of dialect data as seen in Figure 10.

Like tree diagrams to visualize linguistic relatedness, geospatial representation of dialect data has a long-standing tradition in linguistics (see Figure 11 and Figure 12).

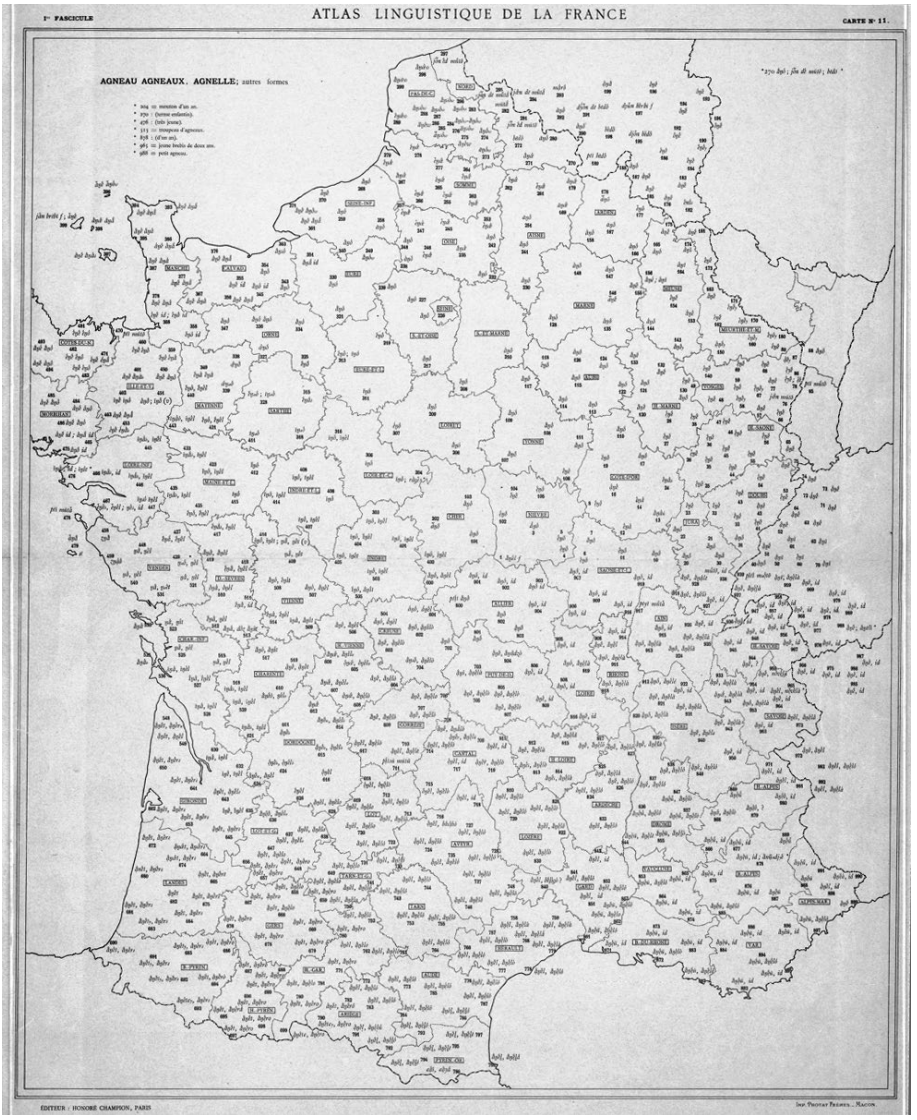


Figure 11: Gilliéron and Edmont's (1902–1910) *Atlas linguistique de la France*, fascicle 1, map no. 11 "AGNEAU, AGNEAUX, AGNELLE; autres formes" (http://cartodialect.imag.fr/cartoDialect/seadragon.jsp?carte=CarteALF0011&width=4852&height=5912 *or* http://cartodialect.imag.fr/cartoDialect/download/Carte-ALF0011.tif – both accessed 02 July 2017).

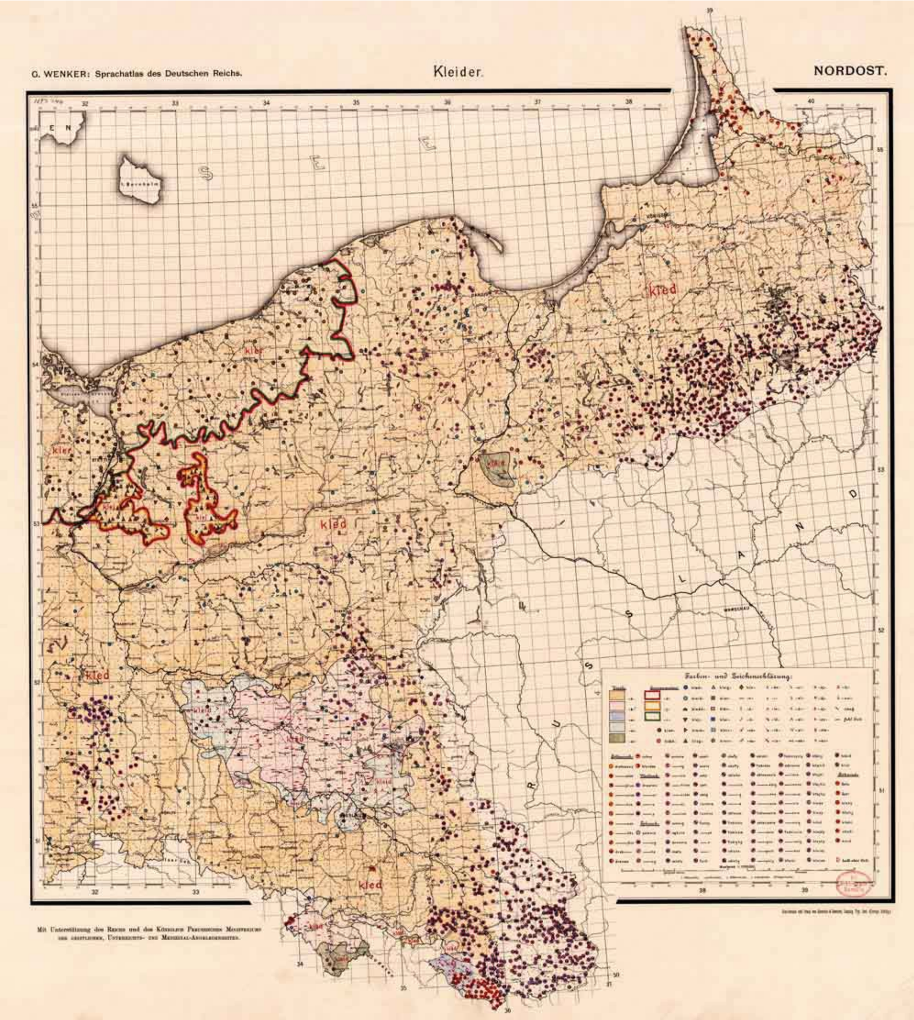


Figure 12: Wenker's (1888–1923) *Sprachatlas des Deutschen Reichs*, northeast sector map for "Kleider" (http://www.graphicscience.de/assets/images/DSA-Kleider_NO_udl-02-1000P.jpg – accessed 09 October 2016).

Just as theoretical models have benefited from advances in visualization tools and techniques, technological advances have provided ever more powerful tools for analysis and interpretation of data, illustrated in Figure 13, again on the example of geospatial mapping.

Technological advances have also made visualizations possible in areas where they were not possible before, where the visualizations themselves actually

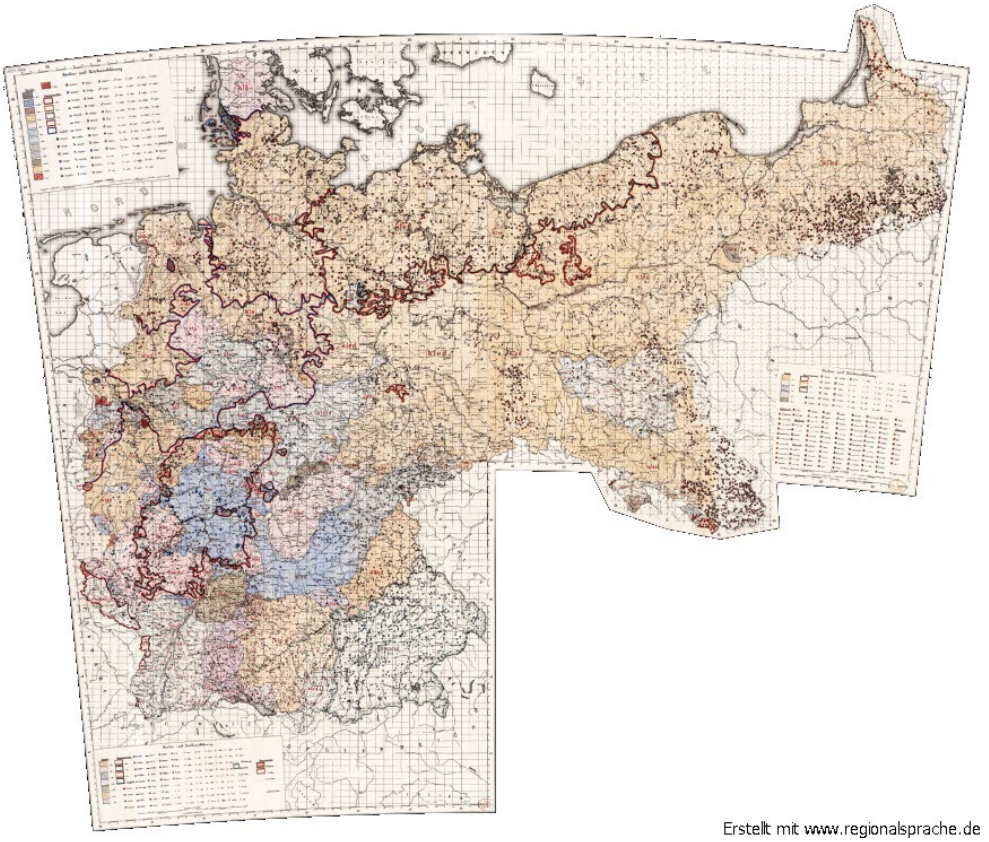


Figure 13: Full information for "Kleider" from Wenker's *Sprachatlas des Deutschen Reichs*, all sectors displayed "stitched together" in the REDE digital environment (www.regionalsprache.de – generated 18 September 2016).

provide for our analyses and interpretations new complementary and supplementary information previously not available (see Figure 14 for an example[6]).

6 As explained by Kevin McGowan (personal communication): "A common problem in phonetics and related fields is the need to visualize many vowel measurements together in a comprehensible way [the actual vowel measurement data also available due to technological advances – MRL]. Simply plotting these measurements as individual points can be uninformative or even misleading. Christian DiCanio of SUNY Buffalo proposes (2013) this novel method of using R (R Core Team 2016) with the ggplot package (Wickham 2009) to instead present the distribution of the vowel, using kernel density estimation to reveal patterns in the measurements that were previously difficult or impossible to discern. [In Figure 14] the distribution density plot reveals bimodal distributions for several vowel quality categories suggesting that a dimension other than simply the F1 (height) or F2 (backness) vowel formant measures

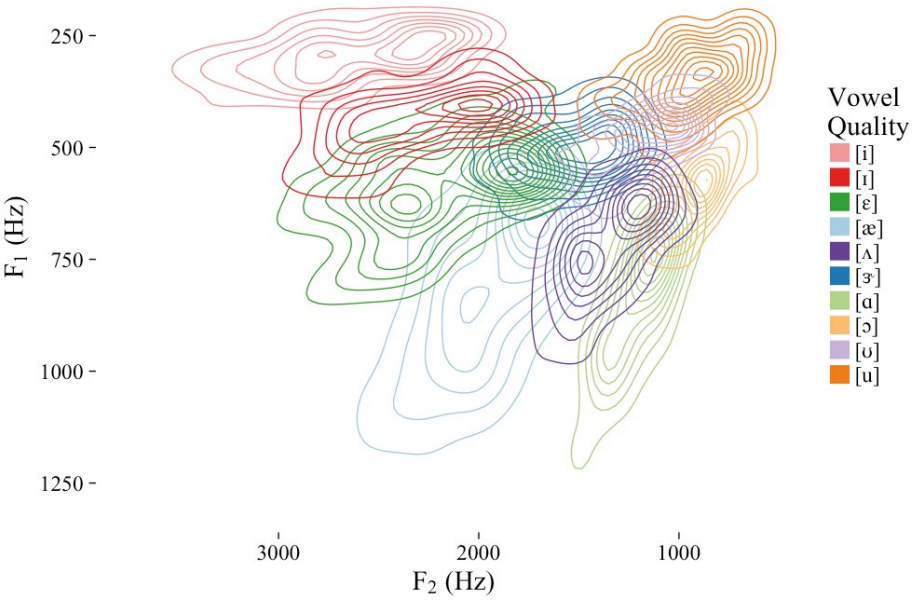


Figure 14: The classic vowel measurements of Peterson and Barney (1952), replotted using kernel density estimation to enhance clarity (image from Kevin McGowan, personal communication).

Summarization and presentation of research outcomes

A third general area in which visualizations have become commonplace in linguistic research is as a means of summarizing for presentation the outcomes of data analysis and interpretation, illustrating the results of an investigation. Similar to visualizations of theoretical models, these visualizations generally serve to render tables of numbers, lists of linguistic data, or extended prose into a visually digestible form (see Figure 15 and Figure 16).

A different, but familiar, example of the use of visualizations to summarize research results can be seen in Figure 17.

> can be expected to explain much of the observed variation. Indeed, replotting these data separately by speaker gender results in largely unimodal vowel distributions with much less overlap across vowel quality categories".

morphological unit	region	pattern*	no pattern	insufficient data
1) 1st sg. n-p. thematic verbs I, II, III	MSlk		X	
	WSlk	X		
	CSlk		(X)**	X
	ESlk		X	
2) instr. sg. masc. & neut. nouns	MSlk	X		
	WSlk		X	
	CSlk		X	
	ESlk	X		
3) dat. pl. masc. & neut. nouns	MSlk		X	
	WSlk	X		
	CSlk	X		
	ESlk	X		
4) instr. pl. masc. & neut. nouns	MSlk	(X)**		X
	WSlk		X	
	CSlk	X		
	ESlk	X†		
5) loc. pl. masc. & neut. nouns	MSlk			X
	WSlk			X
	CSlk			X
	ESlk			X

morphological unit	region	pattern*	no pattern	insufficient data
6) gen./dat./loc. sg. fem. hard-stem adjs.	MSlk		X	
	WSlk	(X)††	X	
	CSlk	X		
	ESlk	X		
7) loc. sg. masc. & neut. hard-stem adjs.	MSlk	X†		
	WSlk		X†	
	CSlk	X		
	ESlk	(X)**		X
8) dat./loc. 2nd sg. & refl. pronouns	MSlk	(X)**		X
	WSlk	X†		
	CSlk	X		
	ESlk	(X)**		X
9) 1st sg. pres. of **byti*	MSlk	(X)**		X
	WSlk	X		
	CSlk	X		
	ESlk		X	

**Throughout this table, a parenthetical "(X)" indicates a possible alternative to "X". An explanatory note describes the nature of the alternative.* ***This is an extrapolated interpretation. The data set is too small to mark the result as certain.* *†This result would benefit from verification with additional data.* *††Possible tendency toward region-wide patterning of* -ej *ending.*

Figure 15: Summary table showing the distribution of patterns of morphological variants across geographical space with measures of certainty and type of patterning (Lauersdorf 2010, 160–161).

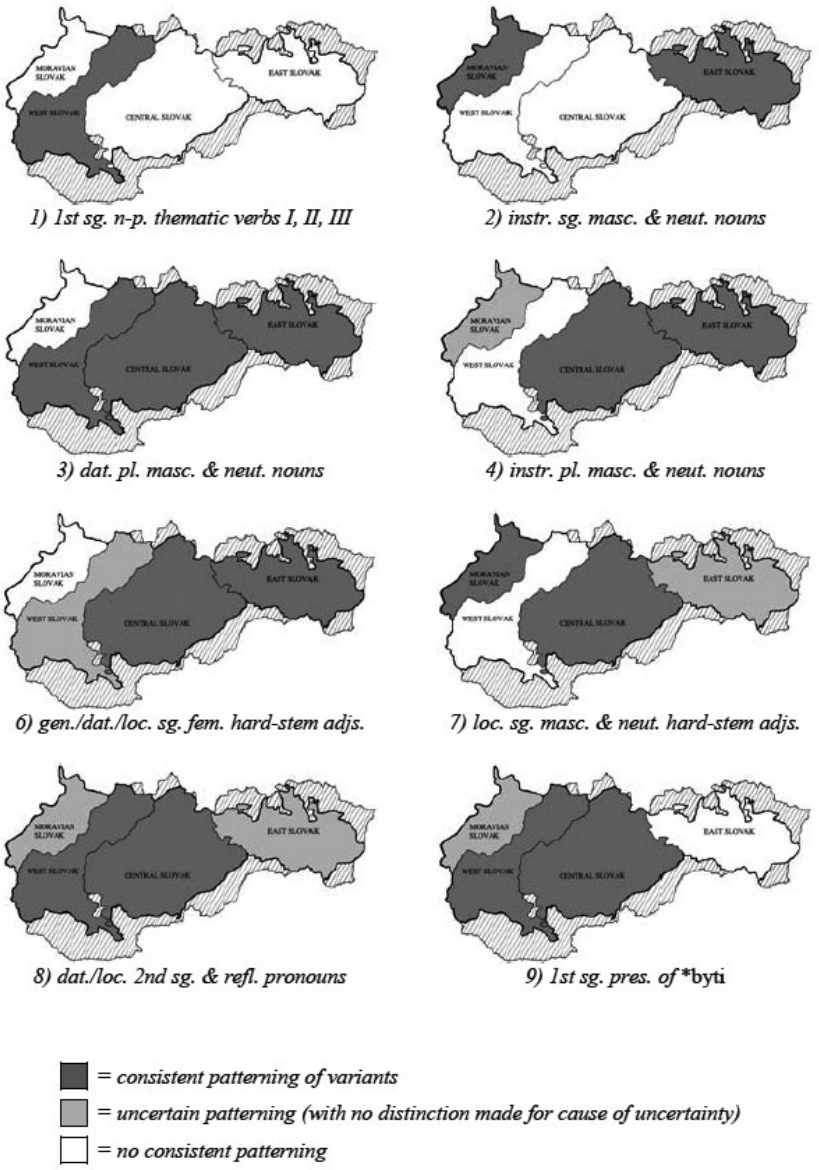


Figure 16: Geospatial visualization of the information in Figure 15 (Lauersdorf 2010, 163).

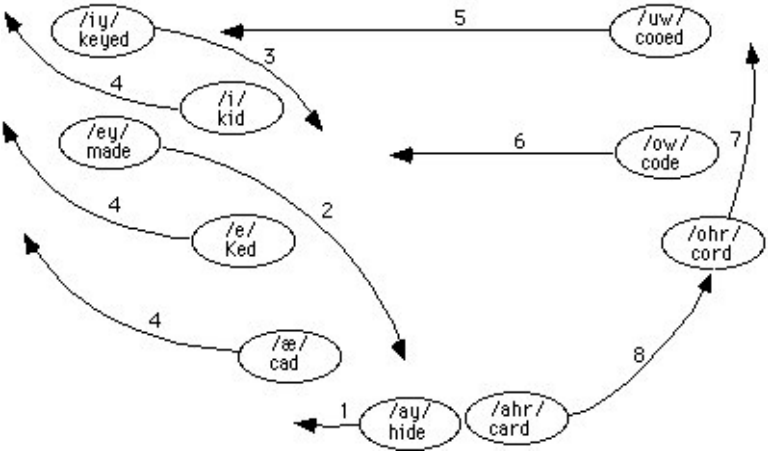


Figure 17: William Labov's schematic overview of the Southern vowel shift in North American English (http://www.ling.upenn.edu/phono_atlas/ICSLP4.html – accessed 09 October 2016).

Figure 18 and Figure 19 provide examples of visualizations that present research outcomes in geospatial and tree form once again.

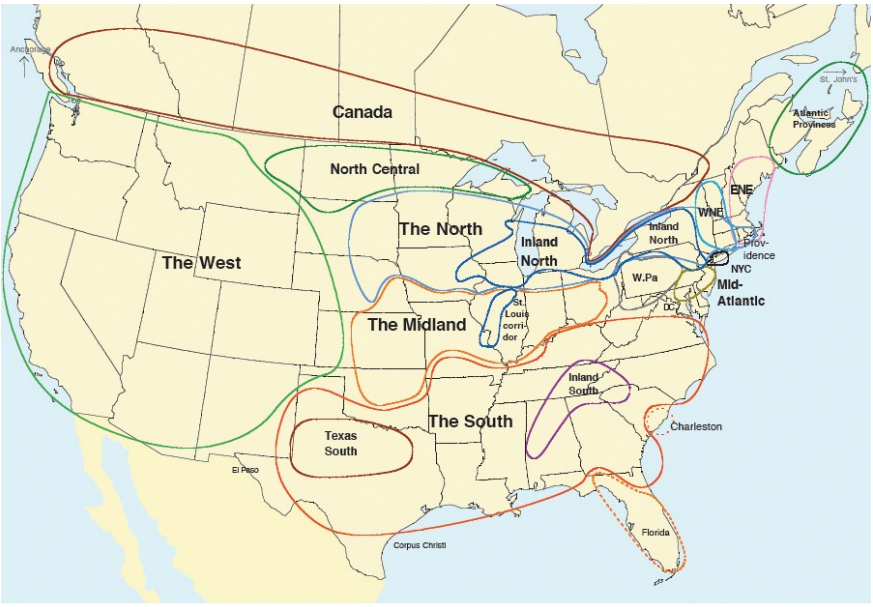


Figure 18: Labov, Ash, Boberg's overview of the major dialect divisions in North American English (2006, 148).

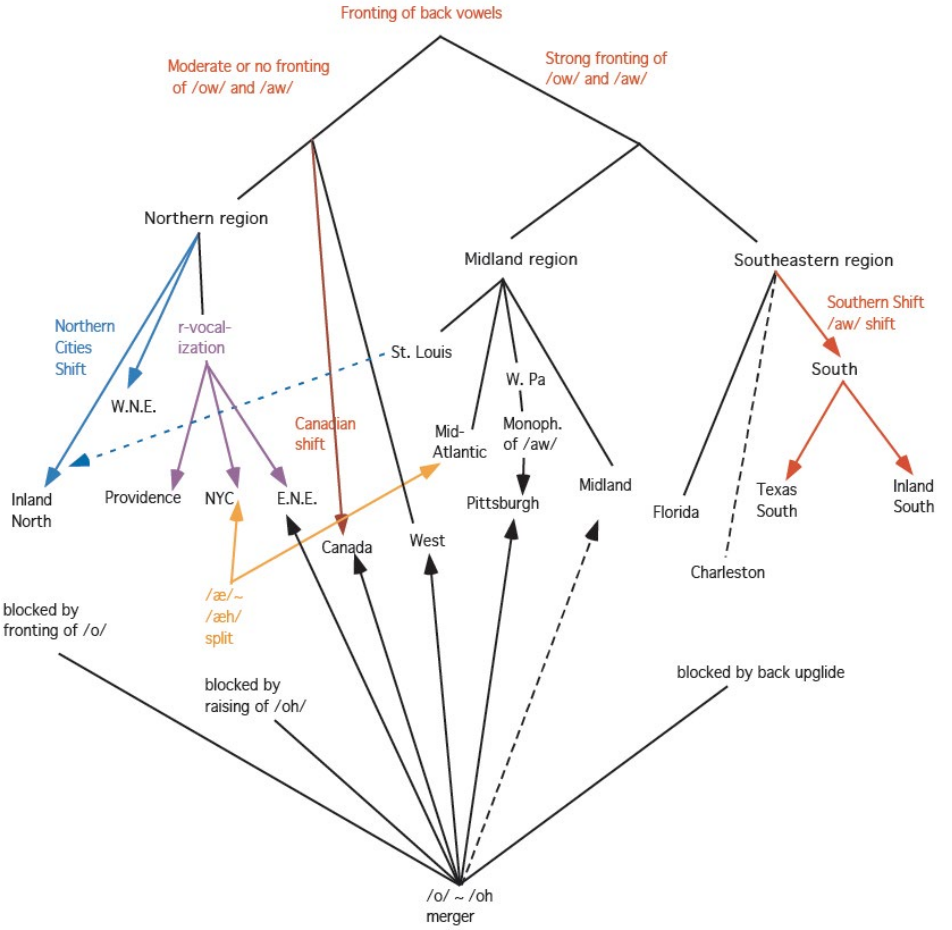


Figure 19: Labov, Ash, Boberg's schematic tree illustrating their "hierarchical structure of North American dialects" (2006, 147).

What visualization accomplishes in these tasks

In each of the three general tasks outlined above, the end goal of the use of linguistic visualization is the same: to gain insight and to facilitate/improve results.

- theoretical models → gain insight into and provide better comprehension and testing of the model
- data analysis → gain insight into and provide better analysis and interpretation of the data

— research outcomes ⟶ gain insight into and provide better comprehension and testing of the outcomes

If there is no additional insight provided by a given visualization, or if there is no facilitation or improvement of the theoretical model, the data analysis, or the research results, then we might question the use of the visualization. We might also question the use of a given visualization in the context discussed earlier where we have perhaps locked in a specific form or type of visualization, considering it immutable/unchangeable, and we then interpret all new data and arguments through that fixed form whereby the visualization becomes an *iconic* representation of the underlying information. Repeating Chvany's cautionary statement, "But the question remains, is the cube [or any other given visualization] the best possible icon of the system?" (1987, 218).

Jakobson's cube as cautionary tale

In the same way that the cube has become *the* visualization of Jakobson's Russian case theory, the different classic visualizations presented above to illustrate theoretical models, data analyses, and research outcomes have become *iconic* for the information that they represent. We all recognize, without any explanation or clarification, how we are to interpret tree diagrams, dialect maps, syntax trees, and vowel shift diagrams because they have become standard visualizations for the information that they illustrate. Chvany states about Jakobson's cube: "The controversies surrounding the cube seem strangely out of proportion to its importance as a theoretical construct. [...] *There is nothing sacred about the cube*" (1987, 218 – emphasis added, MRL). In the same way, we would be wise to question ourselves regarding some of the standard visualizations that have become iconic in our areas of study. Have we objectified these visualizations thereby fixing (locking in) their form and rendering them immutable/unchangeable (iconic)? And do we interpret all new information and arguments through these fixed forms, potentially focusing, in our subsequent analyses, more on the visualization than on the information behind it? And what are we missing if we are, in fact, doing this?

In the iconicity of accepted, standard visualizations:

— we are potentially missing some of the data or some of the relations between the data;
— we are potentially not allowing for all the possible data combinations;
— we are potentially not seeing all the angles;
— we are potentially missing opportunities to try different techniques.

Chvany provides discussion and illustration of several of these points in regards to the Jakobson cube visualization (1987, 218–219), exemplified here in Figure 20 and Figure 21.

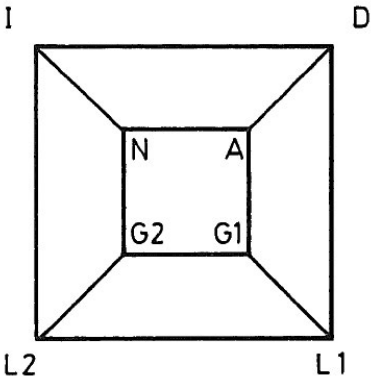


Figure 20: Chvany's alternative visualization of Jakobson's cube (illustrating our notion of viewing *from a different angle*) giving more emphasis to the "central-peripheral distinction" of the theory (1987, 218).

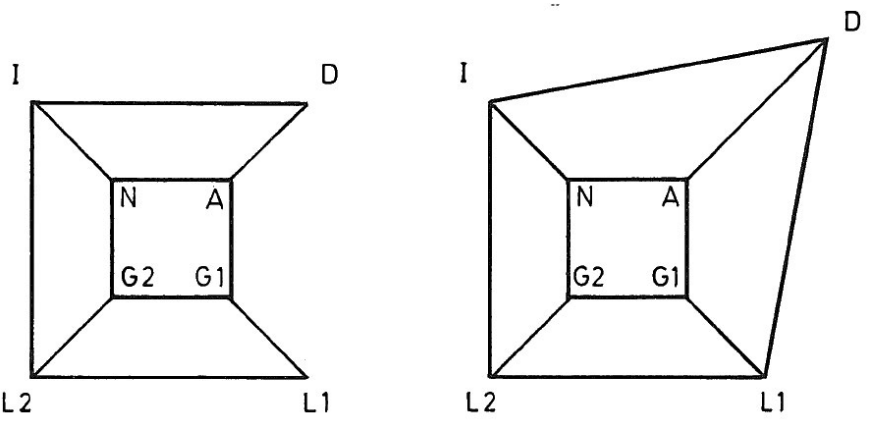


Figure 21: Chvany's further alternatives to Jakobson's cube (illustrating our notions of viewing *from a different angle and showing different relations* between the structures) (1987, 219).

In discussing alternative visualizations of the Russian case system (from different angles or showing different relationships) Chvany states: "The drawing is two-dimensional, and the three-dimensional illusion is not needed to represent the system …" (1987, 218), and "Jakobson actually refers to hierarchization (1958, 175) … . One weakness of the cube figure as an icon of the Russian case meanings is that its dominant reading is ahierarchical. A hierarchy is better represented, and tested, in a tree-shaped model" (1987, 219). This again speaks directly to one of our cautions above that, in locking in one iconic visualization, there is the potential for missing some of the relations between the data.

3. Applying visualizations to linguistics

A contextualization from historical sociolinguistics

Much of the work that I do in the field of historical sociolinguistics involves complex situations of historical language contact with:

— a high number of language varieties in contact;
— no identified standard language or prestige variety;
— a multitude of geographical and political borders;
— quickly changing socio-cultural, socio-political, socio-economic contexts.

What I seek to investigate about those situations is:

— what is the impact of the language contact on the structures of the language varieties in contact?
— what patterning of structural features can be seen across the varieties in contact (is there dialect leveling, koinéization, etc.)?
— if patterning is detected, what type, degree, location, domain, etc. does it demonstrate?
— can specific socio-historical factors be correlated to the structural patterning?

Given that this work is being performed for historical language periods, there is the issue of the so-called "bad data problem", made famous by Labov in his statement that "… [h]istorical linguistics can … be thought of as the art of making the best use of bad data" (1994, 11). Importantly for our discussion here, this is often re-cast by historical linguists as a problem of "imperfect" data (Joseph and Janda 2003, 14), or "making the best use of the data available" (Nevalainen and Raumolin-Brunberg 2003, 26). Given that the available data is "imperfect" (i.e., limited,

fragmentary, or incomplete), it is imperative to gather as much of it as possible for a given investigation, from all interrelated sources, linguistic and socio-historical – in other words, it is imperative to use all the data! This is especially true for the type of investigation that I described at the beginning of this section involving historical language analysis through data-driven pattern identification and correlation with socio-historical factors. At this point it is important to note that, while the discussion in the remainder of this section derives from a specific application in historical sociolinguistics (as described above), I believe that the arguments presented are applicable to any context of visualization in linguistic analysis.

Use all the data!

As stated, historical data (linguistic or otherwise) tends to be "imperfect" data (i.e., limited, fragmentary, incomplete), and generally speaking, the earlier the time period under investigation, the "more imperfect" the data. Thus, if we hope to achieve generalizable results from historical sociolinguistic investigations, it becomes necessary to gather as much of the data as possible from all interrelated sources. Even when dealing with contemporary linguistic data, gathered in the field, in the lab, or in a corpus, it can perhaps never be guaranteed that we are working with "perfect" data, i.e., data that is *not* limited, fragmentary, or incomplete in some respect. In this way, the call to "use all the data" certainly has broader application beyond historical sociolinguistics.

Logically, if you use all the data, you have to process all the data in your analysis. And if you have to process all the data, you will very likely need to use statistics and visualization for data analysis. The need for statistical and visual assistance in analysis can be driven by the size of the dataset, wishing, for example to isolate relevant information in a large dataset or to determine viability and significance in a small dataset. It can be driven by the multifactored nature of the dataset with the interaction of many different data types. It can be driven by the type of information that we wish to extract from the dataset (e.g. correlational information about multiple variables).

In all of this I believe that there is an implied, and very important, set of corollaries regarding data visualization:

→ Use all the data.
→ If you use all the data, *view* all the data.
→ If you view all the data, view all the *combinations*.
→ If you view all the data, view all the *angles*.
→ If you view all the data, use all the *techniques*.

It should be mentioned that I am working through the rationale and the arguments here on the basis of the visualization task of data analysis and interpretation, but the basic tenets of these propositions are also easily transferable to visualization for the elucidation and elaboration of theoretical models and to visualization for the summarization and presentation of research outcomes.

View all the data

Often we decide directly, or the visualization technique that we choose decides indirectly for us, which subsets of the data we end up viewing; and in both cases, the power of the visualization is limited by the decisions made about which data subsets to view. If, in performing our linguistic visualizations, we make *a priori* decisions, for whatever reasons, concerning the subset(s) of the data that should be visualized, we run the risk of potentially missing patterns in the overall dataset. Even in very large datasets, where an initial visualization of the entire dataset could be as dense and opaque to interpretation and analysis as the raw dataset itself, the use of visualization could have the potential to show subdivisions in the dataset that *a priori* pre-visualization decisions would miss. It is necessary to *view all the data*.

A first reaction to Figure 22 might be that, in viewing all the data, the overall amount of data, and the various parameters ascribed to it, create a visualization that is difficult to parse for the purposes of data analysis. However, within the specific framework of the investigation, the authors of the study note, on the contrary, that "At first sight it is apparent that the structure of the city does not only reflect the political dimension discussed earlier: Figure 7 [Figure 22 here] clearly reveals the segregation into different social classes. Some actors only appear in combination with very few events whereas others are highly integrated in the center of the structure. In the center of events we find the craftsmen, the clerks, the merchants, and the educated bourgeoisie, whereas the vintners and the workers are basically linked to the periphery of the system." (Krempel and Schnegg 1999). This conclusion would likely not have been possible without viewing all the data, and it allows for potential subdivision of the dataset for deeper analysis on the basis of having specifically viewed all the data.

Relatedly, if, in performing our linguistic visualizations, we use only the iconic standard visualizations commonly employed for the specific type of data we are analyzing, we might potentially miss some of the data or some of the relations between the data, either through our preconceived notions of what the visualization should show, or through actual restrictions on the data that the visualization can accept, or the type of relations it can show. (We will return to

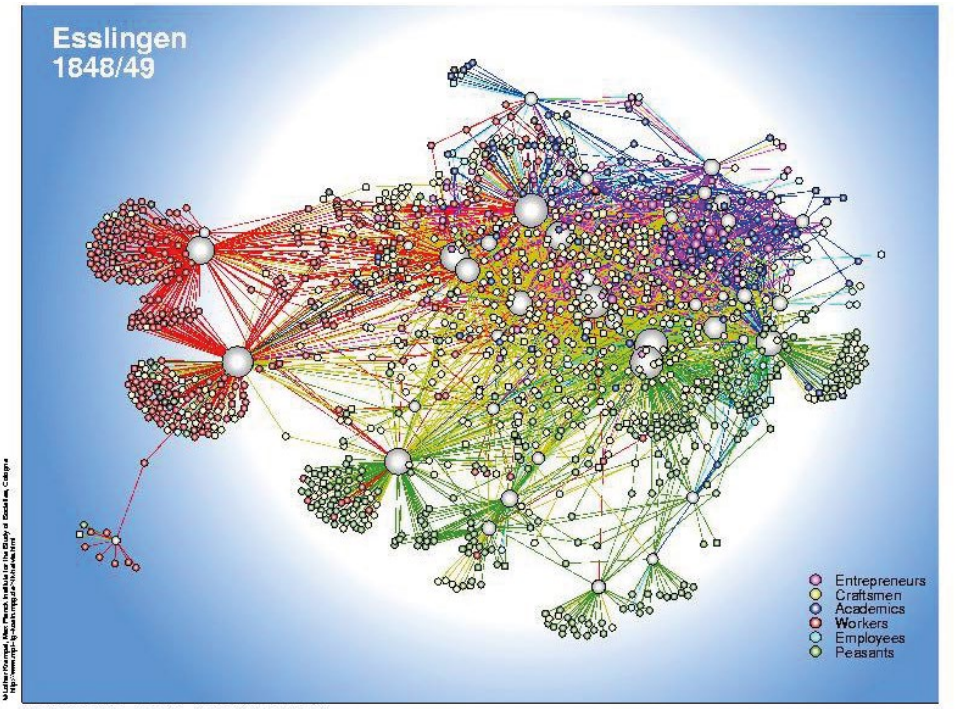


Figure 22: Krempel and Schnegg's (1999) visualization of the social landscape of Esslingen, Germany, 1848/49.

this notion of chosen visualization techniques disallowing a view of all the data in the section on "using all the techniques".)

View all the combinations

As mentioned above, very large datasets have the potential, if viewed with all data points and parameters, to produce highly complex visualizations that may be largely impenetrable to interpretation, so "viewing all the data" at once may not be of much assistance in data analysis – the visual density may be too great, or there may be mixed types of data that are difficult to bring together in a single visualization. On the other hand, in breaking the data down into subsets to assist in visual analysis, *a priori* assumptions about the parts of the data that should be combined and the parts that should be excluded in any given visualization will potentially cause us to miss patterns in our analysis because we perhaps did not bring the appropriate parts of the data together, in our *a priori* selection, to

adequately reveal the correlations that might exist. In cases where it becomes procedurally/methodologically necessary to view the data in subsets, it becomes necessary to *view all the combinations*.

The risk of not viewing all the combinations, and thereby potentially missing patterns in the analysis, also arises if we rely exclusively on the use of iconic standard visualizations. If we always apply only the same standard visualizations to the data we are working with, we are potentially missing some of the combinations of the data either through our preconceived notions of what combinations these visualizations should show, or through actual restrictions on the data combinations the visualization technique can accept, or the types of combinations it can show.

View all the angles

A simple graphic illustration will demonstrate the importance of considering multiple views or angles of the relationship between data points in a visualization. Consider a representation of the connections between data points plotted in a two dimensional square. It is a square when viewed head-on, and a different square (but still a square) from behind (Figure 23).

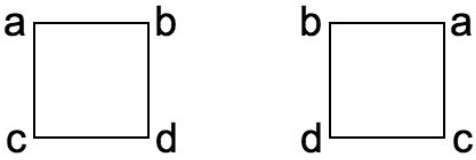


Figure 23: Data points "a", "b", "c", "d" plotted as a square, *viewed from front and back.*

From the top it is a horizontal line, and from the bottom it is a different horizontal line (Figure 24).

Figure 24: Data points "a", "b", "c", "d" plotted as a square, *viewed from top and bottom.*

From one side it is a vertical line, from the other side a different vertical line (Figure 25).

Figure 25: Data points "a", "b", "c", "d" plotted as a square, *viewed from left and right sides.*

From a front angle it is a quadrilateral with two different types of relations between the points, depending on the angle (Figure 26).

Figure 26: Data points "a", "b", "c", "d" plotted as a square, *viewed from left and right front angles.*

We can appreciate already from this simple demonstration (that does not nearly exhaust the possible angles of view on a square) that the interpretation and analysis of data in any given visualization could be considerably affected by the angle of view on the visualization. It thus becomes necessary to *view all the angles*, or at the very least, it is necessary to view more than one angle, in order to be aware of the effect that the angle of view might have on our understanding of the data. An even greater appreciation for this notion of viewing all the angles can be achieved by adding just one additional dimension to the square and considering the cube from multiple angles (Figure 27 – rotated here *only* 90 degrees, *only* on its central vertical axis).

The question must then be posed: what are we missing in the visualizations of our data, (especially in our iconic standard visualizations) if they are static, with no interactive or dynamic component that allows for different views on the data?

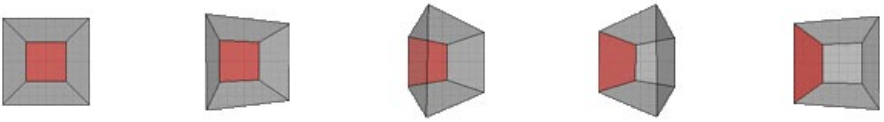


Figure 27: Cube visualization rotated 90 degrees on its central vertical axis (http://www.traipse.com/hypercube/ – accessed 09 October 2016).

Use all the techniques

Earlier in this text, and particularly in each of the last three sections, part of the discussion has revolved around the limitations of using only the visualization techniques that have become commonplace and common practice ("iconic") in the field in which any individual investigation is anchored. One way to escape this *iconicity* problem, where our reliance on iconic standard visualization tools and techniques causes us to potentially not use all the data, not view all the data, not view all the combinations, or not view all the angles, is to *use all the techniques* (or at least *consider more of* the techniques). Of course, there are certain limitations, both methodological and practical, that often prevent us from truly testing *all* of the visualization techniques available. Specific data types match with specific statistical models, and other data types match with other statistical models; and some data types match better with certain visual representations than with others. But trying something out of the ordinary (i.e., a non-iconic or non-typical visualization) may yield extraordinary results.

It should be kept in mind here that the call to *use all the techniques* does not exclude the iconic standard visualizations, but rather begins with them and then goes beyond them. The argument here is that employing non-typical visualizations, in addition to the iconic standard visualizations, provides the potential for additional scientific gains. Bubenhofer (2018) echoes this, stating: "In order, however, to allow for innovation in the area of scientific visualization, this canon of disciplinary visualization practices must constantly be called into question"[7] (S. 45). As just one example of the insights to be gained from using innovative techniques beyond the iconic standard visualizations in a given field, Montgomery (2012) convincingly demonstrates that the visualization of geospatial data in

7 "Um aber Innovation im Bereich wissenschaftlicher Visualisierung zu ermöglichen, muss dieser Kanon disziplinärer Visualisierungspraktiken immer wieder hintergangen werden."

a graph form allows for analysis and interpretation beyond what is possible with the original geospatial representation of the data (see Figure 28 and Figure 29).[8]

Linguistic visualizations as *objets d'art*?

Interestingly, the deliberate use of non-typical visualization tools and techniques to allow the observer to see something not ordinarily seen, or not otherwise perceptible, is a feature of *objets d'art* as well. Different works of art allow us to perceive the world in different ways by presenting the "information" of the world in different combinations, from different angles, using different media and techniques (i.e., different types of "visualizations"), and thus works of art often give us new appreciation, understanding, and insight into the information that they are portraying, even if we have already seen the same information represented in other works of art (i.e., other "visualizations") or in its original form in the world (as "raw data").

Of course, it was *not* the intent of this discussion to examine the viability of a direct equation between linguistic visualizations and artworks, nor was the intent to determine whether linguistic visualizations demonstrate some sort of creative, interpretive, esthetic, or other equivalency with artworks. As stated at the outset, the concept of *objet d'art* served here as a notional category, deployed to draw attention to a range of issues that must be considered in the use of visualizations in linguistic research. The use of visualization tools and techniques has the potential to provide new insights and to facilitate/improve outcomes in linguistic theories, analyses, and results. As we view all the data, all the combinations, and all the angles, using all the techniques, we must simply remain

8 Montgomery's explanation of the maps in Figure 28: "The data gathered for the North-South divide question for each location in Study 1. Each line drawn by respondents is included on the three maps: (a) respondents from Carlisle (67 lines drawn); (b) respondents from Crewe (61 lines drawn); (c) respondents from Hull (72 lines drawn)" (2012, 652). This geospatial data was then converted to graph form (Figure 29): "Consequently, ImageJ (Rasband 2011 [now 2016 – MRL]) was used to interrogate the data more closely. The programme includes a tool that permits analysis of image luminance, which is well suited to investigating the placement of North-South lines as a greater density of lines reduces luminance. ImageJ creates a 3D graph for each map with luminance interpreted on the z-axis of the graph. The luminance value is inverted (so lesser luminance appears as a spike on the graph) and a smoothing technique applied to the data which removes some of the individual variance and permits the investigation of greatest agreement. The 3D image is then rotated in order that a 2D 'slice' of the image viewed from north to south can be captured. This 2D 'slice' allows a user to examine how far north or south North-South lines have been placed by respondents in conjunction with the composite line maps" (2012, 653).

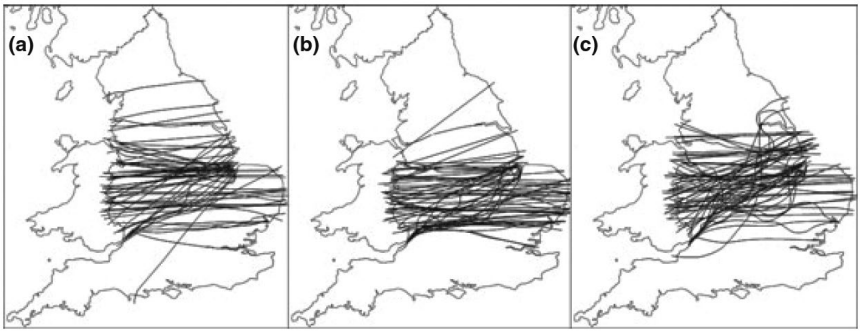


Figure 28: Montgomery's maps showing geospatial locations of north/south dividing lines drawn for a perceptual dialectology study of England (2012, 652).

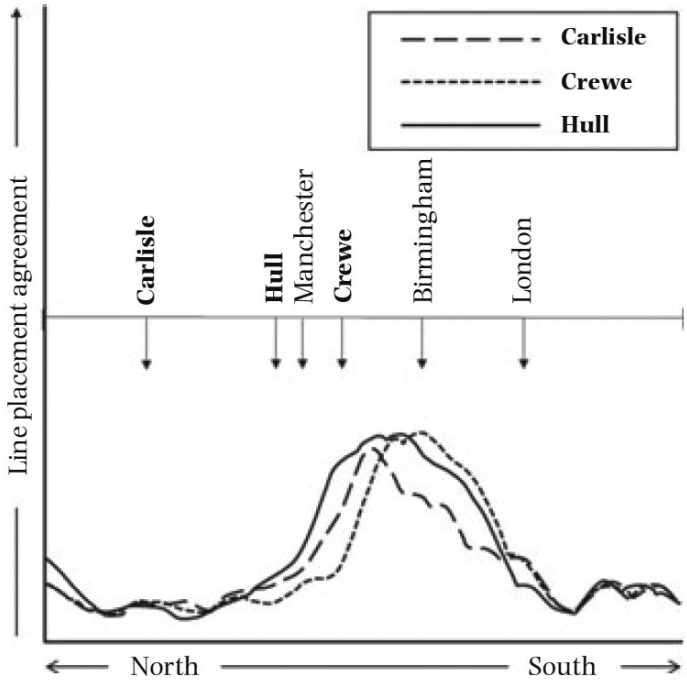


Figure 29: Montgomery's graph showing relative locations and density of geospatial bundling of north/south dividing lines drawn for a perceptual dialectology study of England (2012, 653).

mindful of the different ways in which visualizations have the potential to not only show us more, but also influence what we see.

4. References

Bubenhofer, Noah. 2018. "Visual Linguistics: Plädoyer für ein neues Forschungsfeld." Bubenhofer, Noah. 2018. "Visual Linguistics: Plädoyer für ein neues Forschungsfeld." In *Visualisierung sprachlicher Daten*, edited by Noah Bubenhofer and Marc Kupietz. Heidelberg: Heidelberg University Publishing, 25-62.

Chvany, Catherine V. 1984. "From Jakobson's Cube as *Objet d'Art* to a New Model of the Grammatical Sign." *International Journal of Slavic Linguistics and Poetics* 29: 43–70.

Chvany, Catherine V. 1987. "Jakobson's Cube as *Objet d'Art* and as Scientific Model". In *Language, Poetry and Poetics – The Generation of the 1890s: Jakobson, Trubetzkoy, Majakovskij (Proceedings of the First Roman Jakobson Colloquium, at the Massachusetts Institute of Technology, October 5–6, 1984),* edited by Krystyna Pomorska, Elżbieta Chodakowska, Hugh McLean, and Brent Vine. Berlin: Mouton de Gruyter, 199–230.

DiCanio, Christian. 2013. "Visualizing vowel spaces in R: from points to contour maps." http://christiandicanio.blogspot.com/2013/10/visualizing-vowel-spaces-in-r-from.html (accessed 09 October 2016).

Gilliéron, Jules, and Edmond Edmont.1902–1910. *Atlas linguistique de la France.* 35 fascicles. Paris: Honoré Champion. High resolution scans of the linguistic data maps: http://cartodialect.imag.fr/cartoDialect/ (accessed 09 October 2016). PDF version of all atlas fascicles: http://diglib.uibk.ac.at/ulbtirol/content/titleinfo/149029 (accessed 09 October 2016).

Jakobson, Roman. 1958. "Morfologičeskie nabljudenija nad slavjanskim sklonenie m (sostav russkix padežnyx form)." In *American Contributions to the Fourth International Congress of Slavists.* The Hague: Mouton. [Reprinted in: *Roman Jakobson. Selected Writings II: Word and Language.* The Hague: Mouton, 1971, 154–183. English translation: "Morphological Observations on Slavic Declension (The Structure of Russian Case Forms)." In *Roman Jakobson. Russian and Slavic Grammar: Studies 1931–1981,* edited by Linda R. Waugh, and Morris Halle. Berlin: Mouton, 1984, 105–133.]

Jastrow, Joseph. 1899. "The Mind's Eye". In *Appletons' Popular Science Monthly* LIV (November 1898 to April 1899), edited by William Jay Youmans. New York: Appleton, 299–312.

Joseph, Brian D., and Richard D. Janda. 2003. "On Language, Change, and Language Change – Or, Of History, Linguistics, and Historical Linguistics." In *The*

Handbook of Historical Linguistics, edited by Brian D. Joseph, and Richard D. Janda. Oxford: Blackwell, 3–180.

Krempel, Lothar, and Michael Schnegg. 1999. "Exposure, Networks, and Mobilization: The Petition Movement during the 1848/49 Revolution in a German Town." http://www.mpi-fg-koeln.mpg.de/%7Elk/netvis/exposure/mobv5_all.html (accessed 09 October 2016).

Labov, William. 1994. *Principles of Linguistic Change: Internal Factors*. Oxford: Blackwell.

Labov, William, Sharon Ash, and Charles Boberg. 2006. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: Mouton de Gruyter.

Lauersdorf, Mark Richard. 2010. *The Morphology of 16th-Century Slovak Administrative-Legal Texts and the Question of Diglossia in Pre-Codification Slovakia*. München: Sagner (Slavistische Beiträge, 473).

McMahon, April, and Robert McMahon. 2005. *Language Classification by Numbers*. Oxford: Oxford University Press.

Mel'čuk, Igor A. 1983. "Studies of the Russian Language." In *Roman Jakobson: What He Taught Us*, edited by Morris Halle, 57–71 (*International Journal of Slavic Linguistics and Poetics* 27: Supplement).

Montgomery, Chris. 2012. "The effect of proximity in perceptual dialectology." *Journal of Sociolinguistics* 16, no. 5: 638–668.

Necker, L.A. 1832. "Observations on some remarkable Optical Phænomena seen in Switzerland; and on an Optical Phænomenon which occurs on viewing a Figure of a Crystal or geometrical Solid." *The London and Edinburgh Philosophical Magazine and Journal of Science* 1, no. 5: 329–337.

Nevalainen, Terttu, and Helena Raumolin-Brunberg. 2003. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. London: Longman/Pearson Education.

Neidle, Carol J. 1982. "Case Agreement in Russian." In *The Mental Representation of Grammatical Relations*, edited by J. W. Bresnan. Cambridge, MA: MIT Press, 391–404.

Peterson, Gordon E., and Harold L. Barney. 1952. "Control Methods Used in a Study of the Vowels". *The Journal of the Acoustical Society of America* 24, no. 2: 175–184.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf (accessed 09 October 2016).

Rasband, Wayne. 2016. *ImageJ*. Bethesda, Maryland: U.S. National Institutes of Health. http://imagej.nih.gov/ij/ (accessed 09 October 2016).

Schleicher, August. 1861. *Compendium der vergleichenden Grammatik der indogermanischen Sprachen I: Kurzer Abriss einer Lautlere der indogermanischen*

Ursprache, des Altindischen (Sanskrit), Alteranischen (Altbaktrischen), Altgriechischen, Altitalischen (Lateinischen, Umbrischen, Oskischen), Altkeltischen (Altirischen), Altslawischen (Altbulgarischen), Litauischen und Altdeutschen (Gotischen). Weimar: Böhlau.

Stewart, Ann Harleman. 1976. *Graphic Representation of Models in Linguistic Theory*. Bloomington: Indiana University Press.

Wenker, Georg. 1888–1923. *Sprachatlas des Deutschen Reichs*. Hand drawn by Emil Maurmann, Georg Wenker and Ferdinand Wrede. Project site for the *Digitaler Wenker-Atlas* (2001–2009): http://www.diwa.info/titel.aspx (accessed 09 October 2016). Current host project of the digital version of the *Wenker-Atlas*, *Regionalsprache.de (REDE)*: https://www.regionalsprache.de/ (accessed 09 October 2016).

Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Jana Pflaeging

Zur Ästhetisierung linguistischer Wissensvermittlung

Abstract Dieser Beitrag argumentiert dafür, dass die bis dato sprachdominierten Kommunikationsmuster des linguistischen Wissensaustausches nicht zwingend im Forschungsgegenstand selbst begründet liegen, sondern vor allem in der tradierten Annahme, dass sich Sprache besser zur treffenden Formulierung abstrakt-theoretischer Gedanken eigne. Bildliche Sprache und Metaphern waren seit jeher Bestandteil geisteswissenschaftlicher Denkweisen und Rhetorik. Ich möchte hier schrittweise aufschlüsseln, wie sie zum wertvollen Ausgangspunkt der Erzeugung komplexer Visualisierungen werden können, die den Musterbruch in einem bedeutungspotenzierenden – und kommunikativ völlig natürlichen – multimodalen Gesamttext suchen. Um zu zeigen, wie mittels bewusster multimodaler Brüche mit traditionell eher sprachdominanten Kommunikationsmustern der Linguistik *epistemisches* und auch *motivationales* Potenzial entstehen kann, beginne ich mit einem Exkurs zu „Ästhetik" und „Ästhetisierungsprozessen" und führe das Konzept der „wilden Semiose" ein. Anschließend stelle ich meine theoretischen Überlegungen zur Visualisierung linguistischer Theorien vor und stelle ihre Tauglichkeit im Hinblick auf linguistische Vermittlungskontexte am Beispiel der *Theorie der konzeptuellen Metapher*, aber auch in den Bereichen der Textlinguistik bzw. Gesprächsforschung, unter Beweis. Ich hoffe nicht zuletzt, auf diese Weise zeigen zu können, wie gewinnbringend es sein kann, die kommunikativen Praktiken des Forschens und Lehrens im eigenen Fach selbstreflexiv zu hinterfragen und gezielt nach neuen Wegen der Wissensvermittlung zu suchen.

1. Einleitendes

Verengt man den Blick etwas und überfliegt die ersten Seiten dieses Beitrags, so verdichten sich die Zeichen zu einem für Textsorten der geisteswissenschaftlichen Wissensvermittlung typischen Muster: In Kapitel strukturiert verläuft verbale Sprache in den genormten Bahnen einer akademischen Publikationspraxis. Nur selten wird mit diesen logozentrischen Mustern gebrochen, um beispielsweise das kommunikative Potenzial von Visualisierungen und anderen

Erschienen in: Noah Bubenhofer u. Marc Kupietz (Hrsg.): Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

alternativen Darstellungsformen zu erproben. Die Geisteswissenschaften – und die Linguistik im Besonderen – gelten noch immer als „[e]her […] bilderscheu" (Ballstaedt 2012: 17; siehe Pflaeging 2013; Hauser et al. 2016). Ein Grund hierfür ist vermutlich, dass verbale Sprache eine treffende Formulierung gerade *theoretischer*, d. h. abstrakt-konzeptueller Gedanken, scheinbar am besten unterstützt (Ballstadt 2012: 17).

In diesem Beitrag möchte ich dafür plädieren, dass die bis dato sprachdominierten Kommunikationsmuster des linguistischen Wissensaustausches vor allem in diesen tradierten Annahmen begründet liegen, jedoch nicht zwingend im Forschungsgegenstand selbst. Bildliche Sprache und Metaphern waren seit jeher Bestandteil geisteswissenschaftlicher Denkweisen und Rhetorik. Ich möchte nachfolgend aufschlüsseln, wie sie zum wertvollen Ausgangspunkt einer Erzeugung komplexer Visualisierungen werden können, die den Musterbruch in einer bedeutungspotenzierenden – und kommunikativ völlig natürlichen – Multimodalität (siehe Stöckl 2016) suchen. Um zu zeigen, wie mittels bewusster multimodaler Brüche mit traditionell eher sprachdominanten Publikationsmustern der Linguistik epistemisches Potenzial entstehen kann (Kapitel 2.1), scheint es mir nützlich, meine Argumentation mit einem Exkurs zu Ästhetik und Ästhetisierungsprozessen anzubahnen und das Konzept der „wilden Semiose" (Assmann 1988) einzuführen. Anschließend werde ich meine theoretischen Überlegungen zur Visualisierung linguistischer Theorien vorstellen (Kapitel 2.2) und ihre Tauglichkeit im Hinblick auf linguistische Vermittlungskontexte am Beispiel der konzeptuellen Metapherntheorie (siehe Grafiken N° 01 bis 04), aber auch in den Bereichen der Textlinguistik bzw. Gesprächsforschung, unter Beweis stellen (N° 05 und 06).

Ich hoffe nicht zuletzt, mit diesem Beitrag zeigen zu können, wie gewinnbringend es sein kann, die kommunikativen Praktiken des Forschens und Lehrens im eigenen Fach selbstreflexiv zu hinterfragen und gezielt nach neuen Wegen der Wissensvermittlung zu suchen (siehe aber Schmitz 2004; Kuiper 2011; Pflaeging & Schildhauer 2015; Pflaeging 2015a; Pflaeging & Brock 2017; sowie Bestrebungen in anderen akademischen Disziplinen in Huber et al. 2014). Meines Erachtens setzt der vorliegende Band in dieser Hinsicht wertvolle Impulse für eine Linguistik, die das Visuelle (und andere Modalitäten) nicht nur mehr und mehr ins das analytische Blickfeld rückt, sondern die dort offengelegten Mechanismen und Bedeutungspotenziale multimodaler Kommunikate gerade auch für den eigenen *fachinternen* Wissensaustausch zwischen Forschern, Lehrenden und Studierenden erkennt und nutzt. Ich möchte meinen Beitrag als Spielfeld nutzen, um die epistemischen Potenziale von Musterbrüchen in der Linguistik auch mit selbsterstellten Visualisierungen zu erkunden – und hoffe auf ästhetische Erfahrungen und einen *langen verweilenden Blick*.

2. Theoretisches | Ästhetik und eine Visuelle Linguistik

2.1 Ästhetik | Muster und Musterbrüche

Ästhetik sowie das prozessbetonende Derivat *Ästhetisierung* werden heutzutage am ehesten dazu genutzt, um auf „das Kunstschöne" (Stöckl 2013a: 93) als „Gegenstand unserer Hochschätzung und Bewertung" (Welsch 1996: 65) zu referieren. Bei genauerer Betrachtung wird jedoch die enorme Streubreite der Objekte und Phänomene offensichtlich, auf die sich Ästhetik in Alltags- und Fachsprache beziehen kann: Im Bereich der Kunst selbst finden sich mitunter denotationsschwache und konnotationsstarke Verwendungen, wenn beispielsweise auf amazon.de ein Ein-Klecks-Malset für abstrakte Kunst mit einer „Ästhetik-Garantie vom Profi" beworben wird. Darüber hinaus findet der Begriff auch in Bezug auf Musik, Material, Werbung, Warenwelt, und Natur Verwendung oder wird zur Kontrastierung mit Ethik, Funktionalität und Technik herangezogen (Stöckl 2013a: 90–93). Dieses Abweichen des Ästhetikbegriffs von seinem *kunst*-bezogenen Bedeutungskern könnte, so vermute ich, auf ein ursprünglich weites Verständnis des Konzepts als Beschäftigung mit einer „Sinneserfahrung im Allgemeinen" (Allesch 2006: 10) zurückgehen, welches in den Schriften Alexander Gottlieb Baumgartens (um 1850) angelegt ist, sich bis heute in der Verwendungsbreite des Begriffs niederschlägt und zu einer gewissen „begrifflichen Unschärfe und Vagheit" (Stöckl 2013b: 2) geführt hat. Es scheint einzig die in der Natur des Menschen liegende Präferenz für das Schöne gegenüber dem Abstoßenden zu sein, die das Kunstschöne beständig in den Vordergrund rückt und den Blick auf das Wesentliche versperrt: Ein Ästhetikbegriff, der ein Geschmacksurteil als zweitrangig einstuft und genereller auf perzeptuelle Erfahrungen abhebt, birgt deutliche Vorteile gegenüber einer Ästhetik als „Theorie der Kunst" oder „Theorie des Schönen", die zu eng scheinen, weil ästhetische Erlebnisse nicht nur durch Kunstwerke oder schöne Dinge ausgelöst werden können (Reicher 2005: 13). Jedoch greift auch diese Fassung zu kurz, weil nicht jede Sinneswahrnehmung gleich eine ästhetische ist (Reicher 2005: 15).

Ein m. E. außerordentlich nützlicher, weil objektivierbarer Zugriff auf Ästhetik und Ästhetisierungsprozesse ist jedoch innerhalb der Text- und Medienlinguistik (mit Parallelen zum psychologischen Ästhetikverständnis bei Allesch 2006), zunächst von Fix (2001), dann beispielsweise von Stöckl (2013a) ausgearbeitet worden: Obwohl auch hier zumeist kunstschöne Texte im Mittelpunkt der Analyse stehen, wird generell von einem weiten Ästhetikbegriff ausgegangen. Stöckl (vgl. auch Fix 2001: 37ff.) nennt jene Kommunikate ästhetisch, die „Kraft ihrer Materialien und Gestaltungstechniken – d. h. dank ihrer wahrnehmbaren und fühlbaren Hüllen" (Stöckl 2013a: 93) unsere Aufmerksamkeit auf sich ziehen. Wichtiger als ein Geschmacksurteil scheint also, dass ein Artefakt sich „in der

Konkurrenz mit anderen Artefakten durch ästhetische Reize anbietet, die geeignet sind, die Wahrnehmung des Rezipienten so zu lenken, daß er aus der Fülle der Angebote gerade dieses auswählt" (Fix 2001: 39). Hierin deutet sich die m. E. außerordentlich gewinnbringende Anbindung des Konzepts der Ästhetik an den Begriff der *Textsorte* an. Kommunikative Handlungen, die aufgrund ihrer Zweckdienlichkeit im Erreichen kommunikativer Ziele wiederholt ausgeführt werden, verdichten sich im Mentalen zu abstrakten kognitiven Kategorien, d. h. Textsorten, die Orientierung bei Textrezeptions- und -produktionsprozessen bieten. Ihre prototypische Organisation, mit einem merkmalstreuen Kern und von ihm abweichenden peripheren Exemplaren (vgl. Sandig 2000), scheint ein wichtiger Schlüssel zur Modellierung von Ästhetisierung durch Musterbrüche vor dem Hintergrund etablierter Erwartungen erzeugender Muster zu sein. Textexemplare als konkrete Realisierungen von Textsorten, die mit „Wahrnehmungsroutinen" (Stöckl 2013b: 2) und „kulturellen Konventionen" (Fix 2001: 39) brechen, verleiten Textrezipierende womöglich zu einem „langen verweilenden Blick" (Fix 2001: 39), den Aleida Assmann (1988) treffend als „wilde Semiose" bezeichnet hat. Unser Blick bleibt haften, wenn etwas durch die „Routine des Gewöhnlichen und die ewige Wiederkehr des Bekannten hindurchdringt" (Assmann 2015: 18) und „die Zeichen ‚aus dem Ordnungsgefüge konventioneller Beziehungen entlassen' sind, wenn sie ‚neue, unerwartete Beziehungen eingehen' (Assmann 1988: 238). Jeder Ausbruch aus kulturellen Konventionen führt zu den Bedingungen ‚wilder Semiose'" (Fix 2001: 42, vgl. Assmann 1988: 239)

Für die hier geführte Argumentation ist eine Schlussfolgerung essenziell: Ein verweilender Blick, der in der Betrachtung der Form verharrt, einer „Formästhetik" (Fix 2001: 38) nachspürt, findet sich bald in einer Fülle von inhaltlichen Assoziationen wieder, die der schnelle, flüchtige Blick nie zu evozieren vermocht hätte: „Sie [i. e. die wilde Semiose, J. P.] stellt neue, unmittelbare Bedeutung her, sie unterläuft, verzerrt, vervielfältigt, sprengt die vorgegebenen Raster der Sinnbildung" (Assmann 2015: 22). Ästhetisches Denken kann demnach als ein Denken beschrieben werden, „das von Wahrnehmungen lebt […] und zu Einsichten führt […]" (Welsch 1998: 56). Auf ihre erkenntnisbringende Wirkung verweist Welsch auch bei der Beschreibung revolutionärer wissenschaftlicher Entdeckungen (Welsch 1996: 92–93). Hierin deuten sich die Leistungen des Ästhetischen als Folge einer „‚maximalen Konzentration der Aufmerksamkeit auf einen gegebenen Gegenstand'" an (Mukařovský 1982: 32f., zit. in Fix 2001: 38). Daraus möchte ich eine Kernthese meines Beitrags ableiten, die sich sowohl auf Visualisierungen linguistischer Theorien, wie ich sie hier vorstelle, als auch auf Visualisierungen großer empirischer Datenmengen in der Linguistik bezieht, welche ein Gros der weiteren Beiträge dieses Bandes behandeln:

> Ich vermute, dass eine Ästhetisierung von Linguistikvermittlung, d. h. ein bewusstes Aufbrechen momentan typischer logozentrischer Muster linguistischer Wissenskommunikation durch (knapp betextete) Visualisierungen, rezipientenseitig zu einer verlängerten und bewussteren Wahrnehmung der musterbrechenden Elemente führt. Ein langer verweilender Blick kann dann im Rahmen einer wilden Semiose erkenntnisfördernde und motivationale Effekte erzielen.

Das Wissen um kognitiv folgenreiche „Regelverletzungen innerhalb eines Mediums, aber auch [um] Mischungen von Möglichkeiten verschiedener Medien und Grenzüberschreitungen zwischen den Medien" (Fix 2001: 39) ist dann auch für die bewusste Produktion ästhetischer Kommunikate von Belang. Ein Ästhetikbegriff, der nicht vorrangig an Geschmacksurteile gebunden ist, sondern eine (Nicht-)Passung im Hinblick auf Textsortenwissen als Bezugsgröße wählt, entbindet Gestalter davon, zwangsläufig etwas Kunstschönes zu schaffen. Gerade bei eher logozentrischen Textsorten der linguistischen Wissensvermittlung sind es eher die Visualisierungen überhaupt, nicht ihr künstlerischer Duktus, die Musterbrüche erzeugen und Erkenntnis begünstigen. Auch ungeübte Gestalter sollten sich deshalb nicht scheuen, die Potenziale des Visuellen praktisch zu erkunden. Es gilt Visualisierungen zu entwickeln, die in ihrer formalen Gestaltung „[erwartungs-nonkonform]" (Stöckl 2013a: 106) genug sind, um den Blick zu bannen und eine intensivere Beschäftigung mit der Inhaltsebene zu initiieren.

