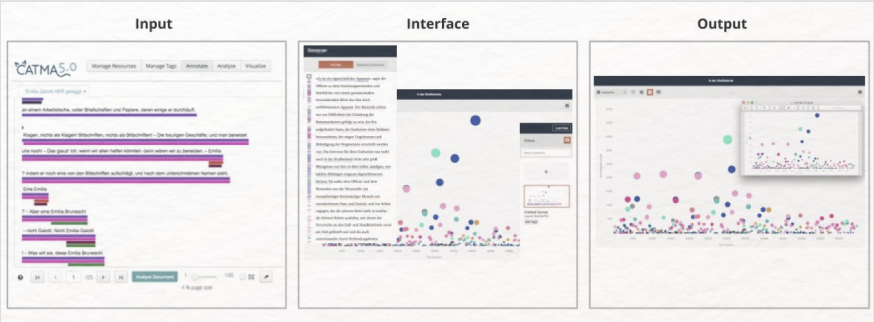
Toolbeitrag: Stereoscope

Mareike Schumacher [1]

1. Universität Regensburg

Thema:	Textvisualisierung	DOI:	10.48694/fortext.3776
Jahrgang:	1	Ausgabe:	5
Erscheinungsdatum:	2024-08-07	Erstveröffentlichung:	2018-12-24 auf fortext.net
Lizenz:	CC BY SA		open access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.



Der Workflow von Stereoscope: Annotationen in CATMA erstellen und in JSON exportieren, Annotationen in Stereoscope diskursiv erkunden und Visualisierungen im JSON-Format oder als Screenshot herunterladen

- **Systemanforderungen:** Webbasiertes (vgl. Webanwendung) Tool, nutzbar über den Browser (z. B. Chrome, Firefox, Safari)
- **Stand der Entwicklung:** Am 13. Dezember 2018 als Prototyp herausgegeben
- **Herausgeber:** 3DH-Team Universität Hamburg
- **Lizenz:** Kostenfrei nutzbar
- **Weblink:** http://stereoscope.threedh.net/
- **Im- und Export:** Annotationen im JSON-Format; Texte als TXT (vgl. Reintext-Version).
- **Sprachen:** Sprachunabhängig: Hebräisch, Arabisch, Deutsch, Englisch, Französisch etc. (Spracheinstellung beim Hochladen des Dokuments. Alle Schriftsprachen stehen zur Auswahl)

1. Für welche Fragestellungen kann Stereoscope eingesetzt werden?

Stereoscope visualisiert Annotationsdaten (vgl. Annotation; Textvisualisierung (Horstmann und Stange 2024)) und kann darum grundsätzlich auf zwei Ebenen eingesetzt werden: Zur inhaltlichen Analyse von Texten oder zur meta-Reflexion der Annotationen (vgl. Metadaten). Die visuelle Annotationsanalyse ist dabei ein erster Schritt zur Systematisierung der Tags (bzw. Taxonomiekategorien) (vgl. Tagset). Eine mögliche Fragestellung wäre: „In welchen Themenfeldern werden ironische oder sarkastische Stilmittel in Kafkas *In der Strafkolonie* besonders häufig eingesetzt und wie sind diese über den Text insgesamt verteilt?“

2. Welche Funktionalitäten bietet Stereoscope und wie zuverlässig ist das Tool?

Funktionen (Auswahl):

- Visualisierung von Annotationsdaten in drei unterschiedlichen Layouts (*Grid*, *Overlap* und *Scatterplot*)
- Enge Verbindung von Primärquelle und Visualisierung
- Anlegen und Vergleichen mehrerer Grafiken
- Filtern, Zoomen und Anpassen der Visualisierung
- Kommentieren, Betiteln und Verschlagworten der Visualisierung

Zuverlässigkeit: Bei Stereoscope handelt es sich um eine Anwendung, die für einen Prototypen hochperformant ist. Bisher sind Im- und Export im JSON-Format vor allem für DH-Einsteiger*innen jedoch noch sehr umständlich. Hinzu kommt, dass die Export-Funktion für Annotationen in CATMA (Schumacher 2024) erst auf E-mail-Anfrage

eingerichtet werden kann. Außerdem gibt es bisher noch nicht die Möglichkeit, mehrere Texte zu analysieren oder Visualisierungen kollaborativ in Echtzeit zu bearbeiten.

