Play(s): Crowdbasierte Anreicherung eines literarischen Volltext-Korpus

,
mathiasgoebel@web.de
Universität Göttingen, Seminar für Deutsche Philologie

,
h-l.meiners@gmx.de
Universität Göttingen, Seminar für Deutsche Philologie

Digitale Korpora entstehen unter bestimmten Voraussetzungen, werden von verschiedensten Institutionen gefördert und haben unterschiedliche Ziele in Bezug auf Qualität und Quantität. In vielen Fällen müssen Forscherinnen weitere Verbesserungen vornehmen, bestehende Daten erweitern und verbessern, im schlimmsten Fall auch neu erheben. In diesem Paper beschreiben wir eine Vielzahl einfach(st)er Probleme, die es zu bewältigen gilt, will man ein Korpus bestehend aus möglichst vielen genuin deutschsprachigen Dramen computergestützt analysieren. Ausgehend von den Beständen des TextGrid Repository (TextGrid Konsortium 2015) soll mittels einer simplen grafischen Oberfläche – abrufbar in jedem Webbrowser – ein Programm zur Verfügung gestellt werden, das sich spielähnlicher Mittel bedient: die Nutzer auffordert, mehrere Level zu durchlaufen, Punkte zu sammeln und mit jeder Eingabe das Korpus zu verbessern, um schließlich Ausgangsmaterial für eine Vielzahl von Fragestellungen zu bieten.

Crowdsourcing, Social Editing und viele verwandte Begriffe sind Konzepte, die innerhalb der Digital-Humanities-Community in den vergangenen Jahren einen kleinen Hype erfahren haben. An Umsetzungen mangelt es, während man sich noch über die Definitionen streitet. Dabei sind die Lösungsansätze sehr vielversprechend – allen voran die von Zooniverse etablierten Projekte, die sich nun auch geisteswissenschaftlichen Themen widmen. Es werden alte Texte transkribiert und jede Person, die über Computer und Internetanschluss verfügt, kann einen aktiven Beitrag leisten und die Forschung unterstützen. Erfahrungen aus bereits gut etablierten Ansätzen, wie dem Projekt DigitalKoot , entwickelt von der Finnischen Nationalbibliothek, zeigen, dass die Anwendung und Umsetzung spielerischer Verfahren durchaus einen Mehrwert für die Wissenschaft generieren können. Darüber hinaus können Aufgaben gemeistert werden, die durch einen Einzelnen oder auch eine kleinere Forschungsgruppe niemals würden selbst bewältigt werden können. Voraussetzung dafür ist das Interesse und die Teilnahme vieler Personen an der zu bewältigenden Aufgabe. Diese so zu isolieren und danach zu vereinfachen, dass auch Laien damit umgehen können, stellt die Herausforderung dar. Viel Resonanz bekommen Projekte wie EyeWire oder GalaxyZoo , beides naturwissenschaftliche Citizien-Science-Vorhaben. Im Projekt GalaxyZoo geht es um die Klassifizierung und Beschreibung von Galaxien. Das Projekt war in der Lage, 50 Millionen Klassifizierungen zu sammeln, und das innerhalb seines ersten Betriebsjahres (2007, vgl. Prestopnik 2011: 2).

Methodisch betrachtet bieten Konzepte, die auf der zunehmenden Beteiligung der sogenannten *Crowd* aufbauen, ein großes Potential zur Weiterentwicklung oder Herausbildung neuer Forschungsfragen und -themen. Das bereits erwähnte *Crowdsourcing* kann dabei als Überbegriff für verschiedene Ansätze gesehen werden, die zwar z. T. deutlich voneinander abzugrenzen sind, in manchen Bereichen jedoch deutliche Ähnlichkeiten aufweisen. In einem groben Kategorisierungsversuch kann eventuell eine Zweiteilung vorgenommen werden, um einen besseren Überblick über diese verschiedenen und doch ähnlichen Methoden und Konzepte zu gewinnen. *Serious Games*, *Games with a Purpose* und *Meaningful Play* wollen das Spiel als Medium nutzen um bestimmte Inhalte oder Absichten zu transportieren und dem Nutzer ein bestimmtes Problem näher zu bringen. Ansätze wie *Gamification*, *Gameful Design*, *Social Editing* oder *Human Computation* nutzen gezielt einzelne Elemente aus dem Spieldesign, um einen eigentlich spielfremden Kontext anzureichern und durch die Schaffung einer neuen Atmosphäre attraktiver für potentielle Nutzer_innen zu gestalten.