Nichtsdestotrotz vermag das Nutzen „künstlerisch[er] Verfahren und handwerklich[er] Vollkommenheit" (Assmann 2015: 24) die Wahrscheinlichkeit einer besonderen Wahrnehmung über die puren Effekte des grafikinduzierten Musterbruchs hinaus zu steigern. Bewusst wähle ich Techniken, die das Künstlerische instrumentalisieren, um ein Kommunikat mit „passenden Konnotationen aufzuladen", wie es Stöckl (2013a: 104) als Ästhetisierungsstrategie von Werbung formuliert. Assmann schreibt: „Durch Farben und Formen, Ornament und Illustration, aber auch durch den Abdruck von Persönlichkeit wird der Weg zum immateriellen Inhalt unterbrochen" (Assmann 1988: 241). Es gilt einzig zu bedenken, dass kommunikative Praktiken nur vor dem Hintergrund eines kontrasterzeugenden Musters musterbrechend wirken. Um dem abgestumpften Blick der „Anästhetik" (Welsch 1998: 10), d. h. dem Umschlagen der „wilden" in eine „milde Semiose" vorzubeugen, müssen gestalterische Entscheidungen wohlbedacht und wohldosiert sein.

Ähnliches gilt auch für die sukzessive Etablierung besagter Musterbrüche. In diesem Zusammenhang möchte ich jedoch genereller betonen: Der Vorschlag zum Erzeugen von Musterbrüchen in momentan noch stark logozentrischen Textsorten der Linguistikvermittlung ist in erster Linie ein zeitgenössischer. Es

ist mein Wunsch und Ziel, dass Visualisierungen künftig einen festen Platz im Musterrepertoire linguistischer Wissenskommunikation erhalten, also bildstarke Musterbrüche zum Muster werden. Auf eine „wilde Semiose" folgt damit eine drastische Erweiterung des Formen- und Funktionsrepertoires nun dezidiert multimodaler Textsorten der Linguistikvermittlung. Hierzu braucht es jedoch noch viele Musterbrüche.

2.2 Visuelle Linguistik | Möglichkeiten des Musterbruchs in der Linguistikvermittlung

Ein Ästhetikbegriff, der zunächst frei von Geschmacksurteilen ist und Wahrnehmung über *alle* Sinneskanäle in den Vordergrund stellt, verweist auf eine Darbietung von Informationen, die den Musterbruch in logozentrischen Textsorten der Geisteswissenschaften durch Integration von Visualisierungen sucht. Das Visuelle scheint aufgrund seiner dichten und räumlichen Zeichenteppiche besonders günstige Bedingungen für „wilde Semiosen" schaffen zu können (Assmann 2015: 24). Im Hinblick auf Vermittlungskontexte liegen die Vorteile jedoch gerade auch in einer medialen und funktionalen Kompatibilität des Visuellen mit der in flächigen Medien realisierten geschriebenen Sprache und den ermöglichten bedeutungspotenzierenden Effekten. Zum gezielten Herbeiführen solch einer Multimodalität in Theorietexten möchte ich zunächst einige theoretische Vorüberlegungen anstellen, welche nachfolgend am Beispiel der konzeptuellen Metapherntheorie praktisch erprobt werden sollen.

Zum speziellen Fall der bewussten Informationsübertragung von einem Zeichensystem in ein anderes gibt es bisher nur wenig Forschung (siehe bspw. Baker 2011: xvii). Wertvolle Anknüpfungspunkte liefert jedoch Jakobsons Konzept der „intersemiotischen Übersetzung" oder „Transmutation", die er als „eine Interpretation sprachlicher Zeichen mit Hilfe von Zeichen nichtsprachlicher Zeichensysteme" (Jakobson 1981: 190) definiert. Weitere Bezüge ergeben sich auch in den medientheoretischen Arbeiten Ludwig Jägers, der sich mit intermedialer Transkription beschäftigt und sie als Verfahren beschreibt, „das mindestens ein zweites mediales Kommunikationssystem zur Kommentierung, Erläuterung, Explikation und Übersetzung […] eines ersten Systems heranzieht." (Jäger 2002: 29) Es betont ein im kommunikativen Austausch nahezu notwendiges *Lesbarmachen* von Zeichen einer Modalität durch die weiterer Zeichensysteme mit dem Ziel der Bedeutungserschließung und -konstitution.

Ausgehend von diesen Konzeptionen möchte ich meine methodischen Überlegungen skizzieren (für einen ausführlicheren Leitfaden siehe Pflaeging 2015a: 385) und nachfolgend praktisch erproben. Für eine Visualisierung linguistischer Theorien ergeben sich aus meiner Sicht folgende zwei Bezugsbereiche:

(1) *Verbalsprachliche Bildlichkeit* | In diesem Bezugsbereich sind m. E. drei Aspekte wesentlich: Der wichtigste scheint mir, erstens, der häufige Gebrauch von bildlicher Sprache in wissenschaftlichen Theorietexten zu sein, wie bereits oben angedeutet wurde. Dies hängt, zweitens, zusammen mit konkreten Denotaten von Wörtern und Phrasen, die sowohl im metaphorischen als auch im nicht-metaphorischen Sprachgebrauch viel Potenzial zur ikonischen Darstellung mitbringen. Eine weitere relevante Dimension ergibt sich, drittens, aus der für verbale Sprache typischen linearen Rezeption der Zeichengefüge. Sprachbasierte Fachtexte führen Inhalte oft Schritt für Schritt ein und bauen so theoretische Modelle sukzessive auf. Dies ist nicht nur nützlich für das Verstehen vielschichtiger Theorien, sondern auch für das schrittweise Visualisieren theoretischer Gedanken, die dann nachfolgend zu einer komplexeren Grafik zusammengesetzt werden können.

(2) *Reflexion modalitätsspezifischer Stärken und Schwächen* | Ein Bewusstmachen der kommunikativen Stärken und Schwächen einzelner Zeichensysteme (siehe auch Stöckl 2016) – oder kontrastierend formuliert: ihrer Gemeinsamkeiten und Unterschiede – erlaubt das Ausgleichen kommunikativer Nachteile im Zusammenklang der Modalitäten. Hier wirken gerade die Differenzen beider Zeichensysteme bedeutungspotenzierend (vgl. „near-analogy"-Prinzip in Brock & Pflaeging 2018). Die räumliche Struktur und Zeichendichte einer visuell-bildlichen Darstellung erlaubt es, verschiedenste Teilaspekte einer Theorie auf kleinem Raum miteinander zu verschränken, profitiert aber von einem parallel dargebotenen Fließtext. Die Tendenz zur Mehrdeutigkeit visueller Zeichenkomplexe kann beispielsweise mithilfe verbaler Labels aufgefangen werden. Holly (2006) beschreibt diesen Effekt als Monosemierung. Ausgehend von grafikstilbezogenen Konnotationen (bspw. durch einen lockeren Strich und farbenfrohes Aquarellieren), lässt sich über das Entstehen expressiver Bildakte spekulieren. Ein bildliches Andeuten von Räumlichkeit durch perspektivisches Skalieren sowie ein Verblassen der Farben könnte zu bildlichen Implikaturen führen. Bei der Finalisierung der auf Sprachbasis erzeugten Grafik spielen all diese Überlegungen eine Rolle und erlauben im Hinblick auf Vermittlungskontexte zielführende Gestaltungsentscheidungen.

Um die Möglichkeit und Nützlichkeit solcher Visualisierungen auszuloten, möchte ich mich nun einigen konkreten Beispielen zuwenden. Dabei werde ich selbst visualisierend und betextend aktiv werden und – hoffentlich anregende – Musterbrüche erzeugen.

3. Praktisches | Zur konzeptuellen Metapherntheorie

3.1 Zum Inhalt

1,80 spontane und 4,08 konventionalisierte Metaphern pro Gesprächsminute (Pollio et al. 1977, zit. in Glucksberg 1989: 126) – konsultiert man quantitative Studien zur Häufigkeit metaphorischer Ausdrücke im Sprachgebrauch, so stößt man auf Zahlen, die Chandlers griffige Feststellung zu unterstreichen scheinen: „there is more metaphor on the street corner than in Shakespeare" (Chandler 2007: 124–125). Und diese Aussage hätte sich problemlos auch schon zu Shakespeares Lebzeiten formulieren lassen, denn obwohl sich die Auswahl zur Metaphorisierung genutzter Konzepte vor dem Hintergrund sich wandelnder soziohistorischer Kontexte neu ordnet, lässt sich mithilfe überlieferter Textquellen belegen, dass „metaphor has been in widespread use during the past 300 years" und früher (Gibbs 1994: 123).

Einer der einflussreichsten wissenschaftlichen Ansätze zur Erklärung dieses Phänomens ist die in den 1980er-Jahren im Zuge der kognitiven Wende von George Lakoff und Kollegen entworfene *konzeptuelle Metapherntheorie* (Lakoff & Johnson 2003; siehe auch Evans & Green 2006: 286–296). Sie verortet Metaphorisierungsprozesse nicht auf der Ebene des sprachlichen Ausdrucks, sondern fasst sie als ein grundsätzlich und naturgegeben *kognitiv* ablaufendes *„understanding and experiencing one kind of thing in terms of another"* (Lakoff *&* Johnson 2003: 5, Hervorhebung im Original).

Metaphorische Übertragungsprozesse weisen im Wesentlichen drei Eigenschaften auf:

(1) *Concretisation* | Der enorme kommunikative Nutzen von Metaphern resultiert aus der Möglichkeit, jenen Gedanken Ausdruck zu verleihen, die sich mit wortwörtlicher Sprache nur schwer beschreiben lassen (siehe „inexpressibility hypothesis", Gibbs 1994: 124–125). Hierzu werden konzeptkonstituierende Aspekte einer konkreten „source domain" (bspw. JOURNEY) genutzt, um eine abstrakte „target domain" (bspw. LIFE) greifbar zu machen (Lakoff & Johnson 2003: 108–109; Lakoff 2006: 232).

(2) *Systematicity* | Metaphorisierungsprozesse sind systematisch, d. h. es wird ein ganzes Bündel an Eigenschaften der „source domain" in fester Konstellation auf die „target domain" übertragen, sodass konzeptuelle Metaphern über verschiedene „metaphorical expressions" kommunikativ realisiert werden können (Lakoff *&* Johnson 2003: 7, 9; Lakoff 2006: 186) – sowohl verbal als auch non-verbal (siehe Forceville 2008).

(3) *Unidirectionality* bzw. *Asymmetry* | Innerhalb der konzeptuellen Metapherntheorie werden Übertragungsprozesse als einseitig gerichtet verstanden. Dies bedeutet, dass „metaphors map structure from a source domain to a target domain but not vice versa." (Evans *&* Green 2006: 296)

Diese sowohl spezifischen und allgemeinen theoretischen Annahmen lassen sich vor dem Hintergrund der hier geführten Argumentation auf einen zentralen Gedanken zuspitzen: Gerade aufgrund der kognitiven Natur metaphorischer Übertragungen und der daraus resultierenden metaphorischen Durchdrungenheit *allen* kommunikativen Handelns, wird die erkenntnisbringende Bildlichkeit der Metapher auch in fachwissenschaftlichen Diskursen ausgiebig genutzt (Gibbs 1994: 171; Drewer 2003). Wie die konzeptuelle Metapherntheorie selbst illustriert, basieren viele Modellierungen mentaler Strukturen, Kategorien und Vorgänge auf der Annahme MIND IS A CONTAINER (Lakoff & Johnson 2003: 148; Lakoff 2006: 196). Bereits in den späten 1970er-Jahren hat Reddy (1979) auf die CONDUIT METAPHOR hingewiesen, auf die nicht nur in der alltagssprachlichen, sondern auch in der linguistischen Kommunikation *über* Kommunikation zurückgegriffen wird. Neben den Grafiken N° 01 bis 04, die sich direkt aus der Metaphorik der konzeptuellen Metapherntheorie ableiten, gründen auch die Grafiken N° 05 und 06 auf der für linguistische Modelle typischen Metaphorik (siehe Kapitel 3.2). Dies zeigt m. E. deutlich: Eine Verbannung der Metapher aus dem wissenschaftlichen Diskurs ist weder möglich noch nützlich (vgl. Chandler 2007: 125). Die über Jahrhunderte kultivierte Ansicht, dass das Nutzen von Metaphern im akademischen Austausch einem „wandering amongst innumerable absurdities" gleicht und zu „contention and sedition, or contempt" (Hobbes 1651; zit. in Lakoff & Johnson 2003: 190) führt, sollte mittlerweile als überholt gelten (Gibbs 1994: 171).

Aus diesem Grund erstaunt es umso mehr, dass die Scheu vor Bildlichkeit als Mittel des geisteswissenschaftlichen Wissensaustausches auf Sprachebene zwar rückläufig ist, sie jedoch in Form einer *Bilderscheu* fortzubestehen scheint – und sich in anhaltend logozentrischen kommunikativen Mustern äußert. Wie diese bereits punktuell und künftig noch bewusster und deutlicher gebrochen werden können, möchte ich nun zeigen.

3.2 Zur Vermittlungspraxis | Musterbrüche durch Visualisierungen

Bevor ich einige eigene Visualisierungsversuche zur konzeptuellen Metapher vorstelle, will ich kurz auf zwei theoriebezogene Grafiken zu sprechen kommen, die bereits viele der oben skizzierten Aspekte berücksichtigen. Sie deuten auch an, dass Theoretisierungen auf ganz natürliche Weise mentale Bilder nutzen, die

hier zu Vermittlungszwecken externalisiert wurden. *Abb. 1* stammt aus einem umfangreichen Einführungswerk zur Kognitiven Linguistik von Evans & Green (2006: 313).

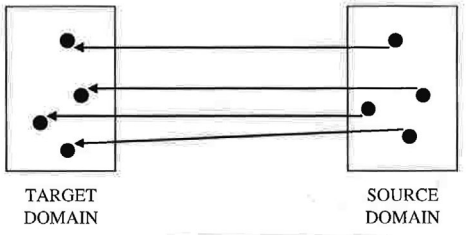


Abb. 1

Ähnlich wie ich es in Grafik N° 01 vorschlage, arbeiten sie mit zwei Flächen als kognitiven Domänen, einem (leider nicht in gleicher Konstellation gedruckten) Punkte-Cluster und mehreren Pfeilen, die den Übertragungsprozess modellieren.

Thematisch verwandte und visuell elaboriertere Grafiken finden sich in Abhandlungen zur „Mental Space Theory" von Gilles Fauconnier und Mark Turner. Ihre Theorie formuliert die Annahme, dass konzeptuelle Metaphern nur *eine Spielart* allgemeinerer Prinzipien konzeptueller Verknüpfung und Integration darstellen (Kövecses 2002: 227; Knowles & Moon 2006: 73). Außerdem betonen sie den dynamischen Charakter eines „blending" kognitiver Domänen, bei dem Eigenschaften von „source" als auch von „target domain" unter Erzeugung eines „generic space" letztendlich in einem bedeutungsangereicherten „blended space" verschmelzen (siehe Knowles & Moon 2006: 74; Fauconnier & Turner 2006). Dies ist in *Abb. 2* visualisiert: Auch hier erscheinen kognitive „spaces" in flächiger

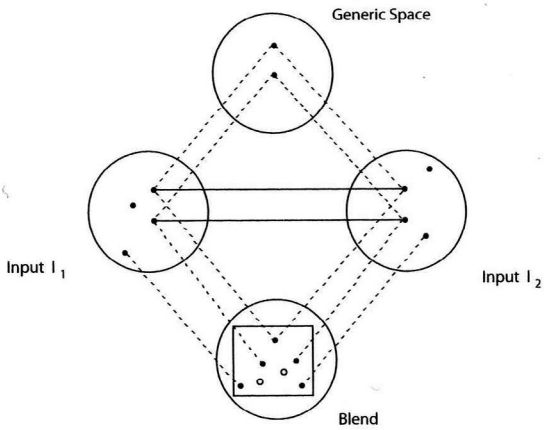
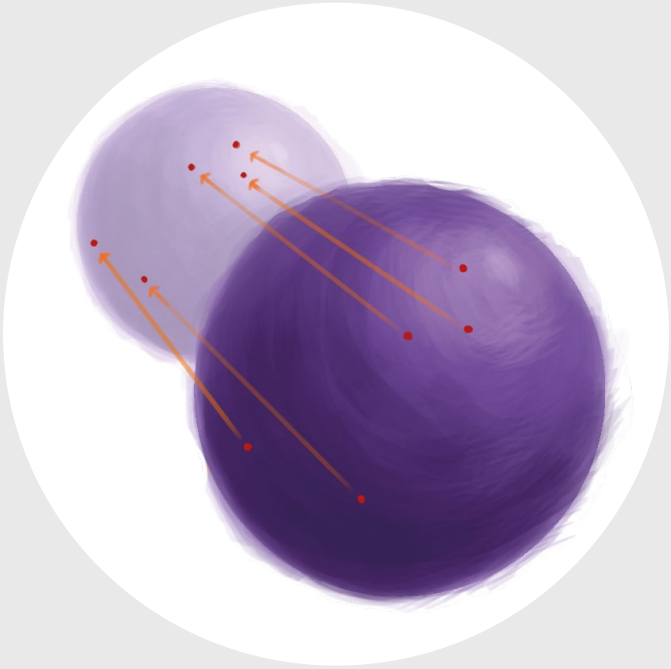


Abb. 2

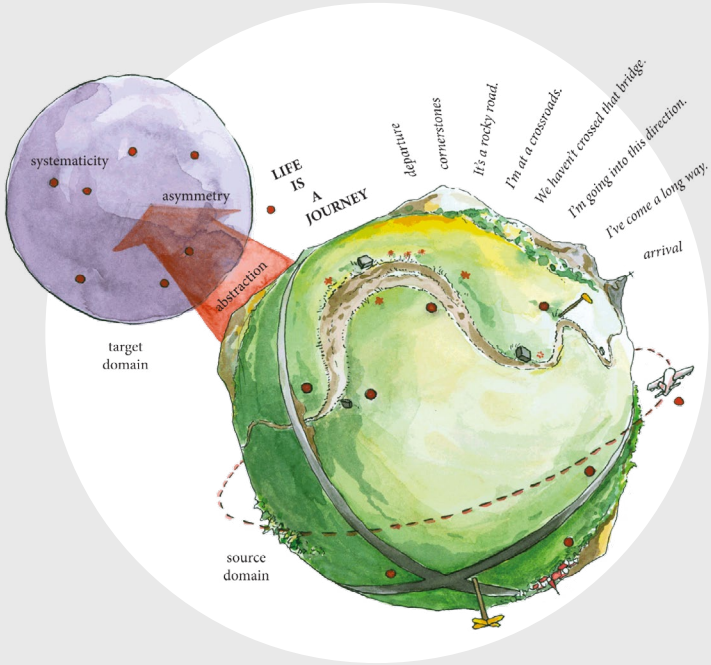
Darstellung; Konstellationen und Relationen von korrespondierenden und übertragenen Eigenschaften werden durch Punkte und verschiedene Linien markiert.

Die hier gezeigten Grafiken stellen einen deutlichen Bruch mit dem etablierten logozentrischen Muster aufgeschriebener linguistischer Metaphernvermittlung dar: Lakoff & Johnson (2003), Lakoff & Turner (1989), Lakoff (2006) sowie beispielsweise Gibbs (1994), Goatly (1997), Glucksberg (2001) und Knowles *&* Moon (2006) nutzen keine Visualisierungen zur Vermittlung der konzeptuellen Metapherntheorie. Obwohl sich hierzu auch bei Kövecses (2002) und Dancygier *&* Sweetser (2014) nahezu keine theoriebezogenen Grafiken finden lassen, wird hier zur Einführung der „Mental Space Theory" ausführlich mit den etablierten grafischen Mitteln (vgl. Fauconnier *&* Turner 2006) gearbeitet. Dies könnte auf theoriebezogene Publikationstraditionen hinweisen und legt aufgrund der formalen und inhaltlichen Parallelen der Darstellung nahe, dass bei der konzeptuellen Metapherntheorie nicht generell ein „Visualisierungsproblem" vorliegt. Welchen Beitrag Visualisierungen bei der Aneignung linguistischen Theoriewissens leisten können, zeigt sich m. E. deutlich: Sie verzahnen Teilaspekte zu komplexen Wirkgefügen, erzeugen damit einprägsame Abbilder theoretischer Konstrukte und offenbaren womöglich – von theoretischem Detail befreit – neue Bezüge und Zusammenhänge komplexerer Theoriegebäude.



N° 01 | Metaphorisierung | Generelle Eigenschaften | *Tablet und Grafiksoftware.*

Die Annahme, dass „metaphors are mappings across conceptual domains" (Lakoff 2006: 185, 196), steht im Zentrum der kognitiven Metapherntheorie und am Anfang meines Visualisierungsvorhabens. Selbst auf der konzeptuellen Metapher Mind Is A Container basierend, verweist der Ausdruck „(conceptual) domains" auf den konkret-bildlichen Eindruck einer nach außen abgegrenzten Räumlichkeit. Diese lassen sich, wie Evans & Green vorschlagen, als flächige Bereiche visualisieren, können m. E. aber noch treffender – gerade im Sinne der Container-Metapher – als dreidimensional anmutende Sphären gestaltet werden. Dem „mapping across" versuche ich durch mehrere, die beiden Sphärenkörper verbindende Linien nahezukommen. Die Charakteristika „tightly structured" und „correspond systematically" können durch eine hohen Liniendichte und duplizierte Markierungen auf der abstrakten Sphäre visualisiert werden. Zur Kodierung der Gerichtetheit der Übertragung „from a source domain [...] to a target domain" (Lakoff 2006: 190) ergänze ich die Linien um Pfeilspitzen. Ohne sich selbst auf ein konkretes Beispiel zu beziehen, nimmt diese Grafik zentrale Aspekte der Theorie auf.

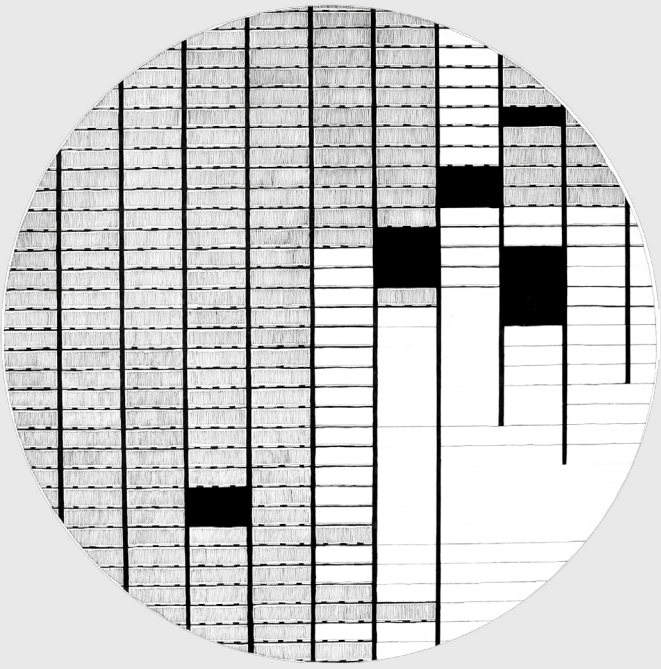


N° 02 | Metaphorisierung | Life Is A Journey | *Aquarell auf Papier.*

Weiterhin auf wesentliche theoriekonstituierende Mechanismen verweisend, illustriert diese Darstellung, dass das kaum fassbare, nahezu enigmatische Konzept Life durch lebensweltlich erfahrbare Dimensionen des Konzepts Journey erschließbar wird. Die „target domain" ist vage umrissen, erhält lediglich auf Basis der erwähnten Container-Metapher eine leicht räumliche, konturierte Form. Der Kontrast zwischen den bildempfangenden und -spendenden Konzepten entsteht, neben der räumlichen Staffelung, gerade durch das Abbilden konkreter Journey-Aspekte, die sich in gängige metaphorische Ausdrücke übersetzen lassen: Die Sphäre umhüllen gewundene, steinige Pfade und gerade, breite Straßen, die sich ihren Weg durch die Landschaft bahnen und in Etappen an Meilensteinen und Kreuzungen vorbei zu individuellen Zielen führen (cf. Lakoff 2006: 189). Die angedeutete Dreidimensionalität der Kugel könnte hierbei implizieren, dass neben den eingezeichneten Übertragungen weitere Beispiele benannt werden können, die, auf der Kugelrückseite liegend, im Verborgenen bleiben. Ein essenzieller Bestandteil sind die verbalen Labels, die der Polysemie vieler Bildelemente entgegenwirken.

N° 03 | Metaphorisierung | Life Is A Journey | *Tusche auf Leinwand.*

Lassen sich die wesentlichen Merkmale der konzeptuellen Metapherntheorie aus den ersten beiden teilweise verbal transkribierenden Grafiken noch recht klar ableiten, so scheint sich die hier abgebildete Visualisierung von Life Is A Journey einer unmittelbaren und eindeutigen Ausdeutung zunächst zu verschließen. Weder die systematischen Korrespondenzen zwischen Quell- und Zieldomäne, noch die Gerichtetheit des Übertragungsprozesses oder die Kontrastierung von Konkretem und Abstraktem finden in der Gestaltung eine explizite Entsprechung. Ein Großteil dieser Aspekte ist jedoch m. E. implizit vorhanden und über Inferenzprozesse erschließbar. Bewusst bleibt die typischerweise über den konkretisierenden Umweg greifbare Zieldomäne in der Gestaltung ausgespart; eine Leerstelle, die ohne Metaphorisierung keine Kontur, keine Körperlichkeit erhält und die metaphorische Übertragung als kommunikative Notwendigkeit offenlegt. Sowohl form- als auch inhaltsästhetisch schafft solch ein Zugriff auf Metaphorik womöglich beste Voraussetzungen für das Verweilen im Bild und das Auslösen einer „wilden Semiose".



N° 04 | Metaphorisierung | Mind Is A Container | *Tusche auf Leinwand.*

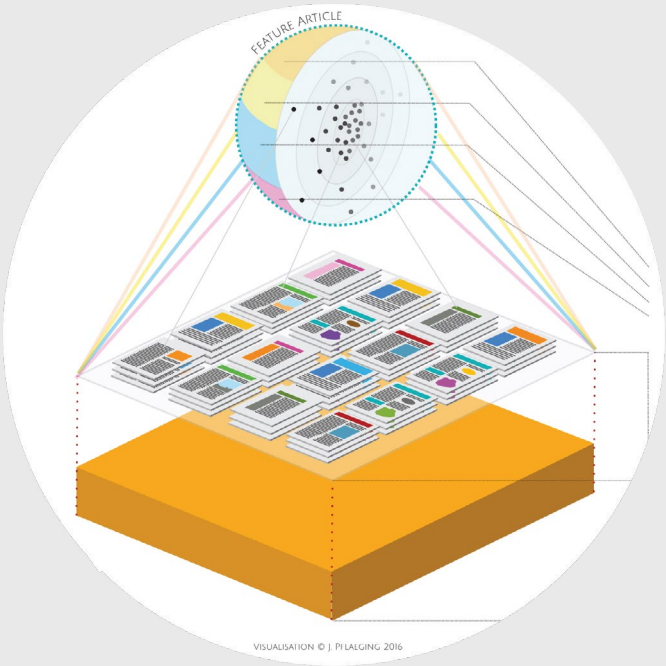
Mit diesen beiden Fassungen der konzeptuellen Metaphern Life Is A Journey und Mind Is A Container möchte ich eine alternative grafische Form anbieten, die in vielerlei Hinsicht, so hoffe ich, eine noch weitreichendere und epistemisch intensivere Annäherung an die Essenz von Metaphorik zu erreichen vermag. Aufgegliederte Ideen von einem Weg; dichte, sich auf- und abbauende Strukturen in mentalen Räumlichkeiten. Ein bildliches Angebot, das trotz allen Bändigens seiner Visualität in eine konkret-ikonische Form formal vage genug bleibt, um individuellen Assoziationen und Interpretationen den nötigen Raum zu lassen – der „wilden Semiose" einen Tummelplatz zu bieten. Selbstverständlich: Ein holistisches Erfassen zentraler Aspekte der konzeptuellen Metapherntheorie lediglich auf Basis dieser Darstellungsform scheint – auch mit metaphernbenennendem Kurztitel – wenig realistisch. Gezielt sprachlich betextet könnten mit solch einer Grafik m. E. jedoch Dimensionen von Metaphorik erkundet werden, die das systematisch-ordnende Raster einer sprachwissenschaftlichen Beschreibung mitunter nicht einzufangen vermag.

4. Geschlussfolgertes & Weiterführendes

Composing an academic Article … Is A Journey. In diesem Beitrag habe ich sowohl theoretisch als auch praktisch dafür argumentiert, dass die Logozentriertheit des zeitgenössischen Wissensaustausches in der Sprachwissenschaft vor allem in der seit jeher stark sprachbasierten Publikationstradition, nicht aber zwingend im Gegenstand linguistischer Forschung begründet liegt. Der linguistische Diskurs lebt gerade auch von mentalen Bildern und metaphorischen Ausdrücken, welche es uns erlauben, die für von Sprache und ihre Erforschung angenommenen abstrakten Mechanismen und vorgeschlagenen theoretischen Zugriffe fassbar zu machen. Am Beispiel der konzeptuellen Metapherntheorie sowie weiterer theoretischer Modellierungen der Linguistik (siehe nachfolgend N° 05 und N° 06) habe ich gezeigt, dass gerade diese verbalsprachliche Bildlichkeit zum Ausgangspunkt für die Erstellung komplexer Visualisierungen linguistischer Theorien werden kann.

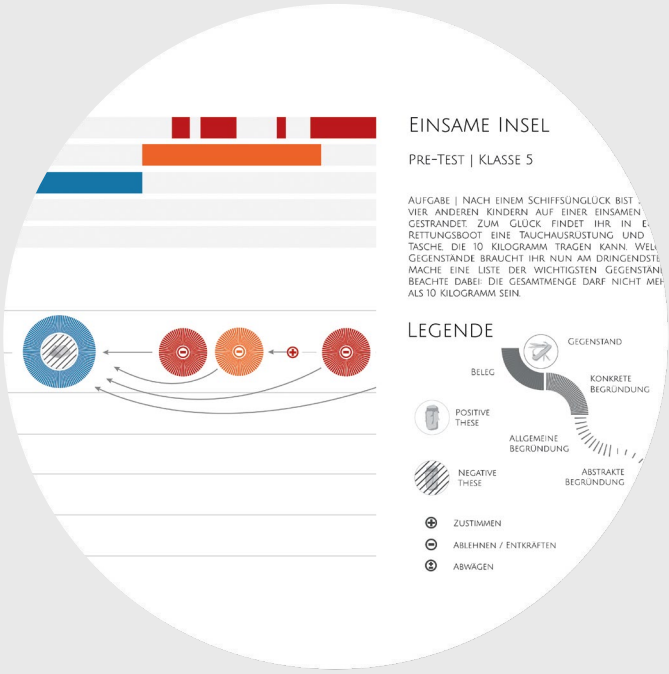
Unterbricht man den logozentrischen Textfluss gängiger Textsorten der Linguistikvermittlung, sei es in Fachaufsätzen (Hauser et al. 2016) oder Einführungslehrwerken (Pflaeging 2013; Pflaeging *&* Brock 2018), so wird es wahrscheinlich, dass sich in der Textrezeption die Aufmerksamkeit gerade bei den musterbrechenden Textteilen bündelt und in einen *langen verweilenden Blick* übergeht. Aufgrund der kommunikativen Eigenschaften mit Texten angereicherter Visualisierungen können bspw. einzelne Aspekte eines komplexen theoretischen Modells erkenntnisschaffend miteinander verschränkt werden. Hierbei entstehen im Rahmen einer „wilden Semiose" möglicherweise auch inhaltliche Assoziationen, die ein verbalsprachlicher Theorietext kaum erzeugen kann. Diese könnten sogar in eine „metakommunikativ[e] Reflexion" über „Form, Inhalt und Wirkungsweise" (Stöckl 2013b: 3) gängiger Praktiken der Linguistikvermittlung münden, die nicht nur die Reflexion fachspezifischer Metaphorik ermöglicht, sondern auch das erkenntnisbringende Potenzial linguistischer Visualisierungen aufzeigt und zu absichtsvollen Musterbrüchen ermutigt.

Ich verfolge mit diesem Beitrag also nicht nur das Ziel, den Weg zu den hier vorgestellten Grafiken transparent zu machen und die Visualisierungsmethodik auch für andere Gebiete der linguistischen und geisteswissenschaftlichen Forschung vorzuschlagen. Ich hoffe nicht zuletzt auch, meine Fachkolleginnen und -kollegen zum Erkunden des Visualisierungspotenzials eigener (und fremder) theoretischer Texte anzuregen – sei es für die linguistische Wissensvermittlung in den öffentlichen Räumen einer *fachinternen* Experten-Experten- und Experten-Laien-Kommunikation oder beim privat-linguistischen Abwandern hermeneutischer Zirkel. Denn nicht zuletzt vermag man dann auch als *Bildproduzent* zu entdecken, was Visualisierung und geisteswissenschaftliche Forschung eigentlich eint:



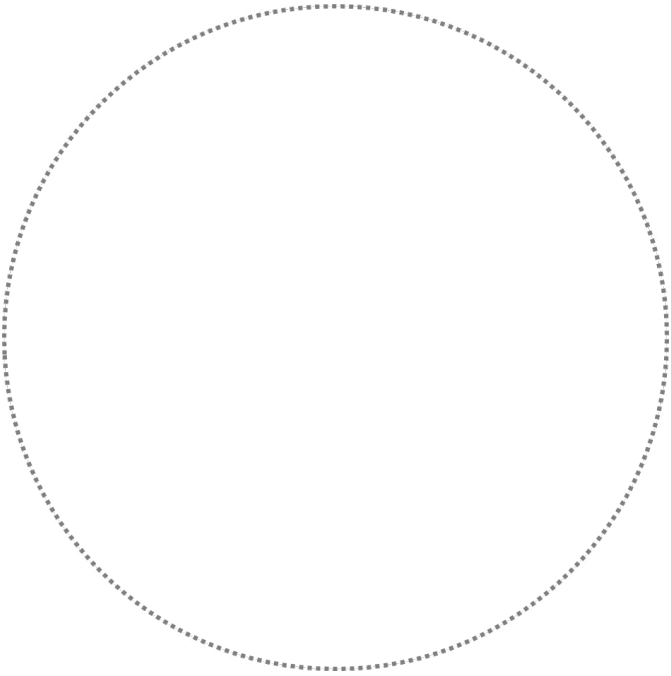
N° 05 | Text- & Medienlinguistik | Mehrebenenmodell (multimodaler) Textsorten.

Diese Grafik versucht innerhalb der Text- und Medienlinguistik gängige Analysekriterien von Textsorten oder *Genres* visuell zu fassen. Durch zielführendes und dadurch wiederkehrendes Produzieren und Rezipieren kommunikativer Handlungen verdichten sich diese zu deutlich umrissenen Mustern, welche Mitgliedern einer Gemeinschaft als prototypisch organisierte, kognitive „devices for sense-making" (Lomborg 2014: 3; siehe auch Swales 1990; Sandig 2000) zur Verfügung stehen. Der Vielschichtigkeit kommunikativer Realität versuchen Text- und Medienlinguisten mittels einer Mehrebenenanalyse Rechnung zu tragen (Heinemann & Viehweger 1991; Brinker et al. 2014), die oft eine Beschreibung der Kommunikations-situation, des Themas, der multimodalen Struktur sowie typischer Textfunktionen umfasst (vgl. bspw. Brinker et al. 2014; Pflaeging 2015b; Stöckl 2016). Konkrete Textexemplare, die eine analytische Rekonstruktion von Textsortenwissen erlauben, existieren jedoch nur auf Basis textsortentypischer medial-technischer Konstellationen (bzw. Kommunikationsformen, Brock & Schildhauer 2017). All diese Aspekte sind tief in kulturellen Kontexten verankert.



N° 06 | Konversationsanalyse | Modellierung von Argumentationskompetenz.

Ebenso wie visuelle Textsortenmodelle eher untypisch für textlinguistische Publikationen sind (siehe aber Pflaeging 2015b; Brock 2016; Schildhauer 2016; Stöckl 2016), finden sich auch in konversationsanalytischen Veröffentlichungen kaum theoriebezogene Grafiken, sodass m. E. auch hier viel epistemisches Potenzial ungenutzt bleibt: Die hier im Ausschnitt abgebildete Grafik deutet an, wie argumentatives Gesprächsverhalten visuell modelliert werden könnte – am Beispiel einer Robinsongeschichte. Während das Erwähnen einzelner Gegenstände (z. B. *Zelt* als Utensil auf einer einsamen Insel) durch Icons markiert wird, ist die Tiefe des vorgebrachten Arguments durch unterschiedlich dicht gestrichelte und breite Linien gefasst. Einzelne „turns" und assoziierte Argumente sind farblich kodiert und räumlich in eine zeitliche Sequenz gebracht. Außerdem werden gegenseitige Bezugnahmen durch Pfeile visualisiert. Die Grafik integriert so nicht nur erstmals (siehe aber Vogt 2007; Grundler 2011) *Sequenzialität*, *Interaktivität* und *Temporalität* von Argumentationen in eine Darstellung, sondern legt auch ihre musterhaften Verläufe offen (Hauser et al. 2016).



Visualisierung … **Geisteswissenschaftliche Forschung …**

… beschreibt eine ganz eigene Art der Auseinandersetzung mit der Welt. Entdeckungen und Erfindungen weltlicher Realität werden durch Textproduzenten mittels Zeichen zu komplexen kommunikativen Angeboten verdichtet und an Textrezipienten vermittelbar. Sie dient der Annäherung an eine vermutete, erhoffte Wahrheit über die Welt, die es zu entschlüsseln gilt.

Ihre bewusste Zusammenführung im Dienste einer visuell(er)en Linguistik kann meines Erachtens nur wünschenswert sein.

5. Dankendes

Mein herzlicher Dank gilt Alexander Brock, Volker Eisenlauer, Michaela Hausmann, Peter Schildhauer und Hartmut Stöckl für unzählige kundige Hinweise und inspirierende Gespräche, durch die die hier vorgestellten Überlegungen weiter an Kontur gewonnen haben.

6. Bibliografie

Allesch, Christian. 2006. *Einführung in die psychologische Ästhetik*. Wien: WUV.

Assmann, Aleida. 1988. „Die Sprache der Dinge: Der lange Blick und die wilde Semiose." In *Materialität der Kommunikation,* herausgegeben von Hans-Ulrich Gumbrecht und K. Ludwig Pfeiffer. Frankfurt am Main: Suhrkamp, 237–251.

Assmann, Aleida. 2015. *Im Dickicht der Zeichen*. Berlin: Suhrkamp.

Baker, Mona. 2011. *Routledge Encyclopedia of Translation Studies*. London: Routledge.

Ballstaedt, Steffen-Peter. 2012. *Visualisieren: Bilder in wissenschaftlichen Texten*. Konstanz: UVK.

Brinker, Klaus, Hermann Cölfen und Steffen Pappert. 2014. *Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden*. Berlin: Schmidt.

Brock, Alexander. 2016. "The Borders of Humorous Intent: The Case of TV Comedies." *Journal of Pragmatics* 95: 58–66.

Brock, Alexander und Jana Pflaeging. 2018. "The Virtues of Near-Analogy in Language and Communication." In *Analogy, Copy, and Representation: Interdisciplinary Perspectives,* herausgegeben von Christoph Haase und Anne Schröder. Bielefeld: Aisthesis.

Brock, Alexander und Peter Schildhauer, Hrsg. 2017. *Communication Forms and Communicative Practices: New Perspectives on Communication Forms, Affordances and What Users Make of Them*. Frankfurt am Main: Lang.

Chandler, Daniel. 2007. *Semiotics: The Basics*. London: Routledge.

Dancygier, Barbara und Eve Sweetser. 2014. *Figurative Language*. Cambridge: Cambridge University Press.

Drewer, Petra. 2003. *Die kognitive Metapher als Werkzeug des Denkens: Zur Rolle der Analogie bei der Gewinnung und Vermittlung wissenschaftlicher Erkenntnisse*. Tübingen: Narr.

Evans, Vyvyan und Melanie Green. 2006. *Cognitive Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Fauconnier, Gilles und Mark Turner2006. "Metal Spaces: Conceptual Integration Networks." In *Cognitive Linguistics: Basic Readings,* herausgegeben von Dirk Geeraerts. Berlin: Mouton de Gruyter, 303–371.

Fix, Ulla. 2001. „Die Ästhetisierung des Alltags: – am Beispiel seiner Texte." *Zeitschrift für Germanistik. Neue Folge* 1: 36–53.

Forceville, Charles. 2008. "Metaphor in Pictures and Multimodal Representations." In *The Cambridge Handbook of Metaphor and Thought,* herausgegeben von Raymond W. Gibbs. Cambridge: Cambridge University Press, 462–482.

Gibbs, Raymond W. 1994. *The Poetics of Mind: Figurative Thought, Language, and Understanding*. Cambridge: Cambridge University Press.

Glucksberg, Sam. 1989. "Metaphors in Conversation: How Are They Understood? Why Are They Used?" *Metaphor and Symbolic Activity* 4, Nr. 3: 125–143.

Glucksberg, Sam. 2001. *Understanding Figurative Language*. Oxford: Oxford University Press.

Goatly, Andrew. 1997. *The language of Metaphors*. London: Routledge.

Grundler, Elke. 2011. *Kompetent argumentieren: Ein gesprächsanalytisch fundiertes Modell*. Tübingen: Stauffenburg.

Hauser, Stefan, Martin Luginbühl und Jana Pflaeging. 2016. "How to Picture an Argument: New Visual Approaches to the Description of Argumentative Competence in Conversation Analysis." Vortrag auf der Tagung *Knowledge Design: Graphic Design and Science Communication*, Universität Tübingen, 07.04.2016.

Heinemann, Wolfgang und Dieter Viehweger. 1991. *Textlinguistik: Eine Einführung*. Tübingen: Niemeyer.

Holly, Werner. 2006. „Mit Worten sehen: Audiovisuelle Bedeutungskonstruktion und Muster transkriptiver Logik in der Fernsehberichterstattung." *Deutsche Sprache* 34: 135–150.

Huber, Ludwig, Arne Pilniok, Rolf Sethe, Birgit Szczyrba und Michael Vogel, Hrsg. 2014. *Forschendes Lehren im eigenen Fach: Scholarship of Teaching and Learning in Beispielen*. Bielefeld: Bertelsmann.

Jäger, Ludwig. 2002. „Transkriptivität: Zur medialen Logik der kulturellen Semantik." In *Transkribieren: Medien/Lektüre*, herausgegeben von Ludwig Jäger und Georg Stanitzek, 19–42.

Jakobson, Roman. 1981. „Linguistische Aspekte der Übersetzung." In *Übersetzungswissenschaft*, herausgegeben von Wolfram Wills. Darmstadt: Wissenschaftliche Buchgesellschaft, 189–198.

Knowles, Murray und Rosamund Moon. 2006. *Introducing Metaphor*. London: Routledge.

Kövecses, Zoltán. 2002. *Metaphor: A Practical Introduction*. New York: Oxford University Press.

Kuiper, Koenraad, Hrsg. 2011. *Teaching Linguistics: Reflections on Practice*. London: Equinox.

Lakoff, George. 2006. "The Contemporary Theory of Metaphor." In *Cognitive Linguistics: Basic Readings*, herausgegeben von Dirk Geeraerts.. Berlin: Mouton de Gruyter, 185–239.

Lakoff, George und Mark Johnson. 2003. *Metaphors We Live by*. Chicago: The University Press of Chicago.

Lakoff, George und Mark Turner. 1989. *More than Cool Reason: A Field Guide to Poetic Metaphor*. Chicago: University of Chicago Press.

Lomborg, Stine. 2014. *Social Media, Social Genres: Making Sense of the Ordinary*. New York: Routledge.

Pflaeging, Jana. 2013. "Promoting the Visualisation of Linguistic Theories." In *Facets of Linguistics: Proceedings of the 14th Norddeutsches Linguistisches Kolloquium 2013 in Halle an der Saale*, herausgegeben von Anne Ammermann, Alexander Brock, Jana Pflaeging und Peter Schildhaue. Frankfurt am Main: Lang, 173–187.

Pflaeging, Jana. 2015a. "How to Visualize Linguistic Theories: Multimodale Linguistikvermittlung in universitären Lehrwerken." *Mitteilungen des Deutschen Germanistenverbandes* 62, Nr. 4: 379–394.

Pflaeging, Jana. 2015b. "'Things that Matter, Pass them on': ListSite as Viral Online Genre." *10plus1 | Living Linguistics* 1, Nr. 1: 156–182.

Pflaeging, Jana und Alexander Brock. 2017. "A Sentence is a Hostel Room: New Approaches to Textbooks for Beginner Students of English Linguistics." In *Exploring the Periphery: Perspectives from Applied Linguistics, Language Teaching, Literary and Cultural Studies*, herausgegeben von Stefanie Quakernack, Till Meister, Diana Fulger und Nathan Devos. Bielefeld: Aisthesis.

Pflaeging, Jana und Peter Schildhauer. 2015. "Generating and Exchanging Knowledge: Rethinking Current Practices in Linguistics." *10plus1 | Living Linguistics* 1, Nr. 1: 1–8.

Reddy, Michael J. 1979. "The Conduit Metaphor: A Case of Frame Conflict in Our Language about Language." In *Metaphor and Thought*, herausgegeben von Andrew Ortony Cambridge: Cambridge University Press, 284–324.

Reicher, Maria E.2005. *Einführung in die philosophische Ästhetik*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Sandig, Barbara. 2000. „Text als prototypisches Konzept." In *Prototypentheorie in der Linguistik: Anwendungsbeispiele – Methodenreflexion – Perspektiven*, herausgegeben von Martina Mangasser-Wahl. Tübingen: Stauffenburg, 93-112.

Schildhauer, Peter. 2016. *The Personal Weblog: A Linguistic History*. Berlin: Lang.

Schmitz, Ulrich, Hrsg. 2004. *Linguistik lernen im Internet: Das Lehr-/Lernportal PortaLingua*. Tübingen: Narr.

Stöckl, Hartmut. 2013a. „Ästhetik und Ästhetisierung von Werbung." In *Werbung – Keine Kunst!?: Phänomene und Prozesse der Ästhetisierung von*

Werbekommunikation, herausgegeben von Hartmut Stöckl. Heidelberg: Winter, 89–116.

Stöckl, Hartmut. 2013b. „Werbekommunikation und Ästhetisierung: Zur Einführung." In *Werbung – Keine Kunst!?: Phänomene und Prozesse der Ästhetisierung von Werbekommunikation*, herausgegeben von Hartmut Stöckl. Heidelberg: Winter, 1–9.

Stöckl, Hartmut. 2016. „Multimodalität: Semiotische und Textlinguistische Grundlagen." In *Sprache im multimodalen Kontext*, herausgegeben von Nina-Maria Klug und Hartmut Stöckl. Berlin: Mouton de Gruyter, 3–35.

Swales, John M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Vogt, Rüdiger. 2007. „Mündliche Argumentationskompetenz beurteilen: Dimensionen, Probleme, Perspektiven." *Didaktik Deutsch* 12, Nr. 23: 33–53.

Welsch, Wolfgang. 1996. *Grenzgänge der Ästhetik*. Stuttgart: Reclam.

Welsch, Wolfgang. 1998. Ästhetisches Denken. Stuttgart: Reclam.

II. Praxis

Armin Hoenen

Recurrence Analysis Function, a Dynamic Heatmap for the Visualization of Verse Text and Beyond

Abstract The Recurrence Analysis Function (ReAF) is a cross-linguistic visualization tool for (historical) verse text, especially handwritten epics. It can also provide a general visualization of various aspects of prose text. It aims to enable intuitive understanding through explorative data analysis of historical, especially bardic-oral texts.[1] The assumption behind this is that bardic/born-oral and non-bardic/born-written texts differ drastically in the way they employ repetition. The ReAF in its first implementation, as presented here, is a language-independent tool that permits the visual exploration of such structures. Firstly, general aspects and formal characteristics of oral verse text are characterized, before the main technical details and some additional applications of the ReAF are explained and illustrated.

1. Preliminaries

Plato warned through one of his characters that 'Those who acquire it [ref. to writing], will cease to exercise their memory and become forgetful' (Coe 2012). Actually, born-oral and born-written texts do have a number of different characteristics. If Plato's assumptions were true, then those differences might be primarily connected to memory, which Ong (2012) implicitly states: "In a primary oral culture, to solve effectively the problem of retaining and retrieving carefully

1 Born-oral refers to the circumstances of composition of a text. More precisely, it refers to all texts, which originated in the pre-literate ages. Born-oral encompasses such texts, which existed before their first written manifestation, that is, texts where composing them in their entirety happened exclusively in the oral medium without any use of writing. The opposite, born-written, encompasses all texts where writing was involved in the text construction processes.

Published in: Noah Bubenhofer, Marc Kupietz: Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

articulated thought, you have to do your thinking in mnemonic patterns, shaped for ready oral recurrence."[2] One striking feature of born-oral texts connected to memorization is repetition. Repetition was used to a larger extent in preliterate text (Lord 1960). This is a good reason for assuming an indication of oral origin of a text. A formal characteristic of the structure of epic texts is that they are most often composed in lines and half-lines, as verse texts. Verse text is found across cultures although rules of rhyme and rhythm are language specific. Here, patterns of verse-text such as rhyme, meter, the (re)use of particular collocates and syntactical structures or more precisely, a high repetitiveness on virtually all linguistic levels, entails a constantly high rate of priming[3]. This ensures the quick and easy retrieval of these items from memory whenever needed. That these structures involve repetition/recurrence in its various forms is an inherent requirement, as Duggan (1973) points out: "The oral poet […] has a good motive for repeating himself with stylized phrases, namely the need to sing verses before a demanding audience at the rate of ten to twenty decasyllabic lines per minute, a pace far too rapid to permit the constant generation of unique word combinations."[4] With the invention of script, the need for memorization became gradually more obsolete, because knowledge could be externalized, that is stored in books and from there be accessed any time. Verse text continued to be used, but became increasingly less repetitive. Comparing modern lyric texts to epics recorded at the beginning of the chirographic age in verse, Lord (1960) remarks about born-written verse texts that there "may be repeated phrases, but the proportion of them to the whole is small". Characteristics of oral literature are enumerated further by Lord (1960) and together with a broader analysis form the so-called Oral Formulaic Theory (OFT). This is the dominant theory for oral texts. The qualitative structural differences between texts which relate to oral literature, according to the OFT, and those which do not can be visualized and are the subject of the ReAF. Finnegan (1992) gives a detailed account of findings about oral literature or oral poetry in all its scope. She stresses the diversity of

2 The difference between recurrence and repetition is contextual; one and the same linguistic unit can be repeated verbatim but take on a different pragmatic function as the local context differs.

3 Priming is a term used in psychology to refer to pre- or coactivation which enables quicker retrieval from memory. A person who has just heard the word cat can produce the word *dog* quicker from memory than without this prior acoustic exposure since both animals belong to a semantically similar category which is activated along with *cat*. *cat* primes *dog*.

4 Reversing this statement sheds light onto the generally more thought-over character of written (especially printed) texts, where authors pause and think, reformulate and rearrange texts to form a coherent outcome. Once spoken, a word cannot be cancelled or rearranged.

this genre and remains skeptical about whether there is one definition that can be put forward explaining all diversity of oral transmission. Goody (1987) shows how even texts we perceive as born-oral might have nevertheless been deeply structurally influenced by writing and the devices, such as lists, it fosters. However, Lord (1960) gives a guideline for how to test whether a text relates to oral rather than written composition.

The ReAF intends to follow Lord (1960), and to provide an aid in classifying texts as born-oral or born-written, but does not remain neutral about Goody's findings or Finnegan's objections. Finnegan (1992) mentions that epics are typically verse: "'Pure' epics like the Iliad and Odyssey are totally in verse"; in this sense the ReAF is primarily designed for the analysis of verse texts as such, whether influenced by the emergence of writing or not.

2. Recurrence in Oral Text

In a world without script, ensuring or controlling for exact repetition of longer texts exceeds the capacity of memory and is therefore impossible. How would one verify, if not from memory, that each and every word in two performances of the 14.000 verse lines of the *Odyssey* had been the same? And why should rigid identity of a text be of importance at all? Obviously, people from oral traditions did not care much about exact text identity of longer texts such as epics. Bards produced a slightly different text each time they performed the same story. However, many factors influenced the kinds of variation that occurred. Instead of a fixed text as in later (after the onset of writing) performance-based genres (e.g. theater) the oral poet could have presumably memorized a series of events ("Hero goes to war" – "Hero gets lost" – "Hero returns"), which when performed were marked with recurrences of various repertories. Such repertories presumably encompassed, for instance, common sub-plots. When a bard had enough time, he would make extensive use of adjectives and sub-plots when retelling events, which increased repetitiveness. If he had less time, he used less *ornamentation* as termed by Lord (1960). Apart from other non-recurrent characteristics of oral poetry, such as inconsistencies, e.g. where a minor character which had already died reappears, Culley (1967), bards would use repetitive structures, so-called formulas, being non-rigid word groups expressing similar content. This is a concept described by Parry & Parry (1987) in close detail, an example being Homeric heroic epithets identified as Regular Expression: (δόλον|νέφος) ἤγαγε (δῖος Ὀδυσσεὺς|Φοῖβος Ἀπόλλων).

The author of a written verse text does have the time to think about the most appropriate word to place in a certain position, instead of being restricted to using the one which comes most quickly to mind. Thus, since bards used

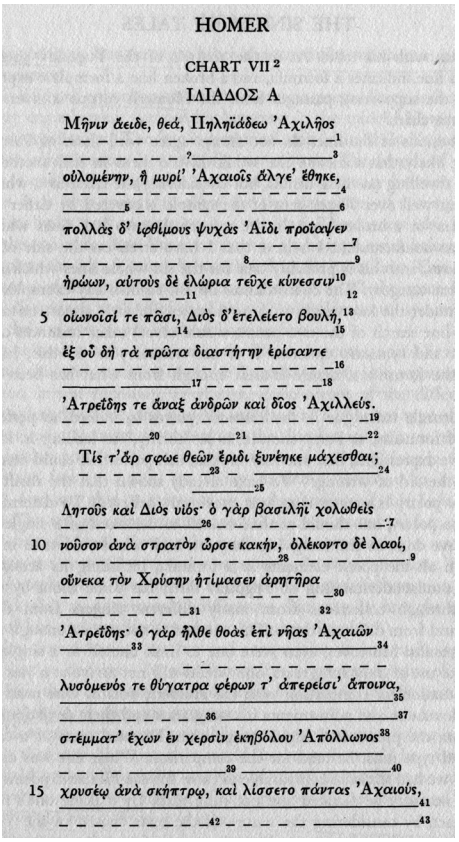
limited repertories and were furthermore forced to use those words which came most quickly to mind, written texts can be considerably less repetitive than oral verse texts.

3. The Visualization History of Verse Text

It has to be mentioned at the start that script itself is a visualization of sound. This is naturally the first visualization applied to verse texts, followed closely by the invention of the first visual prosodic boundary markers, such as dots. For verse texts, line breaks soon often supplemented or substituted the former prosodic boundary markers, which were kept in prosaic texts. The visual representation of poems using line breaks has crossed cultures and writing systems to become the dominant and most wide-spread form for representing poetry. This was stated by Culley (1967): "the unit of composition is usually the single line […] a poem is made by line being added to line. […] The line generally corresponds to a syntactic structural unit in that the end of the line coincides with a natural break such as the end of a sentence or clause."[5] Thus, apart from its other characteristics, from an aesthetic/visualization viewpoint, verse text constitutes a distinguished class of texts. This class can be further subdivided into bardic and non-bardic verse text. Although text itself is already a visualization, only abstract visual transformations allow for a holistic overview of properties of the whole text (at least if it exceeds a certain size). Parry & Parry (1987) and Lord (1960) applied some form of visualization, namely underlining the repeated and near-repeated passages of the beginning of the *Iliad* and other epics, see Figure 1.

Lord related this *extended-scriptural* visualization of verse closely to a test for the classification of oral poetry. His visualization was accepted or slightly modified by many scholars working on orality in subsequent years, for instance Culley (1967), Kailasapathy (1968), Benson (1966), Magoun (1980). A holistic extension of Lord's visualization to an entire epic may have been technically possible, but was never produced. It would have presented another problem with long epic texts: *overview*. In fact, Whallon (1969), Duggan (1973) and later Finnegan (1992) criticize the confinement of this visualization to text excerpts. Finnegan (1992): "Otherwise no overall analysis has been completed, nor any systematic sampling undertaken." Whallon (1969): "They [the visualizations] are not reliable because there is always doubt whether the specimen underlined is typical […]".

5 The invention of the line break as a visual separator interestingly seems to be a feature, which is not necessarily loaned into a new language when a writing system is taken over, but which could have been (re)invented several times. The cross linguistic spread of line breaks underscores their visual effectiveness.



HOMER

CHART VII 2

ΙΛΙΑΔΟΣ Α

Μῆνιν ἄειδε, θεά, Πηληϊάδεω Ἀχιλῆος
οὐλομένην, ἡ μυρί' Ἀχαιοῖς ἄλγε' ἔθηκε,
πολλὰς δ' ἰφθίμους ψυχὰς Ἄϊδι προΐαψεν
ἡρώων, αὐτοὺς δὲ ἑλώρια τεῦχε κύνεσσιν
οἰωνοῖσί τε πᾶσι, Διὸς δ' ἐτελείετο βουλή,
ἐξ οὗ δὴ τὰ πρῶτα διαστήτην ἐρίσαντε
Τίς τ' ἄρ σφωε θεῶν ἔριδι ξυνέηκε μάχεσθαι;
Λητοῦς καὶ Διὸς υἱός· ὁ γὰρ βασιλῆϊ χολωθείς
νοῦσον ἀνὰ στρατὸν ὦρσε κακήν, ὀλέκοντο δὲ λαοί,
οὕνεκα τὸν Χρύσην ἠτίμασεν ἀρητῆρα
Ἀτρεΐδης· ὁ γὰρ ἦλθε θοὰς ἐπὶ νῆας Ἀχαιῶν
λυσόμενός τε θύγατρα φέρων τ' ἀπερείσι' ἄποινα,
στέμματ' ἔχων ἐν χερσὶν ἑκηβόλου Ἀπόλλωνος
χρυσέῳ ἀνὰ σκήπτρῳ, καὶ λίσσετο πάντας Ἀχαιούς,

Figure 1: A. B. Lord, *The Singer of Tales,* 1960, p.143, Scan, (Copyright Harvard University Press).

He enumerates instead all repetitions in long sequences consisting of elements such as "1:13–16 = 1:372–375". A complete list of a work's repetitions in this way is again not easily exploitable in terms of a visual overview. The completeness of it is however superior to that of text excerpts.

To summarize, while a manual visualization line for line, like the one Lord proposed is feasible for smaller sections of text, it would be cumbersome for a complete text. Such a visualization of roughly 14.000 lines of the *Odyssey* would be completely unsatisfactory in terms of an overview.[6]

6 Foley (2002) mentions another individual visual rendering of oral poetry, yet not focussing on repetitions, which is not suitable to generate an overview over many lines.

In the digital age, especially since programmable visualization techniques have become widespread, a holistic analysis/overview is no longer difficult to achieve. ReAF aims at providing consecutive proof of the concept of visualizing repetition, as attempted by Lord (1960). The problem that ReAF will have to overcome is the need to display an overview on the one hand, while on the other hand allowing for examination of the text itself, which was solved by Lord through a confinement to text excerpts.

4. Verbatim Repetition and Bag of Words

In ReAF, verbatim repetitions, that is verbatim repeated verses, are the main feature. They are colored in the same non graded-color. For non-verbatim repeated verses, the words of each verse are used in a bag-of-words representation, that is, an unordered set of the words, where any positioning information is discarded. Depending on the absolute maximum number of shared words with any other verse,[7] each verse's cell is marked with color grading. Consider two lines in two different performances of the same song by the same bard in Lord (1960): "Nit' mu porez ni vergiju daje" and "Nit' mu porez daje ni vergiju" are different in terms of verbatim repetition but the same in terms of bag-of-words repetition. The same scenario is used for repetition within one song. The choice of color for verbatim and bag-of-words should be sufficiently different so as to not confuse the reader.

The number of highly frequent (function) words such as articles, leads to inter-verse repetition even in modern text sentences, especially between verses which are very different. Language itself is redundant when robustly conveying a message. It is repetitive in using the same words for the same references. In other words, there is a base rate of repetition. Since the ReAF is designed to highlight repetition induced by oral formulaic principles, the natural repetition of language can be seen as noise in this context. This noise can be limited by restricting bag-of-words coloring by means of a threshold. Trigrams and other ngrams have been shown to be highly indicative of language (see for instance (Cavnar & Trenkle 1994)). Excluding those verses which share less than three words from a visualization is an ambiguous reduction; it excludes some oral structures such as epithets, but excludes also many arbitrary repetitions. Comparing

7 The more formulaic a verse, the larger this number could get through the same items and protagonists being combined in another verse. This number is also printed on the cell. Another approach would be to color according to the overall number of repeated words or using some measure such as the Jaccard similarity, while keeping the number. Many more approaches thinkable.

the visualizations resulting from a non-restricted bag-of-words repetitiveness visualization with a restricted one, where the boundary is set to above 3, this setting clearly reduces visual overcrowding while higher thresholds result in sparse coloring. Further research employing a usability study will be needed to determine the best measures and thresholds. For the time being, the threshold is heuristically set to three, but the user can modify this. Likewise, the user can choose a color theme which accords to the principles of effective presentation of graded color schemes, as used, for example, in web design.

5. How to Quantify the Oral

We compared various different measures for how indicative they were of an oral composition. One major problem with this is the uncertainty concerning the classification of an epic as born-oral. A gold standard data set for this purpose would ideally contain not only relatively uncontroversial specimens, it would also be balanced in language, text length and other factors. Since it is outside the scope of this article to create such a set, a few well-known examples which were available online (some as base texts of critical editions) have been considered. The main contribution of this article is to demonstrate how overview and details about recurrence can be combined holistically into a visualization. This section is a small side-note on possible measurements for orality (and repetitivity, see for instance Altmann (1988)) which would have to be tested for significance as soon as a suitable data set is compiled. One of the tested measures ranked the texts roughly in accordance with their historically assumed orality; the *verse type/verse ratio* (VVR). This bares a superficial similarity to the Type Token Ratio (TTR). It measures the number of verse types divided by the number of verses. If no verse is repeated the value becomes 1, if there is only one verse type repeated n times, the measure becomes 1/n, which converges against 0, and the value range is thus]0,1]. While the TTR is text length dependent, verses do have different dynamics and distributions. Another measure we tested appears in (Bennet et al. 2003): pairwise comparison of chain-letter texts is achieved by setting the compression size of a text using a compression algorithm. While this measure is readily applicable within one language, the different UTF-8 block sizes and memory sizes lead to them not being readily comparable between languages using different writing systems.

VVR values: **Odyssey** 0.92, **Iliad** 0.94, **Beowulf** 0.98, **Kalevala**: 0.95, **Chanson de Roland** 0.99, **Psalms** 0.97, **Rg Veda** (beginning) 0.85, *Shahname* 0.96, *Parzival* 0.99, *Heliand* 0.97, *Knight With The Tiger Skin* 1, Divina Comedia 1, Faust 0.98, MacBeth 0.99, Luthien 1. For the majority of the most likely born-oral texts (bold), the VVR measure tends to be y low, for born-written texts (underlined)

it is on average higher. There are counter examples, which must be explored visually and qualitatively for the probability of the written medium interfering with the mode of composition. For the time being, measure and visualization are thus meant to be tools for exploration – not classifiers – designed to help the researcher of a specific text to qualitatively evaluate his/her text.

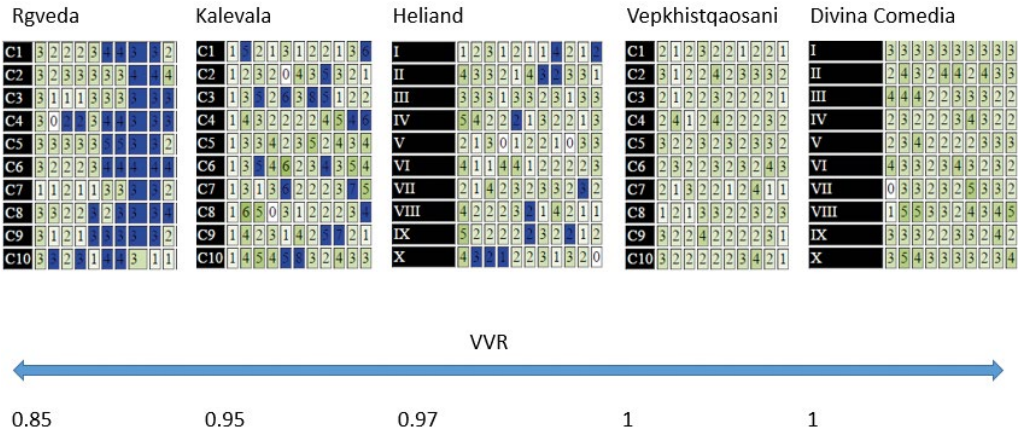


Figure 2: Some selected texts in ReAF overview with VVR value (Copyright A. Hoenen).

6. Technical Details of ReAF

ReAF is a dynamic heatmap with extended features such as text display on demand. While there is software such as R[8] generating heatmaps, for the testing of hypotheses connected with text genesis or text structure, using them without subtle programming intervention is problematic. It requires movement back and forth between the textual and the visual representation, meaning the user must locate certain positions himself each time, jumping from one to the other representation. This is a process prone to errors, especially considering verbatim repetitions which occur close to each other, or thinking about the prevalence of line skips in manuscript copying, or even regressions in eye-movements. The problem is furthermore one of unequal dimension or scaling, where the text must be scrolled in order to be read, whereas the heatmap can only be of use once it presents an overview and therefore displays the whole text or larger portions of it in a small display area (figure 3).

8 http://r-project.org

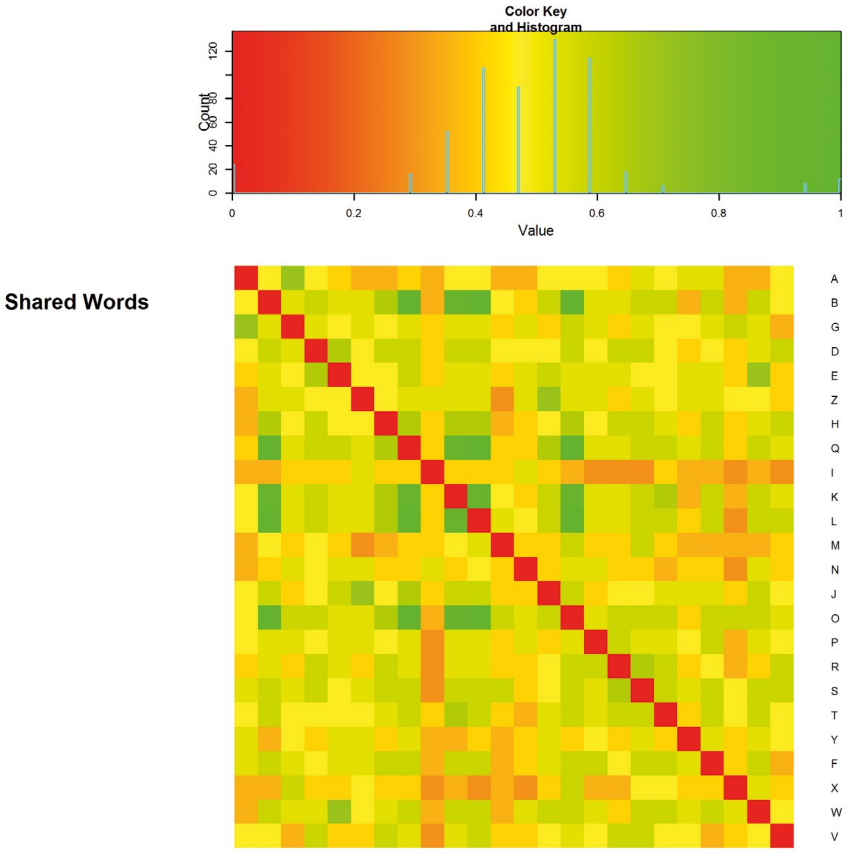


Figure 3: Hoenen ibd. Heatmap generated with R (Copyright A. Hoenen).

