Ressourcenbeitrag: TextGrid Repository

Jan Horstmann [1]

1. Universität Münster

Thema:	Bibliografie	DOI:	10.48694/fortext.3794
Jahrgang:	1	Ausgabe:	11
Erscheinungsdatum:	30-11-2024	Erstveröffentlichung:	2018-08-21 auf fortext.net
Lizenz:			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.

1. Kurzbeschreibung

Das TextGrid Repository ist ein digitales Langzeitarchiv, das Ihnen die wichtigsten kanonisierten Texte aus der germanistischen Literaturwissenschaft von über 600 Autor*innen in zitierfähiger Qualität zur Verfügung stellt.

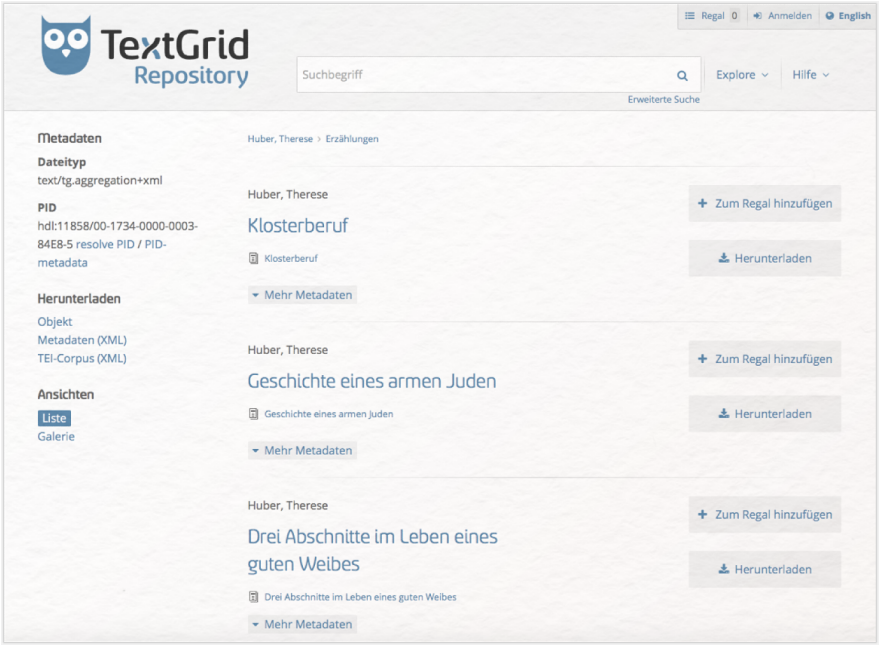

Abb. 1: Benutzeroberfläche des TextGrid Repositorys

Steckbrief

- https://textgridrep.org2
- Volltextsammlung (vgl. Korpus): Texte von Anbeginn des Buchdruckes bis zu den ersten Jahrzehnten des 20. Jahrhunderts von mehr als 600 deutschsprachigen Autor*innen
- Studienausgaben und Erstveröffentlichungen
- Textsorten: Belletristik und Sachliteratur
- Metadaten: Werktitel, Autor, Publikationsdatum, Ort
- Verbundprojekt bestehend aus zehn institutionellen und universitären Partnern (u.a. Berlin-Brandenburgische Akademie der Wissenschaften [BBAW], DAASI International GmbH, Institut für Deutsche Sprache [IDS])
- gefördert vom Bundesministerium für Bildung und Forschung (BMBF) von 2006 bis 2015
- Teil von textgrid.de (und damit der Forschungsinfrastruktur DARIAH-DE), in dem das Repository mit einem downloadbaren Laboratorium und einer Nutzercommunity zusammengebracht wird
- Zielgruppe: Fachwissenschaftler*innen, Entwickler*innen, Forschungsprojekte und -institutionen
- Institutionen wie das Institut für Deutsche Sprache und die Staats- und Universitätsbibliothek Göttingen versprechen die Nachhaltigkeit

- Downloadformate: XML/TEI (und wenige PDF) sowie Bilder als JPEG (und wenige PNG und TIFF)

2. Anwendungsbeispiel

Sie wollen in einem Forschungsprojekt die Erzählungen Therese Hubers miteinander vergleichen. Im TextGrid Repository finden Sie schnell eine Textsammlung dieser Autorin, die auch nach Textsorte klassifiziert sind (hier „Erzählungen“). Ihnen werden sieben Erzählungen angeboten, die inklusive vergleichbarer Metadaten – als kombinierte XML-Datei oder auch einzeln – im standardisierten TEI-Datenformat heruntergeladen oder auch online visualisiert, analysiert oder annotiert werden können.