3. Ist Stereoscope für DH-Einsteiger*innen geeignet?

Checkliste	✓ / teilweise / –
Methodische Nähe zur traditionellen Literaturwissenschaft	✓
Grafische Benutzeroberfläche	✓
Intuitive Bedienbarkeit	✓
Leichter Einstieg	teilweise
Handbuch vorhanden	–
Handbuch aktuell	–
Tutorials vorhanden	–
Erklärung von Fachbegriffen	–
Gibt es eine gute Nutzerbetreuung?	teilweise

Stereoscope ist ideal für die intensive, diskursiv-hermeneutische Textinterpretation und bedient dadurch einen Bedarf der traditionellen geisteswissenschaftlichen Forschung. Derzeit wird der Einstieg durch die Formatvorgabe von JSON-Dateien aus CATMA noch erschwert. Handbücher und Tutorials sind nicht vorhanden und die Nutzerbetreuung läuft über den CATMA-Support, da das 3DH-Projekt, in dem Stereoscope entwickelt wurde, beendet ist. Eine Nachhaltigkeitsstrategie für Stereoscope und eine nutzerfreundliche Pipeline zwischen CATMA und Stereoscope werden derzeit in forTEXT entwickelt.

4. Wie etabliert ist Stereoscope in den (Literatur-)Wissenschaften?

Stereoscope ist erst 2018 erschienen und darum derzeit noch nicht wissenschaftlich etabliert.

5. Unterstützt Stereoscope kollaboratives Arbeiten?

Nein, Stereoscope sieht nicht vor, dass Visualisierungen kollaborativ erstellt werden. Kollaborativ erstellte Annotationsdaten aus CATMA können allerdings verarbeitet werden.

6. Sind meine Daten bei Stereoscope sicher?

Ja. Stereoscope ist ein webbasiertes Tool, das auf inländischen Servern (vgl. Server) läuft. Personenbezogene Daten müssen für die Nutzung nicht angegeben werden. Texte, die Sie in Stereoscope hochladen, werden nicht vorgehalten, sondern müssen bei jeder neuen Session von Stereoscope neu eingespeist werden. Sie sind nicht für Dritte einsehbar. Die Nutzung von Stereoscope ist also aus datenschutzrechtlicher Perspektive vollkommen sicher, unter urheberrechtlichen Gesichtspunkten ist sie relativ unbedenklich.

Externe und weiterführende Links

- Stereoscope: https://web.archive.org/save/http://stereoscope.threedh.net/ (Letzter Zugriff: 19.06.2024)

Bibliographie

Horstmann, Jan und Jan-Erik Stange. 2024. Methodenbeitrag: Textvisualisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 5. Textvisualisierung (7. August). doi: 10.48694/fortext.3772, https://fortext.net/routinen/methoden/textvisualisierung.

Kleymann, Rabea und Jan-Erik Stange. 2018. Towards Hermeneutic Visualization in Digital Literary Studies. http://www.stereoscope.threedh.net/HermeneuticVisualization.pdf (zugegriffen: 20. Dezember 2018).

Schumacher, Mareike. 2024. Toolbeitrag: CATMA. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3761, https://fortext.net/tools/tools/catma.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

CSV CSV ist die englische Abkürzung für *Comma Separated Values*. Es handelt sich um ein Dateiformat zur einheitlichen Darstellung und Speicherung von einfach strukturierten Daten mit dem Kürzel `.csv` , sodass diese problemlos zwischen IT-Systemen ausgetauscht werden können. Dabei sind alle Daten zeilenweise angeordnet. Alle Zeilen wiederum sind in einzelne Datenfelder aufgeteilt, welche durch Trennzeichen wie Semikola oder Kommata getrennt werden können. In Programmen wie Excel können solche Textdateien als Tabelle angezeigt werden.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z.B. HTML, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

POS PoS steht für *Part of Speech* , oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Reintext-Version Die Reintext-Version ist die Version eines digitalen Textes oder einer Tabelle, in der keinerlei Formatierungen (Kursivierung, Metadatenauszeichnung etc.) enthalten sind. Reintext-Formate sind beispielsweise TXT, RTF und CSV.

Server Ein Server kann sowohl hard- als auch softwarebasiert sein. Ein hardwarebasierter Server ist ein Computer, der in ein Rechnernetz eingebunden ist und der so Ressourcen über ein Netzwerk zur Verfügung stellt. Ein softwarebasierter Server hingegen ist ein Programm, das einen spezifischen Service bietet, welcher von anderen Programmen (Clients) lokal oder über ein Netzwerk in Anspruch genommen wird.

Tagset Ein Tagset definiert die Taxonomie, anhand derer Annotationen in einem Projekt erstellt werden. Ein Tagset beinhaltet immer mehrere Tags und ggf. auch Subtags. Ähnlich der Type/Token -Differenz in der Linguistik sind Tags deskriptive Kategorien, wohingegen Annotationen die einzelnen Vorkommnisse dieser Kategorien im Text sind.

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.

Webanwendung Eine webbasierte Anwendung ist ein Anwendungsprogramm, welches eine Webseite als Schnittstelle oder Front-End verwendet. Im Gegensatz zu klassischen Desktopanwendungen werden diese nicht lokal auf dem Rechner der Nutzer*innen installiert, sondern können von jedem Computer über einen Webbrowser „online“ genutzt werden. Webanwendungen erfordern daher kein spezielles Betriebssystem.