Each row and column in Figure 3 represent a chapter of the "Odyssee" and the coloring refers to the maximum proportion of shared words observed in any verse pair of the two chapters (with the diagonal set to 0).

ReAF tries to resolve both the need for such an overview and the need for close reading. It uses HTML, more precisely the rendering of tables in HTML. If in an HTML table, two rows do not have the same number of cells or the cell content differs in length, there is no automatic adjustment of breadth of all rows to one length, although adjusted table cells can easily be generated. This lack of specification allows a presentation of structures with unequal numbers of subunits, such as chapters with unequal numbers of verses. The advantage of this rendering is that differences in length are immediately discernible. ReAF visualizes verse text by representing one verse as one table cell, with a row representing a chapter or paragraph (or in case of plays the speech of one character).

An extension in JavaScript allows the user to read the text by expanding the cell content by double clicking. A second function dynamically colors the sister verses in one color for verbatim repetitions, and in another for bag-of-words-repetitions.[9]In this way not the whole text, but only the passages of interest are readable, allowing the size of the display to be small enough to still exhibit an overview. On mouseover, the sister verse numbers are displayed to facilitate searching. When exploring single cells, the user can choose a color scheme and decide if successive cells should be colored additionally, see Figures 4 and 5.

From the perspective of automated visualization resolving the conflict between text reading or gaining an overview, the ReAF fulfils the requirement formulated by Mazza (2009): "It is necessary for a picture to give the reader as much data as can be processed quickly, using as little space as possible." Furthermore, Shneiderman (1996) postulates a sequence as the Visual Information-Seeking Mantra: "overview first, zoom and filter, then details on demand." This sequence is one that the ReAF complies with, as will be seen in the following example. Furthermore, this is a way to combine distant reading and close reading in one interactive, digital visualization. (Mis)using HTML in this way is however not ideal and more suitable implementations are obviously possible. Text collation in digital scholarly editing, and established visualizations in this field represent possible repositories for more sophisticated technical backends.

7. Application Scenario – Text Exploration of Born-written Text

In the case of a born-written text, it is obvious that a repeated verse line can in principle be copied from a pre-existing instance on paper or from memory. Sometimes, the copied verse text precedes the first text genealogical appearance of the same verse text due to rearrangements made by the author. This is why in philology the term *urstelle* has emerged, referring not to the first *sequential* occurrence of a verse or an ensemble of verses, but to the one place in the text where the *oldest/first authorial version* of the respective verse is located. One example points to an investigation of two similar verses seen through visual inspection using ReAF's rendering of Goethe's *Faust*. This example is simple and does not give new insights into the text genesis of Goethe's *Faust* or any of his

9 To avoid long "loading" intervals, the repetition information is previously computed in the Java programming language and stored directly into the HTML code, which may produce larger files. In this way, the user only has to wait for loading the document upon opening, not at each computational step.

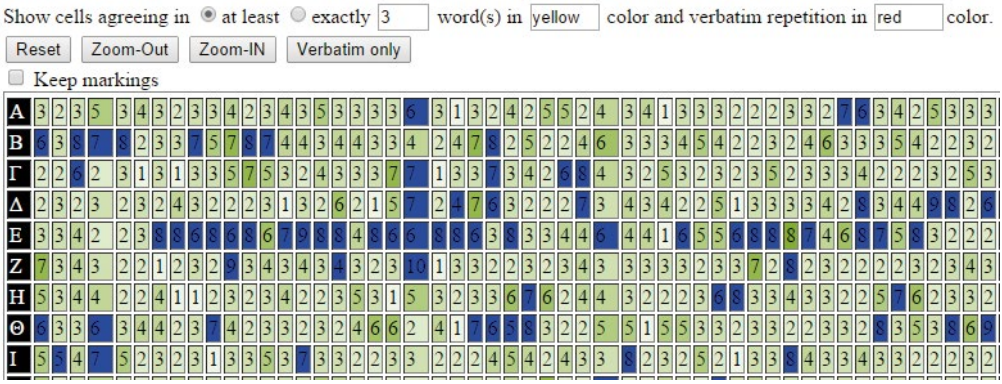


Figure 4: Hoenen ibd. ReAF *Odyssey* overview (Copyright A. Hoenen).

Figure 5: Hoenen ibd. ReAF *The Knight With the Tiger Skin*, close reading (Copyright A. Hoenen).

preferences. Furthermore, the text section can be seen whilst reading without the need for a visualization. *Faust*, Chapter 3, Prologue in the heavens:

Ihr Anblick gibt den Engeln Stärke	*Der Anblick gibt den Engeln Stärke*
Wenn keiner Sie ergründen mag;	*Da keiner dich ergründen mag*
die unbegreiflich hohen Werke	*Und alle deine hohen Werke*
Sind herrlich wie am ersten Tag	*Sind herrlich wie am ersten Tag*

These two passages could have been separately and spontaneously created. However, since Goethe is not an oral poet and the amount of verbatim repetition in his works is fairly low when compared to Homer, an alternative hypothesis would be that the verse has been created only once, and that the other occurrence of the near same sequence has resulted from a copy of the urstelle and an abridgement. Thematically, the subject of the verse is the sun and/or the Lord, the first sequential occurrence being one of three angels singing a praise hymn, and the second occurrence being a chorus of these angels repeating - not verbatim - the last four lines of the first of the three angels. It remains a question for German studies to determine the genesis of the work and which of the two occurrences is the more probable urstelle, drawing from resources such as study notes of the author.

Using the ReAF for born-written verse text such as *Faust*, see Figure 6, 7 and 8, one does not immediately explore places such as the one discussed but spots the obvious preference of Goethe's in Faust for verbatim repetition in close vicinity which works of other composers of written verse text such as Dante's *Divina Comedia* do not show, serving as an entry point to such closer exploration as we have discussed with a simple example.

8. ReAF and Preprocessing

Avestan, an extinct Indo-Iranian language maintained as liturgical language by the Zoroastrians, formed the basis for some individualized renderings for the ReAF, whose production provided the initial impulse for its development. Skjærvoe (2012) connects aspects of the Avestan written witnesses to the OFT. Renderings were based on manuscripts rather than on editions. Lost verses, the extent of which have been determined through numbering, and similar texts, were marked in black. In one rendering, prayers were visualized in order to understand the text structuring function of this special class of recurring elements.

The Avestan manuscripts are mostly written in more than one language, where e.g. Middle Persian commentaries were inserted into the original text. Visualizing the positions of the Middle Persian text in another rendering allowed

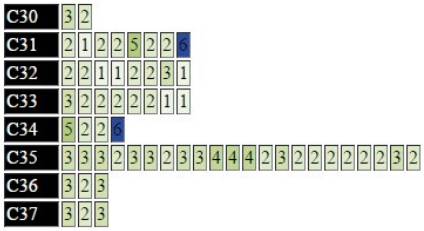


Figure 6: Hoenen ibd. ReAF *Faust* overview (Copyright A. Hoenen).

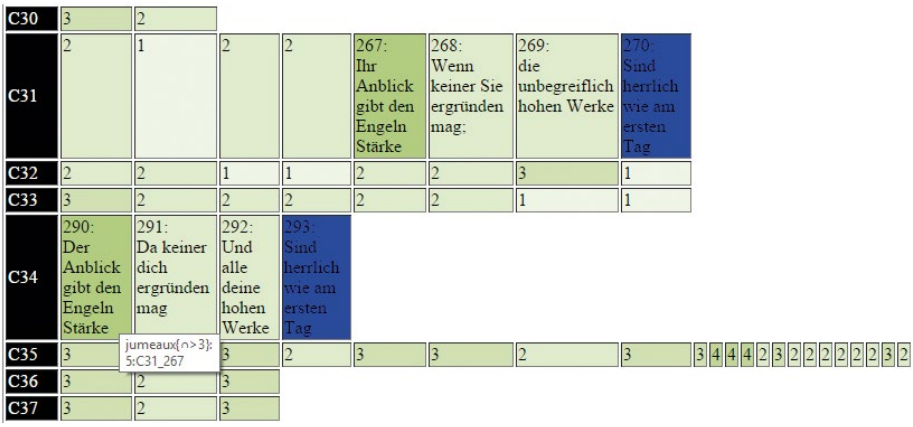


Figure 7: Hoenen ibd. ReAF *Faust*, close reading (Copyright A. Hoenen).

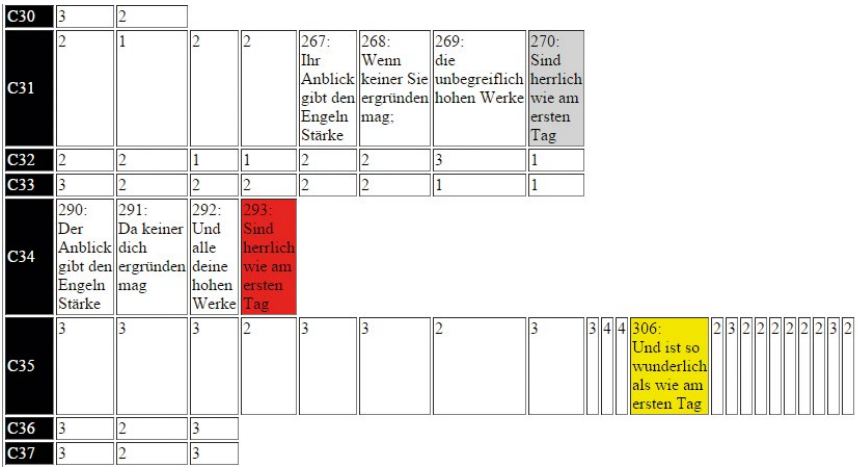


Figure 8: Hoenen ibd. ReAF Faust details, verbatim repeated in red, near verbatim in yellow (Copyright A. Hoenen).

for an intuitive understanding of the text structure, induced or perceived through them, see Figure 9.

One characteristic of the Avestan texts is that the actual performances are so repetitive, that priestly scribes abbreviated multiple repetitions in statements of a meta comment in another language, such as "from here to verse X 3 times"; one rendering of the ReAF spelled out these implied repetitions and colored the relevant cells in light grey, so what the ReAF actually represented was not the manuscript text but the intended oral ritual.

A lemmatic repetition ReAF, where instead of the words, their lemmas have formed the basis for the visualization that was produced. From this, a second visualization highlighted only those verses which were repeated on a lemma but not on a verse basis, which brought forth previously undiscovered principles of formulaic and text genetic alternations, see (Jügel 2015).

These individualized ReAF renderings may serve as examples of what the ReAF can be used for when investigating individual texts, and when preprocessing other data. Other renderings, such as ones displaying syntactic similarity, are among possible additional preprocessing steps.

9. The ReAF principle as a Generalized Visualization Technique for Texts and Language

Instead of visualizing only verse text, the ReAF can be used to visualize text in general. The question arising immediately after breaking free of the "ready-made" chunks called verse, is how to define basic units of visual representation. Unlike verse, where the content is regulated very strictly, as in case of Greek meter which has a very rigid, almost fixed number of syllables, in prose text, sentence length can differ drastically and it is usually only coincidence when two sentences have the same number of syllables. In other words, when the ReAF was applied to verse it did not transform textual information on verse length very much, allocating the same amount of visual space to each cell in the overview, but for prose text this may differ. To summarize, using the ReAF for non-verse text may mean having to compensate for such idiosyncrasies of visualization. In the following its application to non-verse text is outlined.

10. Reference Terms

As a fictional example, the visualization of reference chains is presented. This example is similar to a visualization for reference terms based on trigrams invented by Stede (2007). While Stede's visualization does not supply the close

Figure 9: Hoenen ibd. ReAF for Avestan, showing where the Middle Persian is inserted (Copyright A. Hoenen).

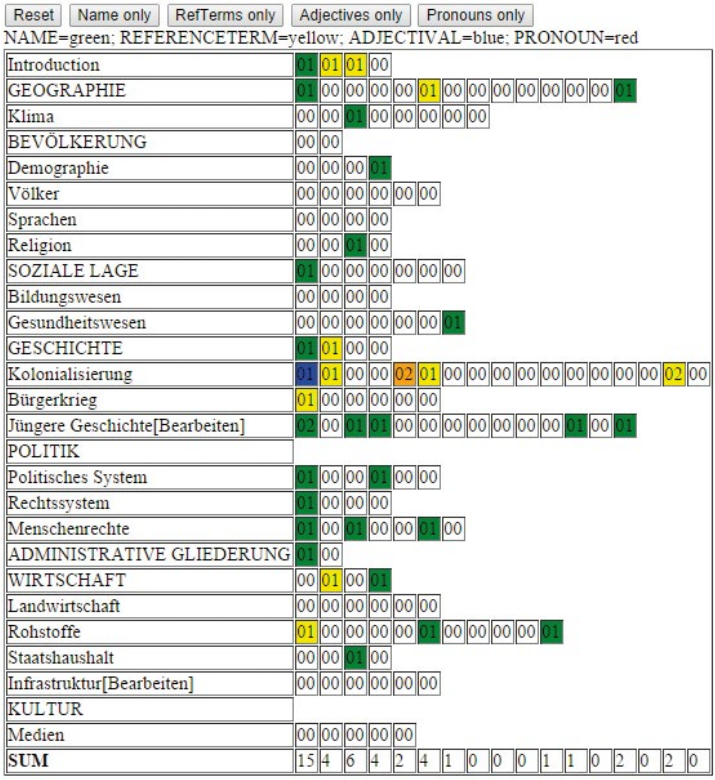


Figure 10: Hoenen ibd. ReAF as general visualization, reference chains (Copyright A. Hoenen).

reading facility, the overview in ReAF looks quite similar. It has, however, been created absolutely independently without prior knowledge whatsoever and any similarity is purely coincidental. Such independent creation of a similar visualization may be a good indicator for its visual effectiveness.

When visualizing reference chains, it is assumed, that a sentence is a basic unit and that sentence length is negligible. A Wikipedia article about Sierra Leone in German from the 3rd of October 2014, is taken and annotated for the occurrence of the title and its (anaphoric) references. Then a visualization is constructed, where the type of reference (noun, pronoun, other) is denoted by color and the number of references within the sentence printed onto the cell, see Figure 10.

With an additional table row summing-up the numeric values at the end, it can immediately be seen that the first sentence of a larger textual unit, in our case in particular and most probably in general, has a higher occurrence of direct reference, which becomes more intuitive via this visualization. Thus, the summary row is one additional customization that can be employed when adapting to a specific research question. An elaboration would be to add a cluster analysis connecting the columns, as is often supplemented in heatmaps (but not in HTML tables).

11. Summary

An initial version of an interactive visualization has been presented (ReAF 1.0). The visualization combines distant and close reading, is platform independent and can be extended through the use of web technology. It can be used to explore, analyze or compare both born-oral and born-written texts. It can also be used for prose text. The ReAF is an interactive visualization which is only feasible in the digital medium and can thus be termed a second generation digital humanities text representation system, not merely imitating print, but extending its capabilities.

12. References

Altmann, Gabriel. 1988. *Wiederholung in Texten*. Bochum: Studienverlag Bockmeyer.

Bennett, Charles, Li, M., and B. Ma. 2003. "*Chain letters and evolutionary histories*." *Scientific American* 32: 76–81.

Benson, Larry D. 1966. "*The literary character of anglo-saxon formulaic poetry.*" *PMLA,* 81 (5): 334–341.

Cavnar, William B., and J. M. Trenkle, J. M. 1994. "*N-gram-based text categorization.*" In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 161–175.

Coe, Michael D. 2012. *Breaking the Maya Code*. New York: Thames and Hudson Ltd.

Culley, Robert. 1967. *Oral formulaic language in the Biblical psalms*. Toronto: University of Toronto Press (Near and Middle East Series).

Duggan, Joseph J. 1973. *The Song of Roland – Formulaic Style and Poetic Craft*. Berkeley: University of California Press.

Finnegan, Ruth. 1992. *Oral Poetry: Its Nature, Significance and Social Context*. Cambridge: Cambridge University Press.

Foley, John. 2002. *How to Read an Oral Poem*. Urbana, Ill.: University of Illinois Press.

Gonda, Jan. 1959. *Stylistic Repetition in the Veda*. Amsterdam: Noord-Holland (Volume 65(3) of Verhandelingen der koninklijke nederlandse Akademie van Wetenschappen, Afd. Letterkunde, N.R. Noord-Hollandsche Uitgevers Maatschappij).

Goody, Jack. 1987. *The interface between the written and the oral*. Cambridge: Cambridge University Press.

Jügel, Thomas. 2015. "*Repetition analysis function (ReAF I)*: Identifying textual units in Avestan." *Indogermanische Forschungen* 120: 177–208.

Kailasapathy, Kanagasabapathy.1968. *Tamil heroic poetry*. Oxford: Clarendon Press.

Lord, Albert B. 1960. *The Singer of Tales*. Cambridge, Mass.: Harvard University Press.

Magoun, Francis P. 1980. *The oral-formulaic character of anglo-saxon narrative poetry. Speculum* 28 (1953): 446–467.

Mazza, Ricardo. 2009. *Introduction to Information Visualization*. London: Springer.

Ong, Walter J. 2012. *Orality and Literacy. The technology of the word*. London: Routledge.

Parry, Millman, and A. Parry. 1987. *The Making of Homeric Verse: The Collected Papers of Milman Parry*. Oxford: Oxford University Press.

Rubanovich, Julia. 2011. "Orality in Medieval Persian Literature." In *Medieval Oral Literature*, edited by Karl Reichl. Berlin: de Gruyter, 653–680.

Shneiderman, Ben. 1996. "*The eyes have it: A task by data type taxonomy for information visualizations.*" In *Proceedings of the 1996 IEEE Symposium on Visual Languages, VL '96*, 336–343.

Skjærvoe, P. O. 2012. "*The zoroastrian oral tradition as reflected in the texts.*" In *The Transmission of the Avesta*, edited by Alberto Cantera. Wiesbaden: Harrassowitz, 3–48 (Iranica 20).

Stede, Manfred. 2007. *Korpusgestützte Textanalyse. Grundzüge der Ebenen-orientierten Textlinguistik*. Tübingen: Narr.

Whallon, William. 1969. *Formula, Character and Context – Studies in Homeric, Old English and Old Testament Poetry.* Cambridge, Mass.: Harvard University Press.

Yamamoto, Kumiko. 2003. *The Oral Background of Persian Epics*. Leiden: Brill.

Online sources of the verse texts, date of last access and traditional classification:

Odyssey: http://titus.uni-frankfurt.de/ 08.09.2014, oral acc. to Lord (1960)
Iliad: http://gutenberg.spiegel.de 08.09.2014 oral acc. to Lord (1960)
Beowulf (halflines): http://www.humanities.mcmaster.ca 26.09.2014 oral acc. to Lord (1960)
Kalevala: http://sacred-texts.com/ 08.09.2014 mainly oral acc. to Foley (2012)
Chanson de Roland: http://www.hs-augsburg.de/ 08.09.2014 oral acc. to Lord (1960)
Psalms: http://www.mechon-mamre.org/ 08.09.2014 oral acc. to Culley (1967)
Rg Veda: http://titus.uni-frankfurt.de/ 08.09.2014 oral, see e.g. Gonda (1959)
Shahname: http://fa.wikisource.org 08.09.2014 acc. to Foley (2002) oral, Yamamoto (2003), Rubanovich (2011) more complex assessment
Parzival: http://titus.uni-frankfurt.de/ 08.09.2014
Heliand: http://titus.uni-frankfurt.de/ 08.09.2014
Knight With Tiger Skin: http://titus.uni-frankfurt.de/ 10.09.2014
Divina Comedia: http://www.filosofico.net 08.09.2014 written
Faust: http://www.gutenberg.org/ 08.09.2014 written
MacBeth: http://www.gutenberg.org/ 08.09.2014 written
Luthien: http://allpoetry.com 08.09.2014 written

Adrien Barbaresi

A Constellation and a Rhizome: Two Studies on Toponyms in Literary Texts

Abstract Although the attention of linguists is commonly drawn to forms other than proper nouns, the significance of place names in particular exceeds the usual frame of deictic and indexical functions, as they encapsulate more than a mere reference in space. In this article, two different examples are presented in order to understand them, along with two practical examples of visualization in literary texts from the beginning of the 20th century. Research on toponym extraction and linkage is discussed from an interdisciplinary perspective, as digital literary studies are not mere numeric accounts: if they deal with detecting, counting, and projecting occurrences, one also ought to describe and criticize the detachment provoked by "blind" computer-based thinking.

The first case consists of a preliminary study of travel literature based on Richthofen's *Travel Journals from China* (1907). The resulting map retraces the path taken by the author in the Shandong province by combining coordinates, sequences, and a sense of time. In order to critically analyze this synthesis the concept of constellation is introduced and discussed. The second study focuses on a complex work, the literary magazine *Die Fackel* (1899–1936) with the particular example of co-occurrences of toponyms. The paths drawn on the map depict chains of thought and lines of flight. The result is understood through the concept of the rhizome, by which heterogeneous information can be connected and displayed. The finality of Visual Linguistics does not reside in an apparatus but rather in the substrate of interpretable representations which put words into perspective.

1. Introduction

Although the attention of linguists is commonly drawn to forms other than proper nouns, the significance of place names in particular exceeds the usual frame of deictic and indexical functions, as they embrace more than a mere reference in space. The value of toponyms is highlighted in testimonies gathered in the field, such as the comment by Ulahi from Papua New Guinea who asks: "I

Published in: Noah Bubenhofer, Marc Kupietz: Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

haven't heard your land names, so who are you?" His study on Bosavi language leads Feld (1996) to the idea that "place significance neither starts nor ends with the linguistic referentiality of placenames", and that "there is considerable variation in how names hold and unleash significance".

In Western tradition, a current of reflection, whose origin can be dated back to the 1960s, has provided the theoretical foundations of the *spatial turn*, whose epitome is the concept of space as emergent rather than existing a priori, and composed of relations rather than structures (Warf 2009). As Foucault described in a seminal lecture:

> "Nous sommes à un moment où le monde s'éprouve, je crois, moins comme une grande vie qui se développerait à travers le temps que comme un réseau qui relie des points et qui entrecroise son écheveau."[1] (Foucault [1967] 1984)

Foucault further explained that "our epoch is one in which space takes for us the form of relations among sites". As a consequence, both the definition and importance of space have been re-evaluated throughout the humanities.

More recently, the emergence of *GeoHumanities* (Dear et al. 2011) or *Spatial Humanities* (Bodenhammer et al. 2010), has prompted the exchange of research objects between disciplines as well as enforcement of the spatial turn in practice through specific methods of analysis. Although some observe that there is "remarkably little overlap between the digital humanities community and the spatial history community" (Mostern & Gainor 2013), the common denominator seems to be the will to open up new spaces and experiment in a transdisciplinary perspective:

> "The parallel disciplinary structure of German-language *Literaturwissenschaft* and linguistics is not merely anecdotal; rather, it is an index of the necessary interchange of methodologies and content which holds beneficial possibilities for both fields." (Domínguez 2011)

Following these premises, I wish to present two studies that center on the visualization of place names in literary texts, with particular emphasis on the concept

1 "We are at a moment, I believe, when our experience of the world is less that of a long life developing through time than that of a network that connects points and intersects with its own skein." http://foucault.info/doc/documents/heterotopia/foucault-heterotopia-en-html

of visualization, that is on the processes and not on the products (Crampton 2001). As a consequence, I feel the need to present them with a much needed critical apparatus, by giving a theoretical perspective on what is being shown and seen: Firstly because digital methods in humanities ought to be criticized (Wulfman 2014) and secondly, because the cartographic enterprise bears both a thrill and a risk: on the one hand "adding more to the world through abstraction", and on the other hand "adding to the riskiness of cartographic politics by proliferating yet more renders of the world" (Gerlach 2014).

The remainder of this article details theory and practice with the following structure: First, research on toponym extraction and linkage is discussed from an interdisciplinary perspective; then the constellation paradigm is presented along with a practical study on travel literature; finally, the notion of the rhizome is introduced and presented by means of a diachronic study.

2. Drawing points and lines out of words

2.1 Distant cartographic reading on relative maps

Progress in fulltext geocoding, also known as "geoparsing" or "geographic information retrieval" (Leetaru 2012), is tightly linked to progress in mapping systems, mostly thanks to a technology-driven evolution (Juvan 2015) as Geographic Information Systems (GIS) and series of tools come from other disciplines. An underlying assumption resides in the belief that understanding language and literature is not accomplished by studying individual texts, but by aggregating and analyzing massive amounts of data (Jockers 2013). Because it is impossible for individuals to "read" everything in a large corpus, advocates of distant reading employ computational techniques to "mine" the texts for significant patterns and then use statistical analysis to make statements about those patterns (Wulfman 2014). One such is Moretti (1999) who pleads for "a geography of literature" and "distant cartographic reading".

There is however a notable difference between literary geography and literary cartography, the first one being essentially anecdotal and auxiliary, if not ancillary (Juvan 2015). Although literary geography has been existing for more than 100 years (Piatti et al., 2011), much work still has to be done to define and establish literary cartography, an interdisciplinary field which has been evolving at an exponential pace over the last decade (Caquard & Cartwright 2014).

Concerning the maps themselves, the consensus in the research community has evolved towards a relativity in construction and uses of maps: "post-representational cartography" (Rossetto 2014), where there is neither a "ground truth", nor a "cartographic truth", and where "the map is not objectively 'above'

or 'beyond' that which is represented" (Crampton 2001). Although the maps seem immediately interpretable, they are not an objective result but a construct, the result of filtering, "a connection made visible" (Moretti 1999). As such, cartography is not the realization of static maps, but rather the description of emergent structures, and there is no single or best map. The paradigm of *geographic visualization* stands in opposition to the tradition:

> "Traditional cartography has emphasized public use, low interactivity and revealing knowns, while visualization emphasizes private use, high interactivity and exploring unknowns." (Crampton 2001)

In literary studies, it is understood that maps are only the beginning of exploratory work, not only following from said paradigm but also out of a defiance against quantitative information of an additive, cumulative nature about potentially heterogeneous phenomena:

> "Whenever literary scholars screen, read, interpret and compare the maps, they do what is regarded as one of their core competences: to consider carefully ambiguities, to compare, to contextualise, to shed light on historical references, to juxtapose several readings, to combine methods and tools." (Piatti et al. 2011)

Counting words, in this particular case place names, does not appear to be enough for the researcher in literary studies. Even if the map in itself is relative, being "less important than the process of making it and using it" (Caquard & Cartwright 2014), it plays an ambiguous part in distant reading, since it has to be flexible enough to adapt to new contexts and analyses, while remaining exact and in this sense trustworthy. The information it contains and reveals cannot be verified on a point-per-point basis, yet it can be the starting point of a comprehensive interpretation. For this reason, the ability to detect and project place names with reasonable accuracy is paramount.

2.2 Placing points: On the extraction of toponyms

In the field of information extraction or information retrieval, named entity recognition is a set of text mining techniques designed to discover named entities, connections and the types of relations between them (Chinchor 1997). The particular task of finding place names in texts is commonly named place names extraction or toponym resolution. It involves the detection of words and phrases that may potentially be proper nouns as well as a second operation classifying them as geographic references (Nouvel et al. 2015). A further step, geocoding, resides in disambiguating and adding geographical coordinates to a place name:

> "At its core, fulltext geocoding involves scanning a body of text to identify potential geographic references and then using an external knowledgebase, called a 'gazetteer', and document context to disambiguate and convert the references to a geospatial form." (Leetaru 2012)

Named entity resolution often relies on named-entity recognition and artificial intelligence (Leidner & Lieberman 2011). However, knowledge-based methods using fine-grained data, for example from Wikipedia, have already been used with encouraging results (Hu et al. 2014). Work by Efremova et al. (2015) on family relationship extraction from historical documents proves that it is possible to use a name dictionary as well as patterns to perform the extraction and remove ambiguity in a robust way, although the documents span around 500 years. Pouliquen et al. (2006) demonstrate that an acceptable precision in the detection and disambiguation of place names (76 % in their study) can be reached by including information such as distance, importance, immediate lexical context, and main places of the text to be analyzed.

The first study below relies on manually annotated data, while the second is based on automatic extraction. One of their common denominators is the necessity to construct appropriate gazetteers, that is mapping historical and possibly not yet standardized variants to a canonical standard from which to derive geographic metadata. Another common denominator consists of the connections that are drawn between the extracted toponyms in the exploratory sense of geographic visualization.

2.3 Connecting and dividing: the ambiguous nature of lines

Maps are mostly projected in Euclidean spaces where two points are connected by a single line. By extension, the word *line* defines a series of connected points on a plane, potentially both linking and separating. Lines can enforce and divide when maps are used as instruments of power, which is another reason why the proponents of post-representational cartography call for a change of perspective. For example, lines that draw state boundaries are increasingly called into question, especially concerning historical states and texts, as they fix an evolving process and convey a sense of immobility that fails to describe the past accurately:

> "We need to recognize that territorial maps of ancient states are an idealized projection of state authority rather than a depiction of the way in which ancient political domains were actually governed." (Smith 2005)

In the framework of geographic visualization, advances have been made towards less static maps, with the wish to foster more flexibility on the map and in the mind of the reader, allowing for reflection out of the box and outside of the boundaries:

> "Recent attention to globalization, diaspora, 'nomadism', and cyberspace is showing us the need for new and powerful theoretical work to replace, rather than simply supplement, the polemics and models produced by an academic collectivity concerned mostly with locatable cultures, bounded nations, and the imperial past" (Campbell 2002).

What is true for the territory also holds true for the texts, linkages can be seen as "mappings and tracing imposed on the data" (Wulfman 2014). The lines may also reveal spatial patterns that would otherwise remain hidden in texts (Bodenhammer et al. 2010), or connect ideas, places, or peoples. As well as lines of force, there can also be lines of thought, traces figuring steps of reflection and analysis. This ambiguous nature calls for a differentiated approach. As long as the traces can be flattened, recombined, and superposed (Latour 1985), they allow for a scientific process of (re-)construction which is open to additions, changes of scale and perspective.

Visualization of linguistic phenomena has to account for their changing nature relative to context and passage of time, thus going deeper than a mere

graph, a general and operative concept which usually implies a series of operations performed upon it. Beyond the immediacy of nodes and edges, the relations may be expressed through the concept of network. However, precisely in the case of connections, the word "network" is to be used with caution as Latour (1999) suggests. Although it is ubiquitous in the terminology of the spatial turn, the now predominant interpretation in the sense of the World Wide Web, suggests an immediacy which is contrary to the status it had before. In most occurrences of the term, "meshwork" would be more appropriated than "network" (Ingold 2007). In the following, two different modes of analysis featuring points and lines are presented. The potential superficiality of networks is carefully avoided as two related alternative concepts which both break the linear model of reflection (Wu 2009) are presented: the constellation and the rhizome.

3. First study: Constellation

3.1 The concept of constellation

According to the words of De Certeau (1990), where the map splits, the narration – as diegesis – traverses. Even if a map remains static in comparison to the movement of the diegesis, an attempt can be made to try and overcome this dichotomy. In this first study, a decision has to be made about whether to trace the line or not, which makes it a prototypical constellation, an assemblage of points which let figures and interpretations emerge:

> "As we make constellations by picking out and putting together certain stars rather than others, so we make stars by drawing certain boundaries rather than others. Nothing dictates whether the skies shall be marked off into constellations or other objects." (Goodman 1983)

Stars and skies are here place names and come from the spatio-temporal frame of the narration in travel literature, a specific case which challenges the idea that "linkages are not in the data" (Wulfman 2014). In fact, it is possible to try to visualize a progression in the narration by bringing phenomena to light with annotation, and drawing lines between certain points to make patterns visible which would remain unnoticed otherwise.

3.2 Modus operandi

The maps with an "indexical function" (Juvan 2015), are part of a preparatory study on travel literature, which includes the annotation of place names from Richthofen's *Travel Journals from China* (1907) in XML format and conform to the guidelines of the Text Encoding Initiative.[2] The outline below focuses on four weeks of travel in 1869 through the Shandong province (East China). Annotation of toponyms is done manually due to difficulties with non-standard transcriptions of historical Chinese names.[3] For each identified name, Wikidata is used as reference. This is a document-oriented, collaboratively edited knowledge base operated by the Wikimedia Foundation (Vrandečić & Krötzsch 2014). As it is international and editable, it allows for the registration of linguistic variants, and metadata such as coordinates can be added. The *type* attribute defines whether the given location is to be represented as a point or as a surface. XML TEI attributes are used to encode direction of travel and sequences of visited places (*next* and *prev*), as well as information about time series (*when* and *notafter*). Finally, toponyms and metadata are extracted from the resulting XML document and combined with information from the authority file before being projected on a map.

3.3 Result

The result combines coordinates, sequences, and a sense of time, which are depicted on Figure 1 by (respectively) points, lines, and a color scheme, projected on a map using the cartographic software TileMill[4] and customized with CartoCSS. For the sake of clarity, current standard names in English are also projected on the map, as well as historical borders of Chinese provinces in the 19th century.[5] Additionally, a color contrast distinguishes the land mass (Mainland China) from the sea (Yellow Sea and East China Sea). The size of the dots is in relation to toponym frequency in the text. Different shades of red illustrate time differences from the first to the fourth week of the trip, while gray points depict places which are named without having been visited., and another constellation, whose lines are not retraced, contrary to the itinerary.

2 http://tei-c.org

3 The preliminary steps have been performed jointly with Benno Wagner (Zhejiang University) and Li Liu (Stuttgart University).

4 https://github.com/mapbox/tilemill

5 The source is a prototypical spatial humanities project, China Historical GIS: http://www.fas.harvard.edu/~chgis/

Fig. 1: Four weeks of travel through Shandong in Richthofen's *Travel Journals from China* (1907). The size of the dots is in relation with toponym frequency in the text. Itinerary and time are expressed respectively with lines and shades of red.

In his chapter about Shandong, Richthofen writes "If you have a map of China, you will be able to follow my way to Chi-Fu [Yantai]."[6] In fact, the constellation restores a feeling for time and space with its lines going from one sea to another, it re-inscribes the narration *in situ*. Interpretation is facilitated, and problems in the manual identification of places and in the coordinates associated with them are much more easily spotted on a map. The lines construct an itinerary, so that improbable constellations are easily singled out. This is particularly useful in a land of the scale of China and can also help manual disambiguation of ubiquitous or past toponyms.

Dating from the time when the German Empire was present in China with the *Tsingtau* (Qingdao) outpost, Richthofen's lines are not only exploratory, his constellations encompass space that is to be conquered if not militarily at least commercially. Lines are an abstract way of figuring the itinerary, but in this first step they do not express the liveliness of the narration or the events related to them:

> "Once a moment of rest along a path of movement, place has been reconfigured in modernity as a nexus within which all life, growth and activity are contained. Between places, so conceived, are only connections." (Ingold 2007)

It could be the task of spatial digital humanities to make these spaces appear and to restore a dimension of wayfaring and uncertainty in travel literature.

4. Second study: Rhizome

4.1 Concept and object of the study

The second study stems from the concept of rhizome as formalized by Deleuze and Guattari (1980).[7] Its main principles are connection and heterogeneity, with no fixed order; "flat" multiplicity defined by "lines of flight"; "asignifying rupture"; and finally "cartography and decalcomania" ("tracing something that comes ready-made"). In order to fully understand the interest of the rhizome in visualizations, it is necessary not to limit it to a particular organic representation

6 „Wenn Ihr eine Karte von China habt, werdet Ihr meinen Weg über I tschóu fu, Tsi Nan fu, Lai tschóu fu nach Tschifu verfolgen können." (Richthofen 1907)

7 This section has been adapted from preliminary studies (see Barbaresi 2017, Barbaresi 2018).

of data, to a mere element in a grammar of visualization as Lima (2011) presents it. In fact, there is a discrepancy between the realm of the visual which can be traced back to the surge in quantification that occurred in the Early Modern Era (Krämer 2010), the "pervasive visualism" (Feld 1996) in Western tradition, with its insistence on visual metaphors to describe and ascribe reason (Barbaresi 2012), and the concept of rhizome. The lines of flight foster a real multiplicity and heterogeneity, and they are recombined, stopped and extended ("asignifying rupture"), which is why the rhizome is to be seen as a mode of experimentation rather than an interpretation, a representation, or an imitation (Antonioli 2010). The rhizome puts an emphasis on several aspects of post-representational cartography as described above; it "pertains to a map that must be produced, constructed, [...] and has multiple entryways and exits" (Deleuze and Guattari 1987 [1980]). For there is no cartographic truth. The map has to be seen as a tool for the multiplication of accesses to reality.

The concept of the rhizome has been used in corpus linguistics by Scharloth et al. (2013) to qualify discourses captured by collocation graphs. The authors performed significance tests to extract relevant collocations over the course of time, isolate clusters, and uncover lingering discourse elements. The concept of collocation and of subterranean word complex is a common one, however, the present study is different in its form and its content because it literally leads to a map, and most importantly because its text base is in itself rhizomatic.

The reality which I wish to depict under the paradigm of the rhizome is a flat version of a diachronic corpus seen through the lens of toponyms (Barbaresi 2017). The text basis for this investigation is the digitized version (AAC-Fackel corpus[8]) of the satirical literary magazine *Die Fackel* ("The Torch"), originally published and largely written by the satirist and language critic Karl Kraus in Vienna from 1899 until 1936. It is a complex work, both from a synchronic and from a diachronic perspective, where turnarounds, quotes, and multiples views play a central part. As such, it carries heterogeneity at its core and contains a considerable variety of toponyms (Biber 2001), which are highly significant due to the multinational nature of the Austro-Hungarian Empire and the later formation of a territorially diminished state. Deleuze and Guattari wrote that we are all "traversed by lines, geodesics, tropics, and zones marching to different beats and differing in nature" (1987 [1980]). Kraus, in his constant balancing act between his origins, his beliefs, his sensibility and criticism, must have been perfectly aware of the acuteness of these lines.

8 http://aac.ac.at/fackel It offers free online access to 37 volumes, 415 issues, 922 numbers, comprising more than 22.500 pages and 6 million wordforms.

4.2 Modus operandi

A treatment of toponyms has been described in Barbaresi & Biber (2016).[9] The texts were digitized, manually corrected as well as manually annotated with respect to the names of persons and institutions, so that most proper nouns which are not place names were excluded from the study. The tokenized files of works to be analyzed were filtered and matched with the database by finite-state automatons. Toponyms were extracted using a sliding window (for multi-word names up to three components). A cascade of filters was used; current and historical states (e.g. Austria-Hungary); regions, important subparts of states, and regional landscapes (e.g. Swabia); populated places; geographical features (e.g. seas, valleys). Wikipedia's API[10] is used to navigate in categories and to retrieve coordinates, which are completed by hand for states and regions. Second, current information is also compiled from the Geonames database[11]: data for European countries are retrieved and preprocessed (variants and place types). Disambiguation being a critical component (Leetaru, 2012), using an algorithm similar to Pouliquen et al. (2006), who demonstrated that an acceptable precision can be reached that way, suggesting the most probable entry based on distance to Vienna (Sinnott, 1984), contextual information (closest-country, last names resolved), and importance (place type, population count).

The main areas of progress in the present endeavor are evolution in the gazetteers in order to bypass the disambiguation for a hand-picked list of places, as well as changes in the visualization. During the 20th century there have been significant political changes in Central Europe that have severely affected toponyms, meaning geographical databases lack coverage and detail. Consequently, the database I develop follows from a combination of approaches, gazetteers are curated in a supervised way to account for historical differences, and current geographical information is used as a fallback.

The visualization below includes a subset of GeoCollocations (Bubenhofer 2014), i.e. collocations of toponyms. Co-occurrences of toponyms are extracted using a sliding window of fixed length and the maximal distance between two extracted place names is fixed at twenty tokens. Lines between co-occurring places are only drawn if the expressions are found within this window, and they follow the sequence by which they are used. This allows for the visualization of paths depicting chains of thought (*Gedankengänge*) as well as their intensity,

9 This extraction of German and Austrian place names in historical texts is part of a cooperation between the Berlin-Brandenburg and the Austrian Academies of Sciences.

10 https://www.wikidata.org

11 http://www.geonames.org

well-trodden or seldom. Surfaces, for instance regions, cannot be represented because of evolution in the course of history but mainly because of the difficulties of linking surfaces without tampering with map readability.

4.3 Result

The concept of rhizome is particularly relevant for Kraus, as the Austrian polemicist was always concerned by the multiple aspects of discourse and reality. In addition, his work as a whole evades distant reading processes due to the number of citations and an ever present and extensive usage of parody. There are resolutely escapes and lines of flight in his work, so that it would be vain to design an authoritative, arguably objective cartography of *The Torch.* For that matter, it can be said that the experiment resulting in Figure 2 depicts a rhizome connecting heterogeneous information about the *Weltanschauung* of Kraus and his contemporaries.

As in the first study, a map was made using the same software and a similar visual grammar. However, not only do the dots vary in color and size, the lines also express more than a simple relation since spatio-temporal information is mostly carried by the latter. The points depict extracted place names classified into categories (yellow: sovereign territories; orange: regions; green: populated places; blue: geographical features). The most frequent place names are printed out. The lines are of two different types with coastlines depicted in gray to give a sense of orientation, and as discussed above there are no borders on the map. The lines picture co-occurrences of extracted place names, the time scale is expressed in a span from light green to deep blue. In accordance with the principles of heterogeneity and "asignifying rupture", they are frequently interrupted.

There are several problems with the mapmaking process. First, it is difficult to place symbolic names (e.g. Europe) and conflicting states of historical nations (e.g. Germany or Austria). Phenomena in the low-frequency range are filtered out due to their size on the map, but clusters may emerge. As regards the extraction in itself, potential conceptual caveats include previous times as well as fictitious places, especially names which can refer to mythological and actual places of Ancient Greece or Rome. However, toponyms can also be seen as lexical entries to particular themes, especially in the high-frequency range ("Austria-Hungary", "German Empire"), as they also convey a sense of chronological evolution.

While *Die Fackel* criticizes mechanical, instrumental language (Hirt 2002), what could be called "well-informed" linguistic instruments assist by materializing dots or sequences, albeit not without "human" intervention. The resulting rhizome is not an authoritative cartography of *Die Fackel,* but rather an indirect depiction of the viewpoint of Kraus and his contemporaries. Drawing on Kraus'

Fig. 2: Experiment on European scale projecting place names types and cooccurrences through time.

vitriolic recording of political life, toponyms in *Die Fackel* tell a story about the ongoing reconfiguration of Europe.

The zoomed-out version of the map conveys the strange impression that Europe is a force field on which points east and west are projected. The lines of force contain European countries and capitals. Their spatial patterns document an inclination for major cultural centers, thus confirming the findings of the previous study (Barbaresi & Biber 2016), whereas the chronological dimension captures a major shift towards the end of publication. The force field intensifies as its range narrows, showing both the interplay of major European powers of the time and the emergence of transatlantic (westwards) and transeuropean (eastwards) relationships. This reconfiguration can be read as an intensification of tensions and a prefiguration of other schemes, this time of a military nature.

5. Conclusion

Diagrammatic inscriptions are a point of linkage between thinking and intuition, between the "noetic" and the "aisthetic" (Krämer 2010). In the sense of the literature mentioned on this topic, the present article is part of a global move to open up a productive joint research field. The maps referenced here are two possible realizations among others, whose goal is to uncover patterns and specificities which do not appear or are not easily retraceable during close reading.[12] The differences in the extraction – manual vs semi- or unsupervised – as well as in the visualization methods – figurative vocabulary, scale, projection – do not account for differences in the data. The first study imposes constellations on a map, which is coherent with the views of the author of the travel journal, who himself relies on maps, abstraction, and underlying structures of power. The second one addresses a complex and heterogeneous work, which is best achieved with the notion of rhizome and the example of co-occurring place names. The resulting lines are neither uniform nor unequivocal. It may be difficult to draw a distinction between lines of thought and lines of force, as the same lines that are an instrument of delimitation and exclusion can also connect, extend, and, entangle. If there are lines of force in the present maps, then it is in a passive way, by yielding through their entanglement an accurate image of the complexity and multiplicity of the context.

Digital literary studies are not mere numeric accounts. If on one hand they deal with detecting, counting, and projecting occurrences, on the other hand they have to include and criticize the alienation provoked by "blind" computer-based thinking. Because the rhizome is in part a structuralist notion, it may be computed or even computerized. Thus it is only superficially paradoxical that blind numeric analyses can be applied in order to make linguistic phenomena visible. The finality of "visual linguistics" is not an apparatus or a concept of an operational nature, but the substrate of interpretable representations which do not follow blindly numeric data but instead put things in perspective.

The "human" interventions on the maps as well as the technical competence to do so replace the studies presented here in the hermeneutic circle of the philological tradition. The difference between a mere data collection project and a research study resides precisely in the number and diversity of filters used, in the perspective chosen and the probable imperfections. In this regard, spatial and digital humanities can hopefully be equipped with the necessary conceptual background to critically observe toponyms resonating in space and time.

12 The code written and the materials gathered for this study are available online: https://www.github.com/adbar/toponyms.

6. References

AAC-FACKEL. *Die Fackel*, edited by Karl Kraus, Wien 1899–1936, AAC Digital Edition No 1, http://www.aac.ac.at/fackel.

Antonioli, Manola. 2010. "Singularités cartographiques." *TRAHIR* (2).

Barbaresi, Adrien. 2012. "La Raison aveugle ? L'époque cybernétique et ses dispositifs." communication at *Les critiques de la raison*, University Paris-Est Créteil (UPEC).

Barbaresi, Adrien, and Hanno Biber. 2016. "Extraction and Visualization of Toponyms in Diachronic Text Corpora." In *Digital Humanities 2016: Conference Abstracts*, Cracow, 732–734.

Barbaresi, Adrien. 2017. "Toponyms as Entry Points into a Digital Edition: Mapping *Die Fackel* (18991936)." In *Digital Humanities 2017: Conference Abstracts,* McGill University & Université de Montréal, 159–161.

Barbaresi, Adrien. 2018. "Toponyms as Entry Points into a Digital Edition: Mapping *Die Fackel*." *Open Information Science*, De Gruyter, to appear.

Biber, Hanno. 2001. „In Wien, in Prag und infolgedessen in Berlin – Ortskonstellationen in der ‚Fackel'." In *Berlin – Wien – Prag. Moderne, Minderheiten und Migration in der Zwischenkriegszeit*, edited by Susanne Marten-Finnis and Matthias Uecker, Bern: Lang, 15–26.

Bodenhamer, David J., John Corrigan, and Trevor M. Harris. 2010. *The spatial humanities: GIS and the future of humanities scholarship*. Bloomington: Indiana University Press.

Bubenhofer, Noah. 2014. „Geokollokationen – Diskurse zu Orten: Visuelle Korpusanalyse." *Mitteilungen des Deutschen Germanistenverbandes* 61 (1): 45–89.

Caquard, Sébastien, and William Cartwright. 2014. "Narrative Cartography: From Mapping Stories to the Narrative of Maps and Mapping." *The Cartographic Journal* 51 (2): 101–106.

Campbell, Mary Baine. 2002. "Travel writing and its theory." In *The Cambridge Companion to Travel Writing*, edited by Peter Hulme, and Tim Youngs, Cambridge: Cambridge University Press, 261–278.

De Certeau, Michel. 1990. *L'invention du quotidien*, vol. 1, *Arts de faire*, Paris : Gallimard.

Chinchor, Nancy, and Patricia Robinson.1997. "MUC-7 Named Entity Task Definition." In *Proceedings of the 7th Message Understanding Conference*.

Crampton, Jeremy W. 2001. "Maps as social constructions: power, communication and visualization." *Progress in Human Geography* 25 (2): 235–252.

Domínguez, César. 2011. "Literary Geography and Comparative Literature." *CLCWeb: Comparative Literature and Culture* 13 (5): 3.

Dear, Michael, Jim Ketchum, Sarah Luria, and Douglas Richardson. 2011. *GeoHumanities: Art, history, text at the edge of place*. London: Routledge.

Deleuze, Gilles, and Félix Guattari. 1987. A Thousand Plateaus: Capitalism and Schizophrenia English translation by Brian Massumi. Minneapolis: University of Minnesota Press.

Efremova, Julia, Alejandro Montes García, Toon Calders, and Jianpeng Zhang. 2015. "Towards population reconstruction: extraction of family relationships from historical documents." In *First International Workshop on Population Informatics for Big Data* (21th ACM-SIGKDD PopInfo '15), 1–9.

Feld, Steven. 1996. "Waterfalls of song: An acoustemology of place resounding in Bosavi, Papua New Guinea." In *Senses of Place*, edited by Steven Feld, and Keith H. Basso. Santa Fe, NM: School of American Research Press, 91.

Foucault, Michel. 1984. "Of Other Spaces, Heterotopias." *Architecture, Mouvement, Continuité*, 5: 46–49.

Gerlach, Joe. 2014. "Lines, contours, and legends. Coordinates for vernacular mapping." *Progress in Human Geography* 38 (1): 22–39.

Goodman, Nelson. 1983. "Notes on the well-made world." In *Methodology, Epistemology, and Philosophy of Science,* edited by Carl G. Hempel, H. Putnam, and Wilhelm K. Essler. Dordrecht: Springer, 99–107.

Hirt, André. 2002. *L'universel reportage et sa magie noire. Karl Kraus, le journal et la philosophie.* Paris: Kimé.

Hu, Yingjie, Krzysztof Janowicz, and Sathya Prasad. 2014. "Improving Wikipedia-Based Place Name Disambiguation in Short Texts Using Structured Data from Dbpedia." In: *Proceedings of the 8th Workshop on Geographic Information Retrieval*, 8–16.

Ingold, Tim. 2007. *Lines: A Brief History.* London: Routledge.

Jockers, Matthew L. 2013. *Macroanalysis: Digital methods and literary history.* Urbana, Ill.: University of Illinois Press.

Juvan, Marko. 2015. "From Spatial Turn to GIS-Mapping of Literary Cultures." *European Review* 23 (1), 81–96.

Krämer, Sybille. 2010. "Epistemology of the line. Reflections on the diagrammatical mind." In *Studies in Diagrammatology and Diagram Praxis,* edited by Alexander Gerner and Olga Pombo. London: College Publications, 13–38.

Latour, Bruno. 1985. "Les 'vues' de l'esprit." *Culture technique* 14, 4–29.

Latour, Bruno. 1999. "On recalling ANT." *The Sociological Review* 47 (S1): 15–25.

Leetaru, Kaley H. 2012. "Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched Wikipedia." *D-Lib Magazine* 18 (9), 5.

Lima, Manuel. 2011. *Visual Complexity. Mapping Patterns of Information.* New York: Princeton Architectural Press.

Moretti, Franco. 1999. *Atlas of the European novel, 1800–1900.* London: Verso.

Mostern, Ruth, and Elana Gainor. 2013. "Traveling the Silk Road on a Virtual Globe: Pedagogy, Technology and Evaluation for Spatial History." *Digital Humanities Quarterly*, 7 (2).

Nouvel, Damien, Maud Ehrmann, and Sophie Rosset. 2015. *Les entités nommées pour le traitement automatique des langues*. London : ISTE Editions.

Piatti, Barbara, Anne-Kathrin Reuschel, Lorenz Hurni. 2011. "A Literary Atlas of Europe-Analysing the Geography of Fiction with an Interactive Mapping and Visualisation System." In P*roceedings of the 25th International Cartographic Conference*, 3–8.

Pouliquen, Bruno, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fuart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, Clive Best. 2006. "Geocoding multilingual texts: Recognition, disambiguation and visualization." *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, 53–58.

von Richthofen, Ferdinand, 1907. *Tagebücher aus China*, edited by Ernst Thiessen. Berlin: Reimer.

Rossetto, Tania. 2014. "Theorizing maps with literature", *Progress in Human Geography*, 38 (4): 513–530.

Smith, Monica L. 2005. "Networks, territories, and the cartography of ancient states." *Annals of the Association of American Geographers* 95 (4): 832–849.

Scharloth, Joachim, David Eugster and Noah Bubenhofer. 2013. "Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn." In *Linguistische Diskursanalyse: neue Perspektiven*, edited by Dietrich Busse and Wolfgang Teubert. Wiesbaden: Springer VS, 345–380.

Vrandečić, Denny, and Markus Krötzsch. 2014. "Wikidata: A Free Collaborative Knowledge Base." *Communications of the ACM* 57 (10): 78–85.

Warf, Barney, and Santa Arias. 2009. "Introduction: the reinsertion of space into the social sciences and humanities." In *The spatial turn: interdisciplinary perspectives*. London: Routledge.

Wu, Jing. 2009. *The logic of difference in Deleuze and Adorno: positive constructivism vs negative dialectics*, HKU Theses Online (HKUTO). http://dx.doi.org/10.5353/th_b4475834.

Wulfman, Clifford E. 2014. "The Plot of the Plot: Graphs and Visualizations." *The Journal of Modern Periodical Studies*, 5 (1): 94–109.

Lucie Flekova / Florian Stoffel / Iryna Gurevych / Daniel Keim

Content-based Analysis and Visualization of Story Complexity

Abstract Obtaining insights into the style and content characteristics of a novel can provide a benefit to a large number of users. Parents and teachers may be interested in finding appropriate books for children. Booksellers may want to assess the fit of a candidate's artwork into their portfolio or determine the target audience for their promotion activities. Literature scholars might discover particular stylistic similarities in writing patterns of different authors. For all of the above, manually reviewing the textual content of the books is a tedious and time-consuming task which can be achieved only to a limited level of detail. The combination of automated data analysis of literature and computer-based visualization techniques proves to be powerful in giving a quick overview as well as providing details of the visualized data.
In this chapter we define the umbrella term *Story Complexity,* and outline the text data analysis required to describe properties of literature contributing to the numerous aspects of this term. We introduce a multi-faceted *model of story complexity* by addressing numerous aspects of writing, which can pose difficulties to human readers attempting to follow a storyline in fictional literature. Approximations of these aspects are computed automatically with state of the art Natural Language Processing methods. We present the corresponding text data analysis methods, as well as giving examples of how the extracted data can be presented visually, so that the results of the data analysis can be perceived more effectively than by examining the extracted properties of text in a numeric way.

1. Introduction

Gaining an overview of a novel in terms of which aspects contribute to the difficulty, or ease, of following its story can be of benefit for multiple user groups. For example, a teacher may be interested in choosing appropriate reading material for the school class and can do so by considering the level of the language used, as well as the number of characters and parallel storylines, and the complexity and appropriateness of each character's behavior. Alternatively, an e-book merchant may consider acquiring a new series of books to their portfolio and

Published in: Noah Bubenhofer, Marc Kupietz: Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

wants to understand how well, based on previous demand, its writing style and content matches their existing customer base. A literature scholar, on the other hand, may be interested in a contrastive analysis of typical patterns in stories written by different authors or in different literary epochs. A common aspect for all these user scenarios is that the users are not necessarily interested in reading each of the novels in detail, but rather have an information need for an aggregated insight.

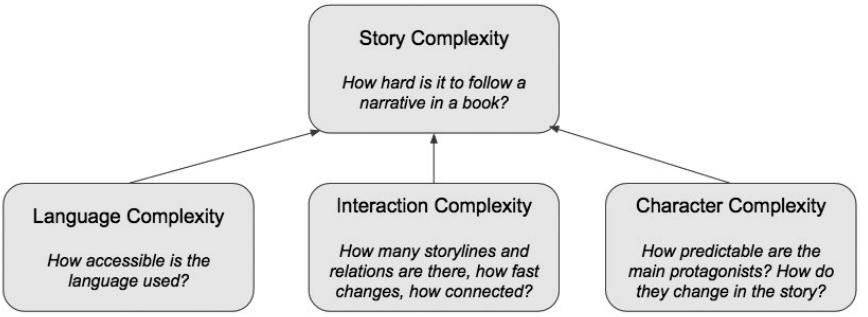


Figure 1: Our understanding of the story complexity and its underlying dimensions.

We approach Story complexity as a covering term that includes numerous aspects which we examine in individual sections of this chapter. In our model, we divide the complexity intuition into three broad areas, as illustrated in Figure 1. The first area is the complexity of the language used. This includes the lexical choice of an author (e.g. if many rare or expert terms are used, increasing the likelihood that the reader won't be familiar with them), the ease of reading on a surface level (e.g. usage of extremely long sentences or words), as well as some syntactic choices such as the dependency types or part-of-speech preferences (e.g. simplification of perception descriptions through interjections).
The second area is the complexity of character interactions. Does the story feature multiple main protagonists? Does it evolve at multiple locations or time periods in parallel? Are the characters in the parallel storylines interconnected in a complex way? How fast do the switches between storylines happen? Is the story linear in time?

The third area of complexity is the behavior of each of the main characters per se. Is the character flat and predictable or does it have both bright and dark sides? Does it develop dynamically as the story progresses? Does it repeatedly show the same emotions?

The overview of which feature types we use to analyze text for each of the three complexity areas is illustrated in Table 1. The list of features should not be treated as exhaustive - there are multiple other options, which can contribute towards understanding the given properties. In this chapter, we aim to provide a broad overview of the possibilities for analyzing a fictional story. A reader can take inspiration from these and expand this framework to meet their own specific analytical needs.

Table 1: Overview of the features we use in each of the complexity areas.

Complexity type	**Natural Language Processing methods used**	**Visualization methods used**
Language complexity	Readability measures, proportion of rare words, proportion of foreign words, density of entities, proportion of part of speech types, occurrence of topics discussed (topic lists, LIWC lexicon), emotions (NRC Emotion Lexicon), sentiment (Stanford Sentiment Analyzer), high-level verb and noun types (WordNet lexicographer files)	paragraph squares bar charts aligned bar charts flow charts
Interaction complexity	Named entity recognition and classification (person, location, organization), entity graph (strength of relationship metrics)	co-occurrence matrices
Character complexity	Speaker identification, emotion development analysis, sentiment development analysis, WordNet lexicographer files for individual character's activities	bar charts flow charts

The content of this chapter is structured as follows. First, we describe previous work related to story complexity in general and previous work related to story visualization (Section 2). Then we discuss each of the three complexity

dimensions, specific work related to these, and how to implement text processing and visualization for each of these dimensions: Section 3 discusses the complexity of the language, Section 4 the complexity of the plot, and Section 5 the complexity of an individual character. In Section 6 we draw conclusions from our work and discuss possible future directions.

2. Related work

This section describes previous work generally related to analyzing story complexity. Specific work related to individual complexity dimensions is discussed at the beginning of each of the following sections.

2.1 Related work in language research and human sciences

In language research, the concept of *narrative complexity* is used to examine the cognitive development of children (Greenhalgh & Strong 2001; Newman & McGregor 2006; Scott & Windsor 2000). Children use narratives to relate events, establish and maintain friendships, and express their thoughts and feelings (McCabe & Bliss 2003). Narratives are defined here as stories about real or imagined events that are constructed by weaving together sentences about situational contexts, characters, actions, motivations, emotions, and outcomes (Gillam & Pearson 2004). Evaluation of children's narratives includes both the overall story grammar, such as characters, setting or events, and detailed aspects, such as pronoun usage or cohesive ties (Petersen et al. 2008). Other measures include the overall story length and artful elaboration (Ukrainetz 2006).

Computer-assisted story analysis for literature has typically occurred at the word level of granularity, suitable for studies of authorial style based on patterns of word use (Burrows, 2004). However, many interesting questions in human sciences lie on a much higher level. Literary scholars explore aspects such as the communal harmony and discord in Russian novels (Lieber 2011), characteristics of fictional portrayals of physicists (Dotson 2009), personality traits of characters in Victorian novels (Johnson 2011), or differences in social interactions in urban and rural English novels (Eagleton 2005). A high-level representation of a story in terms of characters and their interactions is therefore desirable to support such analyses.

2.2 Computational analysis of narratives

Early experiments in representing a story automatically focused on characterizing the plot as a sequence of events. Halpin and Moore (2006) designed an automated system for evaluating a student's ability to rewrite a fictional story. They extracted a chronologically ordered sequence of events in a predicate-argument structure, representing the entities and their actions to predict the plot quality using a supervised machine learning system. They attained up to 56% accuracy in predicting the grades given by the teacher. Chambers and Jurafsky (2009) proposed unsupervised narrative schemas by performing an induction of situation-specific semantic roles and linked them to event chains. Exploiting these schemas, McIntyre and Lapata (2010) created a story generation system. Since it focuses on events, however, it cannot enforce a global notion of how the characters relate to one another, therefore the focus of novelistic plot structure to the level of individual events has been criticized (Elsner 2012).

Recent NLP experiments begin to prioritize the entity-centric models, as proposed by Lehnert (1981). He suggests focusing on the *plot units* as a knowledge structure for representing narrative stories and generating summaries. Plot units are fundamentally different from the story representations that preced them as they focus on the affect states of characters and the tensions between them as the driving force behind interesting and cohesive stories. This theory is followed e.g. by Goyal et al. (2010), who automatically identify the affect states of characters in short fables to represent the story. However, the scheme is very fine-grained and not suitable for larger texts such as novels. Elson et al. (2010) focus on the frequency of dialogue interactions between characters to automatically extract social networks from British 19th-century novels. Kazantseva (2011) suggests an aspect-based summarization model for short fictional stories, focusing on finding the typical attributes of the main characters without revealing the plot of the story. Elsner (2012) uses a set of 19th-century romance novels to identify relationships between the characters. He extracts frequencies of characters in different chapters, and the emotional language with which that character is associated in that chapter, and measures the strength of the relationships between character pairs based on their co-occurrence in a paragraph. He then compares the novels and shows similarities in terms of character emotions and relations. Chambers (2013) improves his previous induction of narrative schemas by learning entity-centric rules (e.g., a victim is likely to be a person). Bamman et al. (2014) and Smith et al. (2013) present latent variable models for unsupervised learning of latent character types in movie plot summaries and in English novels. Iyyer et al (2016) present an unsupervised neural network model for tracking dynamic relationships between fictional characters using latent vectorial features (embeddings) to represent the semantic concepts.

2.3 Related work in story visualization

One of the earliest approaches to applying information visualization techniques to literature based on the contents of documents was developed by Rohrer et al. (1998). Based on a set of principal components extracted from term frequency vectors, a density field is generated and transferred to a three-dimensional visual display using blobs that have directions and are computed from the density field information. The Compus system developed by Fekete and Dufournaud (2000) visualizes lexical and syntactic information from literature, in their case French letters from the 16th century. Tailored to the comparative analysis of different documents, Monroy et al. (2002) propose an information visualization system called ItLv (Interactive Timeline Viewer). Using this tool, the authors demonstrate the visual comparative analysis of books, which ranges from overview like displays down to the page level, thus providing different levels of detail and interactive drill-down capabilities. With the increasing number of input documents to visualize, overview tasks are becoming more important, which is also reflected in corresponding visualization techniques. DeCamp et al. (2005) visualize large document collections using an iconic display built out of the conceptual contents per document. It possible to quickly gain insights into similarities or dissimilarities of the visualized document corpus, as the authors demonstrate for a collection of patents. The work of Chen (2006) also visualizes large collections of documents, but concentrates on revealing patterns as well as connections in the data. Based on node-links diagrams, the author demonstrates the identification and visualization of co-citation networks in scientific literature. While the application is primarily motivated by scientific literature analysis, many of their concepts, in particular, the visualization techniques are also applicable to novels.

Going back to the highest level of detail of literature visualization, the actual text, Weber (2006) introduces a color scheme to visualize text documents that is generated from part of speech information of words. The authors show that the resulting visual display can be used to identify and distinguish different genres of text, as well as insights into the syntactic structure of the documents. Akaishi et al. (2007) propose visualization techniques for the display of narrative structures of a document. They concentrate on the visualization of terms and their relationships, which is demonstrated by the authors to be useful by displaying the structure of the analyzed document in terms of the contained topics.

Keim and Oelke (2007) proposed visualizing a variety of different text features that contain the syntax characteristics, surface properties, and vocabulary metrics. The resulting visualization reveals differences clear enough to characterize and identify authors, and at the same time allows insights into regularities and irregularities of the analyzed book.

A visualization system that reveals common patterns in the analyzed text documents was developed by Don (2007). It allows the exploration of frequent words and n-grams and integrates them in several linked visualizations, which are able to guide users to interesting parts of the explored documents.

Van Ham et al. (2009) developed Phrase Net, a technique to generate overview like visualizations from unstructured text documents. Based on node-link diagrams, it is tailored to relationships, which can be retrieved on the syntactic or lexical level of the input text. The output is suitable for comparing different aspects of the analyzed texts, as well as to give an overview of the contained relationships, for example between characters. The visualization of characters and their relationships is also part of the work by Regan and Becker (2009). Besides that, they also provide insights into the terms connected to characters in order to describe their personality. Noteworthy are also the insights into the design process that produces the amount of text that is included in the visual displays.

Besides the detection of emotions in text, Mohammad (2011) proposes different visualization techniques in order to visualize emotions based on a timeline, as well as using a word cloud to produce words for different emotion categories. The author also proposes techniques for visualizing associated entities with words expressing emotion.

To gain insights into the differences of text documents, Jankowska et al. (2012) use common n-gram classifiers to build up visual signatures. The visualization is compact and therefore suitable for comparing a number of different documents, or parts of documents, by plotting the signatures next to each other, revealing differences in the usage of n-grams over different documents. Continuing with the idea of providing compact signatures or fingerprints, Oelke et al. (2013) demonstrate that matrices of fingerprints can be used to compare character occurrences and co-occurrences, which is suitable for identifying networks of characters as well as their changes over the analyzed text documents.

Weiler et al. (2015) propose visualizations to track different aspects of text data streams, which can also be applied to documents. They identify three properties to track the evolution of topics in documents, being importance, emotion, and context. The authors combine these metrics in a visualization that emulatesa morphing shape over time, effectively communicating topics and their sequential changes.

Besides these feature based visualization techniques, a number of related works are trying to re-create hand-drawn story lines by means of computational methods. Work by Tanahashi et al. (2012, 2015) and Liu et al. (2013) present a set of techniques that provide layout algorithms in order to create a line-based visualization of story progression with respect to characters, events, and locations.

The different entities are also interconnected, which makes important interactions clear in the resulting visualizations.

3. Complexity of language used: analyzing stylistic and content features in book text segments

Each of the following subsections in this and the next two sections is structured in the following way: First, we describe previous research directly related to the specific problem. Next, we explain our methodology for deriving particular features from the story text, which can be helpful for obtaining insights into a story's properties. Finally, we discuss our reasons for implementing a particular visualization of the obtained features, and present cases which enableuser understanding of a story's complexity.

3.1 Expressing the reading ease of a text: readability measures, long and foreign words

Previous work

Traditional readability measures rely on two main features, beingword length and sentence length. They are computed by the average number of characters (or syllables) per word and the average number of words per sentence and are combined with manually determined weights resulting in a grade level as output. The most well known methods of this type are the Flesch–Kincaid Grade Level (Flesch, 1977) formula, which uses the average number of words per sentence and the average number of syllables per word to predict the grade level, the Automatic Readability Index (Smith and Senter, 1967), and the Coleman–Liau Index (Coleman and Liau, 1975). However, they have also been subject to criticism as they only capture surface characteristics of the text and can be misleading (DuBay 2004).

More recently, supervised learning algorithms have been used to automatically combine several text properties extracted from training data and used to associate them with the corresponding readability class. Feng et al. (2010) show that the density of entities (nouns and proper nouns) introduced in a text corresponds to a higher working memory burden for the reader, thus contributing to higher readability level. Pitler and Nenkova (2008) explore discourse level features from the Penn Discourse Treebank (Prasad et al. 2008) and report on their usefulness in predicting text readability.

Our features

Our framework computes the readability measures (Flesch–Kincaid Grade Level, Automatic Readability Index, Coleman–Liau Index) for each paragraph, as well as the ratios of each part of speech type and a proportion of named entities and foreign words in the text, using the OpenNLP (Morton et al. 2005) Tagger and Name Finder.

Visualization

The input data for the readability visualization is computed based on the paragraphs of the analyzed books, since the readability metrics mostly refer to a number of sentences or words, instead of single sentences. For each of the paragraphs, the information from the corresponding chapter is given.