Während Play(s) sich methodisch eher in der zweiten genannten Kategorie wiederfinden soll, ist wichtig zu betonen, dass der Erfolg solcher Projekte eng mit der Entwicklung und dem Design selbst zusammenhängt. Die Oberfläche sollte beispielsweise ansprechend gestaltet bzw. angemessen bezogen auf die Zielgruppe und einfach zu bedienen sein. Als Spielelemente können Levels, das Sammeln von Punkten in Kombination mit einfachen Spielanweisungen dienen. Neben der tatsächlichen Entwicklung einer neuen Anwendung hängt ein großer Teil des Erfolgs von der zu tätigenden Handlung der Teilnehmer_innen ab. Die Aufgabe, die im Rahmen eines Crowdsourcing oder Citizien-Science-Projektes von den Teilnehmer_innen bearbeitet werden soll, ist im besten Fall einfach zu verstehen und in simple Teilbereiche unterteilt. Gleichzeitig unterliegt die Einbindung von freiwilligen, fachfremden Teilnehmer_innen gewissen eher impliziten und wenig ausgesprochenen Regeln. So sollten die Teilnehmenden generell als Partner oder Mitarbeiter_innen betrachtet werden und nicht als günstige Arbeitskräfte. Zudem sollten sie nicht zur Bewältigung von Aufgaben angehalten werden, die eigentlich einfacher und besser von einem Computer ausgeführt werden könnten. Diese Grundethik sollte bei jeder Umsetzung einer

neuen Idee zumindest mitbedacht werden, um künftige Teilnehmer_innen nicht zu verärgern oder zu verschrecken.

Die wenigen bisher gesammelten und verfügbaren Erfahrungen aus Projekten für die Geisteswissenschaft sollen nun ausgewertet werden und in die Umsetzung einer neuen Projektidee eingebracht werden. "Play(s)" ist der Name der Anwendung, die sich damit befassen soll, ein literaturwissenschaftliches Volltext-Korpus anzureichern. Das TextGrid-Repositorium bietet dafür optimale Voraussetzungen: alle Texte sind im TEI-Format erfasst und diese Quelle ist frei zugänglich.

In diesem Projekt knüpfen wir an die von einer Projektgruppe (vgl. Trilcke et al. 2015) bereits herausgefilterten Dramen des Repositoriums an. Dabei wurden bereits in einem manuellen Durchgang allen Sprecherinstanzen im Auswahlkorpus eindeutige Namen (IDs) zugewiesen, um eine Ausgangsbasis für Netzwerkanalysen zu schaffen. In diesen Vorarbeiten wurden die genuin deutschen Texte ausgewählt und dabei aus den insgesamt 666 Dramen auf 465 Werke zurückgegriffen. Um diese Analysen mit einer quantitativ und qualitativ erweiterten Quellenbasis zu vertiefen, bedarf es einer noch genaueren Referenzierung. So sollten zum Beispiel die als Sprecher auftretenden Personengruppen aufgelöst werden und zu diesen die beteiligten Akteure genannt werden. Auch eine Klassifizierung des Geschlechtes, der sozialen Stellung und weitere Features sind denkbar, um differenzierte Analysen tätigen zu können. Hier wird deutlich, dass jeder Text einer bestimmten Aufbereitung bedarf, die aber in vielen kleinen Einzelschritten erfolgen kann, da die Informationen und einzelnen semantischen Anreicherungen in ihren Kategorien unabhängig voneinander sind.

