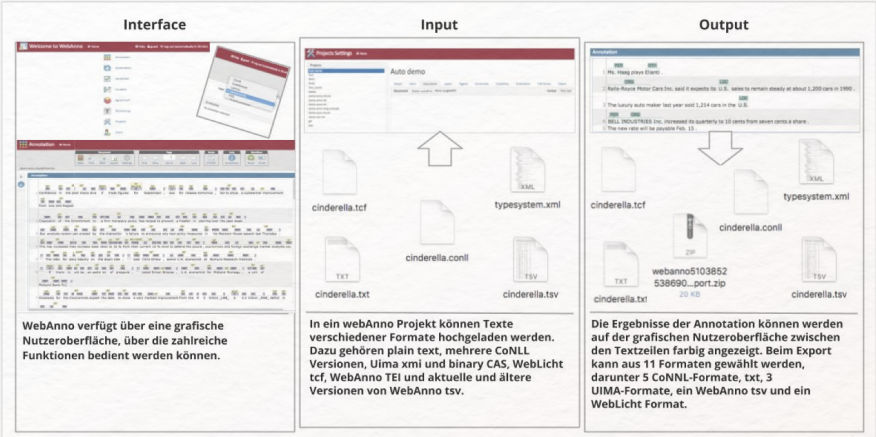
Toolbeitrag: WebAnno

Mareike Schumacher [1]
Sandra Bläß
1. Universität Regensburg

Thema:	Manuelle Annotation	DOI:	10.48694/fortext.3764
Jahrgang:	1	Ausgabe:	4
Erscheinungsdatum:	2024-08-07	Erstveröffentlichung:	2018-10-30 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.



Möglicher WebAnno Workflow: Upload der Texte als TXT, CoNLL, Uima XMI, binary CAS, WebLicht TCF, WebAnno TEI oder WebAnno TSV, annotieren über die grafische Nutzeroberfläche, download der annotierten Texte in einem von 11 Formaten (darunter 5 CoNNL-Formate, TXT, 3 UIMA-Formate, ein WebAnno TSV und ein WebLicht Format)

- **Systemanforderungen:** Webbasiertes (vgl. Webanwendung) Tool, das vom eigenen PC oder einem Server in Safari oder Chrome geladen werden muss; einige Universitäten bieten Demoversionen von WebAnno auf ihren Servern an
- **Stand der Entwicklung:** Version 3.4.5 von 2018, WebAnno wird laufend weiter entwickelt
- **Herausgeber:** TU Darmstadt
- **Lizenz:** Open Source unter Apache 2.0 Lizenz
- **Weblink:** https://webanno.github.io/webanno/downloads/
- **Im- und Export:** Daten können als TXT (vgl. Reintext-Version), CoNLL, Uima XMI, binary CAS, WebLicht TCF, WebAnno TEI oder WebAnno TSV importiert werden. Beim Export kann aus 11 Formaten gewählt werden, darunter 5 CoNNL-Formate, TXT, 3 UIMA-Formate, ein WebAnno TSV und ein WebLicht Format.
- **Sprachen:** Keine Angabe

1. Für welche Fragestellungen kann WebAnno eingesetzt werden?

Mit WebAnno können vor allem Fragestellungen bearbeitet werden, die sich auf klar definierte Einheiten von Texten beziehen. Dazu gehören Fragen wie: Welche Funktionen haben Ortsnennungen in literarischen Texten? Wo finden sich intertextuelle Bezüge von ausgewählten Autor*innen des 20. Jahrhunderts auf die Philosophie Nietzsches? Aber auch Fragen, die sich auf literaturwissenschaftliche Methoden als solche beziehen, können mit WebAnno bearbeitet werden. Ein Beispiel dafür wäre: Wie unterscheiden sich Interpretationen der narratologischen Kategorie „Zeit“, die von drei voneinander unabhängig arbeitenden Literaturwissenschaftler*innen erarbeitet werden?

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

Funktionen (Auswahl):

- Manuelle Annotation (Jacke 2024a) nach eigenen Kategorien oder linguistischen Standards
  - überlappende Annotation
  - widersprüchliche Annotation
- Arbeit an unterschiedlichen Projekten mit unterschiedlichen Team-Mitgliedern (vgl. Kollaboratives literaturwissenschaftliches Annotieren (Jacke 2024b))
- Training automatischer Annotation (vgl. Machine Learning)
- Kurationsfunktionen
  - Vergleich von Annotationen
  - Erarbeiten eines Annotator Agreements

Zuverlässigkeit: WebAnno wird aktuell laufend weiterentwickelt. Es treten bei der Anwendung zum Teil noch Fehler auf, die das Arbeiten mit WebAnno unterbrechen. Diese können aber an das Entwicklerteam zurückgemeldet und die Fehler daraufhin behoben werden.