The visualization of readability metrics for a given book is designed with the following goals in mind:

1. Provide an overview of the changes in readability metrics corresponding to the chapters or paragraphs of a book in a compact way
2. Keep the structure of the book visible, so that interesting values can be correlated to the corresponding unit of text.
3. Clearly indicate areas where the analyzed text is easy or hard to read

Goal one results in two requirements for the general construction of the visualization. First, paragraphs, as well as chapters, have to be indicated visually so that they can be easily seen. Second, the visual design has to be as compact as possible to enable effective communication of the readability metrics, while at the same time providing an overview of as much text (and data) as possible. Visually, these requirements have been met by a matrix-like visual design. Each paragraph is represented by a cell. A chapter, which is composed of a number of paragraphs, is represented by a group of cells. Compact representation also imposes requirements on the data preprocessing, which is done before the visualization is created. If all paragraphs are visualized with a cell, the resulting width and height of the created graphical depiction would be too large to satisfy the compactness constraint. As a consequence, the data is aggregated before it is visualized. The aggregation is based on a fixed window size, of which the arithmetic mean of the contained readability metrics is computed. In this way, outliers in both hard and easy to read directions should stay visible, but at the same time the overall data will reflect the properties of the aggregated paragraphs. In

addition, the windowing follows the logical structure of the text, which is given by chapters. If a window contains a chapter, the window size is reduced so that only paragraphs from the same chapter are aggregated, and finally, the aggregation window is enlarged again and moved to the boundary of the next chapter.

Goal two has in one main requirement, namely that the paragraphs and chapters can be easily perceived as such. Based upon the previous requirements, the structure needs to be resembled by positioning and aligningthe cells to represent a paragraph. One condition of representing the structure is determined by the fact that the cells that are referring to consecutive paragraphs should be placed next to each other. This condition is fulfilled by aligning the paragraphs on a common baseline in the order in which they appear in the single chapters. A visual overflow per row, which could happen when the width of cells exceeds the width of the visualization space, is solved by introducing a line break so that multiple rows can represent the same chapter. Having made sure that the paragraph alignment resembles the structure of the book, the final step in visually representing the structure of the book chapter is to visually indicate the affinity of paragraphs to the corresponding chapter. This is done by introducing a margin between the rows that refer to different chapters, and which is large enough to be perceived easily as the border between chapters. The result is a layout where rows of cells indicate paragraphs, line breaks are used to mitigate the overflow of rows being too wide for the visualization space, and rows referring to different chapters are separated clearly by a wide margin between them.

The requirement resulting from the last desired property, the visual indication of the readability metric, refers directly to the representation of cells. The displayed property is presented per paragraph and is represented by cells, the most prominent visual property of these being their area and color, which are used for the indication of the readability metric. Cells of paragraphs that are, according to the computed readability metric, easy to read, are filled with a light reddish color. In contrast, cells with low readability are filled with a darker red color. Readability values in between are mapped to a number of bins, each represented by a color interpolated between light red and the dark red tone. The result is a color map that assigns readability values to a color starting at light red (easy to read) todark red (hard to read), as well as the colors in between.

In the example given in Figure 2, Flesh Reading Ease is computed from the book *Harry Potter and the Sorcerer's Stone* by J. K. Rowling and visualized with the described technique. The overall impression is quite mixed and shows, except for some single cells (fifth and eleventh row), a mixed picture of the readability score.

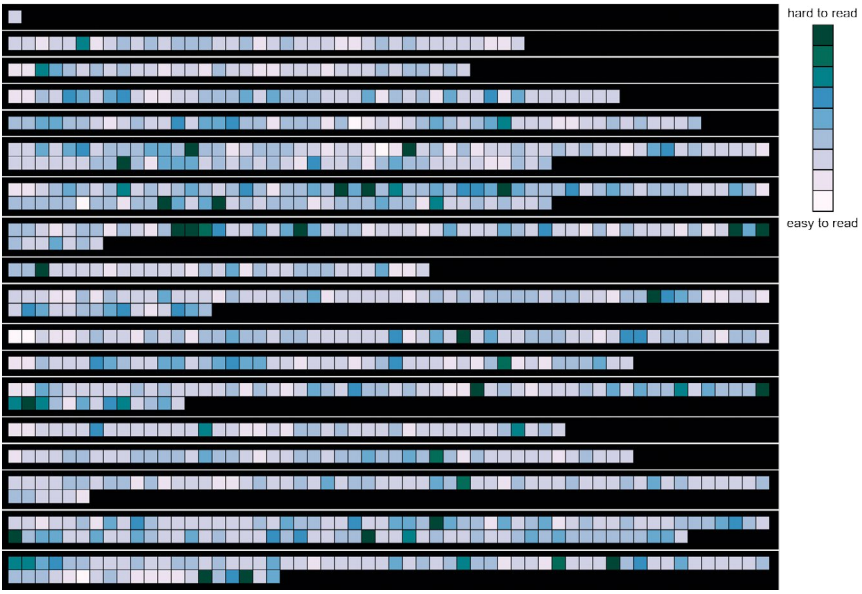


Figure 2: Flesch Reading Ease score per paragraph of *Harry Potter and the Sorcerer's Stone*.

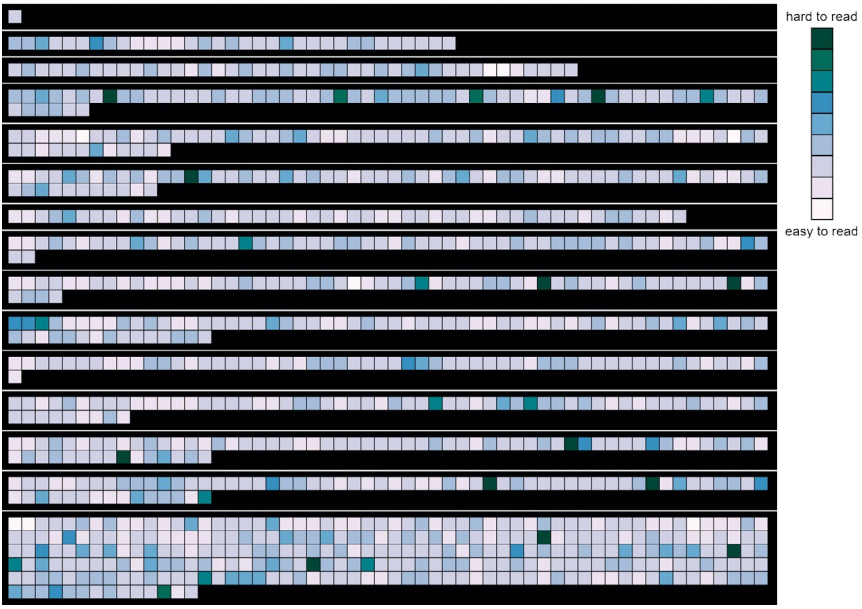


Figure 3: Flesch Reading Ease score per paragraph of *Harry Potter and the Chamber of Secrets.*

Comparing the first two volumes of the Harry Potter series with respect to Flesh Reading Ease, reveals that they are quite similar (Figure 2 and Figure 3). In both books there are very short passages that are hard to read, while the overall impression remains mixed and they are finally judged clearly as easy to read.

3.2 Analyzing the style and content of text

Previous work

Analysis of the style of books has been explored in the areas of authorship attribution (Zhao 2007), author profiling (Rangel et al. 2013) and computational stylometry (Daelemans 2013), mainly with the aim of differentiating between authors and author groups. Ashok et al. (2013) additionally show that stylistic features can predict the book success with high accuracy (84%). They find that less successful books rely on verbs that are explicitly descriptive of actions and emotions (e.g., "wanted", "took", "promised", "cried", "cheered", etc.), while more successful books favor verbs that describe thought-processing (e.g., "recognized", "remembered"). Additionally, less successful books rely more on topical words that could be almost cliche, e.g., "love", typical locations, and involve more extreme (e.g., "breathless") and negative words (e.g., "risk"). They also report that the prepositions, nouns, pronouns, determiners and adjectives are predictive of highly successful books whereas less successful books are characterized by the higher percentage of verbs, adverbs, and foreign words.

Our features

We measure numerous aspects that have proven useful in previous work that captures the style of an author. The following features are based on the OpenNLP part-of-speech tagger (Morton et al. 2005), using the maximum entropy model to annotate tokens with the Penn Treebank POS tagset: frequency of adjectives and adverbs in their comparative and superlative form; the frequency of personal and possessive pronouns, and frequency of exclamation and question marks. Additionally, we measure the contextuality score, which is considered an approximation of how formal or casual a given text is (Heylighen 2002), with the knowledge that some parts of speech types contribute to a more casual style (such as pronouns or adjectives) while others occur more often in a more formal text (such as nouns and determiners). The contextuality score reaches values between 0 and 100 and is calculated as follows:

$$\text{contextuality} = (\text{nouns} + \text{adjectives} + \text{prepositions} + \text{determiners} - \text{pronouns} - \text{verbs} - \text{adverbs} - \text{interjections} + 100)/2$$

where each part of speech type is expressed as its relative frequency compared to all words.

Additional insights into the overall characteristics of a given text can be obtained by exploring the topics that occur in each chapter. Sociolinguists commonly use the Linguistic Inquiry and Word Count (LIWC) lexicons for this purpose (Pennebaker 2003), which we also employ here. LIWC is unique in the sense that it provides not only a set of basic topical categories (such as family, money, friends, work) but also expressions of cognitive processes (insight, tentativeness, uncertainty...) or inner drives (achievement, inclusion ...). There are 69 categories in total. In addition, we use topical word lists from www.enchantedlearning.com, which enrich our set with additional categories such as school, computers, cars, politics or swear words.

Word lists such as those mentioned above are often criticized for being based only on the written form of the occurring expression, without taking into account additional information about its eventual polysemy or morphological variations. Therefore, we attempt to obtain more precise information about the categories of individual words using WordNet (Miller, 1996) semantic categories, sometimes also called lexicographer files or supersenses (Ciaramita and Altun 2006).

Wordnet supersenses are assigned to verbs and nouns on a WordNet synset level, i.e., taking into account the distinction between different senses of the same word. There are 26 categories for nouns, such as animal, person, artifact or process, and 15 categories for verbs, such as communication, motion, cognition or emotion. We retrieve the supersense for each verb or noun in the text by using its lemma and part of speech tag and mapping it to its most frequent WordNet sense.

The visualization of different stylistic features, which can be related, such as the LIWC dictionary words or WordNet senses, is mainly driven by the need to gain an impression abouth whether they occur, and if so, in what relation they stand with each other. To communicate the actual values of features, it must be possible to follow the values over the progression of a book, which is the first property that a visualization needs to have for this kind of data and requirements. Second, to be able to judge the domain of a set of features, as well as a region of concrete values, a comparison must be possible.

A bar chart is capable of adequately fulfilling the requirements resulting from the first property. The different bars are make clear that the displayed data is not coming from a continuously occurring feature, which is being measured at discrete stages. Additionally, the area of the bar, which can be filled with a color,

Table 2: Overview of the lexicons used in our experiments

Lexicon name	Reference	No. of words	No. of categories	Example of categories and content
Linguistic Inquiry and Word Count	Pennebaker 2003	10,555	64	Feeling: hard, press, warm Certainty: Fact, confidence, always
NRC Emotion	Mohammad 2011	8,265	8	Surprise: cheer, inspired, unexpected Joy: amuse, elegant, happily
NRC Sentiment	Mohammad et al. 2013	5,636	2	Positive: mighty, prestige, unconstraint
Hu & Liu Sentiment	Hu and Liu 2004	6,789	2	Negative: annoy, mistaken, worse
In-house emotion list	Wanner et al. 2011	416	16	Anxiety: cautious, fearful, nervous
In-house topic list	Enchanted learning.com	4,735	24	Politics: choice, quorum, voter School: math
WordNet lexicographer files – noun supersenses	Miller 1996	117,798	26	Animal: fish, cat Body: hand, leg Person: teacher
WordNet lexicographer files – verb supersenses	Miller 1996	11,529	15	Motion: fly, walk, swim Communication: talk, scream

can assist in the perception of the feature value. With each bar we represent a number of sentences, which are aggregated with respect to the logical borders of books, namely the chapters. The same reasoning as before also holds here, meaning that the arithmetic mean is the choice of aggregation method for the numeric feature values (since it is sensitive to outliers), which increases the possibility that the aggregate will have a value near to the outliers, as well as providing an effective way of preserving the feature values of the non-outliers. The order of the bars preserves the sequence of the represented aggregates, which allows conclusions based on the position of a bar with respect to the book, e.g. in the beginning, in the first half, or near the end. Also, the distance from neighboring bars is easy to perceive, because this can be done by comparing the different heights, and allows tracking of feature values, trend spotting, as well as tracking outliers during the progression of the book in question.

The second property effectively opens up the design in such a way that it is possible to perform the aforementioned analysis tasks for a set of features. Having the initial design based on a bar chart, a stacked bar chart adds comparison capabilities. However, the direct comparison between two or more features is negatively affected by a classical stacked bar chart, where the single bars are placed on top of each other in a single instance, leaving only one baseline in the chart, which grounds the perception of the lowest part of the bar in the chart. Because of this perception issue, each feature is still represented by a single bar chart instance. To make the charts comparable, they are placed on top of each other, and their scales are normalized accordingly. Their start and end, as well as the window size (resulting in the number of bars), are aligned. The result preserves a baseline for each represented feature, as well as allowing quick, but not exact comparisons of the feature values. An exact comparison is not considered a firm requirement, because of the different origins of the feature set as well as the language use, which may be fundamentally different for the measured properties of the text, already makes exact comparisons hard to interpret. To support navigation in the bar chart, a highlighter covering all charts follows the mouse.

Similar to the previous visualization, the data is aggregated in a window fashion which respects the logical borders of a book, which are determined by chapters. The window can be adjusted in order to give more detail, or aggregate to a high level, so that for very large window sizes the aggregated data corresponds to a whole chapter, while it is still possible to transfer to a high level of detail.

In Figure 4, the LIWC common verb classes “Future Tense”, “Past Tense”, and “Present Tense” are shown. The bar heights are globally normalized, which means they can be compared among themselves. From this viewpoint, it becomes clear that the first Harry Potter novel is written in the past tense, and there are only rare references to either the future or the present tense. Having a story line

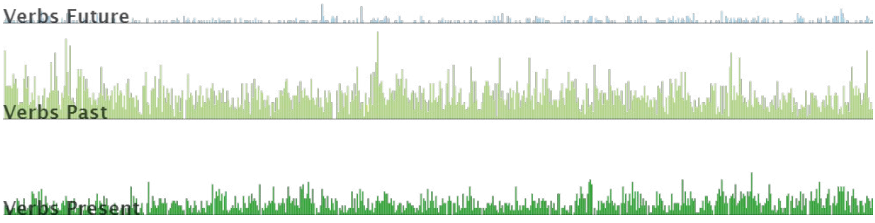


Figure 4: Visualization of the frequency of future, past and present tense in *Harry Potter and the Sorcerer’s Stone.*

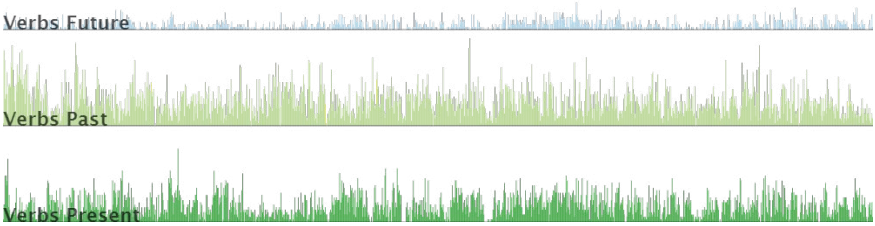


Figure 5: Visualization of the frequency of future, past and present tense in *The Lord of the Rings – The Fellowship of the Ring*.

jumping frequently between tenses can be a sign of a quite complex story with forward and backward references, but this is not the case here. For all other Harry Potter novels, these charts look similar. Comparing them with fictional literature from other authors, for example with the first volume of "Lord of the Rings" by J. R. R. Tolkien (Figure 5), it becomes apparent that a similar picture can be expected for other novels written in the past tense.

The same visualization technique can be used to get an overview of the type of actions in a book. Appropriate for this are the extracted WordNet supersenses of verbs. In Figure 6, a subset of these features and their occurrences in *Harry Potter and the Sorcerer's Stone* is displayed. The feature values are globally normalized, which means that the height of the bars can be compared with each other. While analyzing this visualization, it becomes clear that two of the selected categories are dominating the verbs used, which are "telling, asking, ordering, singing", as well as "walking, flying, swimming". In contrast, verbs from other categories such as "fighting" or "eating and drinking" are only rarely used in the novel. When comparing this with the same data of the last book of the *Lord of the Rings* series (see Figure 7), three observations can be made. At first, there seems to be a much smaller focus on actions in the context of "telling, asking, ordering, singing" in *Lord of the Rings*. The same is true for verbs from the category "touching, hitting, tying, digging", but to a lesser extent (Figure 6).

The third observation seems toprovide the biggest difference between the two books, being is the increased occurence of words from the "eating and drinking" category at the end of the *Lord of the Rings*. This is due to the coronation of the character Aragorn, where the festivities are described. Similar passages are missing from the *Harry Potter* novel, which can be clearly seen when comparing Figure 6 and Figure 7.

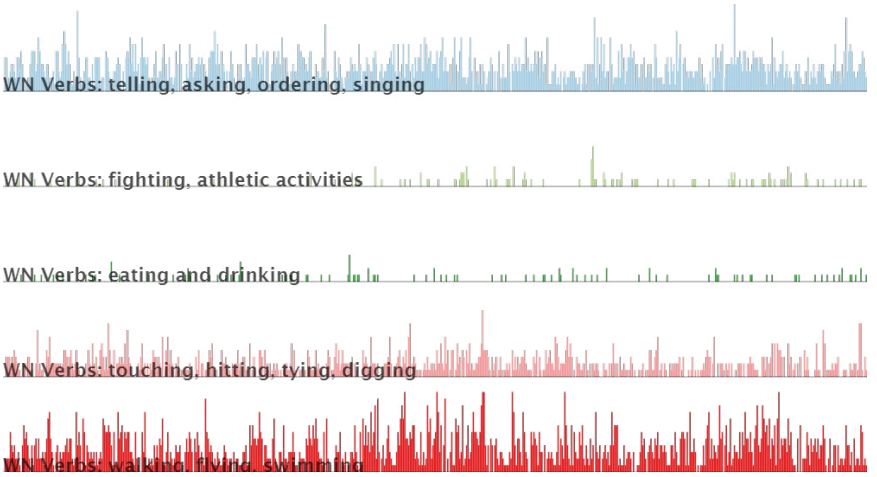


Figure 6: Visualization of a selection of Wordnet verb supersenses in *Harry Potter and the Sorcerer's Stone*.

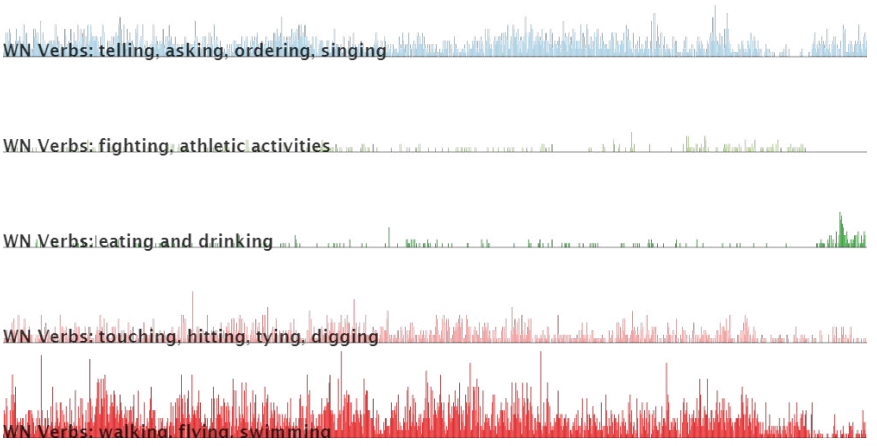


Figure 7: Visualization of a selection of Wordnet verb supersenses in *The Lord of the Rings – The Return of the King.*

3.3 Emotions and sentiment

Previous work

Ovesdotter Alm et al. (2005) set up a system to automatically predict the six basic emotions (anger, disgust, fear, happiness, sadness, surprise) in 22 children's fairy tales on a sentence level. They achieve an accuracy of 63%, and point out that simple bag-of-words models are prone to errors in texts enriched with frequent figurative expressions. Volkova et al. (2010) experiment with the German texts of Brothers Grimm fairy tales. They investigate how emotions are expressed in these stories and how people associate emotions with certain text fragments of these fairy tales. The authors define several positive and negative emotional categories, and then several annotators manually annotate the text passages that convey these emotions. They find that in most texts, positive emotions are expressed more frequently than negative ones. The authors observe a reasonably high inter-annotator agreement for emotions in thetext. Mohammad (2011) uses his *NRC Emotion Lexicon* to explore emotions displayed in fairy tales and novels. He explores how the frequency of words associated with certain emotions differs for different types of literary text and how they change through the course of a narrative. Moreover, he compares distributions of emotional words in novels and fairy tales, finding that fairy tales tend to have higher emotional density.

Our features

First we measured the positive and negative sentiment and the six basic emotions (happiness, sadness, fear, anger, surprise and disgust) using in-house word lexicons inspired by www.psychpage.com. Additional word lists based on the same website measure more fine-grained emotional states such, as anxiety, confusion, helplessness or love.

Sentiment lexicons, while widely used, have been a subject of criticism for capturing only very explicit expressions and, more importantly, out of their syntactic and semantic context. Therefore sentences such as "This movie was actually neither that funny nor super witty" would be incorrectly classified as positive based on the sum of its positive and negative expressions (2 + 0). This problem can be overcome by studying the compositional grammatical structures of the sentences. This has been done in the Stanford Sentiment Analyzer (Socher et al. 2013), using recursive neural tensor networks. In their system, the above-mentioned example is classified correctly as negative. We employ their trained model in our system as well, to predict sentiment score on a 5-point scale on sentence level.

Visualization

The visualization of emotions and sentiment is based on similar reasoning as the visual display of stylistic features. The corresponding visualization should effectively communicate the value of the corresponding feature, as well as its development, and allow comparative findings and insights. Keeping the same visual design as for stylistic features is motivated by the observation that the variety of emotion- and sentiment-related contexts contributes to the perceived degree of story complexity.

For the first property, a bar chart was used to effectively communicate feature values and their development. The feature values are double encoded in the bars by using their height as an indicator of the represented numeric value, as well as the area that is colored uniquely per emotion feature. It is possible to follow the different values, as well as to perceive changes over the progression of a book.

The comparison of different emotion features is enabled by stacking the bars of each feature in a single plot. This different approach is chosen, because in contrast to the stylistic features, the emotional context of a text passage represented by a single bar can be seen as a limited space, whereas a single emotion can dominate the perception of a chapter, for example, if words from a negative emotion context occurr more frequently than any other emotions. This is taken into account by effectively limiting the visual space to the height of one bar chart, and the different emotions, which are about to be compared, are shown as stacked bars in that limited height. Together with the colored area, representing the feature value, this technique ensures that any dominating emotion, and its assigned color, also dominates the perception of the limited area per-visualized text passage. This desirable property sacrifices the exact perception of the feature values and their comparison, but at the same time allows the emotional context and any dominating emotion to be followed effectively.

The visualized data is computed by a window over the sentences that reflects chapter borders and uses the arithmetic mean for feature value aggregation.

Inspecting Figure 8, which depicts the word counts of six different emotion word dictionaries (happiness, sadness, fear, anger, surprise, disgust), reveals that for the first Harry Potter novel, the number of sadness and anger words dominate the emotion categories as computed with the available in-house sentiment word dictionaries. There are outliers of the fear emotion in the beginning and the last third of the book, which are locally quite restricted and therefore describe a drastic, but limited change of the emotional tone of the book. Similar to the stylistic features, this general impression does not change much for further books in the Harry Potter series, but the number of outliers of a specific emotion increases (compare Figure 8 with Figure 9).

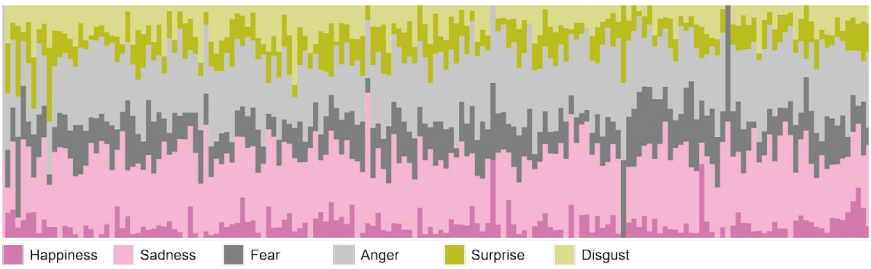


Figure 8: Emotion dictionary word counts from *Harry Potter and the Sorcerer's Stone.*

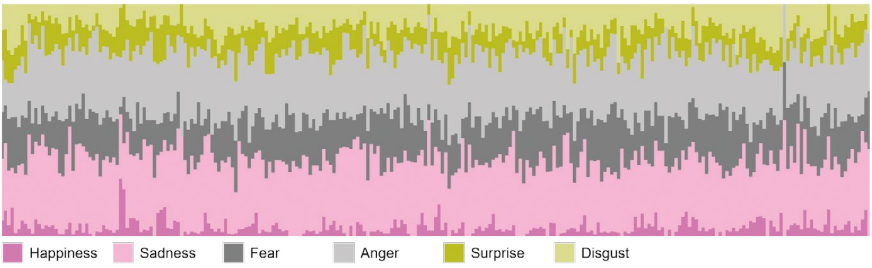


Figure 9: Emotion dictionary word counts of the last Harry Potter novel, *Harry Potter and the Deathly Hallows.*

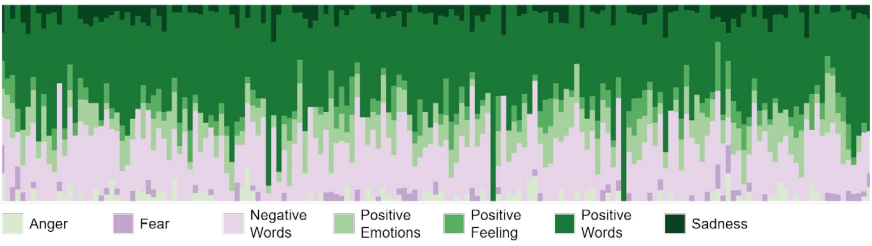


Figure 10: Emotion words counts from enchantedlearning.com for the first Harry Potter novel.

Comparing this with a higher-level abstraction of emotion words, as can be seen in Figure 10, it becomes clear that the aforementioned observation is only true in the analyzed context.

Here, it can be seen that the number of words representing positive emotions, positive feelings, and positive words dominate the chosen emotional context, which is comprised of anger, fear, negative words, positive emotions, positive feelings, positive words, and sadness.

Depending on the chosen words, the visualization will differ and present a different picture of the emotional context.

4. Complexity of a plot: identifying characters in a literary text and relations between them

Beyond linguistic complexity, an important aspect influencing the ability to understand a story is the number of characters appearing in it and the complexity of their interactions. How many literary characters appear in a novel? Despite the seeming simplicity of the question, precisely identifying which characters appear in a story remains a difficult problem in literary and narrative analysis.

Previous work

Characters form the core of many computational analyses, from inferring prototypical character types (Bamman et al. 2014) to identifying the structure of social networks in literature (Elson et al. 2010; Lee and Yeung 2012; Agarwal et al. 2013; Ardanuy and Sporleder 2014; Jayannavar et al. 2015).

Our methodology

In order to identify individual characters in fictional literature, we have developed two character identification methods. A fast one to identify as many mentions of a character in the text as possible, and a precise one to assign direct speech to a particular speaker in the book. The second method is described in section 5. The first method, used here, is a two-phase process, where first a set of candidates is generated using several predefined rules, while in the second step the whole document to be analyzed is scanned for occurrences of candidates, in order to ensure no occurrences are overlooked. The rules contain heuristics based on common salutations (extracted from the English Wikipedia using DBpedia queries for Women's and Men's social titles, Military ranks, Academic

ranks, and Political titles). In addition, grammar based rules are in place that hint at a possible character based on specific parts of speech combinations, such as the identification of possessive constructions indicating a character. These include detection of possessive pronoun constructions (Figure 11, lines six and eight), as well as verbs in the 3rd person singular (Figure n, line ten) usually having a character in context.

Finally, the character candidate tokens are followed to capture full names and titles corresponding to characters (lines 13 to 18, Figure 11). This permits detection of characters such as "Lord Voldemort" or "Professor Dumbledore", while state of the art named entity detection produces results with the titles and salutations typically missing, e.g. with the Stanford Named Entity Recognizer (Finkel et al. 2005). The full attribution of characters, including any titles or salutations, properly reflects the books contents and allows different kinds of references to the same characters to be captured. We performed several experiments with the final set of heuristics and a number of different books written by different authors, to clarify if postprocessing to resolve coreferences is required. Based on the exemplary results and the fact that we kept the set of heuristics small, special treatment of coreferences has been omitted, as we found only very few, if any (below ten) false positives. Since we also identify other types of names, besides animated named entitiessuch as locations, the same visualization can be used

```
   Input  : A sequence of tokens T, a list of salutations and titles S
   Output: sequences T' ∈ T that are likely to denote a character
 1 T' ← {}
 2 for i ← 0 to |T| do
 3 |   t ← T[i], t' ← T[i − 1], t'' ← T[i + 1]
 4 |   if t ∈ S then
 5 |   |   push(T', T[i, follow_character(T, i)])
 6 |   else if POS(t) = NNP and POS(t'') = IN and ends(t, 's) then
 7 |   |   push(T', T[i, follow_character(T, i)])
 8 |   else if is_noun(t) and is_noun(t'') and ends(t, s') then
 9 |   |   push(T', T[i, follow_character(T, i)])
10 |   else if is_uppercase(t'[0]) and POS(t) = VBZ then
11 |   |   push(T', T[i, follow_character(T, i)])
12 return T'

13 define follow_character(T, o)
14 |   for i ← o to |T| do
15 |   |   t ← T[i]
16 |   |   if !is_uppercase(T[i][0]) and !is_hyphen(T[i][0]) then
17 |   |   |   break
18 |   return o
```

Figure 11: Heuristics to detect character names.

to explore the relation between characters and locations in the book. When the secondary characters mostly stay in one location, the book can be considered less complex.

Visualization

Thinking of co-occurrences as a node-link structure is obvious. Nodes represent the characters, and for each co-occurrence a link can be added to the graph, or the weight of an existing link between the two co-occurring character nodes can be increased. Visualizing character co-occurrences directly by means of a graph imposes huge perceptional challenges to the reader, because in fictional literature it can be expected that characters frequently co-occurr with each other, for example, because of interactions. A graph constructed as mentioned before can be visualized by utilizing a number of different graph layout techniques, which all optimize certain criteria, such as keeping the number of edge crossings low, imposing a high degree of visual symmetry, or keeping the average edge length below a certain threshold. For fictional literature it may be expected that the visualization of a graph, based on a suitable graph layout technique, could suffer from an overplotting of the edges or the nodes, making it difficult to perceive frequent co-occurrences or patterns in the co-occurrences.

Using adjaceny matrices is a technique that utilizes the same kind of data, i.e. node-link structures, where the nodes represent characters and links are used to indicate co-occurrences, and where the visualization is well known to scale for large numbers of rows and columns (representing characters), while also supporting the perception of visual patterns. In these matrices, the rows and columns represent nodes from the graphand the cells are used to map the connections of nodes in the graph. There is intentionally no overlap, which preserves any visual patterns and at the same time allows networks in the data to be identified.

To depict character co-occurrences with adjacency matrices (see Figure 12), an undirected, weighted graph is created from the co-occurrence data. Each node represents a character, for each co-occurrence of the character an edge is added to the data structure with the weight of one. In case an edge between two characters already exists, its weight is increased by one. Having constructed the graph, a subgraph representing the co-occurrences between the top n most occurring characters is extracted. The visualization of the adjacency matrix from this subgraph represents each character by a single row and a single column, which makes the result symmetric. For each character c, the co-occurrences with the other characters d are examined, and the corresponding cell in the row belonging to c and columns of characters from d are assigned a color on a color map ranging from light blue (few co-occurrences) to a dark blue (most co-occurrences),

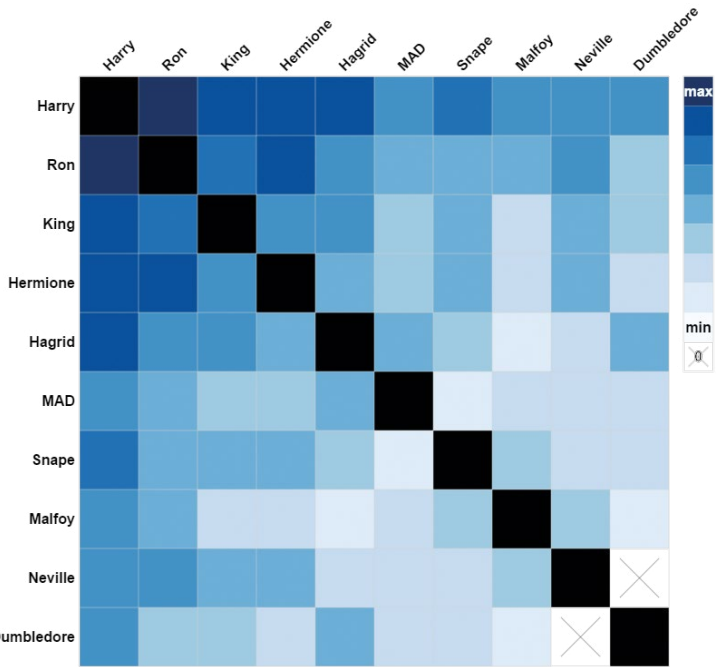


Figure 12: Co-occurrence of the ten most frequently occurring characters in Harry Potter and the Sorcerer's Stone.

that is used to fill the cell area. If a character does not interact with another one, which is part of the adjacency matrix, the visual indication is a cross on white ground. The rows and columns are ordered by the absolute occurrences of characters, ensuring that frequent characters and their co-occurrences are visible together starting at the top left of the matrix visualization. To indicate the symmetry of the matrix, the diagonal, which refers to co-occurrences of characters with themselves, the corresponding cells are marked in black.

Figure 13 shows the 15 characters that occur most often in the first Harry Potter novel. From top left to top right, or top left to bottom left, the characters are ordered descendingly according to their occurrences over the entire book. It can be seen that Harry (first row, first column), occurs together quite frequently with every single of the remaining top ten characters, as is indicated by the dark blue color of the cells. It can also be seen that Dumbledore does not occur at all together with Neville, as is indicated by the black x on white ground of the corresponding cell. Since the matrix is symmetric, the diagonal would indicate co-occurrences of the character with themselves, which is encoded with a cell marked in black. Figure 14 visualizes the same information of the first book of the Lord of the Rings series.

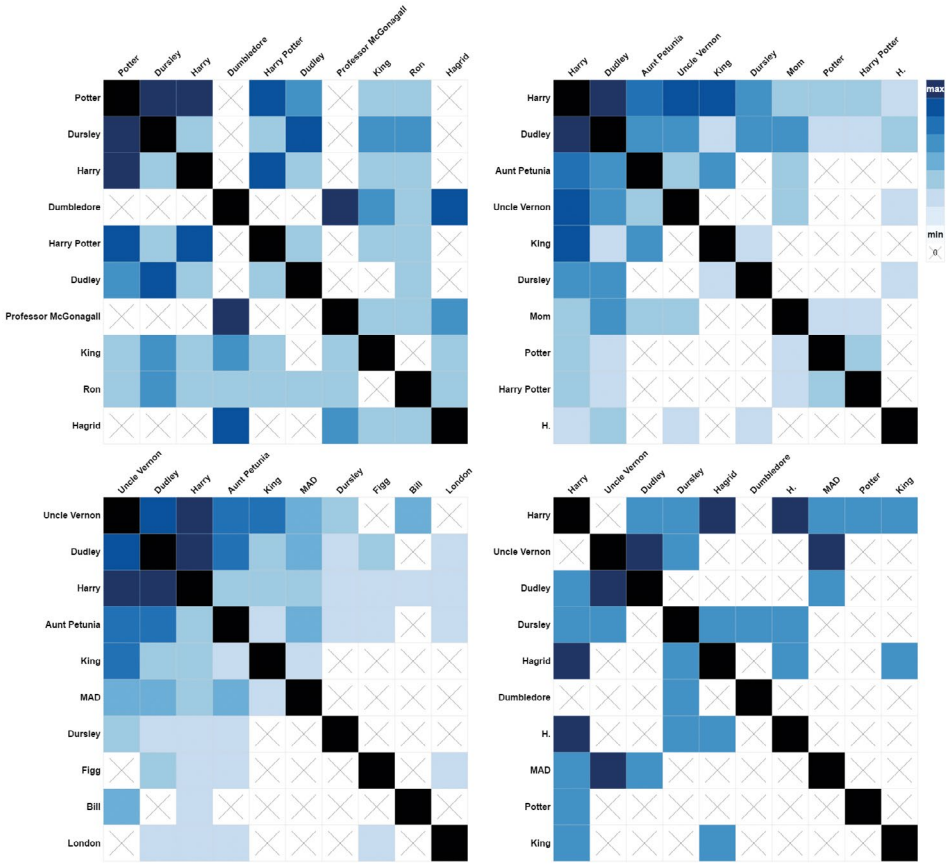


Figure 13: Visualization of the top ten character co-occurrences in the first four chapters of *Harry Potter and the Sorcerer's Stone*.

This matrix-based visualization concept can be applied to a much higher level of detail. In Figure 14, the first four chapters of the first volume from the Lord of the Rings series are shown, the top left matrix depicts co-occurrences of chapter one, top right chapter two, bottom left chapter three, and bottom right chapter four. Simply by examining the names of the top occurring characters per chapter, it can be seen that quite drastic changes in the involved characters also imply a change in the story line. It can also be observed that the number of co-occurring characters in Lord of the Rings is quite different to the Harry Potter novel visualized in Figure 13. In the latter almost every top occurring character has interactions with the others, while in the former this is not the case, as we can observe that certain characters, such as Gandalf, Gollum, or Took have only a limited number of co-occurring characters. For both Harry Potter

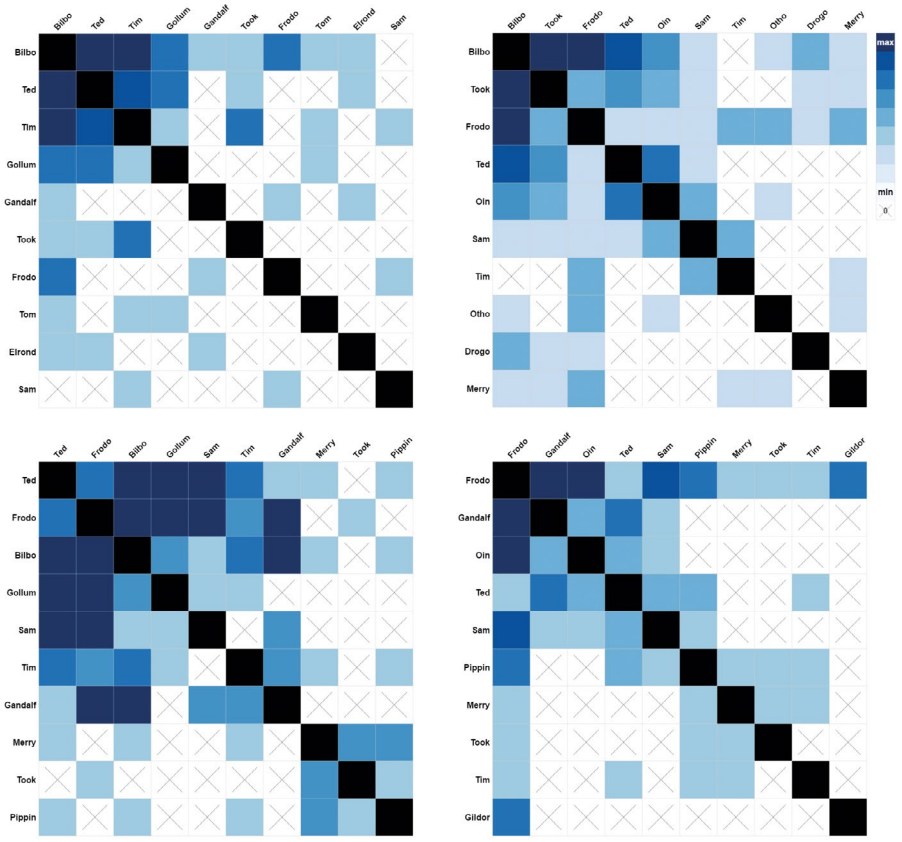


Figure 14: Visualization of the top ten character co-occurrences in the first four chapter of *Lord of the Rings – The Fellowship of the Ring*.

and Lord of the Rings, a group of characters is clearly in focus per chapter. On the top left there is usually a group of four of five characters, which co-occur quite frequently (dark blue) with themselves. In some cases, in particular the main character, they are part of the pattern and co-occur with almost every character of the top ten in the visualized chapter.

5. Complexity of the characters: analyzing the book from an individual character's perspective

5.1 Identifying direct speech and automatically assigning speakers

According to character-driven (as opposed to plot-driven) literature theory, the protagonists in a novel are the central aspect of the story. Characters possess multiple layers of personal traits which are exposed as the story develops. Our aim in this work is to gain an insight into each character's complexity by analyzing the concepts and style in their direct speech produced. In order to explore what the characters are discussing and in which manner, we first need to identify the direct speech segments in the text and assign them to the appropriate speaker. This is a difficult problem since direct speech utterances in modern literature are rarely in the traditional format, such as *"direct speech," said John.*

Related work

Numerous publications study how quotes can be attributed to a speaker, however, only a few of them deal with literary texts. Elson et al. (2010) propose a dialogue attribution method specifically designed for novels. They extract a feature vector for each pair, consisting of a candidate speaker and a quote. They use such information as the distance between the candidate and the quote, and the number of appearances of the candidate in the book. Several classifiers are then used to discriminate between vectors that belong to speakers of a given quote and other characters. Using a corpus of 19th and early 20th-century fiction by six authors, they achieve an overall accuracy rating of 83%. O'Keefe et al. (2012) conduct further experiments using the same corpora. In contrast to Elson et al., they perform the attribution without the use of annotated data. The best result of 53.3% is obtained by a simple rule-based method. He at al. (2013) present another supervised dialogue attribution method, using an unsupervised actor-topic model, which is used to predict likely speakers based on topic distribution of relevant text. Accuracy of between 80% and 86% is achieved.

Elson et al. (2010) further use the dialogue method to construct social networks for book characters. The network is represented as an undirected graph with nodes representing characters and weighted edges describing their relationships. The weight of an edge between a pair of character nodes is set according to the total word length of quotes spoken by one of the characters, in cases where there is a quote by the other character within 300 words. The networks thus extracted are used by the authors to compare the degree of connectedness and structure of the network for books with different settings, providing results that refute popular hypotheses from literary studies. Agarwal et al. (2012)

manually extracted a social network in Alice in Wonderland using social events. Two specific kinds of social events were used: *interactions*, in which both parties are aware of the event, and *observations*, in which only one party is aware of the event. Vala et al. (2015) propose a new character identification technique, bootstrapping characters from seeds of names found with the Stanford Named Entity Recognizer (Finkel et al. 2005) and Stanford coreference resolver (Recasens et al. 2013), or entities denoted as *animated* in WordNet. They achieve an F-score of up to 75%.

Our methodology

The most challenging task in building the direct speech data set is assigning direct speech utterances to the correct speaker. We benefit from the epub format of the e-books, which defines a paragraph structure in such a way that only the indirect speech chunk immediately surrounding the direct speech is considered:

> John turned to Harry. "Let's go," he said.

Given the large amount of text available in the books, we focus on precision rather than coverage and discard all utterances with no explicit speaker (i.e., 30-70% of the utterances, dependent on the book), as the performance of current systems on such utterance types is still fairly low (O'Keefe et al. 2012; He et al. 2013; Iosif and Mishra 2014). Conventional coreference resolution systems, which we tried, did not perform well on this type of data and were therefore not used in the final setup. We adapt the Stanford Named Entity Recognizer (Finkel et al. 2005) to consider titles (Mr., Mrs., Sir...) as a part of the name and to treat the first person "I" as a named entity. However, identifying only the named entity PERSON in this way is not sufficient. In our evaluation sample consisting of a Game of Thrones book "Pride and Prejudice" (the former annotated by us, the latter by He et al. (2013)), 20% of utterances with explicitly named speaker were not recognized. Of those correctly identified as a Person in the adjacent indirect speech, 17% were not the speakers. Therefore, we implemented a custom heuristics (illustrated in Figure 15), which additionally benefits from the WordNet semantic classes of verbs, enhancing speaker detection by recognizing the nouns. With this method, we retrieve 89% of known speakers, of which 92% are assigned correctly. Retrieved names are grouped based on string overlap (e.g. Ser Jaime and Jaime Lannister), excluding the match on the last name, and corrected for non-obvious groupings (such as Margaret and Peggy).

To quickly get an insight into the extracted direct speech of characters, word clouds can be used. They can enable a quick overview of the words used, while

Algorithm 1 Assign speaker

1: *nsubj* ← subjects in adjacent indirect speech
2: **if** *count*(*nsubj*(*i*) = PERSON) = 1 **then** *speaker* ← *nsubj*
3: **else if** *count*(*nsubj*(*i*) = PERSON) ≥ 1 **then** *speaker* ← the nearest one to *directSpeech*
4: **else if** *directSpeech* preceded by VERB.COMMUNICATION **then** *speaker* ← the preceding noun(s)
5: **else if** *directSpeech* followed by VERB.COMMUNICATION **then** *speaker* ← the following noun(s)
6: **else if** *directSpeech* followed by gap & VERB.COMMUNICATION **then** *speaker* ← the noun(s) in gap
7: **else if** *directSpeech* preceded by gap & VERB.COMMUNICATION **then** *speaker* ← the noun(s) in gap
return *speaker*

Figure 15: Our method for assigning a speaker to a direct speech utterance.

at the same time expressing the importance of the words, typically measured by their frequency. To do so, two visual variables are commonly used: the size of the words and their color.

Having a data set with direct speech from the assigned speakers, we decided on an approach that visualized the differences of two characters in terms of their direct speech. To do so, the words have been lemmatized and counted. This is done for two characters, e.g. Harry and Hermione from the Harry Potter novels. To understand where the differences between the characters lie, these two sets of words are then subtracted from each other, which results in two disjoint sets with no overlap.

To construct the visualization, we join the two sets and assign each word an importance score based the number of occurrences. This importance score is reflected in the size of the words, leaving the color to indicate another dimension of the data set. In our case, we decided to indicate the character that spoke the word by their color. Having set the size and colors of the words, the construction of the word cloud follows the classical wordle technique, along a spiral from the origin of the visualization canvas.

The example in Figure 16 shows the differences in direct speech of Harry and Hermione in “Harry Potter and the Sorcerer’s Stone”. White color indicates words attributed to Harry, and red the ones spoken by Hermione. We can see

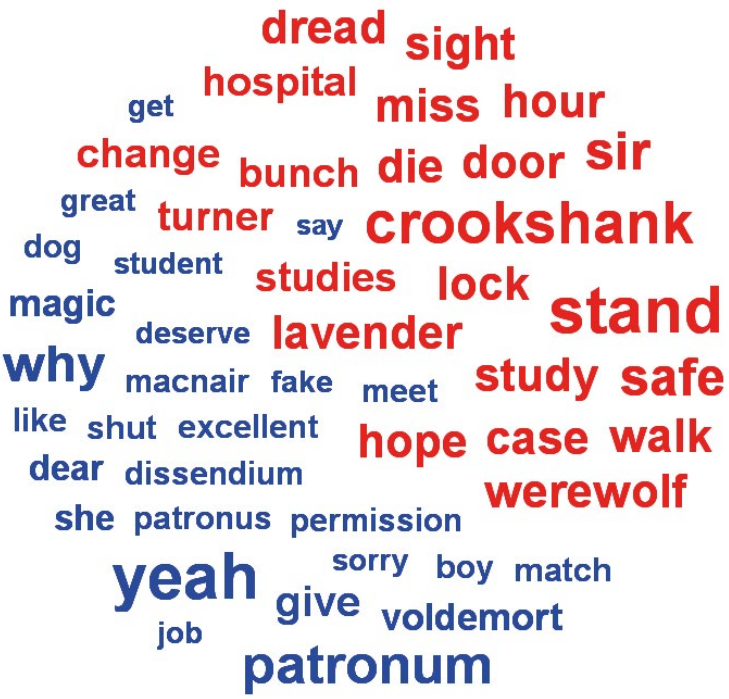


Figure 16: Word cloud displaying the most frequent words of direct speech from Harry (blue) and Hermione (red). Words occurring in direct speech of both characters have been removed from the data set.

that Hermione talks more about the studies and is also more concerned about safety. In contrast to Harry, she also mentions her cat, Crookshanks, multiple times. Harry, as is typical for the book, mentions Voldemort by name more often. He also uses comparatively more spells and often asks 'why'.

5.2 Extracting features on individual level: character's emotions, topics, actions

Related work

Nalisnick et al. (2013) analyze sentiment between characters in Shakespeare's plays. Their method is based on the assumption that a line of speech is directed towards the speaker of the previous line. A sentiment lexicon was then used to extract sentiment from the character's speech and measure the change of

sentiment between pairs of characters throughout the course of a play. Flekova and Gurevych (2015) use text classification techniques to predict personality traits of literary characters. Characters are classified as either 'introvert' or 'extrovert' types. This data is used to train SVM classifiers to predict these traits in unseen characters. Extracting features from the dialogue of the characters, a variety of features is used, including lexical, semantic and stylistic features, and the use of emotional language.

Methodology

For the purpose of analyzing individual characters, we use the same stylometric features as described in section 3, with the difference of applying them only on direct speech utterances of each protagonist separately rather than on the entire text of a book.

Visualization

To follow the emotional context of a character throughout a book, a visualization should provide the following insights into the emotional context of a character: what is the dominating emotion, and which emotions change and how drastic are these changes.

To determine the emotional context, the book is analyzed using a sliding window. For each window, the words from emotional categories, such as negative or positive emotions, the number of occurrences of words from these categories are counted and attributed to each character occurring in the window.

Compared to the overview visualization introduced in section **3.3**, the space-filling idea is discarded in favor of a flow- like visualization metaphor. This eases the task of perceiving the emotional change of neighboring emotional context, since, as well as the estimate of the amount of change, the area occupied by the emotion flow visualization also changes in relation to the overall amount of emotions. Each of the stacks is placed next to each other, as they occur in the book, to reflect the emotional change in the story. The transition between each of the stacks is displayed along a b-spline interpolation, which smoothes radical changes in the data (here: in the emotional changes), but still preserves a truthful transition between the contexts. For the emotional context, the areas corresponding to the different emotion categories are filled with distinct colors, which enables readers to easily follow an emotion category, as well as to effectively estimate the share of emotion categories per context.

In Figure 17, the emotion words connected to the main character, Harry, in the novel "Harry Potter and the Sorcerer's Stone" are shown. It is clearly visible that the number of words with a negative connotation dominate, while there is also a varying, but noticeable amount of "anger" words. In particular, it is striking that positive word classes, such as "joy" or "trust" occur only rarely in Harry's context.

In Figure 18, the emotion words co-occurring with Hermione are displayed. Besides the obvious insight that Hermione is appearing later in the book than Harry, the similarities between hers and Harry's emotion word context is quite clear.

Figure 19 shows a direct comparison of the emotional context (in terms of words from the NRC emotion dictionaries) of the two characters Strider (top) and Aragorn (bottom). For Strider, the context is indicated as being dominated by positive emotions, as well as a quite large extent of fear and negativeness. Aragorn, as the same character is referred to later, is missing a large amount of the positive extent, and his emotional context shows a larger influence of fear. This is in line with the story flow of the first Lord of the Rings book, where Aragorn is joining the fellowship of the ring and encounters, together with them, the Ringwraith. In contrast, Strider has a positive function as he offers help to the Hobbits in Bree, which can be seen in his emotional flow (Figure 19 top).

6. Conclusions and future work

In this chapter, we presented methods for extracting characteristics and features from book chapters, which can be used to approximate the components of story complexity from different aspects. In our model, we suggested that story complexity consists of three core areas: the complexity of the language used; complexity of the plot; and the intrinsic complexity of individual characters. We presented a range of Natural Language Processing techniques that enable initial insights into each of these areas and which can be further built upon.

Information visualization has been introduced as the method to make the different kinds of data visible and intuitively comprehensible to readers. Each of the visualization techniques is designed according to the characteristics of the available data, e.g. stylistic information of character co-occurrences, or the count of emotion words in a reference unit, for example, paragraphs. The visuals shown in this chapter are already highly specialized and tailored to the available set of features that in our opinion contribute to the whole ensemble, which we label as *story complexity*.

The next step in visualizing the different aspects would be the combination of different data, e.g. the co-occurrences of characters together with the emotion

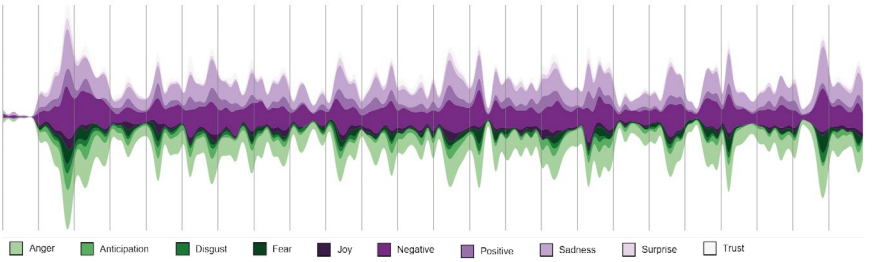


Figure 17: The flow of positive and negative emotions in “Harry Potter and the Sorcerer’s Stone” in the case of Harry. The vertical lines are for orientation purposes only.

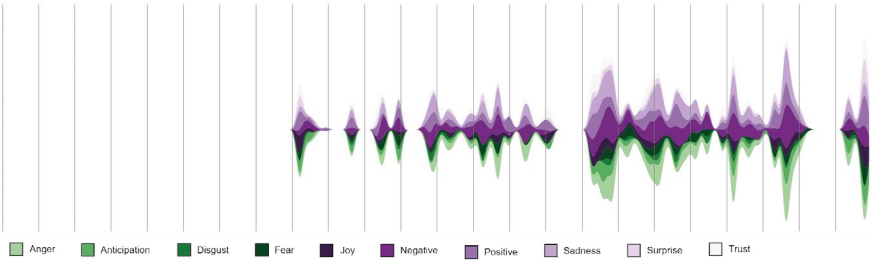


Figure 18: Emotion words in the context of Hermione during the first book of the Harry Potter series. The vertical lines are for orientation purposes only.

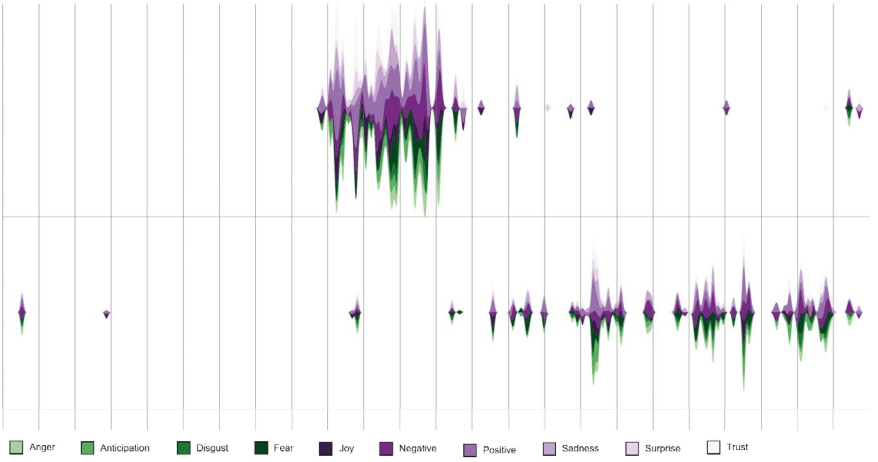


Figure 19: NRC emotion word categories appearing in the context of the character Strider (top) and Aragorn (bottom) in the first book of the Lord of the Ring series. The vertical lines are for orientation purposes only.

words, so that besides the fact that two characters occur together, a qualitative measure can be assigned. This could provide insight into books where a character changes his emotional context, and which gives a hint that his emotional profile might be more complex than that of other characters. In addition, more integrated views can open up new design spaces. More complex information spaces, such as projections based on the extracted features from multiple characters, can also give informative representations of the data, e.g. because of groups of entities or even the shapes formed by the entities.

From the Natural Language Processing perspective, literature still poses many challenges. Most of the text annotation models are focused on modern languages, such as those found in newspaper articles or even social media, and their adaptation to narratives which use notably more figurative language, such as more infrequent word expressions and sometimes an unusual syntactic structure, is challenging. For example, a named entity recognition model trained on Wikipedia is likely to produce very poor results when tried on classical novels. Development of more advanced methods tailored specifically to literature processing is required and exceeds the scope of this chapter.

7. References

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling." In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 363–370.

Akaishi, Mina, Yoshikiyo Kato, Ken Satoh, Koichi Hori. 2007. "Narrative based Topic Visualization for Chronological Data." In *11th International Conference Information Visualization* IV: 139–144.

Ashok, Vikas Ganjigunte, Song Feng and Yejin Choi. 2013. "Success with style: Using writing style to predict the success of novels." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP.

Chambers, Nathanael, and Dan Jurafsky. 2009. "Unsupervised learning of narrative schemas and their participants." In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP:* Volume 2, 602–610.

Chen, Chaomei. 2006. "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature." *Journal of the American Society for Information Science and Technology* (57) 3: 359–377.

Ciaramita, Massimiliano and Yasemin Altun. 2006. "Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger." In

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06), 594-602.

Daelemans, Walter. 2013. "Explanation in computational stylometry." In *International Conference on Computational Linguistics and Intelligent Text Processing*, edited by Alexander Gelbukh, 451–462. Berlin: Springer.

DeCamp, Philip, Amber Frid-Jimenez, Jethran Guiness, and Deb Roy. 2005. "Gist Icons: Seeing Meaning in Large Bodies of Literature." In *IEEE Info Visualization 2005 Conference.*

Don, Anthony, Elena Zheleva, Machon Gregory, Sureyya Tarkan, Loretta Auvil, Tanya Clement, Ben Shneiderman, and Catherine Plaisant. 2007. "Discovering interesting usage patterns in text collections: integrating text mining with visualization." In *CIKM,* 213–222.

Dotson, Daniel. 2005. "Portrayal of physicists in fictional works." *CLCWeb: Comparative Literature and Culture* 11 (2):5.

DuBay, William H. 2006. *The Classic Readability Studies.* Costa Mesa, Cal: Impact Information.

Eagleton, Terry. 2005. *The English Novel: An Introduction.* Blackwell, Oxford.

Eder, Jens, Fotis Jannidis, and Ralf Schneider, Eds. 2011. *Characters in fictional worlds: Understanding imaginary beings in literature, film, and other media.* Berlin: de Gruyter (Revisionen, 3).

Elsner, Micha. 2012. "Character-based kernels for novelistic plot structure." In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics: 634–44.

Elson, David K., Nicholas Dames, and Kathleen R. McKeown. 2010. "Extracting social networks from literary fiction." In *Proceedings of the 48th annual meeting of the association for computational linguistics,* 138–147.

Fekete, Jean-Daniel, and Nicole Dufournaud. 2000. "Compus: visualization and analysis of structured documents for understanding social life in the 16th century." *ACM DL,* 47–55.

Feng, Lijun et al. 2010. "A comparison of features for automatic readability assessment." *Proceedings of the 23rd international conference on computational linguistics: Posters.*

Flesch, Rudolf. 1948. "A new readability yardstick." *The Journal of Applied Psychology,* 32 (3): 221–233.

Flesch, Rudolf. 1979. *How to write plain English.* Harper and Brothers, New York: Harper and Brothers.

Gillam, Ronald B., and N. Pearson. 2004. *Test of narrative language.* Austin, TX: PRO-ED.

Goyal, Amit, Ellen Riloff, and Hal Daumé III. 2010. "Automatically producing plot unit representations for narrative text." In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing,* 77–86.

Greenhalgh, Kellie S. and C. J. Strong. 2001. "Literate language features in spoken narratives of children with typical language and children with language impairments." *Language, Speech, and Hearing Services in Schools*, 32 (2): 114–125.

Halpin, Harry, and Johanna D. Moore. 2006. "Event extraction in a plot advice agent." In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics,* 857–864.

He, Hua, Denilson Barbosa, and Grzegorz Kondrak. 2013. "Identification of speakers in novels." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1312– 1320.

Iyyer, Mohit, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. "Feuding Families and Former Friends: Unsupervised Learning for Dynamic Fictional Relationships." In. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1534–1544. https://doi.org/10.18653/v1/N16-1180

Jankowska, Magdalena, Vlado Keselj, Evangelos E. Milios. 2012. "Relative N-gram signatures: Document visualization at the level of character N-grams." In *IEEE VAST,* 103–112.

Jayannavar, Prashant Arun, Apoorv Agarwal, Melody Ju and Owen Rambow. 2015. "Validating literary theories using automatic social network extraction." In *Proceedings of the NAACL-2015 Workshop on Computational Linguistics for Literature,* 32–41.

John Burrows. 2004. "Textual analysis". In *A Companion to Digital Humanities,* edited by Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell, 324–347.

Johnson, John A., Joseph Carroll, Jonathan Gottschall, Daniel Kruger. 2001. "Portrayal of personality in Victorian novels reflects modern research findings but amplifies the significance of agreeableness." *Journal of Research in Personality* 45 (1): 50–58. https://doi.org/10.1016/j.jrp.2010.11.011

Kazantseva, Anna, and Stan Szpakowicz. 2010. "Summarizing short stories." *Computational Linguistics* 36 (1): 71–109.

Keim, Daniel A., and Daniela Oelke. 2007. "Literature Fingerprinting: A New Method for Visual Literary Analysis." In IEEE Symposium on Visual Analytics Science and Technology, 115–122.

Lee, John, and Chak Yan Yeung. 2012. "Extracting networks of people and places from literary texts." In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.*

Lieber, Emma K. 2011. *On the Distinctiveness of the Russian Novel: The Brothers Karamazov and the English Tradition.* Diss. Columbia University.

Liu, Shixia, Yingcai Wu, Enxun Wei, Mengchen Liu, and Yang Liu. 2013. "StoryFlow: Tracking the Evolution of Stories." *IEEE Trans. Vis. Comput. Graph.* 19 (12): 2436–2445.

McCabe, Allyssa, and L. S. Bliss. 2003. *Patterns of narrative discourse: A multicultural, life span approach.* Boston: Allyn & Bacon.

McIntyre, Neil, and Mirella Lapata. 2010. "Plot induction and evolutionary search for story generation." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 1562–1572.

Mohammad, Saif. 2011. "From once upon a time to happily ever after: Tracking emotions in novels and fairy tales." In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities,* 105–114.

Mohammad, Saif. 2011. "From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales." In *Proceedings of the ACL 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, 105–114.

Monroy, Carlos, Rajiv Kochumman, Richard Furuta, Eduardo Urbina. 2002. "Interactive Timeline Viewer (ItLv): A Tool to Visualize Variants Among Documents." In *Visual Interfaces to Digital Libraries,* 39–49.

Morton, T., J. Kottmann, J. Baldridge, and G. Bierner. 2005. Opennlp: A java-based nlp toolkit.

Newman, Robyn M., and K. K. McGregor. 2006. "Teachers and laypersons discern quality differences between narratives produced by children with or without SLI." *Journal of Speech, Language, and Hearing Research* 49 (5): 1022–1036.

Oelke, Daniela, Dimitrios Kokkinakis, Daniel A. Keim. 2013. "Fingerprint Matrices: Uncovering the dynamics of social networks in prose literature." *Comput. Graph. Forum* 32(3): 371–380.

Pitler, Emil, and Ani Nenkova. 2008. "Revisiting readability: A unified framework for predicting text quality." *Proceedings of the conference on empirical methods in natural language processing (EMNLP): 186–195.*

Rangel, Francisco, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2014. "Overview of the 2nd author profiling task at PAN 2014." In *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers,* 898–927.

Rangel, Francisco, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. "Overview of the author profiling task at PAN 2013." In *CLEF 2013 Conference on Multilingual and Multimodal Information Access Evaluation,* 352–365.

Recasens, Marta, Marie-Catherine de Marneffe, and Christopher Potts. 2013. "The Life and Death of Discourse Entities: Identifying Singleton Mentions." In *Proceedings of NAACL-HLT* 2013: 627–633.

Regan, Tim, Linda Becker. 2010. "Visualizing the text of Philip Pullman's trilogy 'His Dark Materials'" In *NordiCHI,* 759–764.

Rohrer, Randall M., John L. Sibert, and David S. Ebert. 1998. "The Shape of Shakespeare: Visualizing Text using Implicit Surfaces." *INFOVIS,* 121–129.

Scott, Cheryl M. & Windsor, J. 2000. "General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities." *Journal of Speech, Language, and Hearing Research* 43 (2): 324–340.

Senter, R., and E. Smith. 1967. *Automated Readability Index.* Aerospace Medical Research Laboratories.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng and Christopher Potts 2013. "Recursive deep models for semantic compositionality over a sentiment treebank." In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*: 1631–1642.

Tanahashi, Yuzuru, Chien-Hsin Hsueh, and Kwan-Liu Ma. 2015. "An Efficient Framework for Generating Storyline Visualizations from Streaming Data." *IEEE Trans. Vis. Comput. Graph* 21(6): 730–742.

Tanahashi, Yuzuru, Kwan-Liu Ma. 2012. "Design Considerations for Optimizing Storyline Visualizations." *IEEE Trans. Vis. Comput. Graph.* 18(12): 2679–2688.

Ukrainetz, Teresa A., L. M. Justice, J. N. Kaderavek., S. L. Eisenberg, R. B. Gillam, and H. M. Harm. 2005. "The development of expressive elaboration in fictional narratives." *Journal of Speech, Language, and Hearing Research* 48: 1363–1377.

Vala, Hardik, David Jurgens, Andrew Piper, Derek Ruths. 2015. "Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On the Difficulty of Detecting Characters in Literary Texts." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* 769–774.

van Ham, Frank, Martin Wattenberg, Fernanda B. Viégas. 2009. "Mapping Text with Phrase Nets." In *IEEE Trans. Vis. Comput. Graph.* 15(6): 1169–1176.

Wanner, Franz, Johannes Fuchs, Daniela Oelke, and Daniel A. Keim. 2011. "Are my Children Old Enough to Read these Books? Age Suitability Analysis." *Polibits* 43: 93–100.

Weber, Wibke. 2007. "Text Visualization – What Colors Tell About a Text."In *Information Visualization* IV: 354–362.

Weiler, Andreas, Michael Grossniklaus, Marc H. Scholl. 2015. "The Stor-e-Motion Visualization for Topic Evolution Tracking in Text Data Streams." IVAPP: 29–39.

Zhao, Ying and Justin Zobel. 2007. "Searching with style: Authorship attribution in classic literature." In *Proceedings of the thirtieth Australasian conference on Computer science,* 62: 59–68.

III. Tools

Sascha Wolfer / Sandra Hansen-Morath

Visualisierung sprachlicher Daten mit R

Abstract Die ansprechende und geeignete Visualisierung linguistischer Daten gewinnt analog zum steigenden Einfluss quantitativer Methoden in der Linguistik immer mehr an Bedeutung. R ist eine flexible und freie Entwicklungsumgebung zur Umsetzung von statistischen Analysen, die zahlreiche Optionen zur Datenvisualisierung bereithält und sehr gut für große Datensätze geeignet ist. Statistische Analysen und Visualisierungen von Daten werden auf diese Weise in einer Umgebung verzahnt. Durch die zahlreichen Zusatzpakete stehen auch weiterhin zeitgemäße Methoden zur Verfügung, um (linguistische) Daten zu analysieren und darzustellen.
Unser Beitrag vermittelt einen stark anwendungsorientierten Einstieg in das Programm und legt mithilfe von vielen praktischen Übungen und Anwendungsbeispielen die Grundlagen für ein eigenständiges Weiterentwickeln der individuellen Fähigkeiten im Umgang mit der Software.
Neben einer kurzen, eher theoretisch angelegten Einleitung zu explorativen und explanatorischen Visualisierungsstrategien von Daten werden verschiedene Pakete vorgestellt, die für die Visualisierung in R benutzt werden können.