3. Diskussion

3.1 Kann ich das TextGrid Repository für wissenschaftliche Arbeiten nutzen?

Ja. Das TextGrid Repository garantiert die Textqualität folgendermaßen:

- Aufbauend auf einer Analyse der Textdatenstruktur werden Daten in Ordnern nach Wörterbüchern und Enzyklopädien bzw. nach Gebieten (Geschichte, Literatur, Märchen, Musik, Naturwissenschaften, Philosophie etc.) organisiert und jeder Ordner enthält i. d. R. einen Unterordner pro Autor*in, der alle Werke des Autors/der Autorin in einer Datei vereinigt.
- Textdaten werden durch Metadaten angereichert.
- Werkinformationen werden manuell hinzugefügt (bisher für den Literaturordner).
- Die Metadaten ermöglichen eine Filterung der Dateien nach Textsorte.

Zusätzlich sind weitere Qualitätskontrollen in der Planung, wie:

- die Entwicklung eines User-Interfaces zur manuellen Korrektur der Metadaten,
- die Fehleranalyse der TEI-Auszeichnung und Korrekturen,
- die Optimierung der Datenstruktur hinsichtlich der TextGrid-Architektur, sowie
- eine weitere Strukturerschließung der Texte und tiefergehende TEI-Auszeichnung.

Softwarefehler und Feature-Requests können Sie zudem an textgrid-support@gwdg.de melden.

3.2 Wie benutzerfreundlich ist die Arbeit mit TextGrid?

Die Nutzung des TextGrid Repositorys funktioniert auch für Erstnutzer*innen ziemlich intuitiv. Sie können das Repository entweder direkt via textgridrep.org ansteuern, oder zunächst auf die Hauptseite des Projektes textgrid.de gehen.

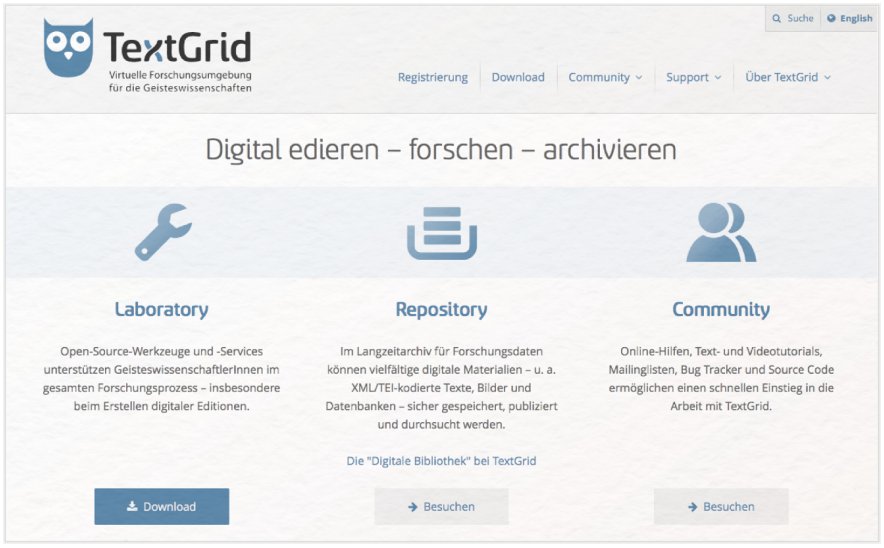

Abb. 2: Startseite von TextGrid

Die drei wesentlichen Teile von TextGrid sind: *Laboratory*, *Repository* und *Community*. Die Menükategorien Regis-