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

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

Das Tool ist besonders für Computerlinguist*innen geeignet, die Übertragbarkeit auf Literaturwissenschaft ist zwar gegeben, aber wohl eher von DH-erfahrenen Wissenschaftler*innen umsetzbar.

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

WebAnno wird von der DFG (Deutsche Forschungsgesellschaft) als Tool für manuelle linguistische Annotation empfohlen (vgl. DFG 2015). Es gibt einige wissenschaftliche Publikationen aus dem Team der WebAnno-Entwickler*innen, die sich allerdings ebenfalls primär an Computerlinguist*innen richten. Das Tool kann also innerhalb dieser Community als etabliert bezeichnet werden. In den (digitalen) Literaturwissenschaften wird WebAnno eher vereinzelt eingesetzt. Hier steht eine Domänenadaption (vgl. Domäneadaption) noch aus.

5. Unterstützt WebAnno kollaboratives Arbeiten?

Ja, WebAnno hat zahlreiche Funktionalitäten, die auf kollaboratives Arbeiten ausgelegt sind. Die Durchführung unterscheidet sich allerdings bei den beiden angebotenen Varianten stark (es gibt WebAnno sowohl zum Herunterladen auf den eigenen PC (Standalone-Variante) als auch als Server-Variante). Bei der Server-Variante ist gleichzeitiges kollaboratives Arbeiten an einem Projekt möglich. Bei der Standalone-Variante kann dieses nur durch einen Workaround aus Ex- und Imports nachgestellt werden.

6. Sind meine Daten bei WebAnno sicher?

Ja. Sowohl bei der Server- als auch bei der Standalone-Variante von WebAnno werden alle Daten auf dem eigenen PC belassen. Bei der Server-Variante ist es möglich, dass sich mehrere Personen über ihren eigenen PC mit einem Projekt verbinden, dieses einsehen und Texte daraus herunterladen können. Dazu muss ein Administrator des Projektes ihnen zwar einen Nutzernamen und ein Passwort zuteilen, diese gelten aber nicht als personenbezogene Daten, da sie keine Rückschlüsse auf Individuen zulassen. Da die Zusammenarbeit in WebAnno in einem geschützten Rahmen stattfindet (nur, wer die entsprechenden Rechte innerhalb eines Projektes hat, kann Personen dazu einladen), ergeben sich keine datenschutz- oder urheberrechtlich bedenklichen Situationen.

Externe und weiterführende Links

- WebAnno: https://web.archive.org/save/https://webanno.github.io/webanno/ (Letzter Zugriff: 03.07.2024)
- Forum zu WebAnno: https://web.archive.org/save/https://groups.google.com/forum/#!forum/webanno-user (Letzter Zugriff: 03.07.2024)

Bibliographie

Deutsche Forschungsgemeinschaft. 2015. *Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora*. http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf (zugegriffen: 25. Juni 2018).

Eckart de Castilho, Richard, Éva Mújdricza-Maydt, Yimam Seid Muhie, Silvana Hartmann, Iryna Gurevych, Anette Frank und Chris Bieman. 2016. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In: *Proceedings of the workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 76–84. Osaka, Japan. https://www.aclweb.org/anthology/W16-4011.pdf.

Jacke, Janina. 2024b. Methodenbeitrag: Kollaboratives literaturwissenschaftliches Annotieren. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3749, https://fortext.net/routinen/methoden/kollaboratives-literaturwissenschaftliches-annotieren.

———. 2024a. Methodenbeitrag: Manuelle Annotation. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3748, https://fortext.net/routinen/methoden/manuelle-annotation.

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.

Domäneadaption Domäneadaption beschreibt die Anpassung einer in einem Fachgebiet entwickelten digitalen Methode an ein anderes Fachgebiet.

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 (Textauszeichung) Die Textauszeichnung (eng. *Markup*) fällt in den Bereich der Daten- bzw. Textverarbeitung, genauer in das Gebiet der Textformatierung, welche durch Auszeichnungssprachen wie XML implementiert wird. Dabei geht es um die Beschreibung, wie einzelne Elemente eines Textes beispielsweise auf Webseiten grafisch dargestellt werden sollen.

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.

TEI Die *Text Encoding Initiative* (TEI) ist ein Konsortium, das gemeinsam einen Standard für die Darstellung von Texten in digitaler Form entwickelt. Die TEI bietet beispielsweise Standards zur Kodierung von gedruckten Werken und zur Auszeichnung von sprachlichen Informationen in maschinenlesbaren Texten (siehe auch XML und Markup).

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.

XML XML steht für *Extensible Markup Language* und ist eine Form von Markup Language, die sowohl computer- als auch menschenlesbar und hochgradig anpassbar ist. Dabei werden Textdateien hierarchisch strukturiert dargestellt und Zusatzinformationen i. d. R. in einer anderen Farbe als der eigentliche (schwarz gedruckte) Text dargestellt. Eine standardisierte Form von XML ist das TEI-XML.