1. Einleitung

Mit dem wachsenden Einfluss quantitativer Ansätze in der Linguistik gewinnt die ansprechende und geeignete Visualisierung linguistischer Forschungsergebnisse kontinuierlich an Bedeutung. Die im folgenden Beitrag vorgestellte freie Statistikumgebung R (R Core Team 2016) bietet eine solche Möglichkeit, statistische Analyse und deren Visualisierung zu verzahnen. Dank der fortlaufenden Entwicklung zahlreicher, den Funktionsumfang von R erweiternder Zusatzpakete ist davon auszugehen, dass Forscherinnen und Forschern auch in Zukunft zeitgemäße Methoden zur Verfügung stehen werden, linguistische Daten zu analysieren und darzustellen.

Im Vergleich zu den anderen Beiträgen in diesem Band fällt dieser Beitrag insofern aus dem Rahmen, als er in einem tutorialartigen Duktus gehalten ist. Dies ist dem Umstand geschuldet, dass der Beitrag in der Tat aus einem Tutorial

Erschienen in: Noah Bubenhofer u. Marc Kupietz (Hrsg.): Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

entstanden ist, das wir auf dem Herrenhäuser Symposium „Visuelle Linguistik“ in Hannover gehalten haben – jener Veranstaltung, aus der dieser Band hervorging. Da es das Ziel des Tutorials wie auch dieses Beitrags ist, einige Visualisierungsoptionen innerhalb von R vorzustellen und anhand linguistischer Beispiele zu illustrieren, gehen wir nicht auf kommerzielle Software (bspw. SPSS, SAS oder Stata) oder quelloffene Alternativen zu R (bspw. bestimmte Anwendungsszenarien der Sprache Python) ein. Zum Verständnis des Inhalts werden bei der Leserin / dem Leser zumindest erste Programmiererfahrungen (idealerweise in R selbst) vorausgesetzt.

1.1 Explorative und explanatorische Visualisierung von Daten

Visualisierungen sind meist entweder explorativ oder explanatorisch. Explorative Visualisierungen sind all jene Schaubilder, Graphen o. Ä., die Forschende einsetzen, um erhobene Daten selbst besser kennenzulernen und/oder Zusammenhänge zu begreifen bzw. zu entdecken. Dabei liegt der Fokus zunächst nicht auf einer ästhetisch ansprechenden Form. Auch auf eine extensive Annotation der Grafiken wird bei explorativen Visualisierungen im Allgemeinen verzichtet. Da die Forschenden die Schaubilder, Graphen etc. selbst entworfen haben, kennen sie ja die Bedeutung bspw. darin vorkommender Achsen oder Datenpunkte. Explorative Visualisierungen sind nicht unbedingt mit Verfahren der explorativen Statistik (z. B. Cluster- oder Hauptkomponentenanalysen) gleichzusetzen. Auch ein simples Streudiagramm (vgl. bspw. Abb. 1) kann explorativen Zwecken dienen. Wie bereits erwähnt: Primäres Ziel explorativer Visualisierungen ist es, die eigenen Daten besser kennenzulernen. In Zielgruppen gesprochen sind explorative Grafiken somit vom forschenden Individuum für sich selbst oder für Kolleginnen und Kollegen, die mit den Daten sehr gut vertraut sind.

Explanatorische Visualisierungen sind dagegen auf eine Zielgruppe ausgerichtet, die mit den zugrunde liegenden Daten nicht oder nur sehr wenig vertraut ist. Daher müssen meist unmissverständliche Achsenbeschriftungen verwendet und gegebenenfalls zusätzliche Annotationen (Legenden, Beschriftungen von Datenpunkten) hinzugefügt werden. Das Ziel explanatorischer Visualisierungen ist es, den Rezipient/innen Sachverhalte bzw. Wissen zu vermitteln oder sie von etwas zu überzeugen. Rezipient/innen explanatorischer Grafiken können Fachkolleg/innen (bei wissenschaftlichen Publikationen) sein, aber auch Laien, beispielsweise bei einer Infografik im Web oder einem Schaubild in einer populärwissenschaftlichen Zeitschrift oder Zeitung.

1.2 Von Daten zu Schaubildern

Bei der Frage, welche Art von Visualisierung sich am besten eignet, sollte immer von den Daten ausgegangen werden, die zu visualisieren sind. Einige einfache Fragen können helfen, die Gruppe von Visualisierungen zumindest einzuschränken. Eine erste Frage könnte lauten: Liegen mir Daten vor, in denen Informationen über *Einzelfälle* vorliegen, oder handelt es sich um auf welche Weise auch immer *aggregierte* Daten? So stellt beispielsweise eine Kontingenztabelle, die Korpushäufigkeiten in Bezug auf verschiedene Kombinationen von Merkmalen enthält, eine aggregierte Datenstruktur dar, weil aus der Tabelle nicht mehr auf den konkreten Einzelfall zurückgeschlossen werden kann (siehe unter anderem Tabelle 2). Eine Tabelle hingegen, in der jede individuelle sprachliche Einheit, sei es ein Wort, eine Phrase, eine Konstruktion oder ein Satz, in einer eigenen Zeile repräsentiert ist und deren Spalten Informationen zu dieser konkreten Einheit (Korpushäufigkeit oder -quelle, Wortart, syntaktische Einbettungstiefe etc.) beinhalten, weist keine aggregierte Datenstruktur auf. Grundsätzlich können auf der Grundlage von nicht-aggregierten, also einzelfallbasierten Datenstrukturen mehr Visualisierungen erstellt werden, da mehr Informationen vorhanden sind, nämlich zu jedem Einzelfall. Ohne diese Einzelfallinformationen sind zum Beispiel keine logistischen Regressionsanalysen möglich, mit denen man das Eintreten oder Ausbleiben von Ereignissen auf der Basis von einem oder mehreren Prädiktor(en) vorhersagen kann. Aggregierte Datenstrukturen können außerdem bei Bedarf aus Einzelfällen jederzeit abgeleitet werden.

Außer der Art der Datenstruktur ist natürlich auch das Ziel der Visualisierung ausschlaggebend für die Wahl des Diagramms. Wenn ich den Unterschied zwischen zwei Gruppen zeigen möchte, könnte ein Balkendiagramm, evtl. mit Fehlerindikatoren, angebracht sein. Möchte ich die Aufmerksamkeit auf den zeitlichen Verlauf lenken, wäre ein Liniendiagramm geeigneter. Wenn ich auf der Suche nach statistischen Ausreißern bin oder einfach einen Eindruck von der Verteilung gewinnen möchte, bietet sich ein Boxplot oder ein Histogramm an.

Als dritter Faktor ist auch die Anzahl der an der Visualisierung beteiligten Variablen relevant: Wie viele Dimensionen müssen im Schaubild abgetragen werden, um das zu visualisieren, was ich zeigen möchte? Ist genau eine Variable beteiligt, handelt es sich meist um Visualisierungen von Verteilungen. Dazu gehören u a. Histogramme, Dichtekurven, Boxplots, Violin-Plots *(violin plots).*[1] Soll die Beziehung zwischen zwei Variablen gezeigt werden, kommt es darauf an, wie die Variablen jeweils skaliert sind. So eignet sich z. B. ein Balkendiagramm

1 Violin-Plots sind Boxplots recht ähnlich. Dabei werden anstatt (oder zusätzlich zu) einer Box links und rechts Dichtekurven angezeigt. In R sind Violin-Plots in den Paketen „ggplot2“ (siehe Abschnitt 3.1) und „vioplot“ verfügbar.

für die Darstellung des Zusammenhangs zwischen einer kategorialen (nominal- oder ordinalskalierten) und einer kontinuierlichen (intervallskalierten) Variable. Gruppiert oder stapelt man Balkendiagramme, kann noch eine weitere kategoriale Variable berücksichtigt werden. Zwei kontinuierliche Variablen und eine kategoriale Variable können in einem Streudiagramm im Zusammenhang betrachtet werden. Dazu werden die beiden kontinuierlichen Variablen auf der x- und y-Achse abgetragen, während die kategoriale Variable beispielsweise durch Farbe oder Form der Datenpunkte dargestellt wird. Tabelle 1 gibt einen Überblick über einige mögliche Kombinationen von Variablen und möglichen Visualisierungen.

Tabelle 1: Anzahl zu visualisierender kategorialer und kontinuierlicher Variablen und einige einfache passende Diagrammarten.

Kategoriale Variablen	**Kontinuierliche Variablen**	**Mögliche Visualisierung(en)**
1	0	Balkendiagramm, Tortendiagramm
> 1	0	Mosaikplot, Assoziationsplot, Gruppiertes/gestapeltes Balkendiagramm
0	1	Dichtekurve, Histogramm, Boxplot, Violin-Plot
0	2	Streudiagramm
0	> 2	3D-Diagramm, Streudiagramm + Radius der Datenpunkte
1	1	Balkendiagramm, Liniendiagramm, Gruppierte Boxplots
1	2	Streudiagramm + Farbe/Form der Datenpunkte
2	1	Gruppiertes Balkendiagramm, Heatmap, Hexbin-Plot

Tabelle 1 enthält längst nicht alle Arten von Visualisierungen, die mit den entsprechenden Kombinationen von Variablen möglich sind. Genannt sind aber die meisten grundlegenden Verfahren, die innerhalb der linguistischen Forschung etabliert sind.

1.3 Gliederung

Bevor wir uns nun konkret mit den Visualisierungsmöglichkeiten in R beschäftigen, sei an dieser Stelle noch ein kurzer Überblick über die Gliederung des folgenden Beitrags gegeben. In Kapitel 2 werden wir uns mit einigen R-Paketen beschäftigen, die für die Visualisierung in R benutzt werden können. In Abschnitt 2.1 gehen wir zunächst auf das in R bereits integrierte Paket „graphics“ ein. Dabei werden wir einerseits Plots selbst erstellen, andererseits aber auch zeigen, wie Elemente in bereits existierende Plots eingefügt werden können. Über diese Funktion können mächtige maßgeschneiderte Visualisierungen erstellt werden. In Abschnitt 2.2 widmen wir uns der Darstellung von kategorialen Daten und führen Mosaik- und Assoziationsplots ein, wie sie im Paket „vcd“[2] (Meyer 2015) verfügbar sind. In Abschnitt 2.3 zeigen wir, wie mit dem Paket „effects“ (Fox 2003) die Schätzer aus bereits vorhandenen statistischen Modellen extrahiert und abgetragen werden können. In Kapitel 3 besprechen wir zwei alternative Plot-Pakete: Mit dem Paket „ggplot2“ (Wickham 2009) können Visualisierungen basierend auf der „Grammar of Graphics“ (Wilkinson 20015) erstellt werden. Dieses Paket stellen wir (kurz) in Abschnitt 3.1 vor. Abschnitt 3.2 schließlich verwenden wir darauf, das R-Paket „rCharts“ (Vaidyanathan 2013) vorzustellen, mit dem interaktive Grafiken erstellt werden können, die auf der JavaScript-Bibliothek d3.js basieren. Ein Schlusskapitel rundet den Beitrag ab.

2. Visualisierung deskriptiver Statistiken und inferenzstatistischer Maße

2.1 Die Basisausstattung: Das Paket „graphics“

Bereits die Grafikfunktionen, die in R ohne jegliche Zusatzpakete zur Verfügung stehen, sind vielfältig anpassbar. Mit diesen „Bordmitteln“ können sowohl explorativ als auch explanatorisch ausgerichtete Visualisierungen erstellt werden. Benutzer/innen können dabei vorgefertigte Funktionen nutzen, um komplexe Grafiken zu erstellen. Da die Logik einer Art Baukastenprinzip folgt, können aber auch beliebige Elemente miteinander kombiniert und so einfache Grafiken sukzessive erweitert und an die eigenen Visualisierungsziele angepasst werden.

Einfache, bereits in Abschnitt 1.2 genannte Schaubilder können mit den Funktionen plot() für x-y-Streudiagramme und Liniendiagramme, barplot() für Balkendiagramme, hist() für Histogramme, pie() für Tortendiagramme und

2 Zusatzpakete können in R über die Funktion install.packages() installiert und über library() eingebunden werden.

boxplot() für Boxplots erstellt werden. Die Arbeitsweise und somit die grafischen Produkte dieser Funktionen können über Parameter angepasst werden. Eine Übersicht über alle Grafikparameter wie Ränder, Farben usw. kann in R über die Eingabe von ?par angefordert werden. Möchte man sich über die Parameter informieren, die jeder dieser Funktionen eigen sind, bspw. die Reichweite der Hinges[3] in Boxplots oder die Anzahl der Bins in Histogrammen, sollte man die Hilfeseite der jeweiligen Funktion (bspw. ?boxplot oder ?hist) zurate ziehen.

Das Interessante am Baukastenprinzip des Basispakets für Grafiken in R ist, dass man zu jedem bestehenden Plot Elemente hinzufügen kann. Alle oben genannten Funktionen starten in ihrer Standardeinstellung ein neues „Grafikgerät“ *(graphics device)*. Funktionen wie abline(), lines(), points(), curve(), arrows(), text() oder legend() fügen den Grafiken die entsprechenden Elemente in bereits geöffneten Grafikgeräten hinzu. Wir demonstrieren dies anhand eines Datensatzes zu einer lexikalischen Entscheidungsaufgabe, der im Paket „languageR“[4] (Baayen 2011) enthalten ist. Der Datensatz besteht aus 1659 Antworten auf eine lexikalische Entscheidungsaufgabe. Wir visualisieren den Zusammenhang zwischen Worthäufigkeit, semantischer Klasse und Reaktionszeit in der lexikalischen Entscheidungsaufgabe, d. h., wie lange jede Teilnehmerin/jeder Teilnehmer des Versuchs benötigt hat, um ein englisches Wort in Abhängigkeit von seiner semantischen Klasse (Tier/Pflanze) und seiner Vorkommenshäufigkeit als solches zu verifizieren. Die Nichtwörter sind nicht im Datensatz enthalten.

```
> library(languageR)
> lexdec$plot.col <- ifelse(lexdec$Class == "animal",
                            "#FF000099", "#00FF0099")
> plot(lexdec$Frequency, lexdec$RT, col = lexdec$plot.col,
       pch = 19, yaxt = "n", xaxt = "n",
       xlab = "Häufigkeit“, ylab = "Reaktionszeit (ms)",
       main = "Häufigkeit und Reaktionszeit in 'lexdec'")
> axis(side = 1, at = 2:8,
       labels = round(exp(2:8)))
> axis(side = 2, at = c(6, 6.5, 7, 7.5),
```

3 Die Höhe der Hinges (der Linien, die oben und unten an der Box angelegt werden) definiert sich durch den höchsten bzw. niedrigsten zulässigen Wert, der tatsächlich auftritt. Im Allgemeinen ist diese Grenze durch 1,5-mal die Höhe der Box (= Interquartilabstand) definiert.

4 Zusatzpakete werden in R über den Befehl library(<Paketname>) eingebunden. Diese Befehle finden Sie in den Code-Beispielen. Pakete müssen vor dem Einbinden installiert werden. Sollten diese Pakete noch nicht auf Ihrem Rechner installiert sein, können Sie das mit dem Befehl install.packages(„<Paketname>“) tun. Für das Paket „languageR“ lautet der Befehl somit install.packages(„languageR“).

```
       labels = round(exp(c(6, 6.5, 7, 7.5))))
> lines(lowess(lexdec$Frequency, lexdec$RT), lwd = 3)
> legend(x = "topright", col = c("red", "green"), pch = 19,
         legend = c("Tier", "Pflanze"), bty = "n",
         inset = 0.05, y.intersp = 1.5)
```

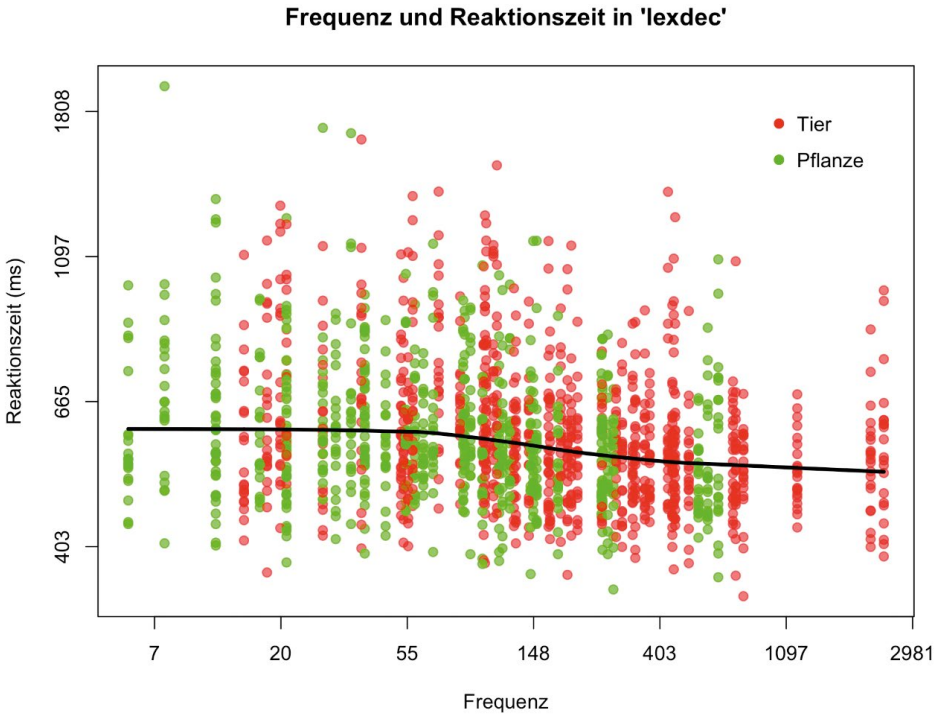


Abb. 1: Streudiagramm für die Reaktionszeit in Abhängigkeit der Häufigkeit im Datensatz „lexdec" mit Lowess-Anpassungskurve (schwarz) und Aufteilung nach semantischer Klasse.

Abb. 1 zeigt das Ergebnis nach Evaluation des abgedruckten Codes. Zunächst wird die zusätzliche Variable plot.col definiert, die je nach Ausprägung der Variable Class (semantische Klasse) einen Wert für ein transparentes Rot[5] für

5 Die Farben sind hier über RGB-Codes mit Alpha Channel angegeben. Die ersten zwei Werte definieren den Rotanteil, die zweiten zwei Ziffern den Grünanteil, die dritten zwei Ziffern den Blauanteil. Die Werte variieren zwischen oo (aus) und FF (komplett). Die letzten zwei Ziffern (Alpha Channel) definieren die Transparenz (variierend zwi-

Tierwörter („animal“) oder ein transparentes Grün für Pflanzenwörter („plant“) enthält. In der nächsten Zeile wird der eigentliche Plot erstellt, zunächst noch ohne Achsen (Parameter xaxt und yaxt beide „n“). Dann werden die beiden Achsen hinzugefügt, die anstatt der in der Variable enthaltenen logarithmierten Werte die zurücktransformierten Werte der Variablen enthalten (exp() ist die Umkehrung von log()). Als Nächstes wird über den Aufruf von lines() eine Lowess-Anpassungslinie (vgl. Cleveland 1981) dem augenblicklich verwendeten Grafikgerät hinzugefügt. Zuletzt wird die Legende konfiguriert und ebenfalls in den Plot eingefügt. Die verwendeten Parameter der Funktionen können jeweils über ?<Funktionsname> in R eingesehen werden (bspw. ?legend für die Hilfeseite von legend()).

Wenn wir versuchen, die Visualisierung aus Abb. 1 in das Schema in Tabelle 1 einzuordnen, so finden wir den vorliegenden Variablentyp in der vorletzten Zeile beschrieben. Zu visualisieren sind nämlich zwei kontinuierliche Variablen – Häufigkeit und Reaktionszeit (abgetragen auf x- und y-Achse) – sowie eine kategoriale Variable, die semantische Klasse (durch Farben symbolisiert).

Einige Zusatzpakete benutzen und erweitern die Basisfunktionalität von R. So ist das Paket „sciplot“ (Morales 2012) nützlich für die schnelle Exploration von Daten, die in einem faktoriellen Design[6] gesammelt wurden oder in einem solchen abbildbar sind. Mit den Funktionen lineplot.CI() und bargraph.CI() können Linien- oder Balkendiagramme mit Fehlerindikatoren erzeugt werden. Der Aufruf ist dabei recht einfach. Es muss eine (vorzugsweise kategoriale) Variable angegeben werden, die die x-Achse definiert sowie eine binäre oder kontinuierliche Variable, die die y-Achse definiert. Optional kann eine (kategoriale) Gruppierungsvariable angegeben werden. Die Berechnung der gruppenspezifischen Standardfehler übernimmt dann lineplot.CI() oder bargraph.CI(). Für die Exploration der eigenen Daten (sofern sie auf diese Weise abgebildet werden können) reicht das Ergebnis meist schon aus. Mit ein wenig Mehraufwand können auch publikationsreife Grafiken erstellt werden. Ein weiteres Beispiel aus dem Datensatz „lexdec“ verdeutlicht das Vorgehen. Wir tragen den Zusammenhang zwischen der Muttersprache der/des Befragten, der semantischen Klasse des Worts und der Reaktionszeit ab.

schen 00 = voll transparent und FF = voll deckend). Die Funktion alpha() aus dem Paket „scales“ erlaubt eine einfachere Definition von Transparenz in Kombination mit Farbnamen (bspw. alpha(„red“, 0.5) für halbtransparentes Rot).

6 Die Funktionen im Paket „sciplot“ eignen sich am besten dazu, den Zusammenhang zwischen zwei gekreuzten kategorialen unabhängigen Variablen und einer kontinuierlichen abhängigen Variable zu analysieren.

```
> library(sciplot)
> lineplot.CI(NativeLanguage, RT, Class, data = lexdec,
       xlab = "Muttersprache", ylab = "Reaktionszeit")
```

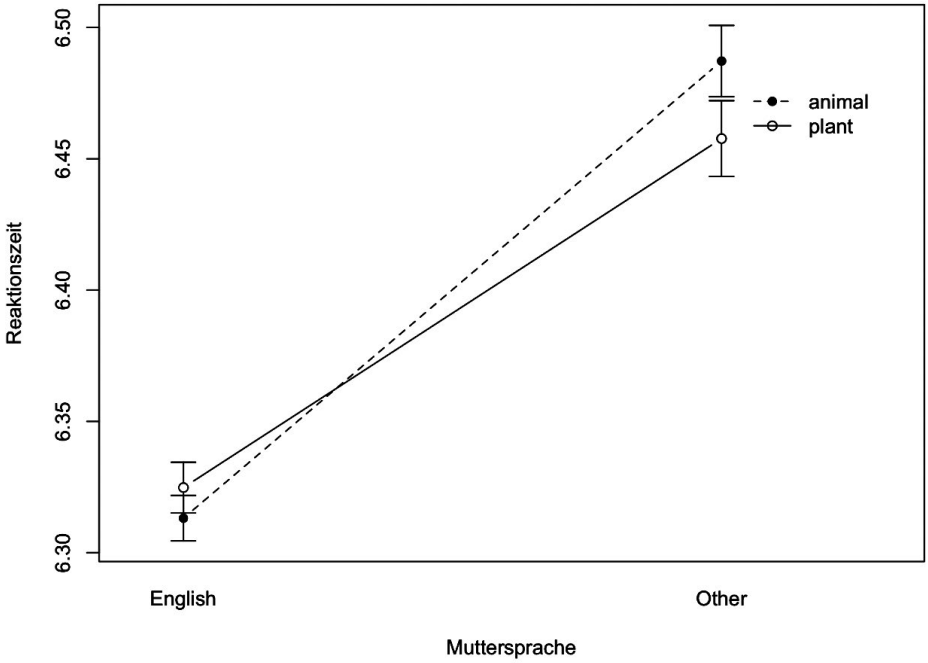


Abb. 2: Beispiel für einen Interaktionsplot mit dem Paket „sciplot“. Abgetragen ist die Beziehung zwischen der Muttersprache der Teilnehmenden, der semantischen Klasse des Worts und der Reaktionszeit in einer lexikalischen Entscheidungsaufgabe. Die Fehlerbalken symbolisieren 1 Standardfehler.

Abb. 2 zeigt das Ergebnis des Aufrufs. Die Grafik ist in dieser Form wohl noch nicht publikationsreif, insbesondere auch, weil die Linien zwischen den Datenpunkten zwar den potenziellen Interaktionseffekt (der noch anhand eines linearen Modells zu überprüfen wäre) gut hervorheben, im Grunde aber irreführend sind, da zwischen den Stufen „English“ und „Other“ des Faktors NativeLanguage keine Werte vorhanden sind. Für die Publikation wäre es daher eher angemessen, die Linien wegzulassen und die Datenpunkte leicht gegeneinander zu verschieben, damit die (Nicht-)Überlappung der Fehlerbalken leichter abgelesen werden kann.

Insbesondere für explanatorische Visualisierungen, die zur Veröffentlichung bestimmt sind, brauchen wir R-Grafiken in Bild-Dateien (.jpg/.png/...). Um ein solches Dateiformat zu erhalten, können wir entweder die Grafikexport-Funktion von RStudio[7] verwenden oder aber vor dem Erstellen der Grafiken mit den Funktionen jpeg(), png() usw. ein spezielles Grafikgerät starten. R wird dann den Grafikoutput nicht auf dem Bildschirm darstellen, sondern direkt in die Datei schreiben. Nachdem die Grafik in die Datei geschrieben wurde, müssen wir das Gerät wieder mit dev.off() schließen. Der Vorteil dieser Exportart liegt in den damit verbundenen Anpassungsmöglichkeiten. Über verschiedene Parameter wie bspw. height, width und res können wir die genauen Abmessungen und die Auflösung der zu erstellenden Grafik beeinflussen.[8] Der folgende Beispielaufruf verdeutlicht dieses Vorgehen:

```
> png("VisLing-testplot.png", width = 2000, height = 1600, res = 250)
> plot(rnorm(1000))
> dev.off()
```

Das Arbeitsverzeichnis (abrufbar mit der Funktion getwd()) enthält nun eine Datei „VisLing-tesplot.png“ mit einer Breite von 2000 Pixeln, einer Höhe von 1600 Pixeln und einer Auflösung von 250 ppi. Der Plot enthält 1000 zufällig normalverteilte Datenpunkte.

Sollen Grafikdateien in vektorbasierte Formate wie bspw. EPS (Encapsulated Postscript) oder SVG (Scalable Vector Graphics) exportiert werden, bietet R auch diese Möglichkeit. EPS-Dateien können über den Befehl postscript() erstellt werden, SVG-Dateien werden mit svg() erzeugt.

Aus den vorangegangenen Anwendungsbeispielen wurde deutlich, dass bereits das Basispaket in R äußerst vielfältige Visualisierungsmöglichkeiten bereithält. Selbstverständlich kann in diesem Rahmen nur ein kleiner Teil der Möglichkeiten präsentiert werden. Die Logik ist aber fast immer dieselbe: Eine Funktion beginnt entweder einen neuen Plot oder fügt Elemente in das augenblicklich geöffnete Grafikgerät ein. Auf diese Weise kann jeder Plot nach eigenen Wünschen konfiguriert und erweitert werden.

7 RStudio ist die wohl am weitesten verbreitete Entwicklungsumgebung für R. RStudio ist frei verfügbar unter www.rstudio.com/products/RStudio (letzter Zugriff am 13.05.2016). RStudio bietet viele Unterstützungen bei der Arbeit mit R an, auf die wir an dieser Stelle aber nicht umfassend eingehen können. Nützliche Funktionen sind unter anderem die Syntaxhervorhebung und -vervollständigung, viele Tastenkombinationen, die die Arbeit mit R unterstützen, Fehlerprüfungen und Versionskontrollfunktionen.

8 Für weitere Parameter siehe die Hilfeseite zu ?png.

2.2 Visualisierung kategorialer Daten: Das Paket „vcd"

Mit dem Paket „vcd" ist es möglich, kategoriale Daten auf verschiedene Art und Weise zu visualisieren. Unter kategorialen Merkmalen versteht man Größen, die endlich viele Ausprägungen besitzen und höchstens ordinalskaliert sind (Fahrmeir et al. 2007). Beispiele für kategoriale Variablen sind Muttersprache oder Wohnort eines Autors. Mithilfe des Pakets „vcd" sind eine Vielzahl an verschiedenen Plots möglich. Zu nennen sind hier beispielsweise der Mosaikplot, der Sieveplot[9] und der Assoziationsplot. Im Rahmen des vorliegenden Artikels werden wir zeigen, wie Mosaik- und Assoziationsplots erstellt und gelesen werden können. Hierzu nutzen wir ein Beispiel aus einer Analyse von Fürbacher (2015), die in einer Studie zu Genitivallomorphen im Deutschen bei starken Maskulina und Neutra zeigt, dass für die Variation der Genitivallomorphe *-es* und *-s* neben sprachimmanenten Faktoren auch regionale Einflüsse eine Rolle spielen können. Da kategoriale Daten üblicherweise durch Kontingenztabellen analysiert und visualisiert werden, bilden sie die Grundlage für Analysen mit dem Paket „vcd". Tabelle 2 zeigt eine solche Kontingenztabelle mit den absoluten Trefferzahlen für die Lemmata mit den Genitivendungen *-es* und *-s* in den Regionen Nordost, Südost (inklusive Österreich), Nordwest und Südwest (inklusive der Schweiz) aus der Analyse von Fürbacher (2015).[10]

Tabelle 2: Absolute Trefferzahlen für Lemmata mit den Genitivendungen *-es/-s* in Abhängigkeit von der Sprachregion nach Fürbacher (2015)

	-es	***-s***
Nordost	425	369
Südost	1435	573
Nordwest	239	235
Südwest	435	438

Wir erstellen in einem ersten Schritt die Kontingenztabelle in R. Hierzu generieren wir zunächst für jede Region einen Vektor mit den absoluten Häufigkeiten für die Genitivendungen *-es* und *-s*.

9 In Sieveplots werden Rechtecke geplottet, deren Flächeninhalte proportional zu den erwarteten Häufigkeiten der entsprechenden Tabellenzelle sind. In die Rechtecke werden Quadrate gezeichnet, die die beobachteten Häufigkeiten kennzeichnen. Die Dichte der ausgefüllten Rechtecke symbolisiert somit die Abweichung zwischen erwarteten und beobachteten Häufigkeiten.

10 Für eine detailliertere Beschreibung der Daten und Auswertungen der Studie vgl. Fürbacher (2015) und Hansen & Wolfer (i. Vorb).

```
> Nordost <- c(425, 369)
> Südost <- c(1435, 573)
> Nordwest <- c(239, 235)
> Südwest <- c(435, 438)
```

Über die Funktion rbind() werden die Vektoren zu einer Tabelle zusammengefügt und in der Variable tab gespeichert. Anschließend werden die Spaltennamen über die Funktion colnames() vergeben.

```
> tab <- rbind(Nordost, Südost, Nordwest, Südwest)
> colnames(tab) <- c("es", "s")
> tab
           es   s
Nordost   425 369
Südost   1435 573
Nordwest  239 235
Südwest   435 438
```

Auf Basis der vorliegenden Kontingenztabelle können nun die Analysen durchgeführt werden. Als Visualisierungsform wählen wir einen Mosaikplot (vgl. Abb. 3). Die Häufigkeiten werden über die Größe der Flächeninhalte der einzelnen Rechtecke dargestellt (vgl. Friendly 1994; Meyer et al. 2006). Darüber hinaus werden in dem Plot auch die signifikante Pearson-Residuen[11] (standardisierte Abweichungen der beobachteten von den erwarteten Häufigkeiten), angezeigt. Der Plot wird mit der Funktion mosaic() erstellt, die Schattierungen werden gemäß den Residuen über den Parameter shade = TRUE eingefügt.

```
> library(vcd)
> mosaic(tab, shade = TRUE)
```

Der Mosaikplot ist wie folgt zu lesen: Jedes Rechteck im Plot steht für eine Zelle der Kontingenztabelle. Der Flächeninhalt des Rechtecks links oben bezieht sich zum Beispiel auf die Häufigkeit der Genitivendung *-es* in der Region Nordost des deutschen Sprachraums. Es wird erkennbar, dass die meisten Tokens (*-es* und *-s*) im Korpus im Südosten vorkommen, denn die zweite Zeile an Rechtecken ist am höchsten. Die Einfärbung der Rechtecke symbolisiert die Pearson-Residuen. Sind die Pearson-Residuen signifikant (Betrag größer 1,97), werden sie im Plot eingefärbt: Die Farbe Blau steht für signifikante Abweichungen nach oben, die

11 Pearson-Residuen berechnen sich aus der Differenz der beobachteten und erwarteten Häufigkeiten geteilt durch die Wurzel der erwarteten Häufigkeiten.

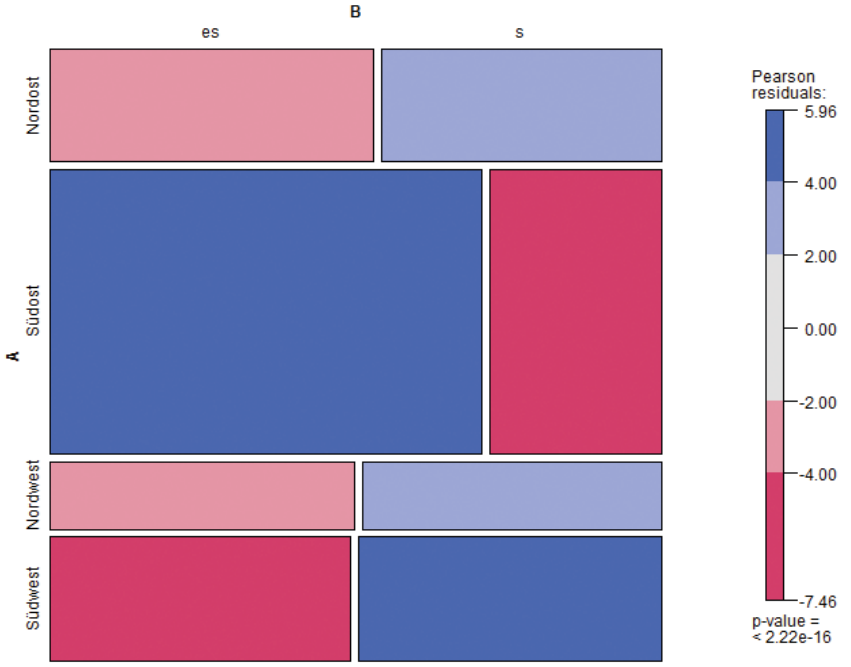


Abb. 3: Mosaikplot für die Genitivallomorphe *-es* und *-s* in Abhängigkeit von Regionen.

Farbe Rot für signifikante Abweichungen nach unten. Wird die Schwelle von +/- 4 überschritten, werden die Balken in einem stärkeren Ton eingefärbt, da hier von einer besonders starken Abweichung ausgegangen werden kann. Die Zuordnung von Einfärbungsgrad zu Residuenhöhe wird rechts neben dem Plot durch eine Legende dargestellt. Unterhalb der Legende wird außerdem der *p*-Wert eines Chi²-Tests ausgegeben, mit dem überprüft wird, ob die Ausprägungen der Variablen voneinander unabhängig sind (vgl. Bortz 2005: 168ff.).[12]

Bezogen auf das Beispiel zeigt der Mosaikplot, dass im Südosten die Genitivendung *-es* signifikant überrepräsentiert ist, während sie in den anderen Regionen signifikant unterrepräsentiert ist. Im Südwesten dominiert sogar die kürzere Endung *-s*.

Durch eine Anpassung über den Parameter labeling im Aufruf werden die genauen Werte für die standardisierten Residuen für jedes Rechteck im Plot angezeigt (vgl. Abb. 4):

12 Laut Konvention spricht man ab einer Irrtumswahrscheinlichkeit kleiner 5 % ($p < 0.05$) von einem signifikanten, ab 1% ($p < 0.01$) von einem hochsignifikanten und ab 0,1 % ($p < 0.001$) von einem höchstsignifikanten Unterschied.

```
> mosaic(tab, shade = TRUE,
       labeling=labeling_values(value_type=c("residuals")))
```

Eine weitere Möglichkeit, die standardisierten Abweichungen zwischen beobachteten und erwarteten Häufigkeiten darzustellen, bietet der Assoziationsplot (vgl. Cohen 1980; Friendly 1992; Meyer et al. 2006). Hierzu kann folgender Aufruf verwendet werden:

```
> assoc(tab, shade = TRUE,
       labeling = labeling_values(value_type=c("residuals")))
```

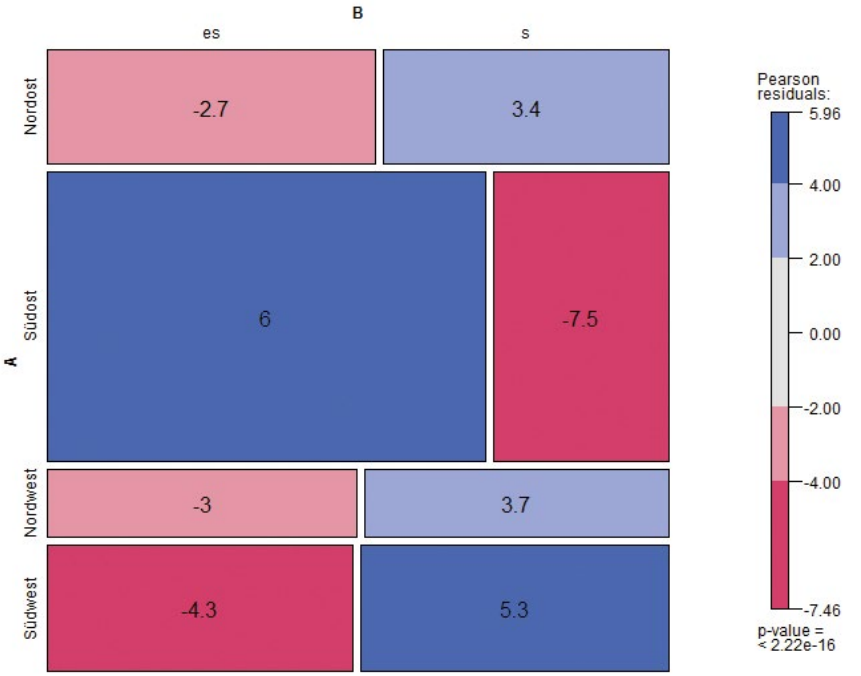


Abb. 4: Mosaikplot für die Genitivallomorphe *-es* und *-s* in Abhängigkeit von Regionen mit der Angabe der Pearson-Residuen für jedes Rechteck.

Abb. 5 zeigt den Assoziationsplot für das vorliegende Beispiel. Abgetragen werden die standardisierten Residuen der beiden Genitivendungen in Abhängigkeit von der regionalen Verteilung. Die Höhe der Balken entspricht der Höhe der

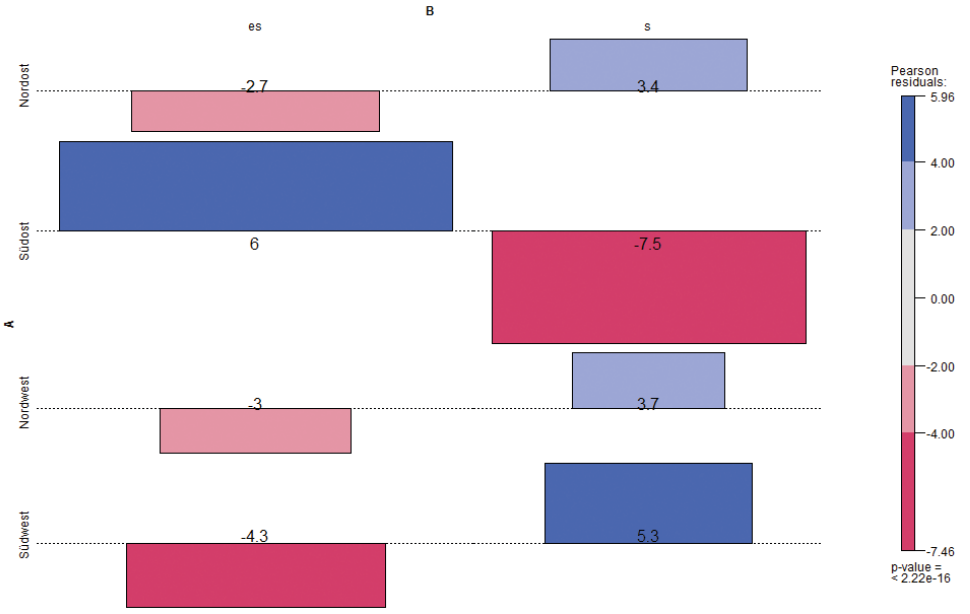


Abb. 5: Assoziationsplot für die Genitivallomorphe *-es* und *-s* in Abhängigkeit von Regionen mit Angabe der Pearson-Residuen für jedes Rechteck.

Abweichung: Balken oberhalb der gepunkteten Linie bedeuten höhere, Balken unterhalb der Linie niedrigere Werte als erwartet. Die genauen Werte der Pearson-Residuen werden hier ebenfalls durch die Anpassung des Parameters labeling angezeigt. Die Breite der Balken spiegelt die erwartete Häufigkeit der Realisierungsvarianten wider. Ist der Betrag des entsprechenden Pearson-Residuums größer 1,97, wird der Balken im Plot eingefärbt (vgl. Hansen-Morath et al., i. Vorb.).

Wie auch im Mosaikplot abzulesen ist, weist der Wert der Residuen in unserem Beispiel im Südosten mit 6,0 für *-es* bzw. -7,5 für *-s*, aber auch im Südwesten mit -4,3 für *-es* bzw. 5,3 für *-s* auf eine besondere Bedeutung der Abweichung für die Signifikanz der Unterschiede hin (vgl. Fürbacher 2015). Laut Fürbacher könnte der signifikant häufigere Gebrauch der *es*-Endung im Südosten durch eine kompensierende Funktion gegenüber der gesprochenen Sprache bedingt sein, da in dieser der Schwa-Laut gemieden wird (vgl. ebd.; Tatzreiter 1988: 79).

2.3 Visualisierung von Modellschätzern: Das Paket „effects“

In einigen Fällen ist es unerlässlich, nicht Rohdaten, sondern von statistischen Modellen *geschätzte* Daten zu visualisieren. Schon bei einfachen linearen Modellen[13], die mehr als einen Prädiktor enthalten, können die Schätzwerte von den Rohwerten abweichen. Unter Umständen kann sich das Verhältnis zweier Gruppen im Vergleich zwischen Modellschätzern und Zellmittelwerten sogar umkehren. Das kann vorkommen, da bei der Berechnung des Einflusses von Prädiktor x_1 Effekte anderer Prädiktoren x_2, x_3 ... x_n berücksichtigt werden und vice versa. Tragen wir also Zellmittelwerte ab, um die Ergebnisse eines linearen Modells zu visualisieren, haben wir das Problem, dass unsere Visualisierung nicht dieselben (und evtl. sogar in ihrer Aussage entgegengesetzte) Werte darstellt, wie wir im statistischen Modell berechnet haben.

Ein Minimalbeispiel anhand des Datensatzes „lexdec“ illustriert das Problem.

```
> library(languageR)
> library(effects)
> mod1 <- lm(RT ~ Class + Frequency, data = lexdec)
```

Mit der letzten Code-Zeile sagen wir die logarithmierte Reaktionszeit (RT)[14] aus der semantischen Klasse (Class, „plant“ vs. „animal“) und der Häufigkeit (Frequency) des Wortes mithilfe eines linearen Modells vorher. Obwohl die Interaktion der beiden Variablen nicht in der Vorhersage beachtet wurde, zeigt sich bereits, dass die Modellschätzwerte leicht von den Zellmittelwerten für Pflanzen und Tiere abweichen.

```
> tapply(lexdec$RT, lexdec$Class, mean) # Zellmittelwerte
  animal    plant
6.387746 6.381751
> Effect("Class", mod1) # Extraktion der Modellschätzer
Class effect
Class
  animal    plant
6.406457 6.358228
```

13 Mit linearen Modellen (bspw. einer linearen Regression) sagt man eine Kriteriumsvariable aus einer oder mehreren Prädiktorvariablen vorher (zur Einführung vgl. Bortz (2005: 483ff). In linearen Modellen können auch kategoriale Variablen als Prädiktoren verwendet werden. Im Beispiel in diesem Abschnitt werden eine kategoriale Variable (Class) und eine kontinuierliche Variable (Frequency) als Prädiktoren verwendet.

14 Reaktionszeiten sind typischerweise rechtsschief verteilt. Um diese Verteilung in eine Normalverteilung zu überführen, ist es zulässig, die Variable einer Log-Transformation zu unterziehen.

Der Unterschied zwischen beiden Gruppen ist bei den Schätzwerten größer: Die Differenz der geschätzten Mittelwerte beträgt dort 0,0482. Im Vergleich dazu beträgt die Differenz für die rohen Mittelwerte lediglich 0,00599. Durch die Einbeziehung der Häufigkeit als Prädiktor wird offenbar so viel Rauschen in den Daten kontrolliert, dass der Effekt des anderen Prädiktors (semantische Klasse) stärker hervortritt. Da wir das auch in der Visualisierung der Datenauswertung beachten wollen, sollten wir also nicht die Zellmittelwerte abtragen.

Durch eine Kombination von plot() mit den Funktionen des Pakets „effects" können wir die vom Modell geschätzten Werte visualisieren. Wir demonstrieren die Mächtigkeit des Pakets anhand eines Beispiels, in dem wir die Korrektheit der Antwort auf lexikalische Entscheidungsaufgaben vorhersagen. Da Korrektheit eine binäre Variable („correct" vs. „incorrect") ist, ist die logistische Regression (Pampel 2000) ein geeignetes statistisches Instrument. Eine solche fordern wir mit der Funktion glm() für „generalisiertes lineares Modell" und dem Parameter family = „binomial" an (vgl. Field 2012: 329f). Als Prädiktoren nehmen wir die Häufigkeit sowie die semantische Klasse des Wortes auf sowie die Interaktion zwischen den beiden Variablen. In den ersten beiden Code-Zeilen werden zunächst die Variablen Correct[15] und Frequency[16] vorbereitet.

```
> lexdec$Correct <- lexdec$Correct == "correct"
> lexdec$c.Frequency <- scale(lexdec$Frequency, scale = F)
> mod2 <- glm(Correct ~ c.Frequency * Class,
              data = lexdec, family = "binomial")
> plot(allEffects(mod2),
    multiline = T, ci.style = "bands",
    xlab = "Häufigkeit",
    ylab = "Geschätzte Wahrscheinlichkeit für korrekte Antwort",
    main = "Interaktionsplot Häufigkeit X Semantische Klasse")
```

Abb. 6 zeigt die grafische Ausgabe des letzten Aufrufs. Zuerst werden über die Funktion allEffects() alle Effekte höchster Ordnung[17] aus dem statistischen Modell mod2 extrahiert. Durch die Verschachtelung der Funktionen wird das

15 Da R die Stufen einer Variable alphabetisch sortiert, ist für die Variable Correct die erste Stufe „correct" und die zweite Stufe „incorrect". In der Vorhersage würde ein positives Vorzeichen somit bedeuten, dass sich die Wahrscheinlichkeit einer falschen Antwort erhöht. Da das nicht gerade intuitiv ist, drehen wir im ersten Aufruf die Stufen der Variable um („correct" wird zu TRUE, logisch wahr, und „incorrect" wird zu FALSE, logisch falsch).

16 Wann immer eine kontinuierliche Variable in eine Interaktion eingeht, sollte diese zentriert werden (Verschieben des Mittelwerts auf 0). Das geschieht mit dem zweiten Aufruf.

17 „Höchster Ordnung" bedeutet hier, dass bei interagierenden Variablen nur der Interaktionseffekt extrahiert wird, nicht die isolierten Effekte der Variablen. Würde das

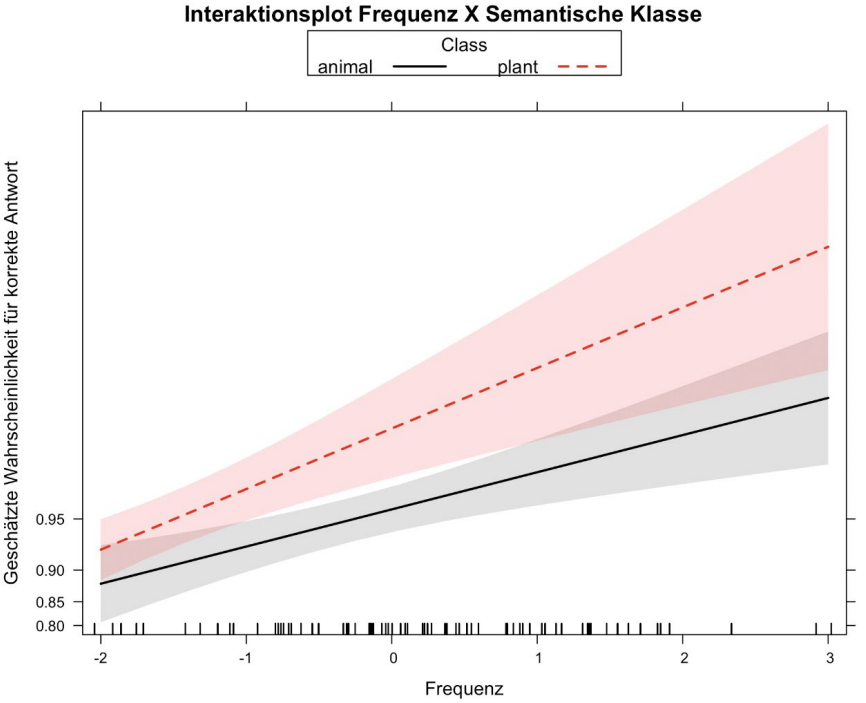


Abb. 6: Interaktionsplot für die geschätzte Wahrscheinlichkeit einer korrekten Antwort in einer lexikalischen Entscheidungsaufgabe in Abhängigkeit von Worthäufigkeit und semantischer Klasse. Fehlerindikatoren symbolisieren 95 %-Konfidenzintervalle.

Ergebnis dieser Extraktion direkt an plot() übergeben.[18] Dieser Aufruf enthält einige zusätzliche Argumente. xlab, ylab und main steuern lediglich die Achsen- und Titelbeschriftungen. multiline = T sorgt dafür, dass beide Linien für die Stufen der interagierenden Variable Class in einem Plot abgetragen werden und nicht – wie es in der Standardeinstellung der Fall wäre – in zwei separaten Panels. ci.style = „bands“ bewirkt, dass um diese Linien Konfidenzintervalle gezeichnet werden. Auffällig ist die Beschriftung der y-Achse, auf der die tatsächlich geschätzten Wahrscheinlichkeiten abgetragen werden. Durch die Transformationen der Kriteriumsvariable, die einer logistischen Regression inhärent

Modell noch einen dritten Prädiktor enthalten, der nicht in eine Interaktion eingeht, würde dieser ebenfalls extrahiert.

18 Intern erkennt R, dass es sich bei dem an plot() übergebenen Objekt um ein effects-Objekt handelt, und benutzt daher plot.eff(). Die Hilfeseite zu dieser Funktion (?plot.eff) enthält Erklärungen zu den übergebenen und einigen anderen Parametern.

sind, müssen die Abstände zwischen den Skalenstrichen, die die geschätzte Wahrscheinlichkeit angeben, entsprechend angepasst werden. Diese Anpassung sowie die Rücktransformation in einfacher interpretierbare Wahrscheinlichkeiten werden vom Zusatzpaket „effects“ übernommen.

Die Interaktion der beiden Variablen Class und c.Frequency ist nicht signifikant. Dies kann man über einen Aufruf von summary(mod2) sehen: Die Werte für die Interaktion lauten $\beta = 0{,}342$; $SE = 0{,}242$; $z = 1{,}413$ und $p = 0{,}158$. Der Unterschied der beiden Geraden-Steigungen in Abb. 6 ist also nicht groß genug, um anzunehmen, dass die Häufigkeit bei Tiernamen einen anderen Einfluss auf die Korrektheit hat als bei Pflanzennamen.

Abb. 6 enthält außerdem einen sogenannten „rug plot“. Die kleinen Striche am unteren Rand des Plotbereichs geben an, an welchen Stellen auf der x-Achse (also bei welchen Häufigkeiten) sich tatsächlich Datenpunkte befinden. Daraus gewinnen wir einerseits einen Eindruck von der Häufigkeitsverteilung der im Experiment verwendeten Wörter. Andererseits hilft es uns einzuschätzen, ob dem statistischen Modell über das komplette Spektrum hinweg ausreichend Datenpunkte zugrunde lagen, um eine Schätzung vorzunehmen. Würde sich im Rug-Plot beispielsweise zeigen, dass nur sehr seltene und sehr häufige Wörter in der Datenbasis vorhanden sind, könnte das auf ein Problem hindeuten. Im Falle von Abb. 6 scheinen über das komplette Häufigkeitsspektrum hinweg Datenpunkte vorhanden zu sein.

Wir fassen also für diesen Abschnitt zusammen: Wollen wir die Ergebnisse eines statistischen Modells (bspw. eines linearen Modells) visualisieren, sollten wir nicht auf Rohmittelwerte zurückgreifen, sondern die Schätzwerte aus dem Modell selbst extrahieren. Verschiebungen, die sich durch die Aufnahme anderer Prädiktoren ergeben, werden so auch in die Visualisierung übernommen. Das Paket „effects“ stellt Funktionen bereit, die die Extraktion und Visualisierung der Modellschätzer übernehmen. Dies ist unter anderem auch für die komputational aufwändigeren gemischten Modelle der Fall. Die Plots sind auf vielfältige Weise über Parameter anpassbar. Wenn für eine Veröffentlichung ein ganz bestimmtes Format an Schaubildern nötig ist, kann man auch die extrahierten Werte in eigens erstellte Plots übernehmen.

3. Alternative Plot-Pakete

Mitte Mai 2016 waren allein auf dem Comprehensive R Archive and Network[19] (CRAN) ca. 8300 Zusatzpakete für R verfügbar. Darin sind Pakete aus anderen Repositorien wie bspw. Bioconductor[20] noch nicht mit eingerechnet. Unter dieser großen Menge befinden sich natürlich auch einige Pakete, die speziell auf die Erzeugung von Grafiken, Schaubildern und Visualisierungen ausgerichtet sind. Der CRAN Task View zu „Graphics"[21] ist ein sehr guter Ausgangspunkt, um sich einen Überblick zu verschaffen. Wer sich für die Visualisierung von räumlichen Daten (bspw. auf Landkarten) interessiert, findet außerdem ausführliche Informationen auf dem Task View zu „Spatial Data".[22] In diesem Abschnitt werden wir uns auf die Vorstellung zweier Pakete beschränken. Das erste, „ggplot2", ist relativ prominent und breit genutzt. Das zweite, „rCharts", befindet sich noch in der Entwicklungsphase und wird noch nicht von vergleichbar vielen Personen benutzt.[23] Es scheint uns aber insofern interessant zu sein, als es eine Verbindung herstellt zwischen R und der JavaScript-Visualisierungsbibliothek d3.js, „Data Driven Documents"[24], die in jüngster Zeit an Bedeutung gewonnen hat.

3.1 „Grammar of Graphics": Das Paket „ggplot2"

Hadley Wickham erklärt die Grundidee jeglicher Grafikerstellung folgendermaßen: „[A] statistical graphic is a *mapping* from *data* to *aesthetic attributes* (colour, shape, size) of *geometric objects* (points, lines, bars)" [Hervorhebungen SW und SHW] (Wickham 2009: 3). In diesem Zitat sind alle relevanten Konzepte, die in „ggplot2" zum Einsatz kommen, bereits angesprochen. Die *Daten* setzen sich aus unseren Beobachtungen, den Variablen im Datensatz sowie deren Merkmalsausprägungen zusammen. Über den Abbildungsprozess *(mapping)* werden diese Daten dann auf die ästhetischen Attribute der *geometrischen Objekte* (in „ggplot2"-Terminologie „geoms") abgebildet. Wie Eugster und Scheipl in einer Präsentation[25] anmerken, ist die *Grammar of Graphics* von Wilkinson (2005)

19 https://cran.r-project.org/web/packages (letzter Zugriff am 13. Mai 2016).
20 https://www.bioconductor.org (letzter Zugriff am 13. Mai 2016).
21 https://cran.r-project.org/web/views/Graphics.html (letzter Zugriff am 13. Mai 2016).
22 https://cran.r-project.org/web/views/Spatial.html (letzter Zugriff am 13. Mai 2016).
23 Dies sind unsere Einschätzungen. Die tatsächliche Nutzung eines Pakets lässt sich nur schwer operationalisieren. Insbesondere wenn das Paket nicht auf CRAN verzeichnet ist, wie im Falle von „rCharts".
24 https://d3js.org/ (letzter Zugriff am 13. Mai 2016).
25 http://www.statistik.lmu.de/~scheipl/downloads/grafiken-05-ggplot2.pdf (letzter Zugriff am 13. Mai 2015).

keine „Anleitung, welche Grafik zu erzeugen ist, um eine konkrete Fragestellung zu untersuchen“, und keine „Spezifikation, wie eine statistische Grafik ausschauen sollte“. Vielmehr ist diese Grammatik als ein „formales Regelwerk“ zu verstehen, „welches die Zusammenhänge zwischen allen Elementen einer (gängigen) statistischen Grafik beschreibt“ (siehe Folie 6 der Präsentation von Eugster und Scheipl).

Die Funktionsweise von „ggplot2“ soll anhand eines einfachen Beispiels erläutert werden. Wir benutzen dazu den Datensatz „ratings“ aus dem Paket „languageR“, der subjektive Einschätzungen (z. B. das Gewicht) zu 81 Items (bspw. *ant*, *apple*, *woodpecker*) enthält (für nähere Erläuterungen siehe ?ratings).

```
> library(languageR)
> library(ggplot2)
> ggplot(data = ratings) +
        aes(x = Frequency, y = meanWeightRating, col = Class) +
        geom_point() +
        geom_smooth() +
        geom_rug()
```

Ergebnis des Aufrufs ist die in Abb. 7 abgedruckte Grafik. Man erkennt im Aufruf die verschiedenen Elemente des obigen Zitats wieder: In der ersten Zeile wird der Datensatz spezifiziert. In der Funktion aes() werden die ästhetischen Abbildungen vorgenommen. Dabei wird die x-Achse mit der Korpushäufigkeit, die y-Achse mit der durchschnittlichen Einschätzung des Gewichts und die Farbe der Datenpunkte mit der semantischen Klasse („animal“ vs. „plant“) belegt. Zuletzt werden noch drei geeignete geoms hinzugefügt. Die Datenpunkte selbst mit geom_point(), die Anpassungslinien mit geom_smooth() sowie die Rug-Plots mit geom_rug(). Zur Interpretation von Rug-Plots, siehe Abschnitt 2.3. Die verschiedenen Elemente werden mit dem Zeichen + verknüpft[26].

„ggplot2“ übernimmt einige Dinge selbst: Die Legende wird automatisch erzeugt und auch die Anpassungslinien werden automatisch für die beiden farblich unterschiedenen Gruppen getrennt erzeugt. Würden wir nicht nach Farbe gruppieren, wäre nur eine Anpassungslinie für die komplette Punktwolke zu sehen. Die geoms übernehmen die jeweils geeigneten Zuordnungen aus dem Aufruf von aes() – deshalb sind auch die Linien der Rug-Plots nach Gruppen eingefärbt. So kann man relativ leicht erkennen, dass sich die beiden Gruppen auf der y-Achse kaum mischen. Das Gewicht von Tieren wird offenbar systematisch als höher eingeschätzt als das von Pflanzen/Früchten. Die Anpassungslinien

26 Die Bedeutung von + ist hier nicht als arithmetische Operation (Addition) zu verstehen und auch nicht im Sinne eines Formeloperators wie im lm()-Aufruf in Abschnitt 2.3.

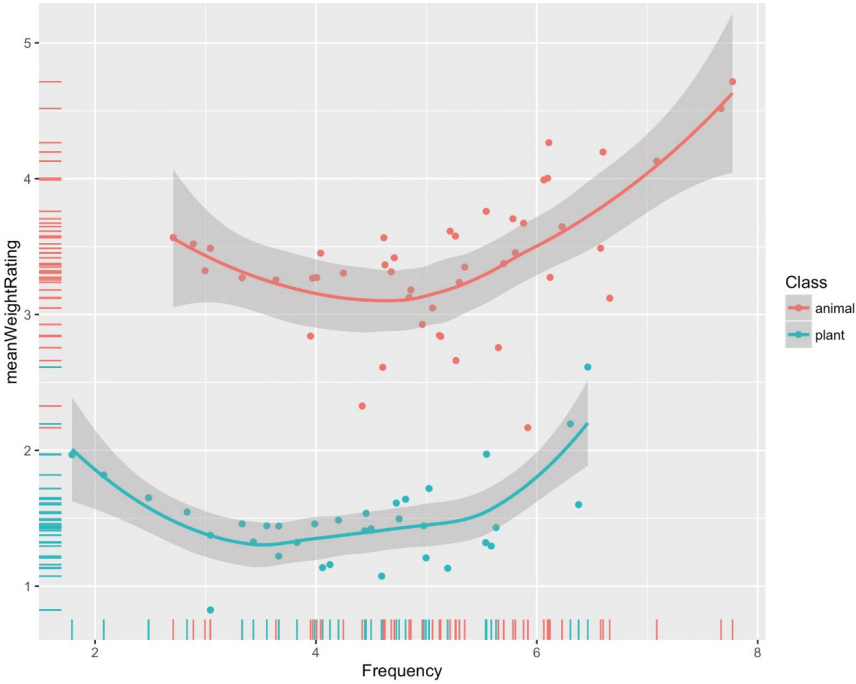


Abb. 7: Mit dem Paket „ggplot2“ erstellte Grafik, die den Zusammenhang zwischen der Häufigkeit eines Wortes, dessen semantischer Klasse und der durchschnittlichen Einschätzung des Gewichts des bezeichneten Tieres bzw. der Pflanze visualisiert. Eingefügt wurden außerdem Anpassungslinien für beide Gruppen sowie Rug-Plots für die x- und y-Werte.

deuten auf einen nicht-linearen, u-förmigen Zusammenhang zwischen Worthäufigkeit und durchschnittlicher Schätzung des Gewichts hin, was natürlich noch statistisch abzusichern wäre.

Grafiken, die mit „ggplot2“ erstellt wurden, sehen meist schon in der Standardausführung sehr ansprechend und „aufgeräumt“ aus und sind insbesondere zu explanatorischen Zwecken gut geeignet, bspw. zur Verwendung in Veröffentlichungen. Eine Grafik wie Abb. 5 mit dem Basisgrafikpaket von R zu erstellen, ist zwar nicht unmöglich, es bräuchte aber sicherlich deutlich mehr Code-Zeilen als mit der Erstellung mit „ggplot2“.

Wir konnten hier nur einen kleinen Eindruck vom Paket „ggplot2“ vermitteln und versuchen, die Grundidee zu verdeutlichen. Es gibt eine Reihe sehr guter Online-Tutorials für „ggplot2“. Empfohlen sei hier ein ausführliches Tutorial, das vom Institute for Quantitative Social Science der Harvard University

bereitgestellt wird[27]. Außerdem bietet die mit dem Paket assoziierte Dokumentationsseite[28] u. a. einen Überblick über alle geoms, die von dem Paket unterstützt werden.

3.2 Interaktive Grafiken mit d3: Das Paket „rCharts“

Ein weiteres Paket, das wohl eher explanatorisch als explorativ ausgerichtet ist, ist das Paket „rCharts“[29]. Es scheint geeignet zu sein, in Zukunft die Lücke zwischen R und der JavaScript-Bibliothek Data Driven Documents (d3) schließen zu können. Interessant ist d3 unter anderem aufgrund der Möglichkeit, interaktive Grafiken zu erstellen. „rCharts“ ist momentan noch nicht über den zentralen R-Paketverteiler CRAN verfügbar, sondern nur über GitHub.[30] Die Installation ist nicht sonderlich schwierig, wenn man zuerst das R-Paket „devtools“ installiert. „devtools“ enthält eine Funktion, mit dem man von GitHub R-Pakete installieren kann.[31] Mit der folgenden Sequenz von Befehlen sollte „rCharts“ auf dem eigenen Rechner verfügbar sein:

```
> install.packages("devtools")
> library(devtools)
> install_github('rCharts', 'ramnathv')
> library(rCharts)
```

Ein warnendes Wort vorweg: Mit „rCharts“ können zwar ansprechende interaktive Grafiken erstellt werden, da sich das Paket noch in der Entwicklung befindet, ist die Dokumentation jedoch noch recht lückenhaft. Das heißt, dass man entweder grundlegende Kenntnisse in JavaScript besitzen muss, um mit dem Paket gute Ergebnisse zu erzielen, oder etwas experimentieren muss, bis man zum gewünschten Ergebnis kommt. Wer es aber schließlich geschafft hat, einen Plot zu erstellen, wird mit einem visuell ansprechenden interaktiven Plot „belohnt“. Die Ergebnisse der folgenden Aufrufe sind lediglich über Links verfügbar, da „rCharts“ HTML-Dateien erstellt, deren interaktive Elemente im gedruckten

27 http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html (letzter Zugriff am 13. Mai 2016).

28 docs.ggplot2.org (letzter Zugriff am 13. Mai 2016).

29 Die mit dem Paket assoziierte Website ist unter rcharts.io (letzter Zugriff am 13. Mai 2016) erreichbar.

30 GitHub ist ein Webhosting-Service, auf dem man u. a. Code für R-Pakete bereitstellen und versionieren kann.

31 Da sich, wie erwähnt, „rCharts“ noch in der aktiven Entwicklung befindet, kann sich die Installationsprozedur eventuell ändern.

Text selbstverständlich nicht dargestellt werden können. Wir werden mit einem Beispiel beginnen, mit dem wir das in Abschnitt 3.1 in Abb. 7 gezeigte Diagramm replizieren werden.

```
> library(rCharts)
> library(languageR)
> pl <- nPlot(x = "Frequency", y = "meanWeightRating", group = "Class",
data = ratings, type = "scatterChart")
> pl$params$width <- 1000
> pl$params$height <- 600
> pl$yAxis(axisLabel = 'Mittleres geschätztes Gewicht')
> pl$xAxis(axisLabel = 'Häufigkeit')
> pl$chart(showDistY = 'true')
> pl$chart(showDistX = 'true')
> pl$chart(color = c("orange", "blue"))
> pl$chart(tooltipContent = "#! function (a, b, c, data) {
    return data.point.Word
           } !#")
> pl
```

Das Ergebnis dieses Aufrufs ist eine HTML-Datei, die ein interaktives Diagramm enthält und unter https://doi.org/10.11588/data/6SO6TG verfügbar ist (Datei „pl.html"). Was genau geschieht im obigen Aufruf? Zunächst laden wir die nötigen Pakete. Dann wird ein Plot mit der Funktion nplot() erstellt. Die Funktion nplot() nutzt die d3-Bibliothek „NVD3"[32] und kann mehrere Arten von Visualisierungen erstellen. Wir wählen hier den Typ „scatterChart". Außerdem müssen wir die Variablen für x- und y-Achse sowie den Datensatz angeben. Die Gruppierungsvariable ist fakultativ und kontrolliert hier u. a. die Farbe der Datenpunkte. Im Folgenden werden einige Eigenschaften des in der Variable „pl" gespeicherten Plots verändert: zuerst die Abmessungen (width, height), dann die Beschriftungen der Achsen (axisLabel). Über die beiden nächsten Zeilen (showDistX/Y) aktivieren wir die Option, dass die Verteilung der Datenpunkte an der x- und y-Achse über Rug-Plots dargestellt wird. Die Kolorierung der Datenpunkte für die beiden semantischen Klassen wird in der nächsten Zeile in Orange und Blau geändert. Die vorletzte Zeile enthält einen Aufruf, der den Inhalt des Tooltips steuert. Dieser wird angezeigt, wenn die Benutzerin/der Benutzer mit der Maus über einen Datenpunkt fährt. In diesem Fall soll das konkrete Wort angezeigt werden. Diese Zeile enthält schon ein wenig JavaScript-Code. In der letzten Zeile wird der Plot nur noch „ausgeführt" und im HTML-Viewer von RStudio angezeigt.