trierung und Download beziehen sich auf das TextGrid Laboratory – eine Software, die verschiedene Textanalysetools zur Verfügung stellt und neben dem Repository das andere wichtige Standbein von TextGrid ist.
Unter den Punkten Community, Support und Über TextGrid finden Sie beispielsweise Informationen über Projekte, die TextGrid nutzen, Möglichkeiten zur Unterstützung durch Online-Hilfen oder Tutorials und die Geschichte und Zielsetzungen von TextGrid. Einen umfangreichen Überblick über die einzelnen Gebiete des Langzeitarchivs (Literatur, Märchen, Geschichte, Philosophie) und eine Darstellung des Korpus für Literatur (Aufbereitung, Metadaten, Download) finden Sie unter „Die Digitale Bibliothek bei TextGrid“ (Kategorie Über TextGrid).
Die Leitlinien des Projektes *Interoperabilität, Homogenität, Open Source* und *Offene Standards* betonen den Nutzungsaspekt: Ein vollständiger wissenschaftlicher Arbeitsablauf inklusive des Austauschs von und über Tools und Texte soll innerhalb der Forschungsumgebung stattfinden können. Das TextGrid Laboratory bietet beispielsweise Software für die kollaborative Erstellung und Publikation digitaler Editionen auf XML/TEI-Basis an.Typische Abläufe bei der Arbeit im TextGrid Laboratory werden in den Tutorials (unter Support) demonstriert: Man findet eine verständliche Beschreibung der Software und lernt den Umgang mit den zentralen Analysewerkzeugen. Die Anzahl der Eulenikone zeigt dabei den Schwierigkeitsgrad des jeweiligen Tutorials an.
In der TextGrid-Community gibt es Informationen über stattfindende Nutzertreffen und Veranstaltungen – und auch die Präsentationen bereits in der Vergangenheit stattgefundener Veranstaltungen können Sie dort herunterladen. Schließlich bietet Ihnen die Mailingliste textgrid-user@gwdg.de die Möglichkeit, sich mit anderen Nutzer*innen von TextGrid auszutauschen.

4. Wie funktioniert die Textsuche im TextGrid Repository?

Auf der Startseite des TextGrid Repositorys können Sie im Suchfeld (vgl. Query) frei nach Texten suchen. Neben der Freitextsuche können Sie unter „Explore“ vordefinierte Suchen nach Autor*innen, Genres, Dateiformaten oder Projekten starten. Tipp: Um in der langen Autor*innenliste schnell die von Ihnen gesuchte Autorin zu finden, bietet es sich an, die Browsersuchfunktion zu nutzen.
Sollten Sie Erfahrung mit der Abfragesprache Lucene haben, können Sie diese im Freitextsuchfeld verwenden und kombinierte Abfragen direkt eingeben.
In der „Advanced Search“ (unter dem Freitextsuchfeld) können Sie beispielsweise nach mehreren Autor*innen gleichzeitig suchen. Mit dem „+“-Button rechts neben der dortigen Suche können Sie weitere Parameter bestimmen oder auch eine Parametersuche mit einer Wörtersuche innerhalb der Texte (unter „Fulltext“) kombinieren und Ihre Suche so verfeinern.
Eine Suche nach „‚Goethe‘ or ‚Schiller‘“ unter „Author“ und dem Genre „Drama“ liefert Ihnen z. B. eine Textsammlung mit 40 Dramen – 28 von Goethe, 12 von Schiller – die Sie entweder einzeln oder unter „Download all“ (rechts oben) als kombinierte XML-Datei herunterladen können. Angemerkt sei hier jedoch, dass häufig auch Teile von Dramen (z. B. von Schillers *Wallenstein* oder Goethes *Faust. Eine Tragödie*) als einzelne Dateien aufgeführt werden und so die genannte Anzahl der gefundenen Dramen irreführend sein kann.
Empfehlung: Legen Sie sich innerhalb des Repositorys ein eigenes virtuelles Bücherregal an! Oberhalb des Downloadbuttons zu jedem einzelnen Text finden Sie hierfür den Button „Add to shelf“. Dies gibt Ihnen die Möglichkeit die Ergebnisse von mehreren Suchdurchläufen zu kombinieren und anschließend diese individuell zusammengestellte Textsammlung als kombinierte Datei herunterzuladen: als XML/TEI-Datei, als komprimierten ZIP-Ordner (vgl. ZIP) oder als E-Book.
Schließlich haben Sie die Möglichkeit, einzelne Texte visuell zu explorieren oder mit einem Tool von DARIAH zu annotieren. Klicken Sie hierzu auf den Titel eines Textes. Links neben dem nun dargestellten Text finden Sie die Kategorie „Werkzeug“ und von dort Verlinkungen zum Visualisierungstool Voyant (Flüh 2024) (in dem dann der jeweilige Text direkt, und ohne dass eine Anmeldung vonnöten wäre, visualisiert wird) und zum DARIAH-Portal zur Annotation, bei dem allerdings zunächst ein Nutzungsprofil erstellt werden muss.
Ebenfalls links vom Text erscheint zudem ein Inhaltsverzeichnis, das die Navigation im jeweiligen Dokument erleichtert.