32 http://nvd3.org/ (letzter Zugriff am 17. Mai 2016).

Wo liegen nun die interaktiven Elemente dieser Visualisierung? Zunächst können oben links mit einem Klick auf den Farbpunkt neben „plant“ und „animal“ die Gruppen aktiviert und deaktiviert werden. Deaktiviert man eine der beiden Gruppen, verschwinden die Datenpunkte und die Skalierung der Achsen wird neu auf die dargestellten Datenpunkte ausgerichtet. Über die Aktivierung der „Magnify“-Option oben links wird in einen „Fischaugenmodus“ umgeschaltet, in dem über die Maus gesteuert werden kann, welche Bereiche des Graphen vergrößert dargestellt werden. Die Achsen werden hier dynamisch skaliert.

Einen weiteren Mehrwert der interaktiven Visualisierung erkennt man, wenn man mit der Maus über einzelne Datenpunkte fährt. Hier geschehen mehrere Dinge. Erstens wird ein Tooltip eingeblendet, in dem das tatsächlich bewertete Wort dargestellt wird. Das ist insbesondere hinsichtlich der Menge der im Schaubild dargestellten Informationen interessant. Man stelle sich eine statische Visualisierung vor, in der zu jedem Datenpunkt das Wort annotiert ist (auch das ist selbstverständlich in R möglich). Das Schaubild droht dann recht schnell unübersichtlich zu werden. Durch das interaktive Element des Tooltips kann die/der Betrachtende selektiv diese Information anfordern.

Fährt man mit der Maus über einzelne Datenpunkte, werden außerdem senkrechte und horizontale Geraden zu den Achsen gezeichnet und die Werte des Punktes auf den Achsen dargestellt. So lässt sich für jeden Punkt der genaue Wert ablesen. Alle interaktiven Elemente werden mit einer Übergangsanimation dargestellt, was es visuell etwas ansprechender macht.

Ein zweites Beispiel soll eine Visualisierungsart zeigen, die in diesem Kapitel zwar schon angesprochen, aber noch nicht anhand eines Beispiels illustriert wurde: ein gruppiertes Balkendiagramm. Wir bedienen uns dazu des Beispieldatensatzes „dative“ aus dem Paket „languageR“. Dieser Datensatz enthält Informationen zu 3263 Realisierungen der englischen „double object“-Konstruktion, in der das Phänomen der „dative alternation“ wirkt („John gives Mary a book“ vs. „John gives the book to Mary“). Im ersten Fall ist der Rezipient der Geben-Handlung, Mary, als eine Nominalphrase (NP) realisiert, im zweiten Fall als eine Präpositionalphrase (to Mary, PP). Wir wollen die Verteilung der Fälle in diesem Datensatz bezüglich der semantischen Klasse des Verbs visualisieren. Diese ist 5-stufig codiert: a = abstract („give it some thought“ / „give some thought to it“), c = communication („tell me your name“ / „tell your name to me“), f = future transfer of possession („promise her some money“ / „promise some money to her“), p = prevention of possession („deny him the book“ / „deny the book to him“) und t = transfer of possession („give him a ring“ / „give the ring to him“). Wir visualisieren also zwei kategoriale Variablen in ihrem Zusammenhang: die Realisierung des Rezipienten (NP vs. PP) und die semantische Klasse des Verbs (a, c, f, p oder t) und interessieren uns für die Häufigkeit, mit der jede der

Kombinationen auftritt. Wir codieren zunächst die Variable SemanticClass um, um die Interpretierbarkeit des Schaubilds zu erleichtern.

```
> levels(dative$SemanticClass) <- c("abstract", "communication",
                "future transfer",
                "prevention of transfer", "transfer")
> table(dative$SemanticClass, dative$RealizationOfRecipient)
                            NP   PP
  abstract                1176  257
  communication            355   50
  future transfer           47   12
  prevention of transfer   225    3
  transfer                 611  527
```

Wir kennen nun schon die Häufigkeitsverteilung, die uns interessiert, und wollen diese jetzt mit rCharts interaktiv visualisieren.

```
> dative2 <- as.data.frame(table(dative$SemanticClass,
                      dative$RealizationOfRecipient))
> names(dative2) <- c("SemanticClass", "RealizationOfRecipient",
                "Freq")
> pl2 <- nPlot(Freq ~ RealizationOfRecipient,
               group = "SemanticClass", data = dative2,
               type = "multiBarChart")
> pl2$yAxis(tickFormat = "#! d3.format(',.0f')!#")
> pl2
```

In der ersten Zeile wird zunächst eine Datentabelle (ein „Dataframe“) erstellt, mit dem das Paket „rCharts“ umgehen kann. Das ist lediglich eine andere Darstellungsform der oben abgedruckten Kontingenztabelle. In diesem Dataframe legen wir zunächst geeignete Spaltenüberschriften mit der Funktion names() fest, dann folgt der eigentliche Plotaufruf, in dem wir die Häufigkeit in Abhängigkeit der Realisierung des Rezipienten gruppiert nach der semantischen Klasse abtragen. Der Typ des Plots ist in diesem Fall ein multiBarChart. Der Plot wird zunächst in der Variable pl2 gespeichert. Im letzten Schritt passen wir das Format der y-Achse so an, dass keine Nachkommastellen angezeigt werden, denn diese sind bei ganzzahligen Häufigkeiten unnötig und verwirren eher. Mit der letzten Zeile wird der Plot schließlich „ausgeführt“.

Das Ergebnis des Plotaufrufs ist wiederum eine HTML-Datei, die das interaktive Schaubild enthält und unter https://doi.org/10.11588/data/6SO6TG abrufbar ist (Datei „pl2.html“). Zunächst ist kein bedeutender Unterschied zu statischen

Balkendiagrammen feststellbar. Doch auch hier bieten die interaktiven Elemente einen Mehrwert: Wiederum kann man rechts oben in der Farb-Legende einzelne Gruppen an- und ausschalten, was gegebenenfalls zu einer Neuskalierung der y-Achse führt. Wenn man mit der Maus über einzelne Balken fährt, wird außerdem ein Tooltip angezeigt, der die genaue Anzahl der Fälle sowie die Kategorie enthält (bspw. „communication – 355 on NP"). Besonders hervorzuheben ist die Möglichkeit, von einem gruppierten auf ein gestapeltes Balkendiagramm umzuschalten. Hierzu muss lediglich oben links „stacked" statt „grouped" aktiviert werden. In der Visualisierung der „dative alternation"-Daten lässt sich mit einem gestapelten Balkendiagramm beispielsweise die Gesamtzahl der Fälle NP vs. PP einfacher vergleichen. Ein gruppiertes Balkendiagramm eignet sich dagegen besser für den Vergleich einzelner Gruppen.

Mithilfe des Pakets „rCharts" können wir noch weitere auf NVD3 basierte Grafiken direkt in R erstellen. Hierzu gehören gestapelte Flächendiagramme (stackedAreaChart), horizontale Balkendiagramme (multiBarHorizontalChart) sowie Tortendiagramme (pieChart) und Liniendiagramme ohne (lineChart) und mit einer speziellen Auswahlfunktion (lineWithFocusChart). Außerdem können noch andere auf d3 basierte Bibliotheken angesprochen werden (bspw. Polychart, Morris, xCharts und HighCharts). Einen relativ umfassenden Überblick über die Fähigkeiten von „rCharts" bietet die Projekthomepage[33] sowie diverse web-basierte Tutorials[34].

4. Schluss

Wir haben einen kleinen Ausschnitt der Visualisierungsmöglichkeiten gezeigt, die R bietet – sowohl mit explorativer als auch mit explanatorischer Ausrichtung. Unseren Visualisierungen lagen statistische Berechnungen zugrunde, die mit einer unterschiedlichen Anzahl und Art von Variablen arbeiteten. Überlegt man sich bereits im Vorfeld der Grafikerstellung, wie viele und welche Variablen visualisiert werden sollen, fällt die Wahl einer geeigneten Visualisierungsart meist deutlich einfacher.

Mit der Basisausstattung von R stehen den Benutzer/innen schon viele Visualisierungsmöglichkeiten zur Verfügung. Für Spezialaufgaben sind bestimmte Pakete jedoch besser geeignet, da sie der Benutzerin/dem Benutzer unter Umständen eine Menge Kodierungs- oder Programmierarbeit abnehmen. Wir konnten

33 http://ramnathv.github.io/rCharts (letzter Zugriff am 19. Mai 2016).

34 Unter http://www.rpubs.com/dnchari/rcharts (letzter Zugriff am 20. Mai 2016) ist eine Sammlung von Programmcode für einige Schaubilder verfügbar, die mit rCharts erstellt werden können.

hier nur einen kleinen Ausschnitt der verfügbaren Pakete vorstellen, haben aber versucht, eine Auswahl zu treffen, die in vielen Bereichen der linguistischen Forschung relevant sein können. Dabei ist die Visualisierung kategorialer Daten besonders interessant (Paket „vcd“). Aber auch die Möglichkeit, mit statistischen Modellen gewonnene Schätzer auf einfache Weise zu extrahieren und ohne Mehraufwand zu visualisieren, ist äußerst relevant – spätestens in Veröffentlichungen, in denen die Ergebnisse der statistischen Modelle und nicht einfach die Rohwerte dargestellt werden sollen (Paket „effects“). Zuletzt haben wir zwei Pakete vorgestellt, die als komplette Umgebungen für die Erstellung von Grafiken gesehen werden können und somit das Portfolio von R als mächtiges Visualisierungswerkzeug erweitern. Das Paket „ggplot2“ tritt dabei mit dem Anspruch an, eine „Grammar of Graphics“ zu implementieren. Das noch in der Entwicklung befindliche Paket „rCharts“ schließt die Lücke zur JavaScript-Bibliothek d3 und verknüpft R somit mit einer anderen mächtigen Visualisierungsumgebung.

5. Bibliografie

Baayen, R. Harald. 2011. languageR: “Data sets and functions with ‘Analyzing linguistic data: A practical introduction to statistics’” [Computer software manual]. Abgerufen von http://CRAN.R-project.org/package=languageR (R package version 1.4)

Bortz, Jürgen. 2005. *Statistik für Human- und Sozialwissenschaftler.* 6. Aufl. Heidelberg: Springer.

Cohen, Ayala. 1980. “On the graphical display of the significant components in a two-way contingency table.” *Communications in Statistics. Theory and Methods* 9: 1025–1041.

Cleveland, William S. 1981. “LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician* 35: 54.

Fahrmeir, Ludwig, Rita Künstler, Iris Pigeot und Gerhard Tutz. 2007. *Statistik: Der Weg zur Datenanalyse.* Berlin: Springer.

Field, Andy, Jeremy Miles und Zoë Field. 2012. *Discovering statistics using R.* Los Angeles: Sage.

Fox, John .2003. “Effect displays in R for Generalised Linear Models.” *Journal of Statistical Software* 8: 1-24.

Friendly, Michael. 1992. “Graphical methods for categorical data. SAS User Group International Conference Proceedings” 17: 190–200. http://www.math.yorku.ca/SCS/sugi/sugi17-paper.html (letzter Zugriff am 19.05.2016).

Friendly, Michael. 1994. “Mosaic displays for multi-way contingency tables.” *Journal of the American Statistical Association* 89: 190–200.

Fürbacher, Monica. 2015. "Variation zur starken Genitivmarkierung. Spezialstudie: Regionale Verteilung." https://dgd.ids-mannheim.de/korpusgrammatik/5087 (letzter Zugriff am 19. Mai 2016.).

Hansen, Sandra und Sascha Wolfer. 2016. „Standardisierte statistische Auswertung von Korpusdaten im Projekt ‚Korpusgrammatik' (KoGra-R)". In: *Grammatische Variation. Empirische Zugänge und theoretische Modellierung,* herausgegeben von Marek Konopka und Angelika Wöllstein. Berlin: de Gruyter, 345–356.

Hansen-Morath, Sandra, Roman Schneider, Hans-Christian Schmitz und Sascha Wolfer. In Vorbereitung. „KoGra-R: Standardisierte statistische Auswertung von Korpusrecherchen." In: *Grammatik im Korpus,* herausgegeben von Eric Fuß, Marek Konopka und Angelika Wöllstein.

Meyer, David, Achim Zeileis und Kurt Hornik. 2007. „The strucplot framework: Visualizing multi-way contingency tables with vcd." *Journal of Statistical Software* 17 (3): 1–48. http://www.jstatsoft.org/v17/i03/ und verfügbar als Vignette („strucplot", package = „vcd").

Meyer, D., A. Zeileis, A. und K. Hornik. 2015. "vcd: Visualizing categorical data" [Computer software manual]. (R package version 1.4-1).

Morales, M. und R Core Team. 2012. "sciplot: Scientific graphing functions for factorial designs" [Computer software manual]. Abgerufen von https://CRAN.R-project.org/package=sciplot (R package version 1.1-0).

Pampel, Fred C. 2000. *Logistic regression: A primer.* London: Sage.

R Core Team 2016. "R: A language and environment for statistical computing" [Computer software manual]. Vienna, Austria. Abgerufen von http://www.R-project.org/

Tatzreiter, Herbert. 1988. „Besonderheiten der Morphologie in der deutschen Sprache in Österreich." In: *Das österreichische Deutsch,* herausgegeben von Peter Wiesinger. Wien: Böhlau, 1988, 76.

Vaidyanathan, Ramnath. 2013. "rCharts: Interactive charts using javascript visualization libraries" [Computer software manual]. (R package version 0.4.5).

Wickham, Hadley. 2009. *ggplot2: Elegant graphics for data analysis.* Dordrecht: Springer.

Wilkinson, Leland. 2005. *The grammar of graphics.* 2. Aufl. New York: Springer.

Jan Oliver Rüdiger

CorpusExplorer v2.0 – Visualisierung prozessorientiert gestalten

Abstract Der CorpusExplorer v2.0 ist eine frei verfügbare Software zur korpushermeneutischen Analyse und bietet über 45 unterschiedliche Analysen/Visualisierungen für eigenes Korpusmaterial an. Dieser Praxisbericht gibt Einblicke, zeigt Fallstricke auf und bietet Lösungen an, um die tägliche Visualisierungsarbeit zu erleichtern. Zunächst wird ein kurzer Einblick in die Ideen gegeben, die zur Entwicklung des CorpusExplorers[1] führten, einer korpuslinguistischen Software, die nicht nur vielfältige Forschungsansätze unterstützt, sondern auch mit einem Fokus auf die universitäre Lehre entwickelt wird. Der Mittelteil behandelt einen der vielen Fallstricke, die im Entwicklungsprozess auftraten: Effizienz-/Anpassungsprobleme – bzw.: Was passiert, wenn Visualisierungen an neue Begebenheiten angepasst werden müssen? Da diese Lösung Teil des CorpusExplorers v2.0 ist, wird abschließend darauf eingegangen, wie unterschiedliche Visualisierungen zu denselben Datensätzen sich auf die Rezeption/Interpretation von Daten auswirken.

1. Von der Idee zur Anwendung

Die eigentliche Idee zur Entwicklung des CorpusExplorers entstand während meiner Magisterarbeit 2013. Zunächst wurden Mitarbeiter/innen und Doktorand/innen des *Instituts für Germanistik* an der *Universität Kassel* befragt, ob/wie sie korpuslinguistisch arbeiten, welche Tools eingesetzt werden und wie aktuelle Lösungen in konkreten Projekten aussehen. Fast alle nutzten korpuslinguistische Tools, mal mehr, mal weniger intensiv. Überraschend war die Vielzahl an Softwaretools, die in der Forschung kursieren. Auf Basis dieser Befragung wurden mehrere Workflows entwickelt, um die folgende Frage beantworten zu können: *„Was muss erledigt werden, um vom einfachen Text zu einem visuellen Endergebnis zu gelangen?"* Die einzelnen Arbeitsschritte lassen sich durch verschiedene Softwaretools erledigen. Das Endergebnis ist ein sogenannter *„Toolchain"* – also eine

1 Der CorpusExplorer kann über http://corpusexplorer.de kostenfrei bezogen werden.

Erschienen in: Noah Bubenhofer u. Marc Kupietz (Hrsg.): Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

Kette von Softwaretools, die in einer bestimmten Reihenfolge genutzt werden, um den Workflow vollständig abzudecken. Ein Workflow-Beispiel: Rohtexte bereinigen, annotieren, Frequenzen auszählen und abschließend als Grafik darstellen. Der exemplarische *Toolchain* könnte wie folgt aussehen: Texte werden mit *JEdit* (per Hand) bereinigt, danach mit dem *TreeTagger* annotiert, Frequenzen werden mit *AntConc* ausgezählt; für die Umsetzung als Diagramm wird schließlich *Microsoft Excel* verwendet[2]. Dieser *Toolchain* ist nur einer von vielen denkbaren Möglichkeiten, manche mögen kürzer, einfacher, detaillierter oder komplexer sein. Für alle ergeben sich aber ähnliche Probleme:

— Für alle Programme müssen Lizenzen erworben werden, insofern diese nicht kostenfrei verfügbar sind. Der Kostenfaktor spielt besonders dann eine Rolle, wenn die Software z. B. in einer Seminargruppe eingesetzt werden soll. Der oben aufgezeigte Toolchain ist in der Regel mit geringen Kosten realisierbar, da die Programme JEdit, TreeTagger und AntConc kostenfrei sind und Excel auf vielen Rechnern vorinstalliert/vorlizenziert ist. Ganz anders gestaltet sich aber der Fall, wenn z. B. Produkte wie MAXQDA, SPSS oder Tableau zum Einsatz kommen sollen – hier können schnell mehrere Hundert/Tausend Euro an Lizenzkosten fällig werden. Das Ausweichen auf kostenfreie Open-Source Lösungen ist erstrebenswert, erfordert aber eine sachkundige Auswahl und kann zur Verstärkung der in den folgenden Punkten aufgelisteten Problemen führen.
— Die Programme müssen untereinander kompatibel sein. Die Ausgabe des einen Programms ist wiederum die Eingabe des anderen. Wie die Umfrage ergab, ist dies eine häufig wahrgenommene Fehler-/Problemquelle. Oft wird diese durch selbstentwickelte/selbsterdachte Lösungen beseitigt. Diese Lösungen sind jedoch fragil, da z. B. im Falle der Aktualisierung eines der beiden Programme die individuelle Anpassung neu konfiguriert werden muss. Daher kann es notwendig sein, dass man in einer Projektgruppe / während einer Projektphase – genau überlegt, welche Programmversionen zum Einsatz kommen, dies abspricht und ggf. ältere Installationspakete bereithält, um ggf. auf die ältere Programmversion zurückwechseln zu können, wenn das Update Probleme bereitet. Im Idealfall wird der Toolchain in regelmäßigen Abständen vollständig auf einem Rechner getestet und erst dann auf alle Rechner des Projektteams/der Seminargruppe übertragen.

2 Auf die einzelnen Tools wird im Folgenden nicht weiter eingegangen. Die Bezugsquellen lauten: JEdit: http://bit.ly/1TOp23W TreeTagger: http://bit.ly/1bsE7eE AntConc: http://bit.ly/1VrkiQY Microsoft Excel: http://bit.ly/1sXVQM4 MAXQDA: http://bit.ly/2qJ1bbT SPSS: https://ibm.co/2rK6hmJ Tableau: http://tabsoft.co/2rb1fCF .

— Die Anzahl der benötigten Softwareprogramme wird schnell zweistellig. Wie bei einer Kette so gilt auch für den Toolchain – das Endresultat ist nur so stark wie das schwächste Glied. Wird z. B. eine Software nicht mehr weiterentwickelt, muss ggf. der gesamte Toolchain umstrukturiert werden. Dieses Problem betrifft sowohl kommerzielle als auch Open-Source-Software. Kleinere/Kurzfristige Projekte können dieses Problem vernachlässigen, größere/mehrjährige Projekte sind gut beraten, alle Softwareprodukte im Toolchain auf Ausfallmöglichkeiten zu überprüfen (z. B. keine Open-Source Software zu nutzen, die bereits bei Projektbeginn seit mehr als einem Jahr nicht mehr aktiv entwickelt wird), abzusichern (z. B. eine Kopie des Quellcodes anzulegen) oder Alternativen bereitzuhalten.
— Durch die große Anzahl an Software wird es zudem schwerer, dieses Wissen in die Lehre zu übertragen. Für jede zusätzliche Software müssen extra Handreichungen für die Studentinnen und Studenten geschrieben werden. Alle Programme müssen auf den Rechnern in der gleichen Version vorliegen. Unterschiedliche Programmversionen führen gelegentlich zu ganz anderen Ergebnissen (siehe oben – Updateproblem). Dozentinnen und Dozenten müssen zudem für eine Vielzahl an Programmen Hilfestellungen leisten können.

Diese Problemfelder, gerade auch was den Einsatz in der universitären Lehre anbelangt, lassen sich z. B. auch bei Bubenhofer (2011), Dipper (2011) und Zinsmeister (2011) wiederfinden. Daher ist davon auszugehen, dass diese Probleme nicht spezifische Probleme der Umfrageteilnehmer/innen sind, sondern einen weit größeren Nutzerkreis betreffen. Darauf aufbauend wurde eine Lösung, der CorpusExplorer, entwickelt. Der CorpusExplorer nutzt ausschließlich freie Software (Freeware/Open Source) – Installation und Konfiguration erfolgen vollautomatisch. Die Softwareauswahl, auf die der CorpusExplorer v2.0 zurückgreift, erfolgt nach den oben angegebenen Kriterien, d. h. Nutzung von Open-Source-Software (die aktiv weiterentwickelt wird), Toolchain-Test, bevor die Software verteilt wird und es stehen mehrere alternative Möglichkeiten bereit (z. B. existiert noch eine ganze Reihe alternativer Tagger, wenn einer der Tagger nicht mehr weiterentwickelt werden sollte. Dies erlaubt es zusätzlich, auch auf individuelle Nutzerpräferenzen einzugehen). Damit sind die Lizenz-, Konfigurations- und Kompatibilitätsprobleme aus Nutzersicht weitestgehend gelöst und die Forderung Zinsmeisters (2011, 72), dass „[für] den Einsatz in der Lehre ‚Download and Run'-Ressourcen, bei denen die Ressource als nutzungsfähiges Gesamtpaket bezogen werden kann […] vorzuziehen" sind, ist erfüllt.

Der CorpusExplorer bietet einen denkbar einfachen Toolchain, da er alle automatisierbaren Aufgaben mittels einer intuitiven Programmoberfläche löst: Programm starten, Korpus importieren und über 45 Visualisierungen nutzen.

Aufbereitung, Bereinigung, Chunking, Annotation, Abtrennen von Metadaten uvm. werden im Importprozess vollautomatisch durchgeführt. Dadurch wird es möglich, ein weiteres Problem zu lösen, den Einsatz in der Lehre – ein Programm, eine Erklärung – Dozentinnen und Dozenten können sich wieder auf die Linguistik fokussieren.

2. Plädoyer für dreistufige Visualisierungsprozesse

Visualisierungen sind meist das Sahnehäubchen einer wissenschaftlichen Publikation. Sie schmücken ein Paper aus oder fassen die detaillierten Ergebnisse optisch klar und verständlich auf einem Poster zusammen. Manchmal sind sie auch selbst Gegenstand der Forschung, dann geht es um Fragen der Ästhetik oder Wahrnehmung. Diese Aspekte der Informationsvisualisierung werden aber im Folgenden bewusst ausgeklammert, zum einen, weil sie im Gesamtkontext des Tagungsbandes mehrfach diskutiert werden, und zum anderen, weil dies den Artikelumfang sprengen würde. Diese Diskussion soll daher um einen völlig neuen Aspekt erweitert werden: Die Effizienz des Visualisierungsprozesses. Konkret bedeutet dies, Abläufe zu schaffen, die die Visualisierung erleichtern, da sie prozessorientiert entweder unterschiedliche Daten gleichartig aufbereiten, um so identische Visualisierungen zu befüllen, oder für gleichförmige Daten eine Vielzahl an Visualisierungen bereitstellen.

Der einfachste denkbare Ablauf ist ein zweistufiger Prozess. Damit es im späteren Verlauf zu keiner Verwirrung kommt, benenne ich beide Prozessschritte gleich so, wie sie später benötigt werden. Die beiden Prozesskomponenten lauten: „*View*“ und „*Model*“[3]. Dabei stellt das Model die eigentlichen Daten bereit. Das Model kann z. B. eine einfache CSV-Tabelle sein, die die Daten in strukturierter Weise enthält, oder ein Korpus, das ausgewertet wird. Die *View* ist die Visualisierung. In dieser simplen Form greift die *View* direkt auf das *Model* zu. Dieser zweistufige Prozess ist zweifelsfrei der einfachste und schnellste Weg, Daten in eine visuelle Form zu pressen. Der wesentliche Nachteil dieser Lösung ist die geringe Nachhaltigkeit. Alle Abfragen, Aggregationen und Funktionen wie Sortieren, Filtern oder Gruppieren werden der einen oder anderen Seite zugeschlagen. Verändern sich die Daten im Model strukturell, muss die *View* zwingend angepasst werden. Umgekehrt verhält es sich ähnlich – *View* und *Model* sind direkt miteinander verwoben. Dadurch können beide nur mit hohem Aufwand gegeneinander ausgetauscht werden. Doch dies ist wichtig, wenn gleichartige Visualisierungen auf unterschiedliche Daten oder

3 Beide Begriffe rekurrieren auf die Disziplin, aus der sie stammen, die Softwaretechnik. Für eine Einführung sei insbesondere Eilebrecht und Starke (2013) empfohlen.

unterschiedliche Visualisierungen auf gleiche Daten angewendet werden sollen. Dieses Problem hatte die erste Version des CorpusExplorers. Daher erfolgt das Plädoyer für einen mehrstufigen Visualisierungsprozess nicht aus einer Laune heraus, sondern basiert auf der Erkenntnis, der eine schmerzliche Erfahrung zugrunde liegt.

Der wesentliche Änderungsschritt ist das Hinzufügen einer zusätzlichen Zwischenebene. Softwaretechnisch könnte man darüber streiten, ob diese als *„Controller"*, *„Presenter"* oder *„ViewModel"* usw. bezeichnet wird[4]. Essenziell muss man aber nur begreifen, dass eine zusätzliche Ebene einige erhebliche Vorteile mit sich bringt. Wie bereits ausgeführt, ist bei einem zweistufigen Prozess nicht immer klar, wo und wie die eigentlichen Abfragen zu realisieren sind. Erst Abfragen machen aus den Daten eine visualisierbare Menge. In einem dreistufigen Prozess lässt sich diese Frage jedoch wesentlich klarer beantworten. Das Model stellt ausschließlich die reinen Rohdaten bereit, die View sorgt ausschließlich für die Darstellung, alles andere wird in die Zwischenebene verlagert. Dadurch wird es möglich, aus unterschiedlichen *Models* Daten zu beziehen, diese zu verbinden und in neuer, aggregierter Form an die View weiterzureichen.

Bei der Entwicklung des CorpusExplorer fiel letztlich die Entscheidung auf das sogenannte *„MVVM"*-Entwurfsmuster[5] *(„Model, View, ViewModel").* Überraschenderweise konnte ich bisher keine konkreten Aussagen in der Literatur finden, die sich mit Effizienzgewinnen einer derartigen Umstrukturierung befassen. Daher habe ich versucht, eine Abschätzung auf Basis der Historie des CorpusExplorers zu entwickeln. Zum jetzigen Zeitpunkt verfügt der CorpusExplorer über exakt 47 Visualisierungen. Ein zweistufiger Prozess würde bedeuten, dass 47-mal alles neu implementiert werden müsste. Durch die Umstellung auf den dreistufigen Prozess benötigt der CorpusExplorer jedoch nur 29 Views, um denselben Funktionsumfang abzudecken. Daher konnte allein durch diese simple Umstrukturierung 39,3% des anfallenden Programmieraufwands vermieden werden.

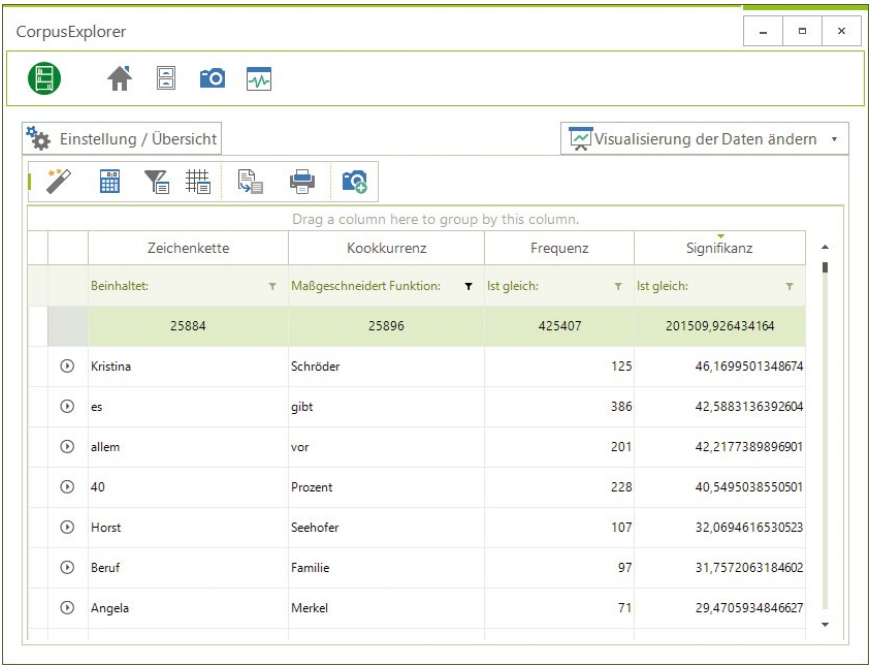
4 Konzeptionell gibt es zwischen diesen Zwischenschichtarten marginale Unterschiede. Diese hier zu diskutieren würde jedoch den Rahmen des Artikels sprengen.

5 Eine genau Definition des MVVM-Entwurfsmusters sowie eine Beispielimplementierung in C# findet sich bei Wegener und Schwichtenberg (2012, 583–609).

3. Identische Daten – Unterschiedliche Visualisierung

Kurz vorab: Für alle Beispiele wurde ein Korpus (1,28 Mio. Token) aus 2113 zufällig ausgewählten deutschsprachigen Zeitungsartikeln der Jahre 2010 bis 2015 zu den Stichworten Frauenquote/Quotenfrau verwendet. Alle Beispiele visualisieren das Resultat einer Kookkurrenz-Analyse. „Statistisch signifikante Kookkurrenzen sind Wortverbindungen, die überzufällig oft in einer bestimmten Datenbasis auftreten“ so Lemnitzer und Zinsmeister (2006, 147). Die erste und einfachste Form für die meisten Auswertungen ist die Darstellung in einer Tabelle. Im CorpusExplorer erlaubt die Tabellen-Ansicht das Sortieren, Filtern und Gruppieren der Daten. Abb. 1 zeigt die Tabellen-Ausgabe der Kookkurrenzanalyse.

Auch wenn Tabellen einen geringen visuell-ästhetischen Wert haben und keine Informationen raffen – im Vergleich zu anderen Visualisierungsformen haben sie jedoch einen eigenen Stellenwert im Visualisierungsprozess verdient. Tabellen sind der Überblick auf die Datengesamtheit. Durch Interaktivität können sie zudem sehr schnell zum gewünschten Analyseziel führen. Im Gegensatz zu anderen Programmen ermittelt der CorpusExplorer die Kookkurrenzen aller Token. Für die Signifikanz kann zwischen Poisson-Verteilung (Programmstandard), Chi-Quadrat-Test und Log-Likelihood gewählt werden. Nicht signifikante



Zeichenkette	Kookkurrenz	Frequenz	Signifikanz
Beinhaltet:	Maßgeschneidert Funktion:	Ist gleich:	Ist gleich:
25884	25896	425407	201509,926434164
Kristina	Schröder	125	46,1699501348674
es	gibt	386	42,5883136392604
allem	vor	201	42,2177389896901
40	Prozent	228	40,5495038550501
Horst	Seehofer	107	32,0694616530523
Beruf	Familie	97	31,7572063184602
Angela	Merkel	71	29,4705934846627

Abb. 1 Kookkurrenz-Tabelle im CorpusExplorer.

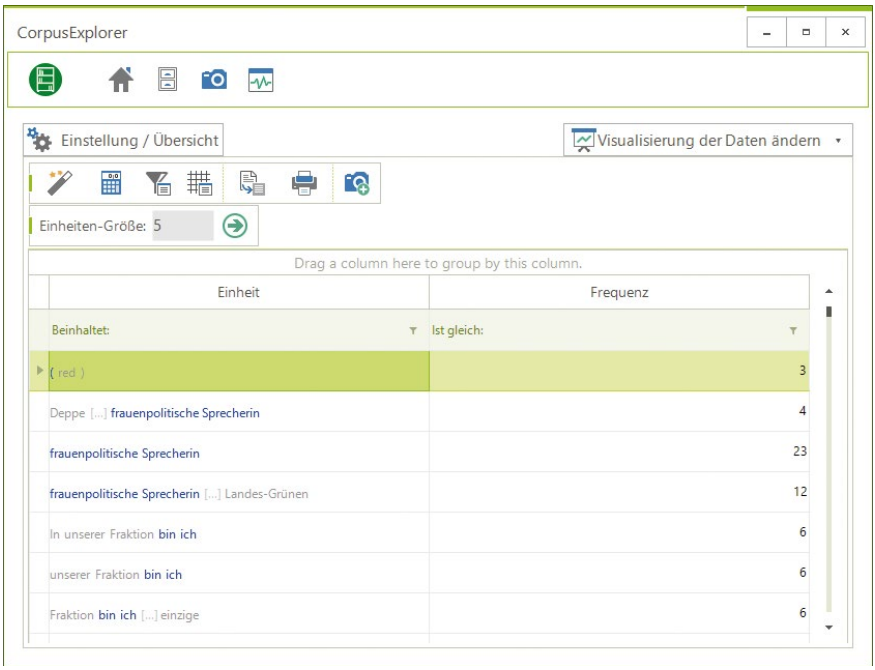


Abb. 2 N-Gramme mit Signifikanzwerten.

Kookkurrenzen werden automatisch gefiltert. Nutzerinnen und Nutzer können so entweder nach den signifikantesten Kookkurrenzpartnern suchen oder mithilfe der Filterfunktion Kookkurrenzen definierter Begriffe herausgreifen. Mit einer geringen Anpassung konnte aus dieser Auswertung eine weitere, weitaus differenziertere Darstellung entwickelt werden.

Abb. 2 zeigt eine weitere Tabelle. Die Idee entstammt dem Versuch, die Kookkurrenzanalyse aus COSMAS II[6] mit möglichst einfachen Mittel nachzubauen. Um diese Visualisierung zu realisieren, wurden zwei bereits vorhandene Auswertungen kombiniert. Hier spielt das *MVVM*-Konzept eine seiner größten Stärken aus, da bereits existierende Datenquellen leicht miteinander verknüpft werden können. Auf die Auswertung von N-Grammen[7] folgt eine Kookkurrenz-Analyse, diese bewertet, wie signifikant die Verbindungen einzelner Token innerhalb des

6 Online abrufbar über: http://www.ids-mannheim.de/cosmas2/

7 N-Gramme sind in der Zusammenfassung von Bubenhofer (2009, 118) wie folgt erklärt: „Das n steht für eine beliebige Zahl > 0; die Bezeichnung leitet sich von den Namen für Ein-, Zwei- oder Dreiwortausdrücke, ‚Unigramme', ‚Bigramme', ‚Trigramme', ab. Normalerweise werden n-Gramme nur als eine Reihe von direkt aufeinander folgenden Wörtern verstanden."

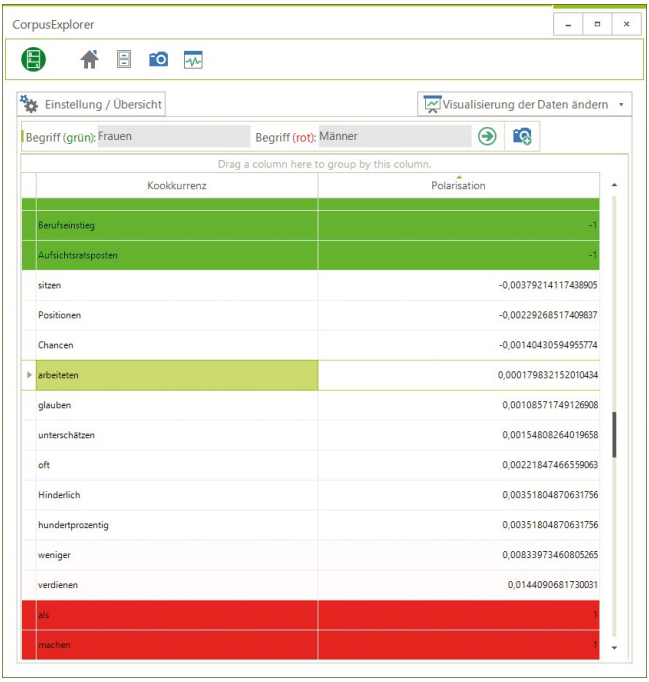


Abb. 3 Kontrastierte Kookkurrenzen. Grün = Frauen / Rot = Männer

N-Gramms sind. Je nach Signifikanzklasse erfolgt dann die Kolorierung. Blau steht für überdurchschnittlich signifikant, Schwarz für normal signifikant, Grau – unterer Grenzbereich der Signifikanz. Graue Auslassungszeichen stehen für nicht signifikante Token.

Abb. 3 zeigt die Kookkurrenzen zu lediglich zwei gewählten Begriffen, die gegeneinander kontrastiert werden.

Der erste Begriff (grün) lautet ‚Frauen', der zweite (rot) ‚Männer'. Diesmal werden jedoch anstelle der Gesamttabelle nur die zwei gewählten Begriffe in der View angezeigt. Dabei geht es um individuelle und überlappende Kookkurrenzen. Die Signifikanzwerte werden auf eine normierte Skala umgerechnet – die Normierung erfolgt im ViewModel. Alle Kookkurrenzen mit dem Wert gleich -1 gehören ausschließlich zum grünen Begriff – also ‚Frauen' – und sind folglich grün eingefärbt. Alle Begriffe mit dem Wert gleich +1 sind ebenso ausschließliche Kookkurrenzen zu ‚Männer'. Neben den ausschließlichen und damit individuellen Kookkurrenzen sind diejenigen besonders interessant, die zwischen diesen beiden Werten liegen – also im Wertbereich von -1 bis +1. Diese Begriffe sind signifikant zu beiden Begriffen. Durch das Vorzeichen ist die Tendenz zu einem der beiden Begriffe – man könnte auch sagen Pole – schnell identifizierbar.

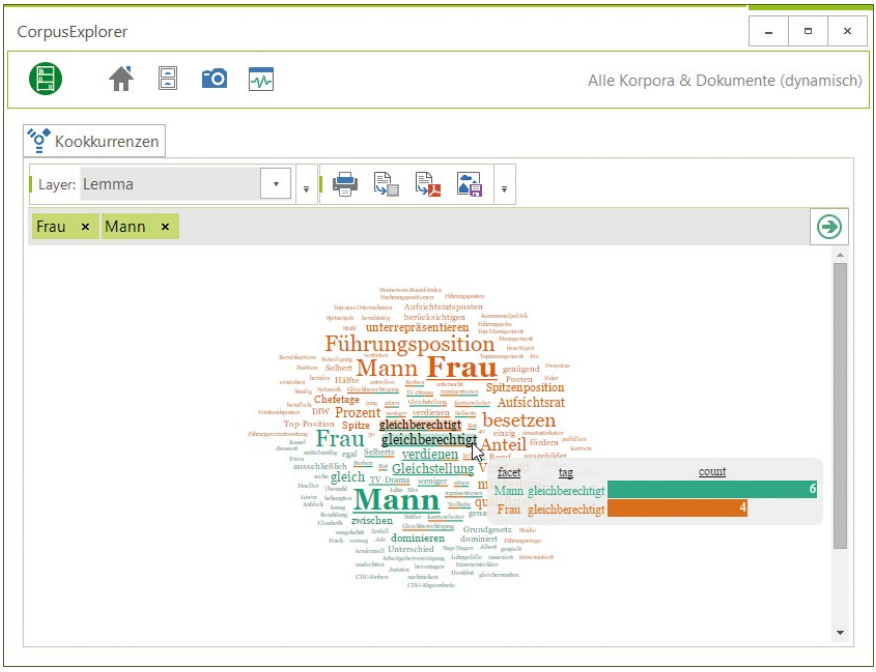


Abb. 4 Kookkurrenz-Wolke zu ‚Mann' und ‚Frau', visualisiert mit TagPies.

Jenseits von Tabellen bieten sich für Kookkurrenzanalysen natürlich noch weitere Darstellungsmöglichkeiten an. Abb. 4 zeigt die Visualisierung mittels TagPies, entwickelt von Jänicke et. al (2015), integriert über die HTML5-Schnittstelle des CorpusExplorers.

Die Datengrundlage ist absolut identisch mit der zuvor gezeigten Kookkurrenz-Kontrastierung. Die Möglichkeiten, die View/Visualisierung, sind aber um einiges umfangreicher, da diese auch mehrere ‚Pole' unterstützt, d. h. mehr als zwei Begriffe gegeneinander kontrastiert werden können. Sehr spannend ist an dieser Visualisierung auch die gegebene Interaktivität. Begriffe lassen sich anwählen und dadurch den einzelnen Polen zuordnen. Im Beispiel: ‚gleichberechtigt' tendiert eher zu ‚Mann' (6) als zu ‚Frau' (4).

Was wäre ein Artikel über Visualisierung in der Linguistik ohne Baumgraphen? – Abb. 5 zeigt einen solchen.

Die Abfrage ist identisch mit den vorherigen, Model und ViewModel konnten erneut eins zu eins wiederverwendet werden. Es erfolgt die Eingabe eines Begriffs, dann der Klick auf die Schaltfläche „Kookkurrenzen" und schon wird ein Baumgraph erstellt. Begriffe werden als farbige Knoten gezeichnet, die Kanten stellen die Relation (wer mit wem) her und geben Auskunft (siehe Werte)

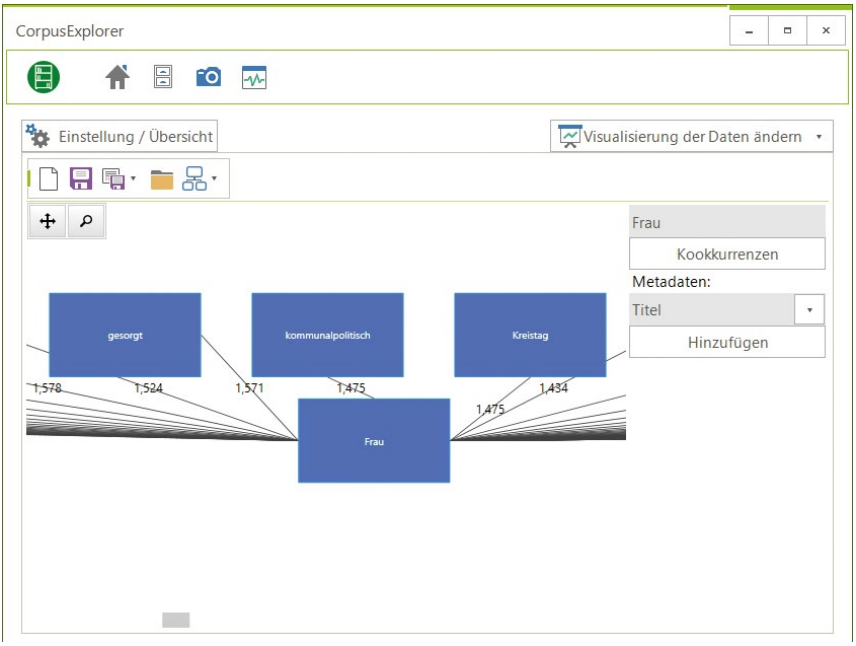


Abb. 5 Ausschnitt Kookkurrenz-Baum zu ‚Frau'.

über die Signifikanz der Verknüpfung. Würden weitere Begriffe wie z. B. ‚Kreistag' eingegeben, bekäme der Baum eine zusätzliche Tiefe – neue Begriffe werden mit existierenden verknüpft. Über die zusätzlichen Schaltflächen lässt sich der Graph layouten und einzelne Knoten/Kanten können zur besseren Darstellung entfernt werden (View-Funktionen). Zu den Begriffen lassen sich noch weitere Ressourcen einblenden wie etwa Metadaten zu den Dokumenten, so können z. B. Autoren oder Verlage als Knoten eingebunden und automatisch mit den Begriffen/Kookkurrenzen verknüpft werden. Auf diese Weise lässt sich in dieser Visualisierung mehr zeigen als ein bloßer Kookkurrenzgraph. Vielmehr erlaubt es Autoren-/Verlagstypiken auf der Mehrwortebene zu analysieren.

4. Fazit

Abschließend möchte ich zuerst festhalten: Der CorpusExplorer stellt keinesfalls eine *Ultima Ratio* für die Korpuslinguistik dar. Vielmehr geht es um einen kritikfähigen Beitrag, der einladen soll, über technische, theoretische und methodische Fragestellungen zu diskutieren. Wie gezeigt werden konnte, macht es Sinn, über Effizienz und Flexibilität im Visualisierungsprozess nachzudenken. Fast die

Hälfte an Ressourcen, hauptsächlich Zeit, konnte durch den dreistufigen Ansatz eingespart werden. Die gezeigten Visualisierungen machen außerdem recht deutlich, wie vielfältig sich die Ergebnisse ein und derselben Methode (Kookkurrenz-Analyse) darstellen lassen. Durch Verknüpfungen mit anderen Daten, auch innerhalb desselben Datensatzes – sowie durch Perspektivwechsel (zeige nur einige wenige, dafür aber spezifische Ergebnisse oder zeige das große Ganze) können Bedeutungsrelationen erzeugt werden, die die Datenrezeption beeinflussen. Aus meiner eigenen Seminarerfahrung heraus kann ich berichten, dass der Einsatz des CorpusExplorers nicht nur Freude in der Anwendung bereitet, sondern auch teilweise überraschende Ergebnisse in einem ansonsten unübersehbaren Berg aus Textmaterial zu Tage fördert. Anfangs herrscht Skepsis darüber, welchen Mehrwert solche Programme liefern können. Aus der Skepsis wird schnell Überraschung darüber, wie simpel doch manche Methoden sind und wie einfach es ist, diese Auswertungen selbst zu produzieren. Dies genügt meist, um Eifer zu entfachen hinter die Dinge sehen zu wollen – Belegstellen zu erkunden und Abfragen in immer komplexerem Maße zu kombinieren. Am Ende eines Seminars mit dem CorpusExplorer stehen motivierte Studierende, verwundert darüber, selbst empirische Forschung betrieben zu haben.

5. Bibliografie

Biemann, Christian. 2003. „Extraktion von semantischen Relationen aus natürlichsprachlichem Text mit Hilfe von maschinellem Lernen." LDV Forum – GLDV Journal for Computational Linguistics and Language Technology 18 (1/2): 12–25.

Bubenhofer, Noah. 2009: Sprachgebrauchsmuster. Berlin: de Gruyter 2009, 1:404. http://www.zora.uzh.ch/id/eprint/111287/1/BubenhoferSprachgebrauchsmusterPub.pdf.

Bubenhofer, Noah. 2011. „Korpuslinguistik in der linguistischen Lehre. Erfolg und Misserfolge." Journals for Language Technology and Computational Linguistics 26 (1): 141–156.

Dipper, Stefanie. 2011. „Digitale Korpora in der Lehre. Anwendungsbeispiele aus der Theoretischen Linguistik und der Computerlinguistik." Journals for Language Technology and Computational Linguistics 26 (1): 81–95.

Eilebrecht, Karl und Gernot Starke. 2013. Patterns kompakt: Entwurfsmuster für effektive Software-Entwicklung. 4. Aufl. Berlin: Springer Vieweg (IT kompakt). E-Book.

Heyer, Gerhard, Uwe Quasthoff und Thomas Wittig. 2006. Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse. Herdecke: W3L-Verl. (IT lernen).

Jänicke, Stefan, Judith Blumenstein, Michaela Rücker, Dirk Zeckzer und Gerik Scheuermann. 2015. Visualizing the Results of Search Queries on Ancient Text Corpora with Tag Pies. http://docplayer.net/23646659-Visualizing-the-results-of-search-queries-on-ancient-text-corpora-with-tag-pies.html (letzter Zugriff am 01. Dezember 2017)

Lemnitzer, Lothar und Heike Zinsmeister. 2006. Korpuslinguistik eine Einführung. Tübingen: Narr (Narr-Studienbücher).

Moreaux, M. A. 1997. „Computerlinguistik für Philologen." LDV Forum – GLDV Journal for Computational Linguistics and Language Technology 14 (2): 22–27.

Runkler, T. A. 2010. Data Mining. Methoden und Algorithmen intelligenter Datenanalyse. Wiesbaden: Vieweg+Teubner.

Wegener, Jörg und Holger Schwichtenberg. 2012. WPF 4.5 und XAML: Grafische Benutzeroberflächen für Windows inkl. Entwicklung von Windows Store Apps. München: Hanser. E-Book.

Zinsmeister, Heike. 2011. „Chancen und Probleme der Nutzung von Korpora, Taggern und anderen Sprachressourcen in Seminaren." Journal for Language Technology and Computational Linguistics (JLCL) 26 (1): 67–79.

Alexander Hinneburg / Christian Oberländer

Getting the Story from Big Data: Interaktive visuelle Inhaltsanalyse für die Sozialwissenschaften mit dem TopicExplorer am Beispiel Fukushima

Abstract Immer mehr Menschen wollen öffentlich gehört werden und ihre Meinung einem breiten Publikum mitteilen. Dieser Trend lässt sich u. a. an der steigenden Beteiligung an sozialen Netzwerken (SNS) ablesen, auf denen folglich große Textmengen in digitaler Form entstehen. Sozialwissenschaftler/innen sind an den Inhalten und Prozessen gesellschaftlicher Interaktion und Meinungsbildung interessiert. Diese auf den SNS entstehenden riesigen Datenmengen (Big Data) in Textform können sie mit herkömmlichen Methoden jedoch bisher kaum auswerten. Als Lösung bieten sich Topic Models an. Sie werden in der Informatik schon länger angewendet und ermöglichen die explorierende Analyse großer Textsammlungen, die persönliche Äußerungen enthalten, indem sie eine Analyse nach häufig verwendeten Themen *(topics)* ermöglichen. Um Sozialwissenschaftlern Zugang zu diesen neuen Forschungsmöglichkeiten zu eröffnen, wurde in Zusammenarbeit von Informatik und Japanologie das Werkzeug TopicExplorer entwickelt. Als Pilotprojekt wurde ein Japan- und sozialwissenschaftlich aktuelles Thema gewählt, nämlich ein Ausschnitt aus der japanischen Internet-Debatte über die Atomkatastrophe von Fukushima und ihre Folgen. Anhand japanischer Blogs über radioaktiv verseuchtes Rindfleisch infolge der Atomkatastrophe von Fukushima im Jahr 2011 wurde demonstriert, wie mit dem TopicExplorer die Wahrnehmung eines gesellschaftlichen Ereignisses im Internet widergespiegelt, sinnvoll strukturiert und für eine vertiefte Analyse aufbereitet werden kann.

1. Topic Models für die Sozialwissenschaften

Das Kernproblem bei der Analyse großer Textmengen mit vielen Autoren ist, dass diese Texte zwar aufgrund ihrer Vielfalt interessant für Geistes- und Sozialwissenschaftler sind, aber ihre vollumfängliche Verarbeitung durch Menschen nicht durchführbar ist, weil diese zu zeitaufwendig bzw. nicht zumutbar wäre.

Erschienen in: Noah Bubenhofer u. Marc Kupietz (Hrsg.): Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

Außerdem wäre das Durcharbeiten aller persönlichen Äußerungen uninteressant, weil viele von ihnen entweder sich nur graduell vom Mainstream unterscheiden oder, im anderen Extrem, abstruse Meinungen vertreten. Dagegen macht explorierendes Text-Mining mit probabilistischen Topic Models große Textmassen spannend für Geistes- und Sozialwissenschaftler, weil mit diesen Methoden ohne manuelle Vorverarbeitung und Annotation, allein durch die Analyse der Wortverteilungen in den Dokumenten eine gegebene Textsammlung verdichtet und zu oft unerwarteten Themen strukturiert werden kann (Blei 2012). Die ermittelten Themen bieten Geistes- und Sozialwissenschaftlern einen fokussierten Zugang zu den informationshaltigen und somit tatsächlich interessanten Texten, die zwischen Mainstream und absurden Einzelpositionen verortet sind. Erst mit der Hilfe von Topic Models können sie sich also ein differenziertes Bild von den Inhalten großer Textsammlungen machen (Evangelopoulos und Visinescu 2012).

Trotz des großen Potenzials von Topic Models besteht eine signifikante Lücke bei ihrer Anwendung. Fernando Pereira, Forschungsdirektor von Google, stellt fest: "While they [topic models] are intriguing, we haven't yet gotten to the point that we can say, 'Yes, this is a practical tool.'" (Anthes 2010). Diese Lücke entsteht dadurch, dass Informatik-Instrumente fehlen, die den Transfer von probabilistischen Aussagen über Wortverteilungen in Dokumenten hin zu inhaltlichen Argumenten – in unserem Hallenser Projekt in geistes- und sozialwissenschaftlichen Kontexten – erlauben (Chang et al. 2009). Denn Topic Models berechnen die latenten Themen nicht nach semantischen Gesichtspunkten, sondern folgen allein informationstheoretischen Kriterien. Es bedarf also geeigneter Auswertungswerkzeuge, die dem Anwender helfen, inhaltliche Informationen und Schlussfolgerungen aus den Berechnungsergebnissen der Topic Models zu ziehen (Blei 2012). Dies wurde auch im Positionspapier (Ramage et al. 2009) des Mimir Projects bestätigt, in dem zwei Barrieren charakterisiert wurden, die in der Zusammenarbeit zwischen Informatikern und Sozialwissenschaftlern bei der Anwendung von Topic Models häufig auftraten: (a) technische Zugänglichkeit der Ergebnisse des Themenmodells für Sozialwissenschaftler und (b) Vertrauen der Sozialwissenschaftler in die Ergebnisse des Themenmodells.

Bei der Entwicklung von Auswertungswerkzeugen für Sozialwissenschaftler muss beachtet werden, dass die Anwender normalerweise über kein Hintergrundwissen zu Text-Mining und probabilistischen Modellen verfügen. Das heißt, die Anwender benötigen Auswertungswerkzeuge, um sich ohne detaillierte Kenntnisse der zugrunde liegenden mathematischen Theorie aus den Ausgangsdaten (der Dokumentensammlung) und einem daraus berechneten Themenmodell eine Plausibilitätsstruktur zur Interpretation der Themen und somit zum Verständnis des Inhaltes der Dokumentensammlung zu erschließen. Dabei dürfen ihnen auch keine falschen Schlüsse suggeriert werden. Deshalb

ist ein weiterer wichtiger Punkt, dass die Anwender/innen über die inhaltliche Auswertung hinaus auch bei der kritischen Überprüfung ihrer Interpretationen durch das Werkzeug unterstützt werden. Beispielsweise hat sich die Möglichkeit, die berechneten Themen zu den relevanten Originaldokumenten zurückzuverfolgen, als besonders hilfreich für die Anwender herausgestellt, denn dies gibt ihnen die Möglichkeit, sowohl ihre inhaltliche Interpretation der probabilistisch berechneten Themen zu überprüfen als auch aussagekräftige Textstellen aufzufinden und gezielt für ihre wissenschaftliche Argumentation zu nutzen.

Die Interaktivität des Auswertungsprozesses stellt dabei hohe Anforderungen an die Effizienz der zu entwickelnden Software-Komponenten. Neben inhaltlichen bestehen also auch hohe technische Anforderungen. Da sowohl die Dokumentsammlungen als auch die daraus berechneten Topic Models sehr große Datenmengen umfassen, müssen außerdem effiziente Such- und Auswertungsalgorithmen entwickelt werden. Diese Algorithmen sollen mit effektiven Visualisierungstechniken für Topic Models (Gohr et al. 2010, Hinneburg et al. 2012, Gohr et al. 2013) zu interaktiven, Web-basierten Benutzeroberflächen verknüpft werden.

2. Der TopicExplorer

2.1 Entwicklung der Analyseplattform

Das Analysesystem TopicExplorer (http://topicexplorer.informatik.uni-halle.de) besteht aus zwei Teilen: der erste Teil ist eine schlanke Software-Plattform, die nicht nur die entwickelten Auswertungswerkzeuge trägt, sondern die es auch ermöglicht, eine Vielzahl von neuen Funktionalitäten zur Exploration von Dokumentensammlungen umzusetzen und in das System zu integrieren, ohne es grundsätzlich im Kern verändern zu müssen. Der zweite Teil baut darauf auf und bildet das eigentliche TopicExplorer-System, das die speziellen Analysefunktionen zur Exploration großer Textsammlungen beinhaltet.

Der TopicExplorer stellt sich für den Anwender als eine interaktiv nutzbare Web-Applikation zur Analyse großer Dokumentsammlungen dar. Die Vorverarbeitung der Texte, die Themenberechnung sowie das Verknüpfen der berechneten Themen mit den Dokumenten sind aufwendige Prozesse, die je nach Größe der Dokumentsammlung zwischen Stunden und Tagen in Anspruch nehmen. Deshalb wurde das TopicExplorer-System in drei Bereiche aufgeteilt: Vorberechnung, Server und Client. Zum Bereich Vorberechnung gehören alle Software-Komponenten mit lang laufenden Aufgaben: Wort- und Satzanalyse der Dokumente, die Themenanalyse der aufbereiteten Dokumente, das Verknüpfen der berechneten Themen mit den Texten und ihren Meta-Daten. Das Ergebnis der

Vorberechnung ist eine Datenbank, die von den Software-Komponenten des Servers mit schnell laufenden Anfragen ausgelesen werden kann, um interaktiv den Client mit den benötigten Daten zu versorgen.

Die Nutzung von Software-Entwurfsmustern erlaubt es zu planen, für welche zukünftige Funktionalität die Software flexibel bleiben soll. Für die Bereiche Vorberechnung und Server-Backend wurde das Software-Entwurfsmuster Filter Pipelines in Kombination mit dem Kommando-Muster verwendet. So bleibt die Software-Architektur flexibel und kann immer wieder neue Verarbeitungsschritte in die Datenaufbereitung und Analyse einzufügen, die während der Entwicklung auftauchen. Dieses kombinierte Software-Entwurfsmuster wurde zum Konzept automatisch konfigurierbarer Workflows weiterentwickelt. Die Software-Module für die Workflows sind in der separaten Bibliothek CommandManager ausgelagert, die Open Source (https://github.com/hinneburg/CommandManager) verfügbar ist.

Als Ergebnis der Nutzung von automatisch konfigurierbaren Workflows sind alle Funktionalitäten des TopicExplorer-Systems in kleinen, übersichtlichen Modulen, sogenannten Kommandos, implementiert. Die Kommandos wiederum werden zu bereichsüberspannenden Plugins gebündelt. Somit können ganze inhaltlich zusammenhängende Analysebereiche an- und abgeschaltet werden. So können aufwendige Workflows entstehen, die sich inkrementell aus den Abhängigkeiten der einzelnen Kommandos ergeben. Abbildung 1 zeigt als Beispiel den automatisch konfigurierten Workflow für den Bereich Vorverarbeitung.

Die meisten Funktionalitäten sind mithilfe eines Datenbanksystems in der Programmiersprache SQL deklarativ programmiert. Ein Vorteil des Datenbankansatzes ist, dass die Datenstrukturen des probabilistischen Themenmodells als relationales Modell durch Tabellen explizit ausgedrückt werden. In zukünftigen Versionen sollen die rechenintensiven Abschnitte der Vorberechnung in SparkSQL (Armbrust et al. 2015) umgesetzt werden.

Der Client ist als Web-Applikation in JavaScript implementiert. Das Benutzer-Interface des TopicExplorers ist mit einem modularisierbaren Web-Client-Programmieransatz auf der Basis von KnockOutJs umgesetzt. Die entwickelte Lösung ist flexibel genug, um Javascript-Bibliotheken z. B. für Informationsvisualisierungen mit D3js, auf einfache Weise einzubinden.

2.2 Visuelle, interaktive Benutzerschnittstelle

Das TopicExplorer-System umfasst in der derzeitigen Ausbaustufe folgende Funktionen: (1) Themen-Rankings von Dokumenten, (2) Dokumentansichten mit Themenzuordnungen der einzelnen Wörter, (3) durchgängige Themenzuordnungen in allen Ansichten durch Farbe und symbolische Nummerierung,

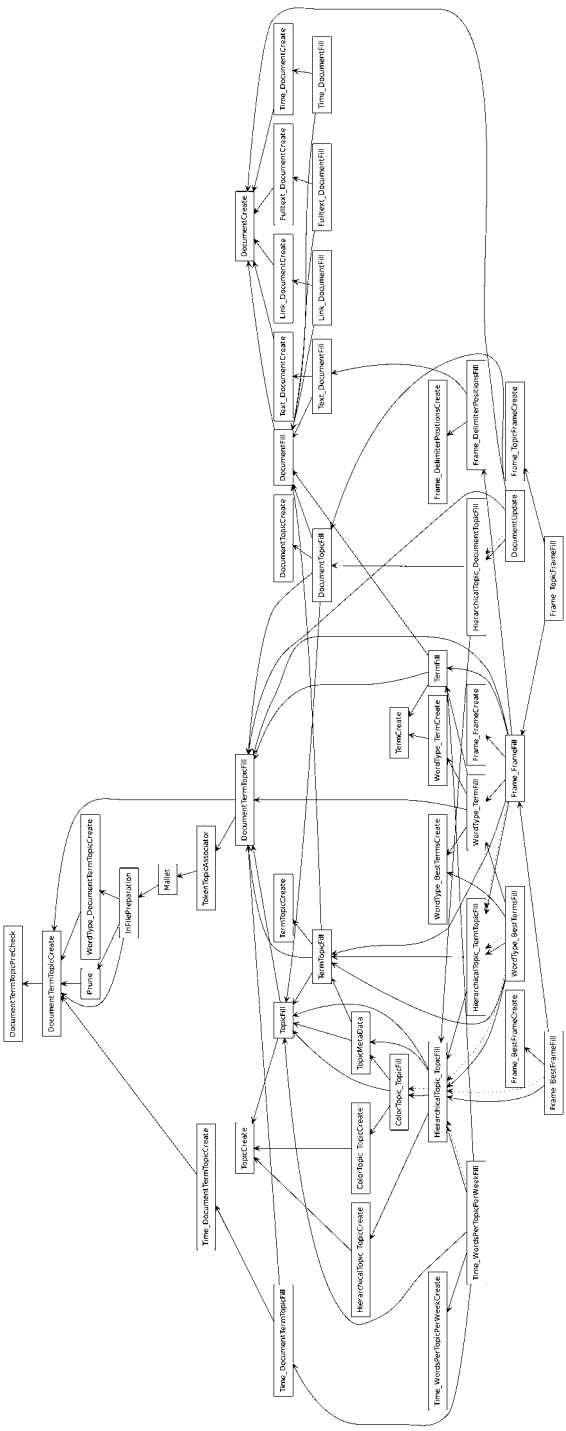


Abbildung 1: Automatisch konfigurierter Workflow-Graph für den Bereich Vorverarbeitung.

(4) automatische Vervollständigung von Suchwörtern mit thematischer Zuordnung, (5) ähnlichkeitsorientierte lineare Anordnung der Themen, (6) thematische Wort-Rankings, (7) zeitliche Analyse und Entwicklung von Themen, (8) Themendarstellung durch bestimmte Wortarten und TopicFrames, (9) themenbasierte Stichwortsuche und (10) interaktives Zusammenfassen und Aufspalten von verwandten Themen.

Basisfunktionen des TopicExplorer-Systems sind Stichwortsuche, Anzeige von Dokument-Rankings sowie die Themendarstellung und -navigation. Abbildung 2 zeigt die Benutzeroberfläche des TopicExplorers am Beispiel einer Dokumentsammlung der 10.000 längsten englischsprachigen Wikipedia-Artikel. Die Themen werden als Wortlisten dargestellt. Ähnliche Themen werden benachbart angeordnet – so gehören beispielsweise in Abbildung 2 die vier gelben nebeneinander liegenden Themen von links alle in den Bereich Unterhaltung. Mittels des horizontalen Sliders kann der Anwender das Spektrum der Themen durchsehen, ohne dass inhaltliche Brüche auftreten. Die Farbe kann aus diesem Grund als Farbverlauf von links nach rechts den Themen zugeordnet werden. Der Farbcode dient dem schnellen approximativen visuellen Referenzieren der Themen in den weiteren Ansichten des TopicExplorers. Das Klicken auf den fett gedruckten Titel eines Themas öffnet einen Browser-Tab mit einem Dokument-Ranking, das wichtige Dokumente des Themas zeigt, z. B. Thema 48 zu Musik im mittleren Teil von Abbildung 2. Ein Dokument wird im Browser-Tab als Kasten dargestellt, der den Titel, den Textbeginn und mittels farbiger Kreise die vier dominanten Themen des Dokuments zeigt. Die farbige Referenz eines solchen Kreises kann genauer untersucht werden, wenn man mit der Maus darüberfährt. Durch Klicken auf einen Kreis kann das entsprechende Thema nachgeschlagen werden; es wird dann in der unteren Themenansicht mittig fokussiert.

Durch Klicken auf den Titel eines Dokumentes, z. B. auf „Queen (Band)“ in Abbildung 2, wird ein neuer Tab mit einer Dokument-Ansicht geöffnet (Abbildung 3), der den Inhalt des Dokuments zeigt, wobei die Wörter durch farbige Unterstreichungen die Themenzuordnungen anzeigen, die bei der automatischen Themenanalyse ermittelt wurden. Somit kann das Ergebnis der Themenanalyse direkt an den Dokumentinhalten verifiziert werden. Ebenso wie in der Dokument-Browser-Ansicht können die farbigen Themenzuordnungen durch Darüberfahren mit der Maus konkretisiert werden. Ein Klick auf ein farbiges Wort fokussiert das Thema in der Themenansicht.

Mit Klicken auf das Zeitreihen-Icon eines Themas (in der farbigen Box des Themas rechts oben) wird die Zeitansicht des TopicExplorers für dieses Thema in einem neuen Tab geöffnet. Ein Beispiel wird in Abbildung 4 für die Cäsium-Rindfleisch-Daten dargestellt, die Teil der japanischen Blog-Sammlung zum Stichwort „Kernkraft“ *(genpatsu)* sind. Zeitreihen zu weiteren Themen sowie die Durchschnittszeitreihe aller Themen können als Vergleich mittels Checkboxen

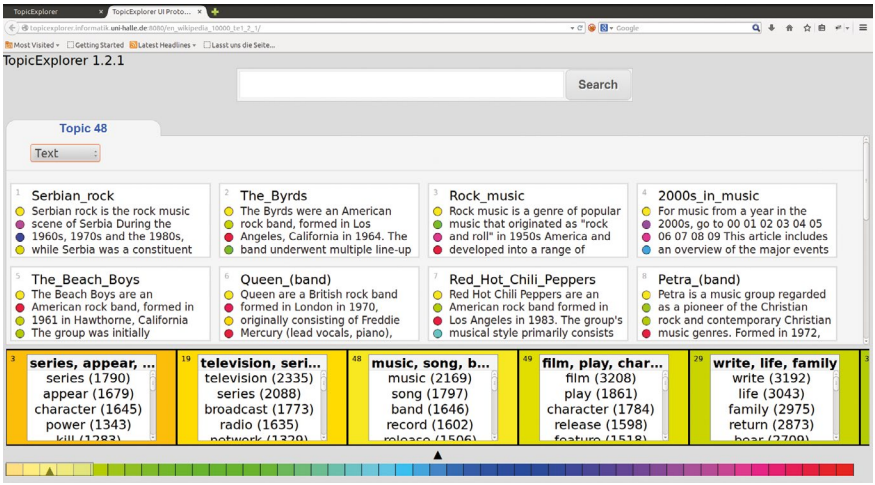


Abbildung 2: Überblick TopicExplorer mit Bedienelementen zur Suche, zum Dokument-Ranking und zur Themendarstellung und -navigation (von oben nach unten).

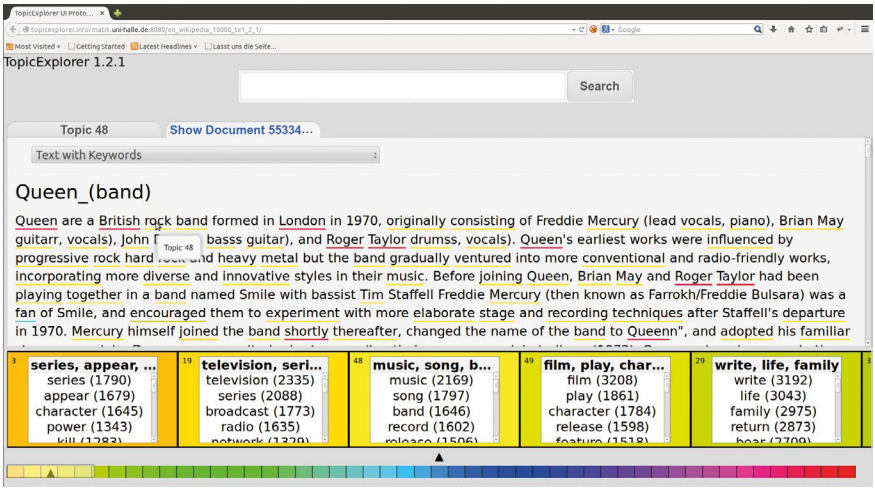


Abbildung 3: Dokumentansicht des TopicExplorer: Die farbigen Unterstreichungen der Wörter zeigen die Themenzuordnungen. Die nicht unterstrichenen Wörter wurden als Stopp-Wörter (sehr häufig) oder als zu seltene Wörter aus der Themenanalyse ausgeschlossen.

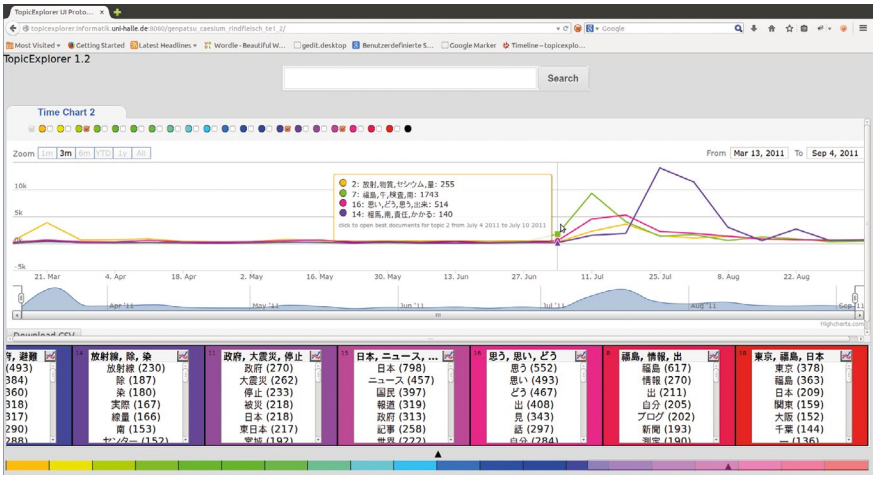
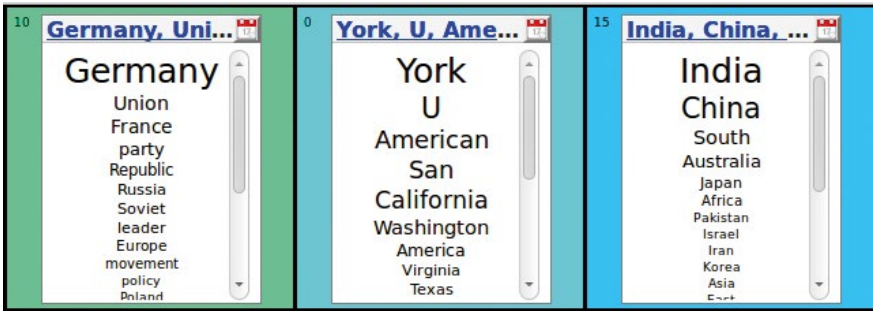


Abbildung 4: Zeitansicht des TopicExplorer: Die Entwicklung mehrerer Themen über die Zeit wird anhand der Anzahl der ihnen zugeordneten Wörter durch die farbigen Zeitreihen dargestellt.

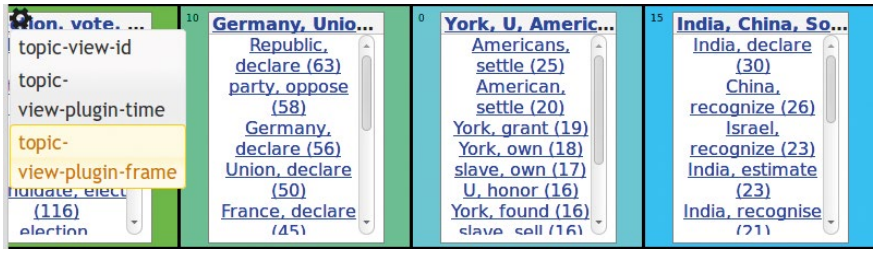
interaktiv eingeblendet werden. Die Legende zeigt die wichtigsten Wörter der gewählten Themen für die Woche, über der sich die Maus befindet. Somit kann die Entwicklung der Themen interaktiv verfolgt werden. Ein Klick auf eine Woche in der Zeitreihe öffnet ein Dokument-Ranking in einem neuen Tab, das die wichtigsten Dokumente zu dem zuerst gewählten Thema in der mit der Maus gewählten Woche zeigt. Die thematischen Zeitreihen geben z. B. Auskunft, ob, wann und in welcher Abfolge Themen einen Höhepunkt erreichen.

Eine zentrale Herausforderung bei der Gestaltung des TopicExplorers ist die Darstellung der Themen, die auf eine Weise erfolgen muss, dass Anwender die Themen interpretieren und von irrelevanten statistischen Artefakten unterscheiden können. Dafür schlagen wir vor, Themen durch Substantiv-Verb-Kombinationen (TopicFrames) darzustellen (Hinneburg et al. 2014). Die Repräsentation eines Themas durch häufige, themenspezifische Substantiv-Verb-Kombinationen hilft dem Benutzer, den Kontext der wahrscheinlichsten Wörter des Themas zu erfassen bzw. einzugrenzen und somit das Thema zutreffend zu interpretieren (Abbildung 5a). Besonders hilfreich sind TopicFrames, wenn die wahrscheinlichsten Wörter eines Themas ausschließlich aus Substantiven bestehen und das Thema aufgrund dieser Wortlisten mehrere Interpretationen haben könnte. Die Kombinationen mit Verben helfen, den Kontext genauer zu bestimmen. Ein TopicFrame besteht aus einem Substantiv und einem Verb, die im gleichen Satz vorkommen und demselben Thema zugeordnet sind. Abbildung 5b zeigt die häufigsten TopicFrames der oben eingeführten Themen. Das Asienthema (rechts)

Abbildung 5: Wechsel der Themenrepräsentation.



(5a) Wahrscheinlichste Wörter.

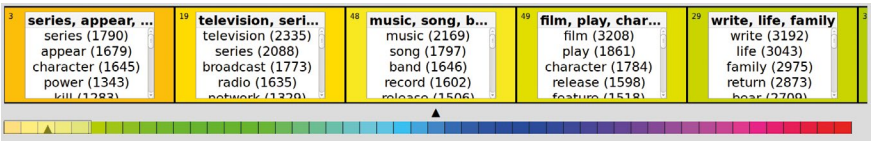


(5b) Häufigste TopicFrames.

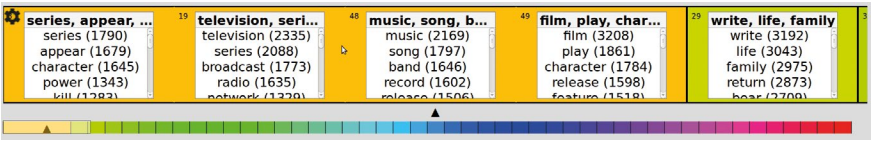
kann durch die Kombination von *declare* bzw. *recognize* mit Ländernamen als asiatische Außenpolitik erkannt werden.

Für das Berechnen der TopicFrame-Repräsentationen führt der TopicExplorer eine Wortartbestimmung („Part-Of-Speech-Tagging) in der Vorverarbeitung durch. In weiteren Schritten wurde für den TopicExplorer das TopicFrame-Konzept dahingehend erweitert, dass neben Substantiv-Verb-Kombinationen weitere frei wählbare Wortarten für die Repräsentation der Themen kombiniert werden können. Diese können bis zu den Dokumenten zurückverfolgt werden, sodass Anwender leicht auf Belegstellen für TopicFrames zugreifen können. Die verschiedenen Repräsentationen der Themen lassen sich interaktiv umschalten, sodass Anwender verschiedene Sichten auf ein Thema erhalten. Mit der Bestimmung der Wortarten für jedes einzelne Wort in den Dokumenten ergibt sich die Möglichkeit, die Themen durch die wahrscheinlichsten Wörter einer bestimmten Wortart zu repräsentieren. Gerade weil die Wahrscheinlichkeiten für die unterschiedlichen Wortarten sehr variieren, ist es bei der Interpretation eines Themas oft hilfreich, eine weniger häufige Wortart in den Fokus zu rücken. So sind Adjektive und Adverbien sind im Vergleich zu Substantiven und Verben seltener. Durch die Repräsentation eines Themas durch Adjektive

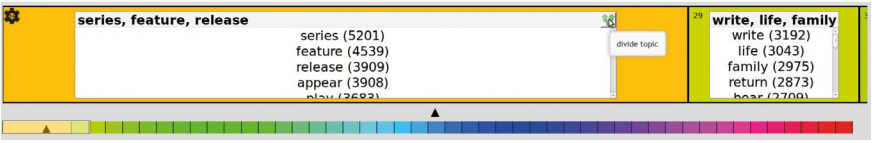
Abbildung 6: Interaktive, hierarchische Themennavigation.



(6a) Durch Themenanalyse berechnete Themen.



(6b) Vorschau des Zusammenfassens von Themen aufgrund der Hierarchie.



(6c) Zusammengefasstes Thema, interaktives Aufteilen ist möglich.

lässt sich außerdem erkennen, ob und in welcher Weise ein Thema emotional gefärbt ist.

Hierarchische Themen ermöglichen Anwenderinnen und Anwendern, die Themengrundlage an jeweils eigene Fragestellungen anzupassen. Dieses interaktive Definieren der Themen durch die Anwender wird durch Vorschauansichten, Ranking- und Dokumentansichten sowie Zeitdiagramme zur Themenentwicklung unterstützt, die sich alle den aktuell gewählten Themen anpassen. Somit sind die Anwender nicht nur passive Konsumenten der Themen, sondern können diese während des Explorationsprozesses selbst definieren.

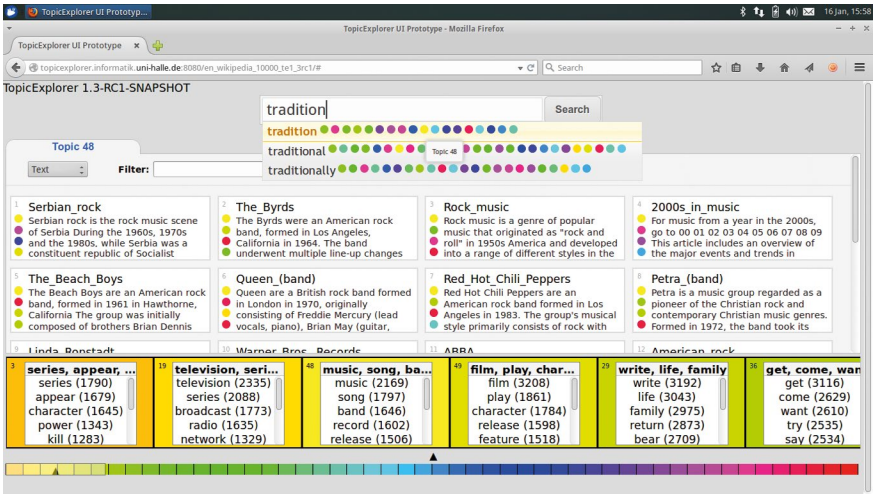
Um dies interaktiv zu ermöglichen, werden die bei der Themenanalyse berechneten Themen während der Vorverarbeitung mittels eines hierarchischen Clustering-Verfahrens aufgrund ihrer Ähnlichkeit in den Wortverteilungen zusammengefasst. Als Ergebnis des Clusterings entsteht ein binärer Baum, dessen Blätter die durch die Themenanalyse berechneten Themen sind. Für jedes Thema, das in der Clusteranalyse durch Zusammenfassen entsteht, werden in der Vorverarbeitung ebenfalls alle relevanten Informationen und TopicFrames berechnet. Die Anwender/innen können im TopicExplorer interaktiv in diesem binären Baum der Themen navigieren, indem sie die Maus auf die Trennlinie zwischen den farbigen Themenspalten positionieren. Dabei ändert sich als Vorschau

die Hintergrundfarbe der Themen, die durch einen Klick auf die Trennlinie zu einem einzigen Thema zusammengefasst würden. Es können zwei oder mehr Themen durch eine solche Interaktion zusammengefasst werden in Abhängigkeit davon, wie weit die beiden Themen rechts und links der Trennlinie im binären Baum auseinanderliegen. Abbildungen 6a und 6b zeigen durch den Wechsel der Themenfarbe an, dass die vier Themen von links über Serien, Fernsehen und Radio, Musik und Filme zu einem Thema über Unterhaltung zusammengefasst würden, wenn zwischen die beiden mittleren Themen geklickt würde. Abbildung 6c zeigt das zusammengefasste Thema. Das Zusammenfassen kann durch Klicken auf den Button rückgängig gemacht werden, der rechts oben bei dem zusammengefassten Thema erscheint. Die Information des Zusammenfassens wird sofort an alle anderen Ansichten des TopicExplorers (Dokument-Browser, Dokumentansicht, Zeitansicht) weitergeleitet und dort dargestellt. Gerade in der Zeitansicht können durch das Zusammenfassen von Themen zeitliche Häufungen sichtbar werden, die sonst in mehrere kleine Themen zergliedert sind. Somit wird die Auswirkung des Zusammenfassens interaktiv erfahrbar.

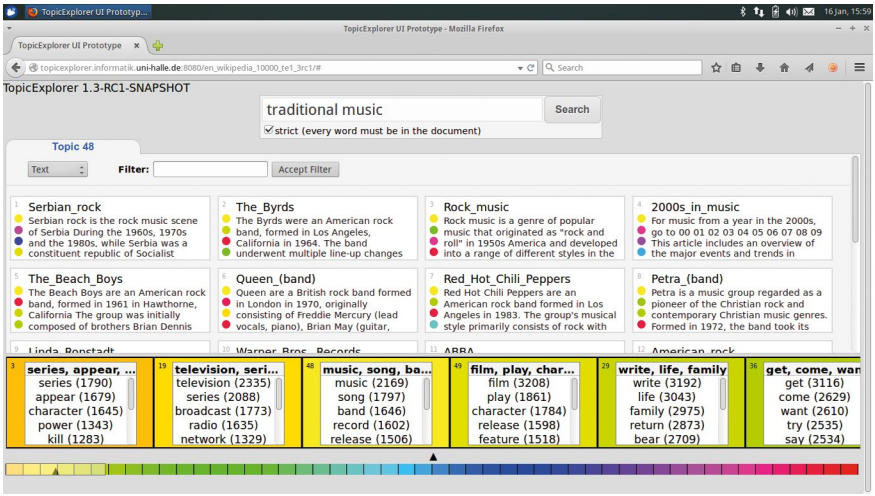
Das TopicExplorer-System bietet weiterhin Volltextsuche sowie Filterfunktionen, um Anwendern den Einstieg in ein Korpus anhand von konkreten Suchwörtern zu erleichtern. Es können einzelne oder mehrere Wörter im Verbund als Suchbedingung angegeben werden. Solange ein einzelnes Wort im Suchfeld steht, werden durch die Auto-Complete-Funktion die Themen mit absteigender Wahrscheinlichkeit angezeigt, die zu diesem Wort passen (siehe Abbildung 7a). So können interessante Themen ausgehend von diesem einen Wort rasch gefunden werden. Wenn mehrere Wörter eingegeben werden, kann zwischen normaler und strikter Suche gewählt werden (siehe Abbildung 7b). Bei strikter Suche müssen alle Wörter in einem Trefferdokument vorkommen, während die normale Suche nur sicherstellt, dass die Trefferdokumente möglichst viele, aber nicht unbedingt alle Wörter der Anfrage enthalten. In der Dokument-Browser-Ansicht wird für jedes Treffer-Dokument angezeigt, welche Wörter der Anfrage im Dokument wie oft vorkommen.

Die thematische Stichwortsuche findet Dokumente, die das gewünschte Stichwort enthalten, jedoch wird zusätzlich geprüft, ob das Wort im Trefferdokument auch dem gewünschten Thema zugeordnet ist. Mit dieser Suchmethode können auch mit allgemeinen Suchwörtern relevante Trefferdokumente gefunden werden, wenn das Thema entsprechend eingeschränkt wird. Beispielsweise findet die allgemeine Stichwortsuche nach *traditional* Dokumente aus sehr verschiedenen Bereichen. Mit einer thematischen Stichwortsuche nach *traditional* für das Thema Musik werden jedoch nur Dokumente zu Volks- und Regionalmusik gefunden, die *traditional* enthalten und dem Thema Musik zuzuordnen sind.

Abbildung 7: Stichwortsuche.



(7a) Autocomplete mit Themendarstellung.



(7b) Mehrere Stichwörter.

Ausgestattet mit den interaktiven Möglichkeiten des TopicExplorers können nicht technisch ausgebildete Anwender ohne Programmierkenntnisse Topic Models nutzen, um Inhalte in großen Dokumentensammlungen mit wenig Aufwand vorzustrukturieren. Weiterhin können die Effekte von Parametereinstellungen wie „Anzahl der Themen", „Auswahl der analysierten Wortarten" und „Weglassen von häufigen und seltenen Wörtern" auf die Themen studiert werden, um deren semantische Qualität zu prüfen und gegebenenfalls zu verbessern.

2.3 Stand der Technik im Vergleich

In den letzten Jahren wuchs das Interesse an Werkzeugen, die automatische Text-Analyse und -Visualisierung miteinander verbinden. Zu diesem Thema wurden mehrere Workshops (z. B. Chuang et al. 2014) veranstaltet sowie einzelne Konferenzbeiträge veröffentlicht. Parallel zum TopicExplorer wurden daher international auch eine Reihe anderer Systeme entwickelt und vorgestellt. Um den TopicExplorer relativ zum Stand der Technik auf diesem Gebiet zu verorten, vergleichen wir ihn – trotz erheblicher Unterschiede in den anwendungsspezifischen Zielsetzungen – mit den anderen Systemen (Tabelle 1). Die Tabelle fasst die wichtigsten methodischen Ansätze zusammen, die sich auf Topic Models stützen, und stellt deren Eigenschaften dar.

Zu den international bekannten Systemen gehört beispielsweise das Serendip-System (Alexander et al. 2014), das Informatiker und Anglisten an der Universität Wisconsin-Madison, USA zur Exploration englischsprachiger Korpora entwickelt haben. Ähnliche weniger ausgebaute Ansätze sind LDAvis (Sievert & Shirley, 2014) und topic-explorer (Murdock & Allen, 2015), ein einfacher Themen-Browser mit gleichem Namen. Der Hallenser TopicExplorer bietet zu fast allen Möglichkeiten hinsichtlich Ranking und Themenrepräsentation ähnliche Funktionen wie Serendip an, doch basieren sie auf anderen Visualisierungs- und Interaktionstechniken. Darüber hinaus hält der TopicExplorer aber auch wichtige zusätzliche Funktionen vor, wie z. B. Themen interaktiv zusammenzufassen und ihren zeitlichen Verlauf darzustellen. Diese zusätzlichen Möglichkeiten, die von den Anwendern hoch geschätzt werden, sind auf dem internationalen Stand der Technik ausgeführt, dessen Aspekte im Folgenden kurz umrissen und mit den Systemen in Tabelle 1 verknüpft werden. Wesentlich sind die Aspekte: (1) maschinelles Lernen der zeitlichen Themenentwicklung, (2) deren Visualisierung, (3) die hierarchische Themengliederung und (4) die Software-Organisation.

Die zeitliche Themenentwicklung wird am einfachsten modelliert, indem die Themen zuerst ohne Berücksichtigung der Zeit gelernt werden und danach deren Entwicklung durch die gemeinsame Analyse der Zeitstempel der Dokumente und der Zuordnungen der Wörter zu den Themen charakterisiert wird. Dieser

Tabelle 1: Überblick der Visualisierungswerkzeuge für Topic Models mit ihren Eigenschaften.

System	Features																			
	Ranking				Representation							Hierarchy		Time			Search		Software	
	Document given Topic	term given topic	topic given document	topic given term	word list	POS	frames	token topic assignment	document groups	topic concentration	document authors	overview	interactive merge/split	overview	evolving relations	interactive zoom	keyword	topic	config. prepocessing	config. user-interface
TopicExplorer (Hinneburg et al. 2012, Hinneburg et al. 2014)	✔	✔	✔	✔	✔	✔	✔	✔				✔	✔	✔		✔	✔	✔	✔	✔
Seredip (Alexander et al. 2014)	✔	✔	✔		✔			✔	✔	✔							✔			
LDAvis (Sievert & Shirley, 2014)	✔	✔			✔							✔								
topic-explorer (Murdock & Allen, 2015)	✔	✔			✔															
TopicPanorama (Liu et al. 2014)	✔	✔			✔							✔	✔							
TIARA, (Liu et al. 2009, Pan et al. 2013)	✔	✔			✔									✔	✔	✔				
EvoRiver (Sun et al. 2014)	✔	✔			✔						✔			✔	✔					
RoseRiver (Cui et al. 2014)	✔	✔			✔							✔		✔	✔	✔	✔			
HierarchicalTopics (Dou et al. 2013)	✔	✔			✔							✔	✔	✔		✔	✔			
Dvita (Derntl et al. 2014)	✔	✔	✔		✔									✔		✔	✔	✔	✔	
Footprint (Isaacs et al. 2014)	✔		✔										✔	✔		✔	✔	✔		
HierarchicalTopicGrid (Jojic et al. 2016)	✔	✔			✔							✔	✔							

Ansatz wird von fast allen Visualisierungssystemen TopicExplorer, TIARA (Liu et al. 2009, Pan et al. 2013), EvoRiver (Sun et al. 2014), RoseRiver (Cui et al. 2014) in diesem Bereich verfolgt. Eine Ausnahme ist Dvita (Derntl et al. 2014), das ein deutlich rechenintensiveres Themenmodell nutzt und die zeitlichen Änderungen von Themenpräferenzen durch Wahrscheinlichkeitsverteilungen modelliert. Bisher zeigte jedoch keiner der beiden Ansätze (Lernen der Themen mit oder ohne Zeitinformation) besondere Vorteile für die Inhaltsanalyse. Neben einer Übersichtsdarstellung der Themenentwicklung und verschiedener Interaktionsmöglichkeiten, die alle genannten Systeme bieten, wurden in TIARA, EvoRiver und RoseRiver prototypisch die Entwicklungen von Beziehungen zwischen Themen visualisiert. Diese Möglichkeit soll ebenfalls in zukünftigen Versionen des TopicExplorers eingebaut werden, um neuartige Verbindungen zwischen Themen zu erkennen.

Die Visualisierungsansätze zur hierarchischen Themengliederung wurden in LDAvis, TopicPanorama (Liu et al. 2014) und HierarchicalTopicGrid (Jojic et al. 2016) prototypisch umgesetzt, um speziell entwickelte Visualisierungs- und Interaktionstechniken zu untersuchen. Die Kombination von Hierarchie und zeitlicher Entwicklung wurde in TopicExplorer, HierarchicalTopics (Dou et al. 2013) und RoseRiver umgesetzt, wobei HierarchicalTopics eine nur prototypische Studie ist. RoseRiver ist dagegen eine spezielle Visualisierung für wenige vorausgewählte Themen, die beide Aspekte verbindet und nach einer Evaluation ebenfalls im TopicExplorer umgesetzt werden kann.

Software-Design, Wartung, Erweiterbarkeit und Konfigurierbarkeit als letzter Aspekt wurden nur im TopicExplorer und Dvita explizit adressiert. Für die anderen Systeme wurde dieser Aspekt in den Veröffentlichungen nicht angesprochen.

Footprint (Isaacs et al. 2014), das interessante Funktionen im Bereich Suche bietet, fällt ansonsten aus dem Vergleichsrahmen heraus, weil es keine Topic Models benutzt. Stattdessen werden hier Dokumente durch einen externen Informationsservice mit Themen annotiert, der nur für Englisch verfügbar ist. Deshalb können dort keine Wortzuordnungen zu Themen gemacht werden und daher enthält das System keine weiteren darauf aufbauenden Funktionen.

Zusammenfassend kann festgestellt werden, dass der Hallenser TopicExplorer im Vergleich mit allen anderen Systemen einen großen Funktionsumfang bietet und alle in der Forschung auftretenden Aspekte der Visualisierung und Interaktion mit Topic Models abdeckt. Einzelne innovative Ideen für weitere Funktionalitäten können durch die erweiterbare Software-Architektur ergänzt werden.

3. Sozialwissenschaftliche Anwendung am Beispiel Fukushima

3.1 Fukushima und neue Medien als Artikulationsraum

Als ein Beispiel für die Anwendung des TopicExplorers werden in diesem Aufsatz exemplarisch japanische Blogs über den Skandal des mit radioaktivem Cäsium verseuchten Rindfleisches, das einige Monate nach dem Atomunfall von Fukushima zunächst in Supermarktregalen in Tokio, später dann auch in anderen Teilen Japans auftauchte, untersucht. Im Falle dieses Skandals spielte das Internet eine zentrale Rolle. Denn die Gefahren, die von Radioaktivität ausgehen, sind für den Einzelnen in der realen Umwelt weder direkt erfahrbar noch sichtbar. Folglich ist das Bewusstsein für die Gefährlichkeit der Lage in der Gesellschaft zunächst nur latent vorhanden. Das Ausmaß der Bedrohung wird den Menschen erst durch gesellschaftliche Instanzen, insbesondere durch die Medien, vermittelt und somit sichtbar und „begreifbar" gemacht. Die Menschen nutzen bei diesem, nach Latour (2004) auch als „Artikulation" bezeichneten Vorgang in erheblichem Maße das Internet als „Artikulationsraum" (Kuchinskaya 2011: 406). Denn durch die neuen Medien kann der Einzelne seine Ansichten, Einschätzungen und Informationen zu gesellschaftlich relevanten Themen im Internet formulieren und rasch verbreiten. So verschafft das Internet dem Einzelnen die Möglichkeit, im Austausch mit anderen ein Verständnis des Unglücksgeschehens zu entwickeln, das bis zu einem gewissen Grade unabhängig von den Interpretationen ist, die ihm die konventionellen Medien anbieten. Um die Rolle des Internets als Artikulationsraum einschätzen zu können, ist u. a. die Frage bedeutungsvoll, in welchem Umfang die Berichterstattung der konventionellen Medien die Inhalte z. B. von Blogs prägt und bis zu welchem Grad Blogs davon unabhängige Themen aufgreifen. Im folgenden Abschnitt werden der Skandal und seine Darstellung in japanischen Blogs anhand eines Blog-Korpus und seiner automatisierten Inhaltsanalyse mit dem TopicExplorer beschrieben. Einem Vergleich der so erzielten Ergebnisse mit den Inhalten der Printmedien folgt eine kurze Diskussion und Zusammenfassung.

3.2 Der Cäsium-Skandal nach Fukushima in der TopicExplorer-Analyse von japanischen Blogs

Nach dem Großen Ostjapan-Erdbeben vom 11. März 2011 wurden aus dem Kernkraftwerk Fukushima-1 große Mengen radioaktiver Substanzen in die Umwelt freigesetzt, die nicht nur die nähere Umgebung, sondern stellenweise auch entfernte Landesteile Japans verseuchten. Die japanische Regierung tat sich schwer damit, die radioaktive Kontamination einzudämmen und die Sicherheit von

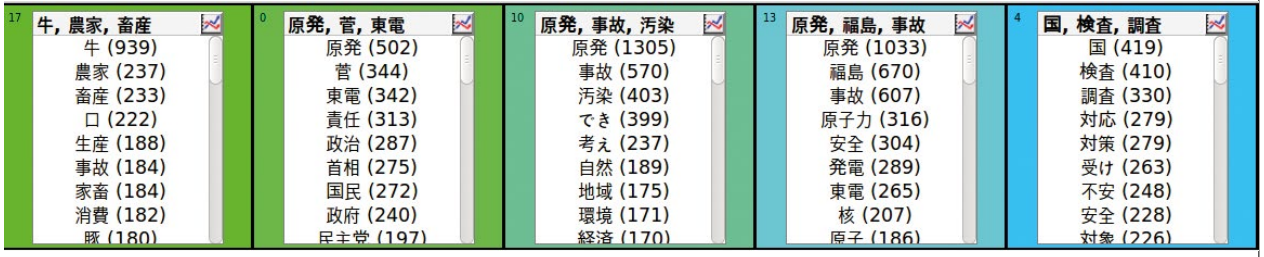
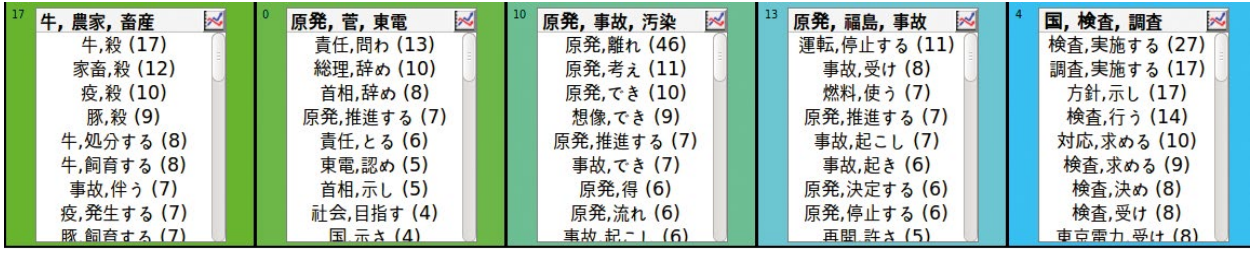
Lebensmitteln, besonders aus der direkt betroffenen Region Tohoku, sicherzustellen. Politiker und Beamte ergriffen nur zögerlich Maßnahmen, um die Verseuchung von Japans Nahrungsmittelversorgung zu verhindern. Unter anderem versäumten sie es, den Vertrieb von Reisstroh aus den kontaminierten Regionen wirkungsvoll zu verbieten. Dieses Reisstroh wurde folglich an Züchter im ganzen Land als Rinderfutter verkauft, sodass auch Rinderbestände außerhalb des ursprünglich betroffenen Gebiets mit radioaktivem Cäsium verseucht wurden. Außerdem verließ sich die japanische Regierung zunächst auf freiwillige Vertriebssperren für Rindfleisch aus radioaktiv gefährdeten Gebieten, nur um dann im Juli 2011 aus Berichten von Kommunen und Medien zu erfahren, dass Rindfleisch mit einem erhöhten Gehalt von radioaktivem Cäsium in den Regalen von Tokioter Supermärkten auslag. Erst bei Bekanntwerden dieses Skandals reagierte die Regierung mit einem Vertriebsverbot für Rindfleisch aus der betroffenen Region. Diese Nachlässigkeiten schadeten der Gesundheit der Konsumenten und störten das Vertrauen, das über Jahre in hochpreisiges japanisches Rindfleisch aufgebaut worden war (Kingston 2012).

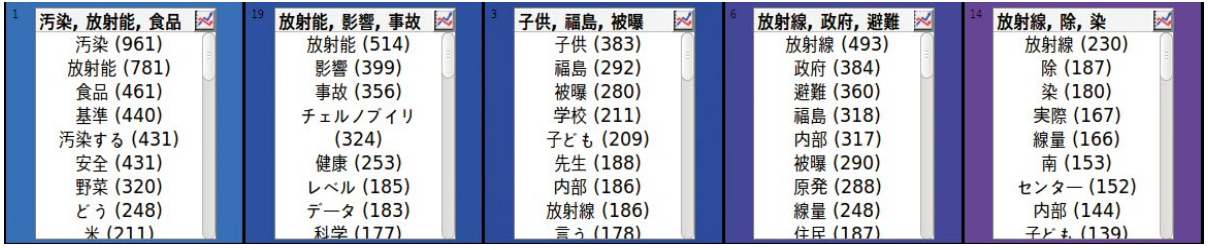
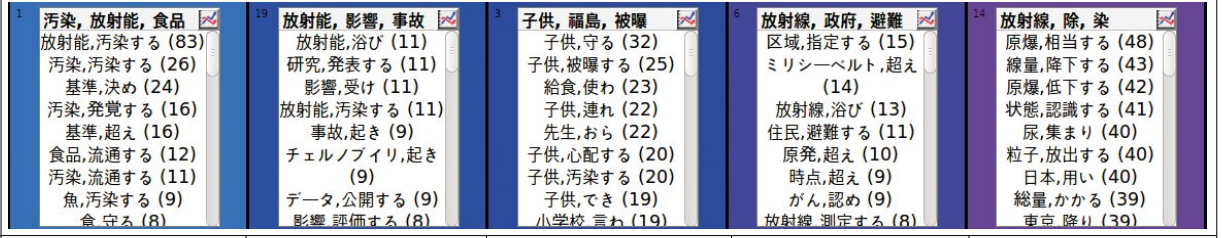
Für die TopicExplorer-Studie wurden Blog-Einträge in japanischer Sprache aus dem Internet maschinell heruntergeladen, die drei Stichworte, nämlich „Rind", „Cäsium" und „Atomkraft" *(genpatsu)*, enthalten. Letzteres Stichwort wurde hinzugefügt, um den thematischen Fokus der Texte sicherzustellen. Für den Untersuchungszeitraum von März bis September 2011 konnten so 1931 Texte mit einer durchschnittlichen Länge von 2402 Worten gesammelt und anonymisiert werden. Das Korpus von Blog-Texten wurde nach linguistischer Vorverarbeitung dann einer automatischen Inhaltsanalyse durch den TopicExplorer unterzogen. Die so ermittelten Themen sind durch eine Liste ihnen zugeordneter Worte charakterisiert, die vom TopicExplorer noch durch eine Aufstellung der häufigsten im jeweiligen Thema vorkommenden Kollokationen ergänzt wird (Tabelle 2). Die anhand von Stichworten und Kollokationen interpretierten Themen, denen je ein Thementitel zugeordnet wird, können dann anhand der Textpassagen, in denen die Themen behandelt werden, überprüft werden. Dadurch lässt sich die Plausibilität der Themenanalyse direkt an den Blog-Inhalten verifizieren, und die wichtigsten Textpassagen zu einem bestimmten Thema können schnell aufgefunden werden. In der separaten Zeitansicht (Abbildung 8) kann die Entwicklung eines Themas im Untersuchungszeitraum oder einem Ausschnitt davon interaktiv verfolgt werden, z. B. auch im Vergleich zu Zeitreihen weiterer Themen oder zur Durchschnittszeitreihe aller Themen.

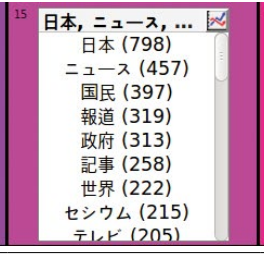
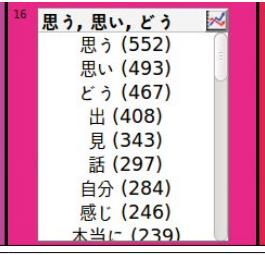
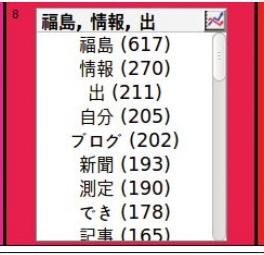
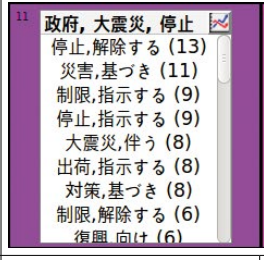
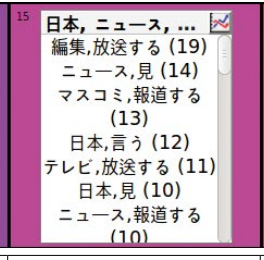
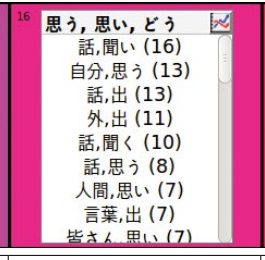
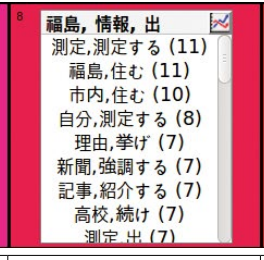
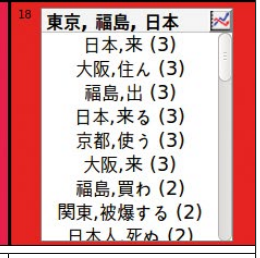
Die zwanzig Themen, die der Algorithmus im Blog-Korpus zum Cäsium-Skandal gebildet hat, sind in Tabelle 2 in der Reihenfolge aufgeführt, in der sie vom TopicExplorer arrangiert werden. In den Spalten sind nach der Nummer des jeweiligen Themas die Spitzeneinträge aus den Listen der häufigsten Stichworte

Tabelle 2: Zwanzig Themen im Blog-Korpus

Thema	2	5	9	7	12
Charakt. Worte	2 放射, 物質, セシ… 放射 (800) 物質 (665) セシウム (456) 素 (333) ヨウ (329) 放射線 (327) 影響 (277) 半減 (208) 体内 (201)	5 放射, セシウム, … 放射 (676) セシウム (615) ベクレル (588) 検出する (578) 基準 (503) 暫定 (418) 物質 (363) 超える (325) 当たり (248)	9 稲, わら, セシウム 稲 (714) わら (682) セシウム (586) 農家 (585) 牛 (542) 与え (411) 肉牛 (365) 汚染 (327) 福島 (309)	7 福島, 牛, 出荷 福島 (702) 牛 (682) 出荷 (502) 検査 (488) 検出する (457) 出荷する (453) 肉 (434) 農家 (381) セシウム (367)	12 牛, セシウム, 食べ 牛 (981) セシウム (979) 食べ (794) 牛肉 (674) 被害 (485) 汚染する (412) 肉 (373) 風評 (316) 汚染 (308)
Charakt. Kollokationen	2 放射, 物質, セシ… 体内,取り込ま (30) 大気,放出する (27) 物質,拡散する (21) 物質,飛散する (21) 物質,放出する (21) 物質,吸収する (20) 物質,付着する (19) 骨,蓄積する (19) 物質,含ま (18)	5 放射, セシウム, … セシウム,検出する (133) ベクレル,検出する (128) 放射,検出する (119) 基準,超える (97) ベクレル,超える (81) 基準,超え (47) 物質,検出する (44)	9 稲, わら, セシウム わら,与え (162) 肉牛,与え (58) 餌,与え (55) 農家,出荷する (45) セシウム,汚染する (38) 飼料,与え (35) 肉牛,出荷する (32) 牛,与え (32)	7 福島, 牛, 出荷 セシウム,検出する (86) 農家,出荷する (80) 相馬,出荷する (71) 放射,検出する (70) 規制,超える (60) 牛,出荷する (53) 暫定,超える (41) 区域,出荷する (36)	12 牛, セシウム, 食べ 牛肉,食べ (103) 牛,食べ (82) セシウム,汚染する (67) セシウム,食べ (59) 肉,食べ (55) ワラ,食べ (32) 牛肉,流通する (30) 藁,食べ (26)
Thementitel: Charakt. Worte (u. a.); Charakt. Kollokationen (u. a.)	**Ausbreitung u. Aufnahme radioaktiver Substanzen:** Strahlung, Substanz, Cäsium (456), Element, Jod; in den Körper aufgenommen werden, [in die] Atmosphäre freisetzen, Substanz verstreuen	**Nachweis von Cäsium:** Strahlung, Cäsium (615), Becquerel, feststellen, Grundlage; Cäsium feststellen, Becquerel feststellen, Strahlung feststellen	**Ursache:** Reis, Stroh, Cäsium (586), Landwirt, Rind (542); Stroh geben, Schlachtvieh [Futter] geben, Futter geben, Landwirte verschicken [Stroh]	**Entdeckung/ Kontrolle von Cäsium-verseuchtem Rindfleisch:** Fukushima, Rind (682), Lieferung, Untersuchung, Cäsium (367); Cäsium feststellen, Landwirte liefern [Fleisch], Strahlung feststellen	**Schäden durch Cäsium-verseuchtes Rindfleisch:** Rind (981), Cäsium (979), essen, Rindfleisch, Schaden, verseuchen; Rindfleisch essen, Rind essen, [mit] Cäsium verseuchen

Thema	17	0	10	13	4
Charakt. Worte					
Charakt. Kollokationen					
Thementitel: Charakt. Worte (u. a.); Charakt. Kollokationen (u. a.)	**Probleme der Rinderzüchter**: Rind (939), Landwirt, Viehzucht, Maul, Cäsium (82); Rinder/Haustiere/Schweine töten, Seuche tötet [Tiere], Rinder züchten, [mit dem Atom] Unfall einhergehen, Seuche auslösen	**Verantwortung für Unfallfolgen**: Atomkraftwerk, [Premiermin.] Kan, TEPCO, Verantwortung, Politik; Verantwortung prüfen, [als] Premierminister zurücktreten, [als] Regierungschef zurücktreten	**Zukunft der Atomkraft in Japan**: Atomkraftwerk, Unfall, Verseuchung; [von] Atomkraftwerk entfernen, [über] Atomkraftwerk nachdenken, Atomkraftwerk [nicht kontrollieren] können	**Atomkraftwerksbetrieb und -wieder -aufnahme**: Atomkraftwerk, Fukushima, Unfall; [Kraftwerks]Betrieb anhalten, Unfall erleiden, Brennstoff einsetzen	**Forderungen**: Staat, Messung/Untersuchung, Maßnahme, Rind (23) Untersuchungen/Nachforschungen durchführen, Plan vorlegen, Maßnahmen fordern

Thema	1	19	3	6	14
Charakt. Worte					
Charakt. Kollokationen					
Thementitel: Charakt. Worte (u. a.); Charakt. Kollokationen (u. a.)	**Verseuchung von Lebensmitteln**: Verseuchung, Radioaktivität, Lebensmittel, Cäsium (90); [mit] Radioaktivität verstrahlen, verseuchen, [Mess]Grundlage festlegen	**Erfahrungen aus Tschernobyl**: Radioaktivität, Unfall, Einfluss, Tschernobyl, Gesundheit; Radioaktivität ausgesetzt sein, Forschung veröffentlichen, Einfluss erleiden	**Schäden durch radioaktives Cäsium bei Kindern**: Kinder, Fukushima, Strahlenschäden, Schule, Cäsium (129); Kinder schützen, Kinder erleiden Strahlenschäden, bei der Schulspeisung [verstrahlte Lebensmittel] einsetzen,	**Messwerte radioaktiver Strahlung und Gefahr eines Aufenthalts**: Radioaktive Strahlung, Regierung, Flucht; [Sperr]Gebiet festlegen, [Werte von X] Milli-Siebert übersteigen, radioaktiver Strahlung ausgesetzt sein, Einwohner flüchten	**Radioaktivität und Dekontamination**: Radioaktive Strahlung, Dekontamination, Strahlenmenge, Cäsium (63); [so und so vielen] Atombomben entsprechen, Strahlungsmenge niedergehen, Atombomben [strahlung] reduziert [sich]

Thema	11	15	16	8	18
Charakt. Worte					
Charakt. Kollokationen					
Themen-titel: Charakt. Worte (u. a.); Charakt. Kollokationen (u. a.)	**Maßnahmen nach Atomunfall, u. a. Rind-fleischlieferstopp**: Regierung, Großes Erd-beben, Stopp, Erleiden einer Katastrophe; Stopp aufheben, Katas-trophe [...] basieren auf, Begrenzung anordnen, Stopp anordnen	**Kommentare/Zsfg. zur Medienberichter-stattung**: Japan, Nachrichten, japanisches Volk, Berichterstattung; Cäsium (215); Edition senden, Nach-richten sehen, Massen-medien berichten	**Persönliche Wahrneh-mung**: Denken, sehen, sagen, selbst, fühlen; Erzählung hören, selbst denken, Erzählung entstehen	**SelbstMessungen von Radioaktivität**: Fukushima, Information, selbst, Blog, Zeitung, Messung; Messung durchführen, [in] Fuku-shima wohnen, [in der] Stadt wohnen, selbst messen	**Atomkraft-werksunfall u. seine Folgen**: Tokio, Fukushima, Japan, Kantō, Osaka; [nach] Japan kommen, [in] Osaka wohnen, Fukushima verlassen

und Kollokationen für das jeweilige Thema in japanischer Sprache entsprechend der jeweiligen TopicExplorer-Ansicht wiedergegeben. Diese Listen haben oft einen Gesamtumfang von einigen Hundert Einträgen und werden von der Analysesoftware für jedes Thema erstellt. In der folgenden Zeile ist eine deutsche Benennung des Themas (=Thementitel) eingetragen, die durch Interpretation der Stichwortliste, der Kollokationen und der repräsentativen Texte des Themas gewonnen wurde. Es folgen die deutschen Übersetzungen ausgewählter Stichworte und Kollokationen. Bei den Themen, die u. a. auch durch die Stichworte „Cäsium" oder „Rind" charakterisiert sind, ist zusätzlich die Häufigkeit dieser Stichworte vermerkt. Beispielsweise sind dem Thema 9 „Ursache", „Cäsium" und „Rind" 586-mal bzw. 542-mal zugeordnet.

Die Themen 7, 9, 12 und 17 sind am engsten mit dem eigentlichen Skandalgeschehen verbunden. Alle vier Themen zeichnen sich durch häufiges Auftreten der Worte „Cäsium" und insbesondere „Rind" aus. Bei Thema 7 weisen die Stichworte und besonders die typischen Kollokationen darauf hin, dass es die Entdeckung und nachfolgende Überwachung von Cäsium-verseuchtem Rindfleisch zum Gegenstand hat. Die genauere Durchsicht der repräsentativen Textstellen bestätigt, dass hier die Aufdeckung des Skandals, dass von radioaktivem Cäsium verseuchtes Rindfleisch aus Fukushima zu Lebensmitteln verarbeitet worden war, thematisiert wird.

Unter den häufigsten Stichworten zu Thema 9 findet sich „Reis-Stroh" an oberster Stelle. Die für dieses Thema repräsentativsten Texte beschreiben die bereits früh erkannte Ursache des Skandals, nämlich dass Reis-Stroh im Freien gelagert worden und dort radioaktivem Fallout ausgesetzt war, bevor es dann an Rinder verfüttert wurde. In Thema 12 sind Textpassagen zusammengefasst, die über die Folgen bzw. Schäden, die der Konsum von radioaktiv verseuchtem Rindfleisch mit sich bringen kann, berichten. Ebenfalls auf den Cäsium-Skandal im engeren Sinne nehmen auch die typischen Textstellen des Themas 17 Bezug, die die Probleme und das Vorgehen der Rinderzüchter behandeln, die teilweise schon vor der Nuklearkatastrophe mit der Maul-und-Klauenseuche zu kämpfen hatten und nun erneut erhebliche Verluste ihrer Bestände hinnehmen müssen.

Unter den Themen 4 und 11 sind Textstellen gebündelt, die die Maßnahmen der Regierung und die Reaktionen darauf wiedergeben. Die typischen Kollokationen zu Thema 4, dem auch das Wort „Rind" an 23 Stellen zugeordnet ist, lauten „Untersuchungen/Nachforschungen durchführen", „Plan vorlegen" und „Maßnahmen fordern". Hier wird u. a. die Lebensmittelkontrolle der Rindfleischbestände und die Forderung nach ihrer Verschärfung, um die Konsumenten zu schützen, Vertrauen wiederherzustellen und dadurch die Auslieferung von Fleisch wieder zu ermöglichen, angesprochen. Unter Thema 11 taucht das Stichwort „Rind" zwar nicht auf, aber wie die Stichworte „Regierung", „anhalten" und „Katastrophe erleiden" sowie die Kollokationen „Stopp aufheben", „Begrenzung

anordnen“ und „Stopp anordnen“ zeigen, thematisieren die hier gesammelten Blog-Auszüge die Maßnahmen im Gefolge der Reaktorkatastrophe, zu denen u. a. auch der Auslieferstopp für bestimmte Lebensmittel gehörte.

Die sechs Themen 1, 2, 3, 5, 14 und 15 haben das Auftreten von „Cäsium“ unter den charakteristischen Stichworten gemeinsam, ohne dass das Wort „Rind“ präsent wäre. Thema 1 zu den Auswirkungen auf „Lebensmittel“ wird u. a. charakterisiert durch die Stichworte „Verseuchung“, „Radioaktivität“, „Lebensmittel“, „Grenzwert“, „verseuchen“, „Sicherheit“, „Gemüse“ und, mit etwas geringerer Häufigkeit, „Cäsium“. Die häufigste Kollokation für dieses Thema lautet „[mit] Radioaktivität verstrahlen“. Die charakteristischen Einträge für Thema 1 illustrieren die Sorgen der Blogger, die in Folge des Rindfleischskandals die radioaktive Verseuchung der Lebensmittelversorgung als persönliche Bedrohung erkennen. Beispielsweise kommentiert ein Blogger am 20. Juli 2011 ausdrücklich: „[Man muss] den eigenen Körper selbst schützen. So ein Zeitalter ist angebrochen“ (*jibun no mi wo jibun de mamoru. so iu jidai ni natta*).

Die Detailanalyse repräsentativer Textpassagen ergibt für Thema 1 außerdem, dass hier vielfach Medienauftritte von Koide Hiroaki (*1949) wiedergegeben werden. Insgesamt wird Koide in 108 Textstellen mit Bezug zu Thema 1 erwähnt. Koide, ein auf Nukleartechnik spezialisierter Ingenieur, war Assistenzprofessor am Kyoto University Research Reactor Institute (KURRI). Er ist einer der wenigen japanischen Nuklearspezialisten, die schon frühzeitig Japans Ausstieg aus der Kernenergie gefordert haben – eine Unabhängigkeit, für die er mit der Stagnation seiner Universitätskarriere auf der Stufe des „ewigen Assistenzprofessors“ (*mannen no jokyō*) bis zu seiner Pensionierung 2015 bezahlte. Am 23. Mai 2011 sagte Koide als Experte vor einem Ausschuss des japanischen Parlaments aus. Obwohl an diesem Tag insgesamt vier prominente Kritiker der Atomenergie angehört wurden, übertrug das japanische Fernsehen, das sonst regelmäßig *live* aus dem Parlament sendete, diese Sitzung nicht, und auch die Zeitungen berichteten darüber nur mit kurzen Zusammenfassungen. Im Blog-Korpus dagegen waren erste Berichte schon am Folgetag der Ausschusssitzung zu finden, die auch danach noch oft zitiert wurden. Besonders häufig wurde Koides Aussage wiedergegeben, dass es auch unter seinen Kollegen viele gebe, denen gesagt worden sei, dass sie alarmierende Messergebnisse nicht veröffentlichen sollten *(watashi no dōryō demo, kenshutsu shita dēta wo kōhyō shinai you iwareta hitotachi ga nannin mo imasu)*, um „Panik“ zu vermeiden.

Thema 15 umfasst Passagen aus 215 Texten, die das Wort „Cäsium“ enthalten, und bündelt Kommentare und Zusammenfassungen zur japanischen Medienberichterstattung über Fukushima und seine Folgen, wie die charakteristischen Stichworte „Japan“, „Nachrichten“, „japanisches Volk“ und „Berichterstattung“ sowie die Kollokationen „Freiheit [von] Berichterstattung“, „Edition senden“, „Nachrichten sehen“ und „Massenmedien berichten“ erkennen lassen. Thema 14

schließlich nimmt nur an 63 Stellen direkten Bezug auf Cäsium. Wie die Stichworte „radioaktive Strahlung", „Dekontamination" und „Strahlenmenge" sowie die Kollokationen „[so und so vielen] Atombomben entsprechen", „Strahlungsmenge niedergehen" und „Atombomben[strahlung] reduziert [sich]" vermuten lassen, behandeln diese Textstellen die dramatische Menge freigesetzter Radioaktivität und die dadurch notwendig gewordenen Dekontaminationen.

Die übrigen Einzelthemen wurden ohne das Stichwort „Cäsium" gebildet. Sie behandeln zentrale Aspekte des Kernkraftwerkunfalls, die teilweise aber auch mit dem Cäsium-Rindfleisch-Skandal in Verbindung gebracht werden. Beispielsweise sind unter Thema 0, das u. a. durch die Wörter „Atomkraft", „[Premierminister] Kan", „TEPCO", „Verantwortung" und „Politik" charakterisiert wird, Fragen nach den Konsequenzen und nach der Verantwortung u. a. für den konkreten Umgang mit dem Kraftwerksunfall erfasst. Darauf weisen auch die am häufigsten verwendeten Kollokationen hin wie z. B. „Verantwortung prüfen", „Premierminister zurücktreten" und „Premierminister [Amt] aufgeben". Dabei werden auch in Verbindung mit dem Cäsium-Skandal Anschuldigungen gegen die damals amtierende Regierung erhoben. So behauptete ein Blogger am 21. Juli 2011, „die Ausbreitung Cäsium-verseuchter Rinder ist ein Verbrechen der Kan-Regierung" *(seshiumu osen gyu no kakudai ha Kan seiken no hanzai da).* Auf die übrigen Themen wird aus Platzgründen hier nicht näher eingegangen.

3.3 Die zeitliche Entwicklung der Themen und der Einfluss einzelner Akteure

Der zeitliche Verlauf der Einzelthemen unterstützt ihre aus Stichworten, Kollokationen und Beispieltexten gewonnene Interpretation. Abbildung 8 zeigt den zeitlichen Verlauf aller Themen, die in den Blogs zum Skandal um das mit radioaktivem Cäsium verseuchte Rindfleisch identifiziert wurden. Insgesamt markieren die Themen 7 „Entdeckung" und 9 „Ursache" deutlich das Bekanntwerden des Cäsium-Skandals in der zweiten Juliwoche. Die früheste Spitze erreicht naturgemäß Thema 7, da es die Entdeckung des Skandals zum Gegenstand hat. Neun der zehn repräsentativsten Textstellen dieses Themas datieren auf den 9. Juli bzw. die Tage unmittelbar danach, als die Medien erstmals über den Skandal berichteten. Auch das Thema 5 „Nachweis…", das die Bestimmung der Cäsiumwerte beschreibt und insofern mit Thema 7 eng verwandt ist, erreicht in dieser Woche seinen Spitzenwert. Die für Thema 9 „Ursache" repräsentativsten Texte datieren hingegen überwiegend auf die zweite Juli-Hälfte, als man Reis-Stroh als „Überträger" von radioaktivem Cäsium auf Rinder erkannte. Es überrascht nicht, dass die Zahl der Textstellen von Thema 9 daraufhin besonders stark abnimmt: Nachdem die Ursache des Skandals erst einmal gefunden worden war, stieß sie bald weniger auf Interesse als ihre Folgen. Folglich treten mit Thema 12

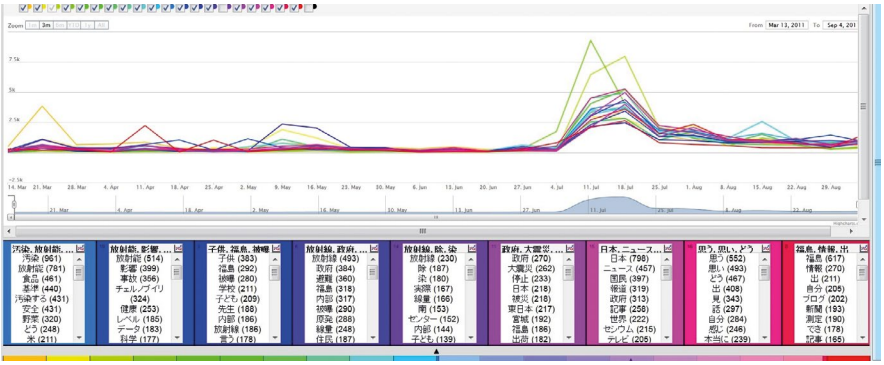


Abbildung 8: 19 Themen im Zeitverlauf (ohne Thema 14).

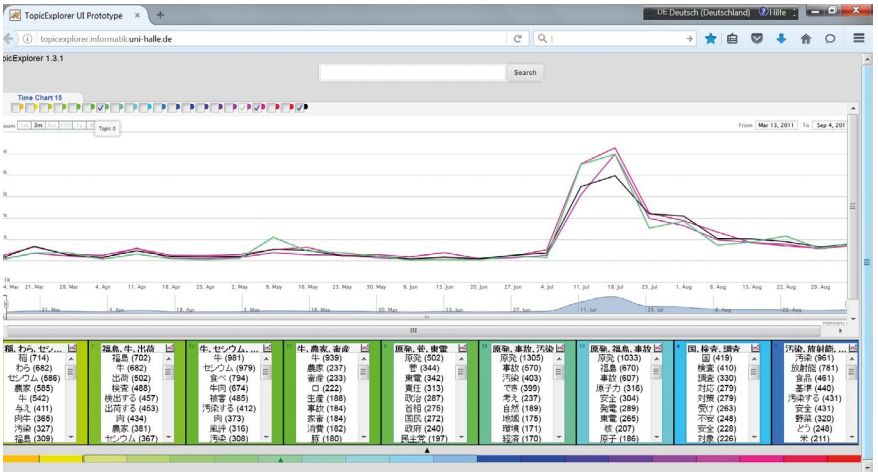


Abbildung 9: Themen 0, 15 und 16 im Vergleich zum Durchschnitt Zeitverlauf.

„Schäden…“ schon in der zweiten Juli-Hälfte Textpassagen, die von den Folgen bzw. Schäden, die der Konsum von radioaktiv verseuchtem Rindfleisch mit sich bringen kann, berichten, besonders häufig auf.

Mit dem Bekanntwerden des Skandals weisen alle übrigen Themen der Blogs einen Anstieg auf, der aber weniger prononciert ausfällt als bei den o.g. vier Themen und in seiner Größenordnung weitgehend dem Durchschnitt der Themen im Gesamtkorpus entspricht. Die Themen 0, 15 und 16 treten dabei noch am stärksten, nämlich überdurchschnittlich, hervor (Abbildung 9). Die Zunahme von Thema 16 „persönliche Wahrnehmung“, das persönliche Eindrücke zusammenfasst, lässt die große emotionale Anteilnahme am Geschehen vermuten. In der darauffolgenden Woche bleibt nicht nur dieses Thema prominent, sondern

auch Thema 15 „Berichterstattung und Kommentare“ schließt dicht auf, das auf eine intensive Auseinandersetzung mit der medialen Berichterstattung schließen lässt. Schließlich steigt Thema 0 „Verantwortung…“ im Zuge des Skandals ähnlich steil an wie Thema 15, sodass zu vermuten ist, dass der Skandal die Bewertung der Rolle des Kraftwerkbetreibers von Fukushima und der japanischen Regierung beeinflusst.

3.4 Schlussfolgerungen: Neue Medien als Artikulationsraum

Yanagisawa (2012) hat die Berichterstattung über den Cäsium-Skandal anhand von ca. 1000 japanischen Zeitungsartikeln für den gleichen Zeitraum detailliert untersucht. Ein Vergleich seiner Ergebnisse mit den Themen, die mit Hilfe des TopicExplorers im Blog-Korpus identifiziert wurden, zeigt, dass sowohl inhaltlich als auch in ihrer spezifischen Abfolge die Themen weitgehend korrespondieren (Tabelle 3). Beispielsweise hat Thema 9 im Blog-Korpus, das die Ursache des Skandals beschreibt, einen ganz ähnlichen Gegenstand wie Cluster 8 bei Yanagisawa. In den Blog-Texten werden sämtliche Themen, die in den Printmedien behandelt werden, ebenfalls aufgegriffen. Die bei der Beschreibung der einzelnen Themen behandelten Beispiele illustrieren, dass die konventionellen Medien auf die Inhalte der Blogs einen großen Einfluss ausüben – schon alleine deshalb, weil mediale Inhalte vielfach zitiert und die darin enthaltenen Informationen in den Blogs oft kommentiert werden.

Im Blog-Korpus finden sich darüber hinaus jedoch auch zahlreiche Gegenstände, die die Printmedien nicht thematisieren. Sie sind in den unteren Zeilen von Tabelle 3 aufgeführt. So findet im Internet eine kritische Auseinandersetzung mit der Berichterstattung der konventionellen Medien statt (Thema 15), es wird über die Notwendigkeit und die Ergebnisse unabhängiger Radioaktivitätsmessungen berichtet (Thema 8), und Erfahrungen aus Tschernobyl (Thema 19) werden ebenso weitergegeben wie Appelle für eine rasche und flächendeckende Dekontamination (Thema 14). An diesen Punkten zeigen sich die Interessen und Sorgen der Bevölkerung, auf die die konventionellen Medien nicht eingehen und die daher im Sinne einer unabhängigen „Artikulation“ im Internet besprochen werden.

Tabelle 3: Entsprechungen von Clustern nach Yanagisawa (2012: 71) mit Topic-Explorer-Themen (Quelle: Eigene Zusammenstellung).

Cluster	**Cluster-Nr.**	**Themen-Nr.**
Schaden	1	12
Atomkraftwerksunfall und seine Folgen	2	6, 10, 13, 18
Forderungen	3	4
Lebensmittel	4	1
Auswirkungen auf den menschlichen Körper	5	2, 3
Gegenwärtige Lage	6	5, 7
Gegenmaßnahmen	7	11
Ursachen	8	9
Verantwortung für Unfallfolgen		0
(Selbst)Messungen von Radioaktivität		8
Dekontamination		14
Medienkritik		15
Persönliche Wahrnehmung		16
Probleme der Rinderzüchter		17
Erfahrungen von Tschernobyl		19

Aber auch innerhalb der Themen, die Blogs und Printmedien gemeinsam haben, zeigen sich unterschiedliche Schwerpunktsetzungen. Dies ist oft auf den besonderen Einfluss zurückzuführen, den bestimmte Persönlichkeiten auf das Internet ausüben. So ist davon auszugehen, dass die Kritik, die z. B. fachliche Autoritäten wie Koide Hiroaki am Handeln der Regierung äußerten und die im Internet weite Verbreitung fand, das Vertrauen der Internetnutzer in die Lebensmittelsicherheit beeinflusst hat. Es ist daher anzunehmen, dass trotz des Einflusses, den die traditionellen Medien besitzen, Personen mit ausgewiesener Expertise ihren Ansichten durch die Verbreitung von Blogbeiträgen zu einer großen gesellschaftlichen Präsenz verhelfen können.

4. Fazit und Perspektiven

Mit dem TopicExplorer wurde ein Werkzeug entwickelt, das Personen ohne Programmierkenntnisse die explorierende Analyse großer Textsammlungen ermöglicht und dadurch einer breiten Gruppe von Anwendern Zugang zu neuen Forschungsmöglichkeiten mit Big Data eröffnet. In zukünftigen Arbeitsschritten soll insbesondere die Benutzerfreundlichkeit weiter erhöht und die Steuerung des Systems vereinfacht werden. Außerdem sollen in den TopicExplorer Qualitäts- und Kohärenzmaße für jene automatisch berechneten Themen eingebaut

werden, die mit menschlichen Bewertungen von Themen hinsichtlich ihrer Interpretierbarkeit eng korrelieren (Röder et al. 2015). Somit können auch Nutzer ohne vertiefte Kenntnisse statistischer Verfahren Topic Models intuitiv bewerten und aussagekräftige Themen identifizieren.

Die Beispielanalyse japanischer Blogs mithilfe des TopicExplorers hat gezeigt, dass viele Japanerinnen und Japaner den Lebensmittelskandal um Cäsium-verseuchtes Rindfleisch in den alten und neuen Medien kritisch verfolgen. Außerdem lässt die Analyse Aspekte in den Blog-Einträgen erkennen, die in den gedruckten Medien nicht in vergleichbarer Weise zum Ausdruck kommen, wie etwa die große Sorge der Bevölkerung vor gesundheitlichen Schäden und das Interesse an Informationen, die als regierungsunabhängig eingeschätzt werden. Der Einfluss der sozialen Medien geht dabei bereits so weit, dass sich das japanische Industrie- und Wirtschaftsministerium METI veranlasst sah, ein Unternehmen zu beauftragen, die Diskussion über Kernenergie auf dem Forum „2channel" und auf Twitter zu überwachen – angeblich, um wirtschaftlichen Schäden durch Verbreitung falscher Gerüchte vorzubeugen.

5. Bibliografie

Alexander, Eric, Joe Kohlmann, Robin Valenza, Michael Witmore und Michael Gleicher. 2014. "Serendip: Topic Model-Driven Visual Exploration of Text Corpora." In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 173–182, https://doi.org/10.1109/VAST.2014.7042493.

Anthes, Gary. 2010. Topic models vs. unstructured data, *Commun. ACM* 53, no. 12 (December): 16–18. https://doi.org/10.1145/1859204.1859210.

Armbrust, Michael, Reynold S. Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K. Bradley, Xiangrui Meng, Tomer Kaftan, Michael J. Franklin, Ali Ghodsi und Matei Zaharia. 2015. "Spark SQL: Relational data processing in spark." In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15)*, New York, NY, USA, 2015, 1383–1394. https://doi.org/10.1145/2723372.2742797.

Blei, David M. 2012. "Probabilistic Topic Models." *Commun. ACM* 55, no. 4 (April): 77–84.

Chang, Jonathan, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang und David M. Blei, 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS 2009)*, Vancouver, British Columbia, Canada , December 07–10, 2009, 288–296.

Chuang, Jason, Spence Green, Marti Hearst, Jeffrey Heer und Philipp Koehn, Herausgeber. 2014. *Proceedings of the Workshop on Interactive Language*

Learning, Visualization, and Interfaces. Association for Computational Linguistics, Baltimore, Maryland, USA, June 2014.

Cui, Weiwei, Shixia Liu, Zhuofeng Wu und Hao Wei. 2014. "How Hierarchical Topics Evolve in Large Text Corpora." *IEEE Transactions on Visualization and Computer Graphics,* 20, no. 12 (December): 2281–2290. https://doi.org/10.1109/TVCG.2014.2346433.

Derntl, Michael, Nikou Günnemann, Alexander Tillmann, Ralf Klamma und Matthias Jarke. 2014. "Building and Exploring Dynamic Topic Models on the Web." In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM '14),* New York, NY: ACM, 2012–2014. http://dx.doi.org/10.1145/2661829.2661833.

Dou, Wenwen, Li Yu, Xiaoyu Wang, Zhiqiang Ma und W. Ribarsky. 2013. "Hierarchical Topics: Visually Exploring Large Text Collections Using Topic Hierarchies." *IEEE Transactions on Visualization and Computer Graphics,* 19, no. 12 (December): 2002–2011. https://doi.org/10.1109/TVCG.2013.162.