Externe und weiterführende Links

- Abfragesprache Lucene: https://web.archive.org//web/20241112164707/https://lucene.apache.org/core/ (Letzter Zugriff: 06.11.2024)
- DARIAH-DE: https://web.archive.org/web/20241106113745/https://de.dariah.eu/ (Letzter Zugriff: 06.11.2024)
- TextGrid: https://web.archive.org/web/20241106115617/https://textgrid.de/ (Letzter Zugriff: 06.11.2024)
- TextGrid Repository: https://web.archive.org/web/20241106113832/https://textgridrep.org/ (Letzter Zugriff: 06.11.2024)

Bibliographie

Flüh, Marie. 2024. Toolbeitrag: Voyant. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 5. Textvisualisierung (7. August). doi: 10.48694/fortext.3775, https://fortext.net/tools/tools/voyant.

Neuroth, Heike, Andrea Rapp und Sibylle Söring, Hrsg. 2015. *TextGrid: Von der Community - für die Community. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften*. Glückstadt: Werner Hülsbusch.

Wegstein, Werner, Andrea Rapp und Fotis Jannidis. 2015. Textgrid – eine Geschichte. In: *TextGrid: Von der Community – für die Community. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften*, hg. von Heike Neuroth, Andrea Rapp, und Sibylle Söring, 23–35. Glückstadt: Hülsbusch.

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.

Browsersuchfunktion Um eine Browser-Suchfunktion durchzuführen, und beispielsweise eine Webseite auf bestimmte Suchbegriffe zu filtern, klicken Sie auf Ihrem Mac „cmd“ + „F“ und auf Ihrem Windows PC „Strg“ + „F“. In das sich öffnende Suchfenster tragen Sie Ihren Suchbegriff ein und die jeweils geöffnete Seite wird darauf hin durchsucht.

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.

Korpus Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.

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.

OCR OCR steht für *Optical Character Recognition* und bezeichnet die automatische Texterkennung von gedruckten Texten, d. h. ein Computer „liest“ ein eingescanntes Dokument, erkennt und erfasst den Text darin und generiert daraufhin eine elektronische Version.

PDF PDF steht für *Portable Document Format* . Es handelt sich um ein plattformunabhängiges Dateiformat, dessen Inhalt auf jedem Gerät und in jedem Programm originalgetreu wiedergegeben wird. PDF-Dateien können Bilddateien (z. B. Scans von Texten) oder computerlesbarer Text sein. Ein lesbares PDF ist entweder ein OCRter Scan oder ein am Computer erstellter Text.

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.

Query *Query* bedeutet „Abfrage" oder „Frage" und bezeichnet eine computergestützte Abfrage zur Analyse eines Textes. Um Datenbestände zu durchsuchen, werden Abfragesprachen eingesetzt, die *Queries* (Anfragen) an den Datenbestand senden. So bilden alle möglichen Queries zusammen die *Query Language* eines Tools.

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.

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.

ZIP ZIP steht für ein Dateiformat (zip = engl. Reißverschluss), in welchem mehrere Einzeldateien verlustfrei, komprimiert zusammengefasst werden. ZIP-Dateien werden beim Öffnen entweder automatisch entpackt oder lassen sich per Rechtsklick extrahieren.