Evangelopoulos, Nicholas und Lucian Visinescu. 2012. "Text-Mining the Voice of the People." *Commun. ACM* 55, no. 2 (February): 62–69.

Gohr, André, Myra Spiliopoulou und Alexander Hinneburg. 2010. "Visually Summarizing the Evolution of Documents under a Social Tag." In *Proceedings of LWA2010 – Workshop-Woche: Lernen, Wissen & Adaptivität,* Kassel, Germany 2010, 85–94, http://users.informatik.uni-halle.de/~hinnebur/PS_Files/kdir2010_TT.pdf.

Gohr, André, Myra Spiliopoulou und Alexander Hinneburg. 2013. "Visually Summarizing Semantic Evolution in Document Streams with Topic Table." In *Knowledge Discovery, Knowledge Engineering and Knowledge Management,* herausgegeben von A. Fred, J. L. G. Dietz, K. Liu und J. Filipe. Berlin: Springer, 136–150 (Communications in Computer and Information Science).

Hinneburg, Alexander, Rico Preiss und René Schröder. 2012. "TopicExplorer: Exploring Document Collections with Topic Models." In *Machine Learning and Knowledge Discovery in Databases,* herausgegeben von Peter A. Flach, Tijl De Bie, Nello Cristianini. Berlin: Springer, 838–841 (Lecture Notes in Computer Science 7524).

Hinneburg, Alexander, Frank Rosner, Stefan Pessler und Christian Oberländer. 2014. "Exploring Document Collections with Topic Frames." In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM '14),* New York, NY: ACM, 2084–2086.

Isaacs, E., K. Damico, S. Ahern, E. Bart und M. Singhal. 2014. "Footprints: A Visual Search Tool that Supports Discovery and Coverage Tracking." *IEEE Transactions on Visualization and Computer Graphics* 20, no. 12 (December): 1793–1802.

Jojic, Nebojsa, Alessandro Perina und Dongwoo Kim. 2016. "Hierarchical learning of grids of microtopics." In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence (UAI'16)*, New Jersey, USA, June 25–29, Arlington, Virginia: AUAI Press, 299–308.

Kingston, Jeff. 2012. "The Politics of Disaster, Nuclear Crisis and Recovery." In *Natural Disaster and Nuclear Crisis in Japan. Response and Recovery after Japan's 3/11,* herausgegeben von Jeff Kingston. London: Routledge, 188–206.

Kuchinskaya, Olga. 2011. "Articulating the signs of danger: Lay experiences of post-Chernobyl radiation risks and effects." *Public Understanding of Science* 20, no. 3: 405–21.

Liu, Shixia, Michelle X. Zhou, Shimei Pan, Weihong Qian, Weijia Cai und Xiaoxiao Lian. 2009. "Interactive, Topic-Based Visual Text Summarization and Analysis." In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (*CIKM '09)*, New York, NY: ACM, 543–552, http://dx.doi.org/10.1145/1645953.1646023.

Liu, Shixia, Xiting Wang, Jianfei Chen, Jun Zhu und Baining Guo. 2014. "Topicpanorama: A Full Picture of Relevant Topics." In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, October 2014, 183–192.

Murdock, Jaimie und Colin Allen. 2015. "Visualization Techniques for Topic Model Checking." In *29th AAAI Conference on Artificial Intelligence (AAAI-15),* 4284–4285.

Pan, Shimei, Michelle X. Zhou, Yangqiu Song, Weihong Qian, Fei Wang und Shixia Liu. 2013. "Optimizing Temporal Topic Segmentation for Intelligent Text Visualization." In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI'13)*, 339–350, New York, NY: ACM, 339–350, https://doi.org/10.1145/2449396.2449441.

Ramage, Daniel, Evan Rosen, Jason Chuang, Christopher D. Manning und Daniel A. McFarland. 2009. "Topic Modeling for the Social Sciences." In *Workshop on Applications for Topic Models (NIPS 2009): Text and Beyond*, Whistler, Canada, December 2009. http://vis.stanford.edu/files/2009-TopicModels-NIPS-Workshop.pdf.

Röder, Michael, Andreas Both und Alexander Hinneburg. 2015. "Exploring the Space of Topic Coherence Measures." In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*, New York, NY: ACM, 399–408.

Sievert, Carson und Kenneth Shirley. 2014. "LDAvis: A Method for Visualizing and Interpreting Topics." *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics, 63–70. https://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf.

Sun, Guodao, Yingcai Wu, Shixia Liu, Tai-Quan Peng, J.J.H. Zhu und Ronghua Liang. 2014. "Evoriver: Visual Analysis of Topic Coopetition on Social Media." *IEEE Transactions on Visualization and Computer Graphics* 20, no. 12 (December): 1753–1762.

Yanagisawa, Toru. 2012. "Keiryō tekisuto bunseki ni yoru media-furēmu no tansakuteki kentō – 'hōshasei seshiumu osen-gyū mondai' no shinbun kiji wo tōshite." (An Exploratory Study of Media Frames in Quantitative Text Analysis. Through the Newspaper Report on the Problems with Radioactive Cesium-contaminated Beef). *Shakai jōhō-gaku* 1, no. 2: 61–76.

Velislava Todorova / Maria Chinkina

Significance Filters for N-gram Viewer

Abstract This paper presents a visualization tool for the analysis of tendencies in language use over time. Given a dated and tokenized corpus, it calculates frequencies of selected n-grams and visually presents them as data points on a line chart in a coordinate system, with time on the x axis and relative frequency on the y axis. It provides the option of smoothing the graph in order to make the general tendency more salient. The user can specify an n-gram as a sequence of tokens, lemmas, and/or POS tags, if the corpus provides these annotations. Along with the original text, the tool also accesses the metadata of the corpus, such as dates and authors' names, allowing for a comparison of the use of n-grams by different authors at different time periods in context. The latest version of our tool introduces a filtering mechanism that indicates the periods of time throughout which the observed values within one or more datasets are significantly different. We used Fisher's exact test of independence because it has the advantage of providing reliable results even for sparse data.

1. Introduction

Exploration of the patterns in language use over time is useful for a number of Natural Language Processing (NLP) tasks such as authorship attribution and topic detection. Google Ngram Viewer[1] is one example of such a tool. While providing the functionality of querying the Google Books corpus[2], its current functionality does not include uploading and processing one's own text corpora. This limitation is overcome by our new Ngram Tendency Viewer *Slash/A*[3], which is more suitable for researchers interested in exploring a specific collection of texts.

Given a dated and tokenized corpus, *Slash/A* calculates and visually presents frequencies of selected n-grams as a line chart in a coordinate system, with time

1 The online Google Ngram Viewer is available here: https://books.google.com/ngrams
2 The Google Books corpus can be accessed here: https://books.google.com
3 The online *Slash/A* tool can be accessed here: https://tinyurl.com/slasha-tool

Published in: Noah Bubenhofer, Marc Kupietz: Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

on the x axis and relative frequency on the y axis and provides the option of smoothing the graph in order to make the general tendency more salient. The user can specify an n-gram as a sequence of tokens, lemmas and/or POS tags, if the corpus provides these annotations. Along with the original text, the tool also accesses the metadata of the corpus, such as dates and authors' names, allowing for a comparison of the use of n-grams by different authors at different time periods in context.

The latest version of our tool introduces a filtering mechanism that indicates whether the observed values in a specified time period are significantly different in terms of (i) lower and higher extremes of n-gram frequencies and (ii) use of the same n-gram by different authors. In these cases, a significance filter can facilitate scientific hypothesis testing. The statistical test that we decided to use is Fisher's exact test of independence (Fisher, 1950). It is very similar to the $\chi 2$ test of independence but has the advantage of providing reliable results even in cases with very little data, e.g., if an n-gram only occurs five times in the whole corpus.

```
<?xml version="1.0" encoding="UTF-8"?>
  <correspondence to="E" from="R"></correspondence>
  <date>
    <written date="1845-01-10"></written>
  </date>
  <text>
    New Cross, Hatcham, Surrey.
    I love your verses with all my heart, dear Miss Barrett, [...]
    R. Browning.
  </text>
  <tokens>
    <token ID="t_0">New</token>
    <token ID="t_1">Cross</token>
    <token ID="t_2">,</token>
    [...]
  </tokens>
  <POStags tagset="PennTB">
    <tag tokenIDs="t_0">NP</tag>
    <tag tokenIDs="t_1">NP</tag>
    <tag tokenIDs="t_2">,</tag>
    [...]
  </POStags>
  <lemmas>
    [...]
    <lemma tokenIDs="t_528">R.</lemma>
    <lemma tokenIDs="t_529">Browning</lemma>
    <lemma tokenIDs="t_530">.</lemma>
  </lemmas>
```

Figure 2.1: A simplified sample file in XML format. Required elements are highlighted blue

2. *Slash/A* N-gram Tendency Viewer

Data retrieval

Slash/A is designed to process corpora in XML format, one text per file.[4] To make use of the full functionality of the tool, each XML file should contain the original text, the date, the author's name, and the annotations for tokens, lemmas and POS tags in the order they appear in the text (see Figure 2.1).

All of the token annotations can then be used to compose a corpus query. The following are examples of valid queries using the Penn Tree Bank POS tag set:[5]

our present	a query for the bi-gram *our present*
/VBP presents/NNS	a query for all bi-grams with a non-third person singular verb in present tense as the first token and the plural form of the noun *presents* as the second one
present/lemma/V*	a query for all uni-grams with any form of the verb *present*

The second example shows that an omitted token in a query leads to matching any token with the specified POS tag. The last example illustrates the use of a combination of all three types of annotations along with the wildcard character (*). The search hits are counted in every document of the corpus and a relative frequency is calculated for each day of the whole time period the corpus covers. These daily relative frequencies can then be smoothed for a more abstract view. We use moving average as a smoothing technique, and the length of the sliding window can be determined by the user. The five preset levels of smoothing correspond to common time intervals: day (no smoothing), week (smoothing parameter p = 3), month (p = 15), three months (p = 45), and year (p = 182). For more detailed information on smoothing as well as the requirements for the corpus format, see Todorova and Chinkina (2014).

4 The Brownings' corpus, our development corpus, the examples from which are used in this paper, can be downloaded here: http://linguistics.chrisculy.net/lx/resources/. A detailed description of the format of the corpus can be found here: http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format

5 The Penn Tree Bank POS tag set can be found here: http://www.cis.upenn.edu/~treebank/

Visualization

The basis of our visualization is a simple graph. Relative frequencies of n-grams are plotted as data points, and the selected level of smoothing is represented by a continuous line fitted to the data points (see Figure 2.2). Following Schneiderman's (1996) taxonomy, we have included functionality that permits the high level tasks of overview, zoom, filter, details-on-demand and history. The interactive visualization allows for non-linear exploration of a dataset: adding and deleting word lines, keeping track of successful and unsuccessful queries within the current session and getting access to original text.

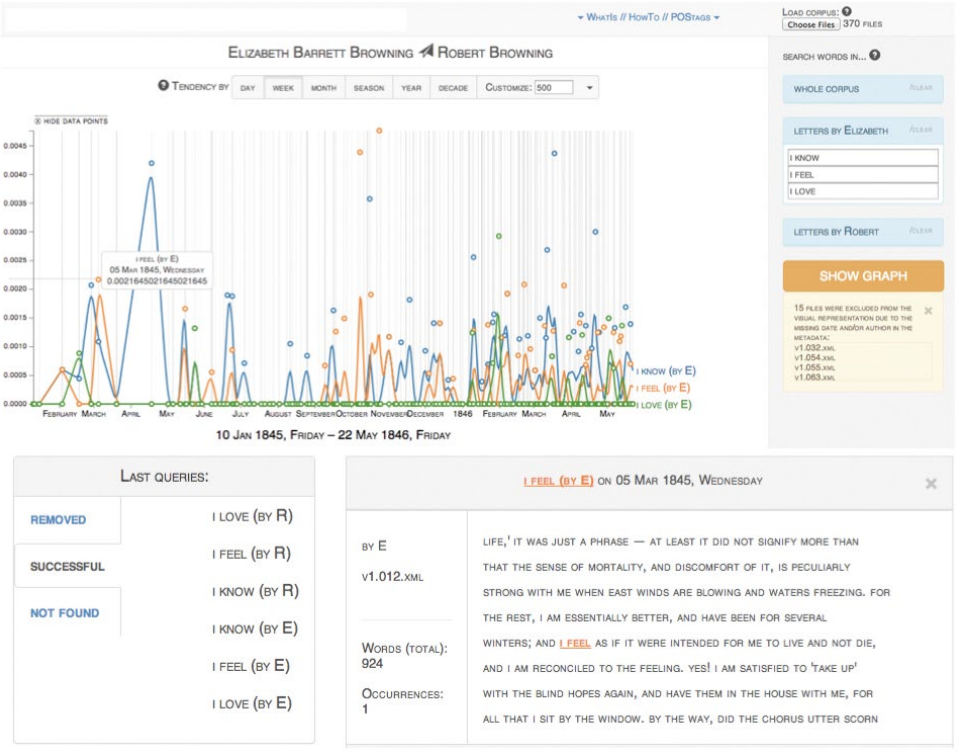


Figure 2.2: *Slash/A* interface: a top panel containing a link to the information about the tool and a button for loading a corpus, a graph showing smoothing by week, a query panel, a history element, and a reading element.

Motivation for Significance Filters

The visualization of relative n-gram frequencies manages to convey a lot of information quickly and efficiently by providing an overview of the general tendency the data follows. However, it can be a source of confusion. Sometimes extreme ups and downs in the use of an n-gram can be perceived as significant even though they are not representing enough data with these high (or low) values. Or, when comparing the frequencies of an n-gram in two subsets of a corpus, some minimal difference might seem insignificant, while it can actually be of significance. To our knowledge, there are currently no n-gram viewers that eliminate the possibility of such confusions.

We try to overcome the problem by suggesting that the inclination of the user to see significance in the visualization should be taken into account, and that various kinds of statistical analysis should be introduced in *Slash/A* for the different tasks that it can be applied to. The following section presents our work on two kinds of statistical analysis and their visualization. We call the technique statistical filtering as its goal is to extract and present portions of data (defined by time intervals) that have the property of obtaining a significant result when subjected to a certain statistical text. The first filter presented here indicates the intervals of time within which the occurrences of an n-gram are significantly higher or lower than in the rest of the period over which the corpus spans. The second filter is to be applied when comparing different subsets of the corpus, and it indicates the time intervals in which the occurrences of an n-gram in the two subsets are significantly different. In what follows, we will present the technical side of the analysis as well as the visual representation of the results and will use various specific cases as illustrations.

3. Filters for significant fluctuations in one data set

Motivation

The first filter that we present filters out the time intervals throughout which a specific n-gram occurs about as much as is expected to occur by chance, given its occurrences in the rest of the documents. This way, the user can focus on the periods with significantly higher (or lower) frequencies than observed in the rest of the corpus. For example, if one is interested in how much a formality marker appears in the first letters of Elizabeth Barrett and Robert Browning's correspondence, one should look for periods at the beginning of

their exchange that contain significantly more occurrences of this marker than the subsequent letters.[6]

Mechanism to detect significant time intervals

It is not possible to automatically identify the interval that is of interest for the user in their current task and only test it for significance. It is also not possible to identify the size or the boundaries of this interval. One possible solution could be to ask the user to provide this information for every query. Alternatively, one might just adopt a strategy to select intervals to test. We decided to go for a compromise: we let the user determine the size of the interval and let an algorithm decide the boundaries.

We have taken advantage of the possibility to customize the length of the tested intervals without complicating the use of the program. The different levels of smoothing are related to time intervals of different length, and the same approach can also be used for the statistical tests. Thus, the user is specifying two things at the same time – the size of the sliding window for the moving average smoothing, and the size of the intervals to be tested.

After the size of the intervals of interest is determined, it is checked if the size is meaningful. Having intervals longer than the half of the whole time period tested is not allowed because it can be confusing when presented visually: The tested interval should appear as an object against the rest of the period as background, but if what has to be perceived as background were smaller, the exact opposite visual effect would be produced. Following the principle of a sliding window, all sub-intervals of the specified length are tested for significant difference from the rest of the data. We have decided to apply the standard one-tailed χ^2 test and one-tailed Fisher exact test of independence (Fisher, 1950), each of which determines if two variables are connected, i.e., if one can be used as a predictor for the other. The two variables tested for independence here possess a certain linguistic feature (e.g., being a noun or containing the consonant complex "np") and belonging to a certain time interval.

In the general case, we use χ^2, which is faster to compute. But whenever there are too few observations, which is usually the case with longer n-grams and with some rare n-grams, we apply the Fisher test of independence as it also provides reliable results in these cases.

6 See the "Examples" section for clarification.

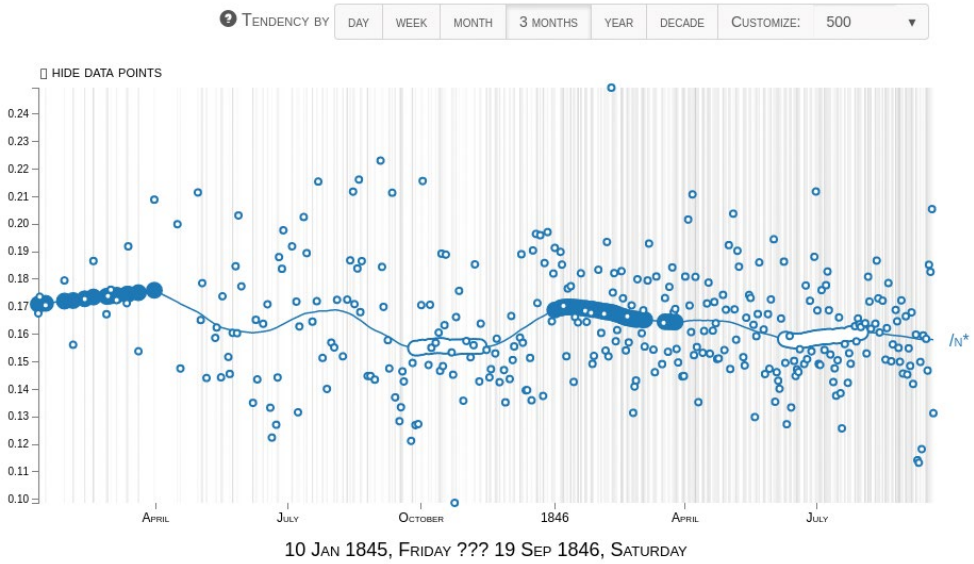


Figure 3.1: Frequency of nouns in the Brownings' corpus (level of smoothing: three months, smoothing parameter p = 45).

Visualization

Visualizing potentially overlapping intervals with either high or low frequency is challenging. First, one needs to make sure that the difference between the indications for significantly higher and significantly lower values is clear. Second, it is important to deal with cases of overlaps without introducing confusion.

The solution we will present here, which is also illustrated in Figure 3.1, is to indicate only the centers of the significant intervals. This way, we avoid the problem of overlaps in the visual presentation. Furthermore, since centers of intervals are points, one can use special marks to indicate them (instead of changing the properties of the part of the line that represents the interval, which is presumably a more confusing approach). These special marks can have two sets of features – one that associates them with the graph they belong to, and the other that makes it clear if this is a period with a lower or higher frequency than in the rest of the documents. We decided to include the stroke color of the mark (which should be the same as that of the graph) in the first set of features and the filling color of the mark in the second set. If the mark has a white center, it indicates that it denotes an interval with a low frequency, and if the mark is

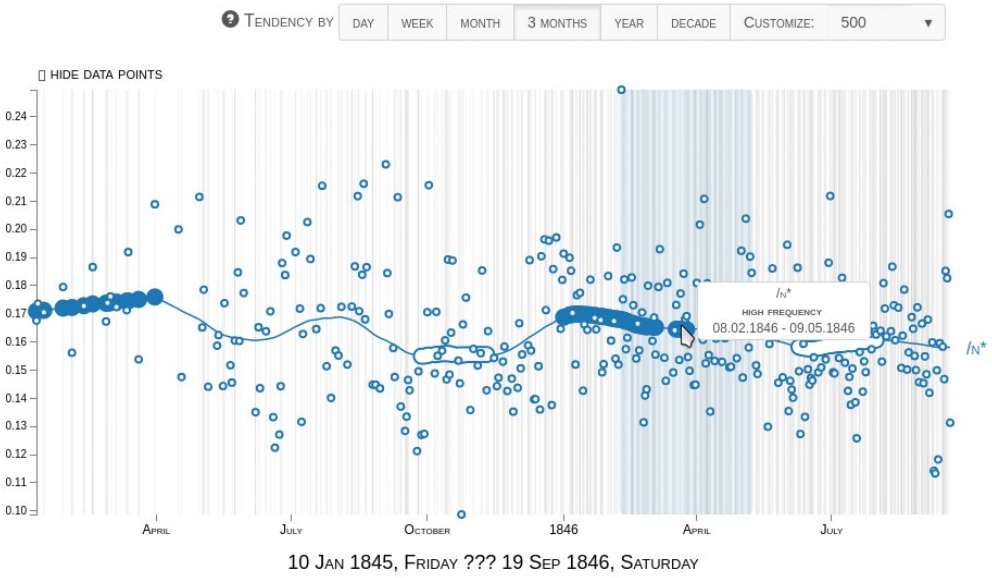


Figure 3.2: Statistical details on demand: Frequency of nouns in the Brownings' corpus (level of smoothing: three months, smoothing parameter p = 45).

filled with a solid color (the color of the graph), it indicates an interval with a high frequency.

We have included functionality of statistical details on demand, which is illustrated in Figure 3.2. To indicate the boundaries of the intervals, we color the background whenever the center of a significant interval is pointed at. For this, we use the color of the selected mark, but with a very low opacity. The details about the significance hypothesis are shown in a tooltip.

4. Filters for significant differences between two data sets

Motivation

The filter described above is applicable to one time series at a time. However, significance filtering can also be very useful in a task that involves comparison between two time series – that is, to compare the frequencies of a certain n-gram in two distinct groups of texts.

Mechanism for detecting significant periods

Detection of significant periods in this filter is similar to that used for the first one, with the important difference that not one interval is compared to the rest of the data, but two subsets of the data are compared to each other at a given time interval. Again, we use a sliding window with the size of the moving average sliding window to go through the whole time series and detect intervals with significant results.

Visualization

The visualization of significant differences between data subsets inherits the idea of only indicating the center of the intervals, and it does so by placing a thin red line connecting the two time series. On mouse-over over a red line, one obtains the indication of the interval length in the background and a tooltip with additional information, as shown in Figure 4.

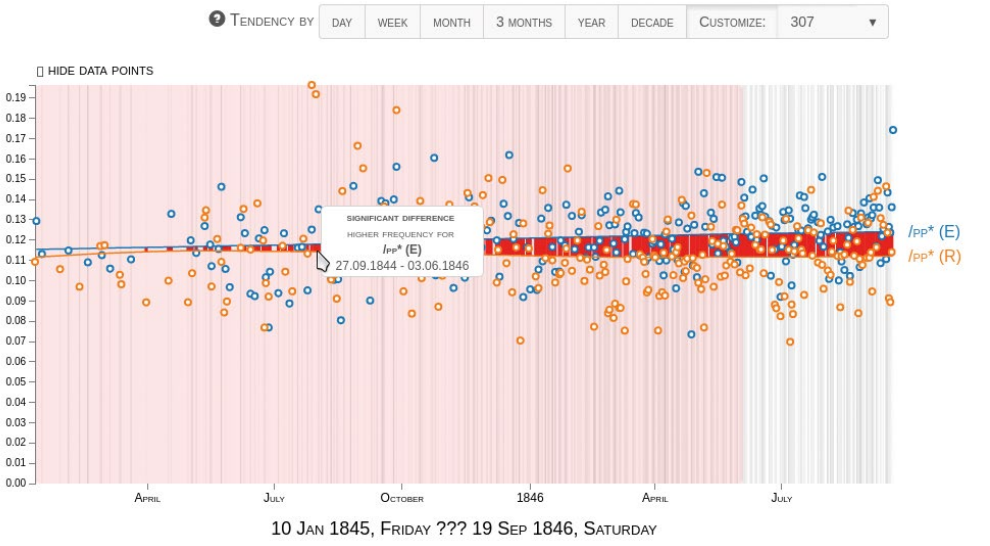


Figure 4: Comparison of the frequency of pronouns in the letters by Robert and Elizabeth (smoothing parameter p = 307).

5. Examples

General notes

The statements that we will use here to demonstrate the statistical filtering functionality of *Slash/A* are from the domains of formality theory and of gender linguistics. More specifically, the statement about nouns as an indicator of language formality is taken from Heylighen and Dewaele (1999), and the statement about the gender specific use of pronouns comes from Koppel et al. (2002) and Argamon et al. (2003). The corpus that will be surveyed is the Brownings' corpus mentioned earlier.

Investigating one data set

Let us explore the statement "Formal language is characterized by higher frequency of nouns" using *Slash/A*. Figure 3.1 shows that the beginning of the correspondence is characterized by a significantly higher frequency of nouns than in the rest of the corpus. The end of the correspondence is characterized by a significantly lower frequency of nouns. Thus, the language in the corpus contains more formal markers (nouns in this particular case) in the first quarter of the exchange and less formal markers in the last quarter. It is interesting that the middle part of the time series doesn't follow the expected pattern of a gradual lowering of the formality level. The rare occurrence of nouns in the second quarter of the period is followed by a frequent use of them, which could be indicative of two things: (i) that the formality of the correspondence does not evolve linearly, but rather goes back and forth; or (ii) that the frequency of nouns is also a marker for something else, and this other thing interferes with the formality of the language producing unclear patterns. To check if (i) holds, we can look at a higher level of smoothing. The expectation would be that when longer intervals are taken into account, not quarters but halves of the whole periods will be marked for significance, and the first half will contain more occurrences of nouns than the second one. Figure 5.1 is obtained by applying the largest smoothing parameter that is accessible for fluctuations testing. One can no longer see anything informative about the beginning and the end of the letter exchange as the differences in noun usage are not significantly different. It appears that the presence of this formality marker does not change linearly over time – at least not for the whole collection of letters.

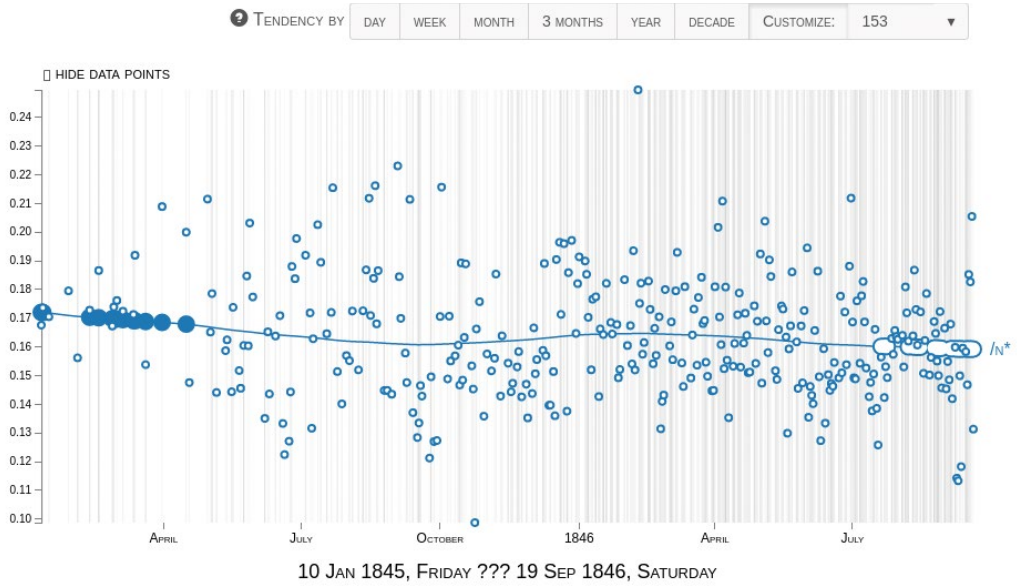


Figure 5.1: Frequency of nouns in the Brownings' corpus (a custom smoothing parameter p = 153)

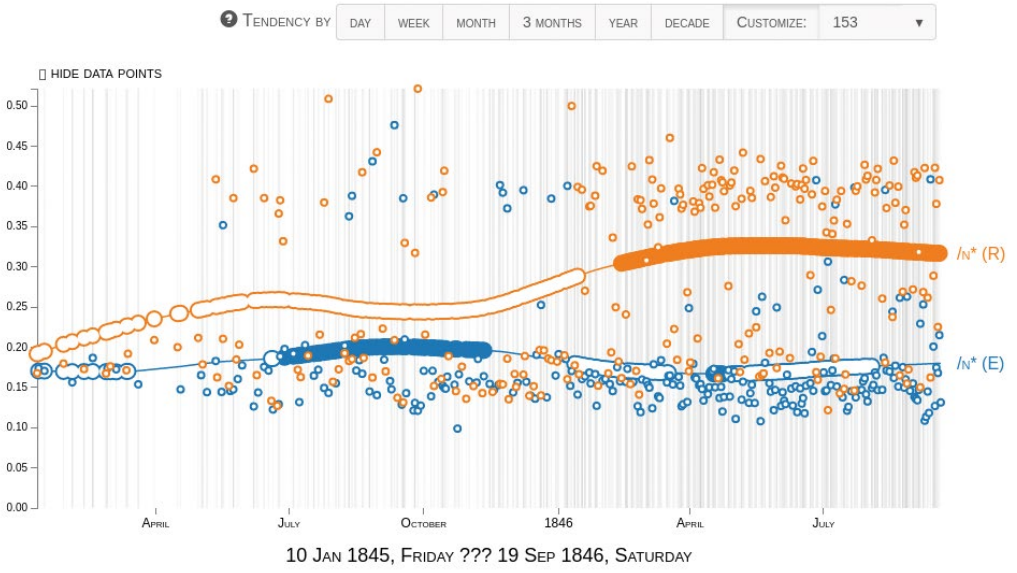


Figure 5.2: Frequency of nouns in the letters by Robert and Elizabeth separately (smoothing parameter p = 153).

To explore the question further, one can look at Elizabeth's and Robert's letters separately.[7] An interesting observation can be made from Figure 5.2: Elizabeth's letters show a linear decrease in the use of nouns. That is, on a larger scale, the formality level of her letters lowers neatly linearly with time. On the other hand, Robert appears to be following a contrary pattern, his use of nouns in the first half of the exchange being lower than in the second one.

Making sense of all these observations would require some exploration of other formality markers. What we can suggest as a possible interpretation is that Robert was quicker in dropping this particular formality marker (there are multiple low frequency marks in the second quarter of Robert's time series in Figure 5.1). Elizabeth held on to her nouns longer and lowered their use more significantly only after the first half of the correspondence period. A process of language adaptation can explain why Robert then raised his use of nouns to a level closer to the one his beloved respondent was sustaining.

Comparing two data sets

For the second type of filtering, we present the results obtained by exploring the statement "Women use more pronouns than men". Figure 4.1 above illustrates the statement – Elizabeth uses more pronouns than Robert in the course of their correspondence.[8] More interesting is what can be seen in a less smoothed view, like the one in Figure 5.3. It again shows that Elizabeth uses pronouns significantly more often, but only in certain periods, and that there are times when the frequencies of pronouns in the letters of the two authors are so close that their differences are insignificant.

6. Conclusion and Future Work

We have presented *Slash/A* N-gram Tendency Viewer that extracts data from a corpus, conducts searches on it, calculates and plots n-gram frequencies and smooths them. We have also discussed the advantages of including a significance filtering functionality and proposed two significance filters to improve the user's

7 Note that the number and distribution of data points are different when we inspect Elizabeth's and Robert's letters separately compared to inspecting the corpus as a whole. In the latter case, each data point represents one day, and the frequency is calculated based on all the letters written on this day.

8 The parameter 307 is chosen because it is the largest parameter value that can be applied to this data.

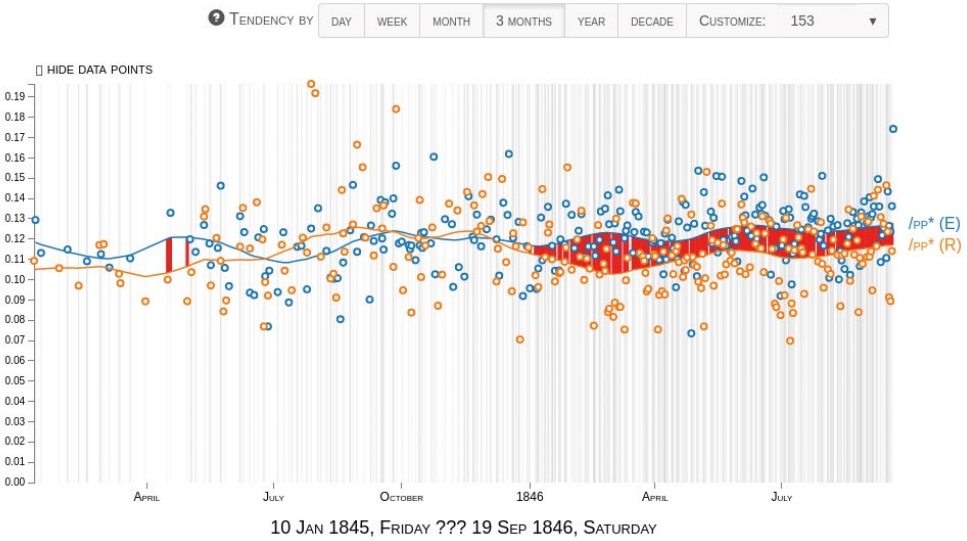


Figure 5.3: Comparison of the frequency of pronouns in the letters by Robert and Elizabeth (level of smoothing: three months, smoothing parameter p = 45).

certainty level in their conclusions. One of the filters shows periods with significantly different values as compared to the rest of the time series, and the other one shows significant differences between two time series.

As future work, other similar filters could be introduced for other kinds of tasks. For example, one that indicates the time intervals throughout which a certain combination of words is used more often than it is expected for these words to appear together by chance. This can be helpful for collocation strength monitoring if collocations are understood simply as words occurring together unexpectedly often (for the definition of collocation, see Dale et al., 2000).

7. Acknowledgements

We would like to thank the supervisor of this project and a former professor at the University of Tübingen, Dr. Christopher Culy, for his continuous support and valuable advice. Velislava Todorova is also grateful to the German Exchange Service (DAAD) for the grant that made it possible for her to study in Tübingen and to work on this project there.

8. References

Argamon, Shlomo, Moshe Koppel, Jonathan Fine and Anat Rachel Shimoni. 2003. "Gender, genre, and writing style in formal written texts." In *Text – Interdisciplinary Journal for the Study of Discourse,* 23, no. 3: 321–346. https://doi.org/10.1515/text.2003.014.

Dale, Robert, Hermann Moisl, and Harold Somers, 2000. Eds. *Handbook of Natural Language Processing*. CRC Press.

Fisher, Ronald Aylmer. 1950. "Statistical methods for research workers." 11th ed. Edinburgh: Oliver and Boyd. https://doi.org/10.1038/123866a0.

Heylighen, Francis and Jean-Marc Dewaele. 1999. "Formality of language: definition, measurement and behavioral determinants." Technical report. Center "Leo Apostel", Free University of Brussels & Birkbeck College, University of London pespmc1.vub.ac.be/Papers/Formality.pdf.

Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. 2002. "Automatically categorizing written texts by author gender." *Literary and Linguistic Computing* 17, no. 4: 401–412. http://dx.doi.org/10.1093/llc/17.4.401.

Shneiderman, Ben. 1996. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations." In *Proceedings of IEEE Visual Languages*, 336–343 https://doi.org/10.1109/vl.1996.545307 (accessed 12 January 2018).

Todorova, Velislava, and Maria Chinkina. 2014. "*Slash/A* N-gram Tendency Viewer: Visual Exploration of N-gram Frequencies in Correspondence Corpora." In *Proceedings of the ESSLLI 2014 Student Session*, 229–239. http://www.kr.tuwien.ac.at/drm/dehaan/stus2014/proceedings.pdf (accessed 12 January 2018).

Manuel Burghardt

Visualization as a Key Factor for the Usability of Linguistic Annotation Tools

Abstract Linguistic annotation is an important means of adding information to corpora of spoken or written language. While some less complex annotation tasks can be performed automatically, a great number of annotation tasks require manual annotation, which is typically very time-consuming and tedious. As a consequence, tools for manual annotation tasks should provide a user-friendly interface that makes the annotation process as convenient and efficient as possible; in other words, *usability* should play an important role in the design of such tools. This article contributes to the field of "visual linguistics" by investigating the role of *visualization* in linguistic annotation tools with regard to good and bad usability practices. While there are several studies that are dedicated to visualizing linguistic results, visualization in the context of linguistic annotation has so far been largely neglected. Accordingly, a heuristic walkthrough evaluation study with 11 annotation tools was conducted to find out about typical usability problems. It showed that many of the usability issues identified during the evaluation are related to aspects of interaction design. However, there are also a large number of usability issues that are directly connected to aspects of visualization and visual design. These aspects of good and bad visualization are discussed by means of existing usability heuristics, which can be used to illustrate and explain how and why visualization influences the usability of linguistic annotation tools.

1. Introduction

Digital annotations are an important means to make the daily flood of information manageable, as they allow us to add "invisible intelligence" (Ruecker et al. 2011, 27) to a text, thus making implicit information explicitly available for computer-based analyses[1]. Linguistic annotation constitutes a specific type of digital annotation. Leech (1997, 2) defines it as "the practice of adding interpretative,

1 The work presented in this article is part of a PhD project finished in 2014 (cf. Burghardt 2014). This article reuses some of the passages from the original PhD thesis. The

Published in: Noah Bubenhofer, Marc Kupietz: Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

linguistic information to an electronic corpus of spoken and/or written language data". Linguistic annotation can be carried out manually, automatically, or semi-automatically (i.e. automatic annotation with manual correction) (McEnery and Hardie 2012, 30). Automatic annotation, however, is limited to fields of manageable degrees of complexity (hence it is also called shallow annotation), including simple text processing tasks such as tokenization and sentence segmentation, or simple tagging and parsing tasks such as part of speech tagging or syntactic phrase detection / categorization (Brants and Plaehn 2000, 1). More sophisticated types of annotation cannot be fully automated, but rather need to be carried out by human annotators (cf. Brants and Plaehn 2000; Dandapat et al. 2009). As manual annotation is a laborious task, computer-based annotation tools need to provide a user-friendly interface that makes the annotation process as convenient and efficient as possible. The important role of *usability*[2] in the domain of linguistic annotation tools is also stressed by a large body of related work (cf. Burghardt and Wolff 2009; Burghardt 2012; Dybkjaer, Berman, Bernsen, et al. 2001; Dipper et al. 2004; Reidsma et al. 2004; Eryigit 2007; Dandapat et al. 2009; McEnery and Hardie 2012, 33; Palmer and Xue 2010; Hinze et al. 2012).

In this article, I will focus on the aspect of *visualization* in linguistic annotation tools and discuss how it influences good and bad usability practices. While there are several studies that are dedicated to visualizing linguistic results (cf. e.g. Wattenberg and Viégas 2008; Culy and Lyding 2010), visualization in the context of linguistic annotation has so far been largely neglected. I present the results from a large-scale usability evaluation study (Burghardt 2014) of existing annotation tools, which illustrate that an adequate visualization is a key requirement for user-friendly annotation tools.

2. Evaluating the usability of linguistic annotation tools

As most of the existing linguistic annotation tools struggle to implement a user-friendly interface, I conducted an evaluation study with 11 annotation tools to find out not only about typical usability problems, but also about positive aspects of the different tools. The evaluated annotation tools are:

focus of this condensed article lies on the aspect of visualization and its implications for the usability of linguistic annotation tools.

2 ISO 9241-11 (1999) definition for usability: The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use.

— Analec (http://www.lattice.cnrs.fr/Telecharger-Analec?lang=fr)
— Brat (http://brat.nlplab.org/)
— CATMA (http://www.catma.de/)
— Dexter (http://www.dextercoder.org/)
— GATE (https://gate.ac.uk/)
— Glozz (http://www.glozz.org/)
— Knowtator (http://knowtator.sourceforge.net/docs.shtml)
— MMAX2 (http://mmax2.sourceforge.net/)
— UAM Corpus Tool (http://www.wagsoft.com/CorpusTool/index.html)
— WebAnno (https://code.google.com/p/webanno/)
— WordFreak (http://wordfreak.sourceforge.net/index.html)

I used the heuristic walkthrough method (Sears 1997) to discover a total of 207 usability problems and 84 positive aspects for the 11 tools. It showed that many of the usability issues identified during the evaluation are related to aspects of interaction design. There are, however, also a large number of usability issues that are directly connected to aspects of visualization and visual design.

In the following section, I will discuss aspects of good and bad visualization by means of existing usability heuristics. Usability heuristics – sometimes also called guidelines, rules, recommendations or best practices – are meant to capture and promote good design in a generic way (Johnson 2010, xi). There are many examples for such generic heuristics[3] and they often seem to overlap or even appear redundant. This is largely because most of these heuristics share a common basis and origin, which is knowledge about human psychology, for instance perception, reasoning, memory, etc. (Johnson 2010, xiii). The following set of 10 usability heuristics is among the most widely used heuristics (detailed descriptions taken from Nielsen 1994, p.30, Table 2.2):

— *H1 Visibility of system status:* The system should always keep users informed about what is going on, through appropriate feedback within a reasonable time.
— *H2 Match between system and the real world:* The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
— *H3 User control and freedom:* Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted

3 Cf. Johnson (2010, xi) for an overview of some of the most prominent guidelines and heuristics in the field of human-computer interaction (HCI).

function without having to go through an extended dialogue. Support undo and redo.
- *H4 Consistency and standards:* Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
- *H5 Error prevention:* Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.
- *H6 Recognition rather than recall:* Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
- *H7 Flexibility and efficiency of use:* Accelerators – unseen by the novice user – may often speed up the interaction for the expert user so that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.
- *H8 Aesthetic and minimalist design:* Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
- *H9 Help users recognize, diagnose, and recover from errors:* Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
- *H10 Help and documentation:* Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

3. Visualization in linguistic annotation tools – A usability perspective

As a result of the usability evaluation of linguistic annotation tools, a large number of usability issues that are related to different aspects of visualization were identified. These issues are structured into the following three subsections: (1) visualization of primary data, (2) visualization of annotation schemes and its items, and (3) visualization of the actual annotations, including parallel annotations as well as relational annotations (e.g. coreference annotation). The different visualization aspects will be discussed from a usability perspective by means of

Nielsen's (1994) ten heuristics, which were introduced in the preceding section. Whenever appropriate, I will also refer to related interaction design patterns by Jenifer Tidwell (2011), which describe generic solutions to recurring usability issues in interface and interaction design.

(1) Primary data

During the annotation process, the primary data is typically not read sequentially from beginning to end, but rather scanned for certain text fragments that can be used as an anchor for a specific annotation. The standard visualization of primary data often does not support such episodic scanning and reading. There are several features that can be implemented on the visualization level to enhance the readability of primary data:

a) The *page* metaphor allows the user to break down very long documents into smaller units that are familiar to the user (cf. Figure 1, left).
   — *Heuristic*: Match between system and the real world
   — *Pattern*: Pagination (Tidwell 2011, 224)
b) The use of two different colors helps to distinguish alternating lines from each other (cf. Figure 1, right).
   — *Heuristic*: Flexibility and efficiency of use
   — *Pattern*: Row striping (Tidwell 2011, 220)
c) Numbered lines facilitate the navigation through the primary data document (cf. Figure 1, right).
   — *Heuristics*: Recognition rather than recall, flexibility and efficiency of use
d) Facilitated orientation in primary data by means of a macro-view and positional syncing (cf. Figure 2): a thumbnailversion of the document (macro-view)

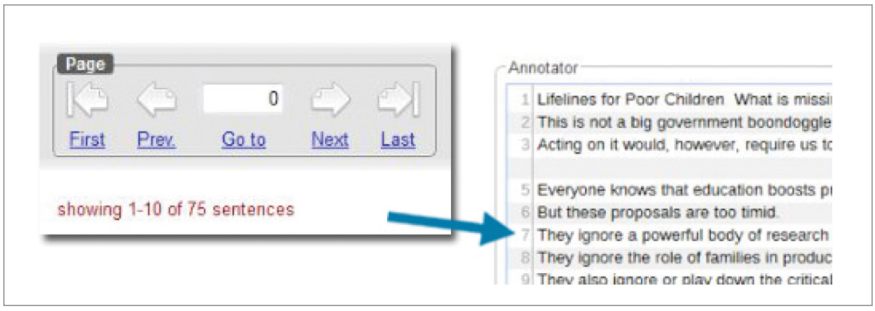


Figure 1: Pagination (left), row striping (right) and numbered lines (right) in the Brat annotation tool.

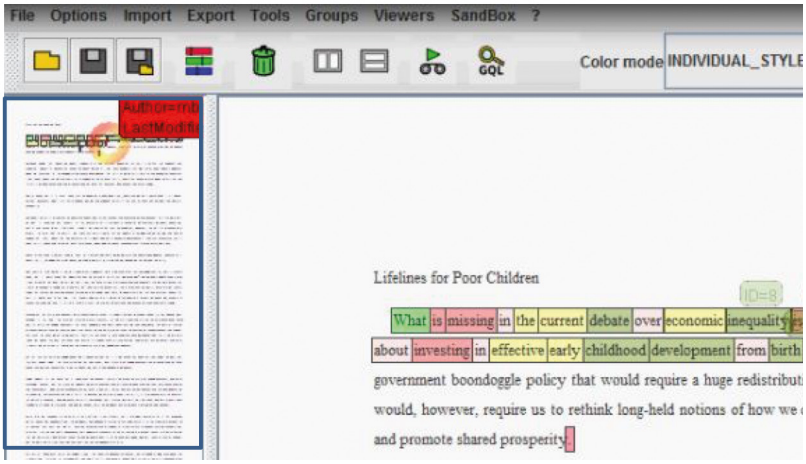


Figure 2: Macro-view of document and positional syncing in the Glozz annotation tool.

allows users to quickly navigate through long documents. The whole text can be accessed via a scrollbar or by clicking into a macro-view of the whole document on the left side. Whenever the mouse cursor is moved somewhere in the document, the position is highlighted in the macro-view (and vice versa). Good visualization of the primary data increases its readability and thus accelerates the overall annotation process

— *Heuristic*: Flexibility and efficiency of use
— *Pattern*: Overview plus detail (Tidwell 2011, 296)

(2) Annotation scheme

The creation of an annotation scheme that defines different levels of annotation as well as concrete annotation items on each level is a crucial task in any annotation project. Typically, annotation schemes are defined by means of document grammars known from markup languages like XML or SGML. Users without technical knowledge about markup languages will have difficulties in creating a scheme in XML syntax. User-friendly annotation tools should provide a visualization of the annotation scheme that can also be understood by markup novices. Adequate visualizations rely on well-known metaphors for the creation of hierarchical structures, e.g. ordered lists and file-trees (cf. Figure 3).

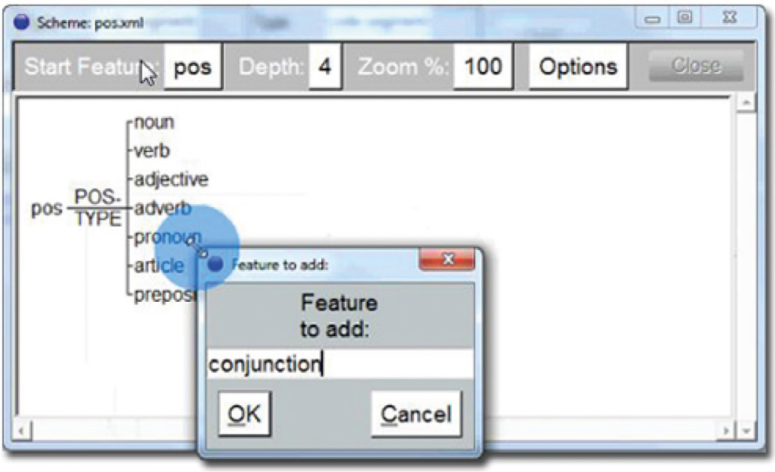


Figure 3: Hierarchical scheme editor in the UAM CorpusTool.

As ad hoc modifications of the annotation scheme are part of the typical annotation process, a good visualization for annotation schemes speeds up the overall annotation process, increases the learnability of the annotation tool and decreases the number of potential errors that may occur when novices are forced to translate linguistic annotation schemes into formal markup languages.

— *Heuristics*: Aesthetic and minimalist design, error prevention

(3) Annotations

Linguistic annotations consist of three basic elements: *body*, *anchor* and *marker* (Marshall 2010, 42ff.). The body of an annotation is the actual content that is added to a text. It is connected to an anchor that denotes the scope of a portion of text an annotation relates to. The marker is the actual visualization of the anchor. For the case of linguistic annotation tools, typical visualizations for anchors are colored underlines or highlights (cf. Figure 4).

In linguistic annotation scenarios, a single anchor is typically annotated with multiple values, which results in parallel annotations (cf. Figure 5).

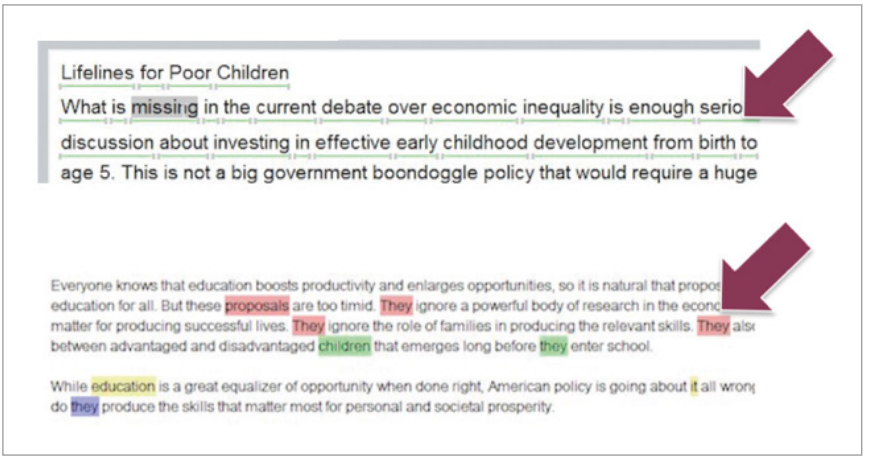


Figure 4: UAM CorpusTool uses underlining (top), GATE uses highlighting (bottom).

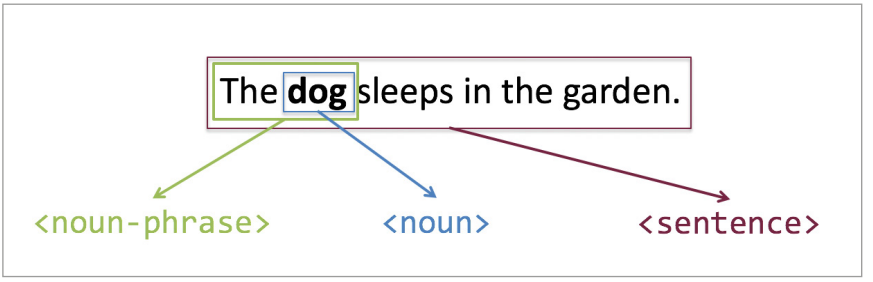


Figure 5: The anchor "dog" is annotated on three different annotation levels.

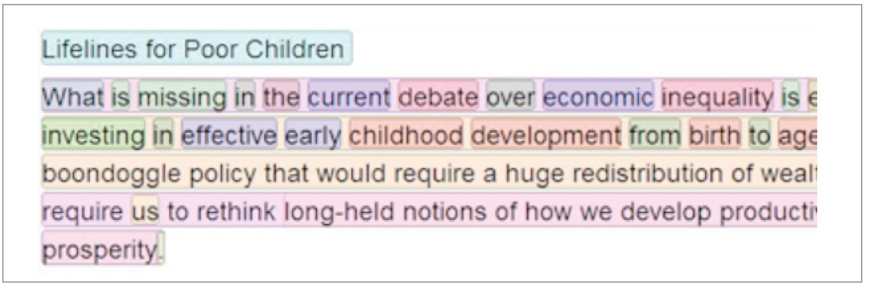


Figure 6: Colored highlights in the Dexter annotation tool overlap one another.

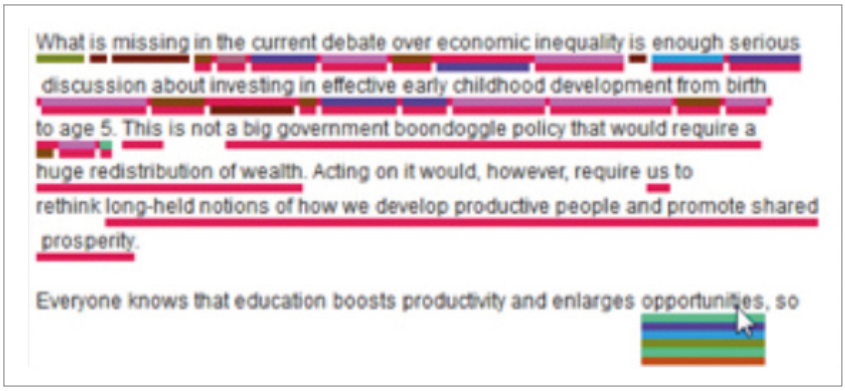


Figure 7: Stacked, colored underlines in the CATMA annotation tool.

It shows that some visualizations are better suited for parallel markers than others. Colored highlights do not work well, as they cannot be stacked, but rather overlap one another (cf. Figure 6).

In contrast, underlines are better suited, as they can be stacked without problems (cf. Figure 7).

— *Heuristics*: Error prevention, aesthetic and minimalist design

Parallel annotation also poses challenges for the visualization of the annotation body, as several competing bits of information – on different levels of annotation – need to be displayed in an adequate way. A user-friendly tool visualizes parallel annotations in a context menu that is displayed next to the respective anchor. The annotation values are displayed as text strings in the context menu (cf. Figure 8). Alternatively, they may be displayed in a separate window or pane rather than in a context menu. Another way to visualize parallel annotation values is by means of a stack view that displays an anchor and (optionally) some of its left and right textual context in the horizontal dimension. In the vertical dimension, parallel annotation values are displayed as a stack of different annotation levels (cf. Figure 9). By visualizing multiple, parallel annotations for one anchor, users have more control about the annotation process and are therefore less likely to produce annotation errors.

— *Heuristics*: Visibility of system status, aesthetic and minimalist design
— *Pattern*: Datatips (Tidwell 2011, 300)

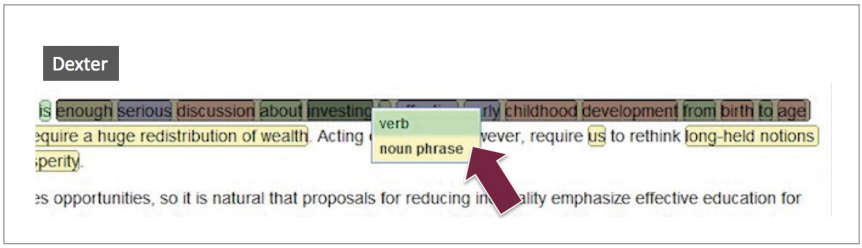


Figure 8: A context menu shows all existing, parallel annotations for a selected anchor in the Dexter annotation tool.

Figure 9: Parallel annotations are displayed in a stack view with different layers in the GATE annotation tool.

Another major challenge, with regard to adequate visualization, is posed by relational annotations, which frequently occur in linguistic annotation scenarios, e.g. for the annotation of coreference relations between two or more anchors. In a related study on human handwritten annotations in a linguistic context, we observed a number of different visualizations for relational annotations (cf. Figure 10). The study participants were asked to create coreference annotations between an antecedent (*ante*) and several corresponding personal pronouns (*pp*). In most cases, lines or directed arrows were used to establish a relation between the separated constituents. The direction of the arrows was mostly pointing

toward the antecedent. The lines and arrows either reached directly from the pronouns to the antecedent, thus creating a tree-like structure (a), or they were connected in some sort of chain, where only the first pronoun pointed to the antecedent and the other pointed to the preceding pronoun (b). In some cases, short arrows were used as deictic devices that indicate the direction and position of the antecedent (c). Some participants chose to draw their arrows and lines directly through the text, while others tried to interrupt the lines so they would not obscure the text (d). One participant even tried to draw the lines around the text using the margins of the page (e). Another way to establish a relation between different constituents is by means of an indexing system (f).

For the case of linguistic annotation tools, relational annotations should be realized by means of arrows or connecting lines (cf. the "chain relation" visualization in Figure 10b and Figure 11) between the participating anchors, as these can be understood by the users in a natural and intuitive way.

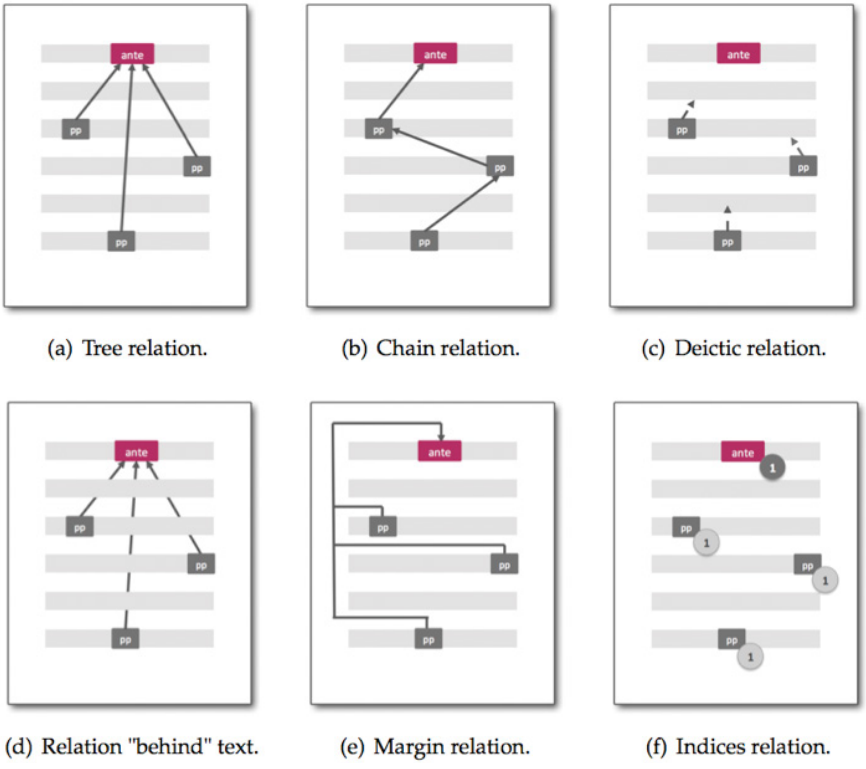


Figure 10: Examples for different realizations of relational annotation.

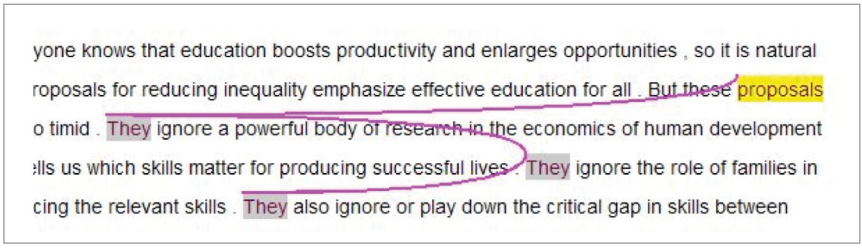


Figure 11: Relational annotation visualization in the *MMAX2* annotation tool.

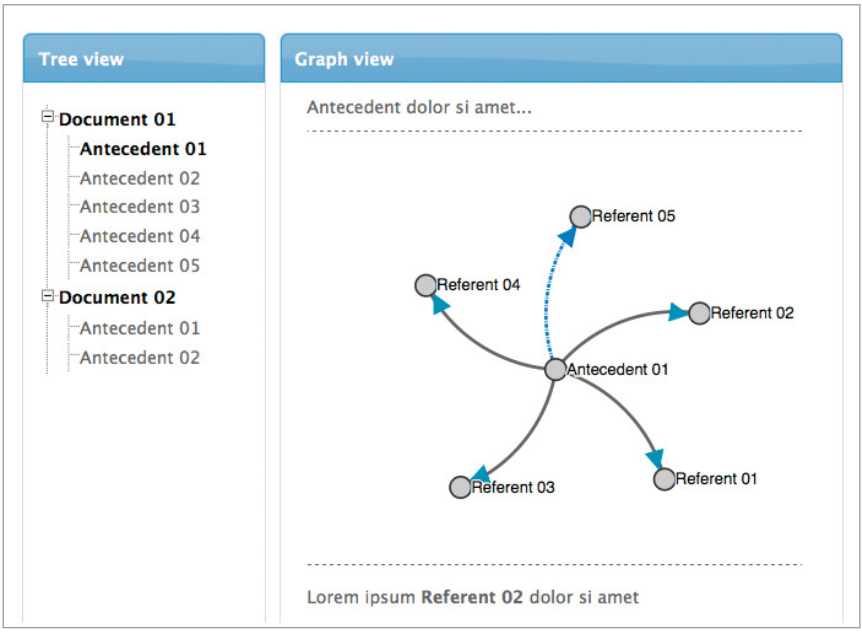


Figure 12: Alternative view for the visualization of coreference annotations according to Witte and Tang (2007).

— *Heuristics*: Match between system and the real world, recognition rather than recall, error prevention

Another good way for the visualization of coreference annotations is described by Witte and Tang (2007). The proposed solution makes use of *Topic Maps* and *OWL ontologies* and can be summarized as follows: relational annotations should be displayed in a separate view that is detached from the primary data view. This view shows all existing relational annotation chains as an integrated graph. Such a graph view even allows users to visualize relational annotations from different documents and thus greatly facilitates navigation in coreference chains and documents (cf. Figure 12).

— *Heuristic*: Aesthetic and minimalist design
— *Pattern*: Alternative views (Tidwell 2011, 66)

4. Conclusion

This article illustrates that visualization plays an important role for the usability of linguistic annotation tools. While there are many competing visualizations, existing usability heuristics can be used to assess and discuss their specific strengths and weaknesses. This kind of assessment is helpful not only for tool developers who design new annotation tools, but also for users of annotation tools, who need to choose from a wide variety of applications and who might want to use "adequate visualization" as a selection criterion.

5. References

Brants, Thorsten, and Oliver Plaehn. 2000. "Interactive corpus annotation." In *Proceedings of Second International Conference on Language Resources and Evaluation (LREC '00)*, 453–459.

Burghardt, Manuel, and Christian Wolff. 2009. "Werkzeuge zur Annotation diachroner Korpora." In *Proceedings of the GSCL-Symposium "Sprachtechnologie und eHumanities"*, edited by Wolfgang Hoeppner. Duisburg: Abteilung für Informatik und Angewandte Kognitionswissenschaft, Universität Duisburg-Essen, 21–31.

Burghardt, Manuel. 2012. "Usability Recommendations for Annotation Tools." In *Proceedings of the ACL 2012, 6th Linguistic Annotation Workshop (LAW '12)*, Stroudsburg, PA: Association for Computational Linguistics, 104–112.

Burghardt, Manuel. 2014. "Engineering Annotation Usability – Toward Usability Patterns for Linguistic Annotation Tools." Doctoral dissertation, Universität Regensburg.

Culy, Chris, and Verena Lyding. 2010. "Double Tree: An Advanced KWIC Visualization for Expert Users." In *14th International Conference Information Visualisation.* Los Alamitos, CA: IEEE, 98–103. doi: 10.1109/IV.2010.24

Dandapat, Sandipan, Priyanka Biswas, Monojit Choudhury, and Kalika Bali. 2009. "Complex Linguistic Annotation – No Easy Way Out! A Case from Bangla and Hindi POS Labeling Tasks." In *Proceedings of the 3rd Linguistic Annotation Workshop (LAW '09).* Stroudsburg, PA: Association for Computational Linguistics, 10–18.

Dipper, Stefanie, Michael Götze, and Manfred Stede. 2004. "Simple annotation tools for complex annotation tasks: an evaluation." In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC '04)*, 54–62.

Dybkjaer, Laila, Stephen Berman, Michael Kipp, Malene Wegener Olsen, Vito Pirrelli, Norbert Reithinger, and Claudia Soria. 2001. *Survey of existing tools, standards and user needs for annotation of natural interaction and multimodal data (Deliverable D11.1). ISLE Natural Interactivity and Multimodality Working Group.*

Eryigit, Gülşen. 2007. "ITU Treebank Annotation Tool." In *Proceedings of the 1st Linguistic Annotation Workshop (LAW '07).* Stroudsburg, PA: Association for Computational Linguistics, 117–120.

Hinze, Annika, Ralf Heese, Markus Luczak-Rösch and Adrian Paschke. 2012. "Semantic Enrichment by Non-Experts: Usability of Manual Annotation Tools." In *Proceedings of the 11th International Semantic Web Conference (ISWC '12).* Springer: Berlin, 165–181.

Johnson, Jeff. 2010. *Designing with the Mind in Mind. Simple Guide to Understanding User Interface Design Rules. Children.* Amsterdam: Morgan Kaufman.

Leech, Geoffrey. 1997. "Introducing Corpus Annotation." In *Corpus Annotation. Linguistic Information from Computer Text Corpora*, edited by Roger Garside, Geoffrey Leech, and Tony McEnery. Harlow, Essex: Addison Wesley Longman, 1–18.

Marshall, Catherine C. 2010. *Reading and Writing the Electronic Book.* San Rafael, CA: Morgan, and Claypool Publishers (Synthesis).

McEnery, Tony, and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice.* Cambridge: Cambridge University Press.

Nielsen, Jakob. 1994. "Heuristic evaluation." In *Usability Inspection Methods*, edited by Jakob Nielsen and Robert Mack. New York: Wiley, 25–62.

Palmer, Martha, and Nianwen Xue. 2010. "Linguistic Annotation." In *Computational Linguistics and Natural Language Processing Handbook*, edited by

Alexander Clark, Chris Fox, and Shalom Lappin. Oxford: Wiley-Blackwell, 238–270.

Reidsma, Dennis, Nataša Jovanović, and Dennis Hofs. 2004. "Designing Annotation Tools Based in Properties of Annotation Problems." http://doc.utwente.nl/49282/ (accessed 26 August 2011).

Ruecker, Stan, Milena Radzikowska, and Stefan Sinclair. 2011. *Visual Interface Design for Digital Cultural Heritage*. Farnham: Ashgate Publishing.

Sears, Andrew. 1997. "Heuristic Walkthroughs: Finding the Problems Without the Noise." *International Journal of Human-Computer Interaction 9* (3): 213–234.

Tidwell, Jennifer. 2011. *Designing Interfaces*. 2nd ed. Sebastopol, CA: O'Reilly Media.

Wattenberg, Martin, and Fernanda B. Viégas. 2008. "The Word Tree, an Interactive Visual Concordance." *IEEE Transactions on Visualization and Computer Graphics*, 14 (6): 1221–1228.

Witte, René, and Ting Tang. 2007. "Task-Dependent Visualization of Coreference Resolution Results." *International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, September 27–29, 2007, Borovets, Bulgaria. http://www.rene-witte.net/system/files/075_witte.pdf (accessed 25 February 2014).

Visualisierungen spielen in den Wissenschaften eine wichtige Rolle im Forschungsprozess. Sie dienen der Illustration von gewonnener Erkenntnis, aber auch als eigenständiges Mittel der Erkenntnisgewinnung.

Auch in der Linguistik sind solche Visualisierungen bedeutend. Beispielsweise in Form von Karten, Baumgraphen und Begriffsnetzen. Bei korpuslinguistischen Methoden sind explorative Visualisierungen oft ein wichtiges Mittel, um die Daten überblickbar und interpretierbar zu machen.

Das Buch reflektiert die theoretischen Grundlagen wissenschaftlicher Visualisierungen in der Linguistik, zeigt Praxisbeispiele und stellt auch Visualisierungswerkzeuge vor.

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

ISBN 978-3-946054-77-1