Innerhalb des TextGrid-Korpus beschränken wir uns auf die Betrachtung der Dramen und innerhalb derer sind es die Strukturinformationen, die auf Grundlage des XML-Codes Netzwerkanalysen auf Basis des gemeinsamen Auftretens in einer Szene ermöglichen. Gemeinsames Auftreten heißt in diesem Fall, dass innerhalb einer Szene alle Sprecher_innen in Verbindung gebracht werden. Dazu gilt es die einzelnen Akteure ausfindig zu machen, da das Korpus selbst keine Information, wie man sie im TEI-Attribut who (TEI Consortium 2015) erwarten kann, mitliefert. Betrachtet man als Beispielfall das Drama "Fraw Wendelgard" von Nicodemus Frischlin, finden sich innerhalb der Sprecherbenennung drei verschiedene Schreibweisen, die alle auf die Gräfin Wendelgard verweisen: Wendelgard, Wendelgart und Wendelgardt. In einem anderen Werk taucht in einem Dialog zwischen Faust und Mephistopheles ein einziges Mal der Sprecher "Mephistoph" auf. Häufig beobachtet man Akteure, die mit einem unbestimmten Artikel eingeführt werden, im Folgenden aber mit bestimmtem oder ohne Artikel angegeben werden, wobei offensichtlich ist, dass es sich um die vorangehende genannte Entität handelt.

Die Ursache kann drucktechnisch bedingt in den Buchausgaben liegen, in denen Sprechernamen abgekürzt werden, um Platz und Papier zu sparen, es können auch schlicht Fehler im Satz auftauchen und eine weitere Fehlerquelle kann der Digitalisierungsprozess sein. In all diesen Fällen ist die Korrekturaufgabe denkbar simpel: man muss jene Sprecher zusammenführen, bei denen es sich offensichtlich um die gleiche Person handelt. Getreu der Buchausgaben handelt es sich dabei nicht um Fehler, das Encoding muss hier schlicht um semantische Information erweitert werden, wie es das Attribute who in den TEI Guidelines vorsieht. Dazu zählen auch Fälle, in denen das Markup innerhalb des TextGrid-Korpus fehlerhaft ist. Das betrifft leere speaker-Elemente, solche, in denen noch Teile der Bühnenanweisung mit einfließen und auch jene, die noch ein leeres Element stellvertretend für zum Beispiel einen Seitenumbruch beinhalten und dadurch als Auswertung des Inhaltes von tei:speaker ein Leerzeichen voran steht.

Man findet außerdem bei gemeinsam sprechenden Personengruppen unterschiedliche Nennungen. In einem Drama Friedrich Kaisers ist eine solche Aggregation mit “HELFER UND ROBERT” benannt, später folgt aber "ROBERT UND HELFER”. Eine bestimmte vom Autor intendierte Hierarchie soll das Datenmodell nicht abdecken und somit gilt es die verschiedenen Zeichenketten als eine Entität zu betrachten. Zudem soll die Tiefenauszeichnung dieser Elemente weiter gehen und jeder Gruppe die einzelnen, sofern bestimmbaren, Akteure zugewiesen werden.

Diesen Beobachtungen folgt die Spielstruktur.

In einem ersten Level gilt es die unterschiedlich benannten aber in der fiktiven Welt gleichen Sprecher zu identifizieren. Dazu werden alle unterschiedlichen Zeichenketten innerhalb der tei:speaker-Elemente eines Dramas zunächst in der Reihenfolge ihres ersten Auftretens gelistet. Mutmaßlich gleiche Namen sind nacheinander auswähl- und abspeicherbar. Ist dies für ein Drama vollständig geschehen, kann dieses Drama als “gelöst” markiert werden.

Weiterhin gilt es Aggregationen ausfindig zu machen (Level 2).

Diese Aggregationen sollen schließlich aufgelöst werden (Level 3). Dazu sind nicht nur die an einer Gruppe beteiligten Akteure zu nennen, sondern auch deren Vollständigkeit zu deklarieren. Es kann zum Beispiel das Volk sprechen und weiterhin einzelne Personen aus dem Volk auftreten. Diese sind Teil des Volkes, die Gruppe selbst ist aber eine weitaus größere und daher unvollständig durch die einzelnen Akteure belegt. Sprechen zwei auch näher bestimmte Einzelpersonen gemeinsam, so kann diese Gruppe vollständig aufgelöst werden.

Die Geschlechter der Akteure sind in Level 4 zu bestimmen. Dabei ist zu wählen aus male, female, both, und unknown. Die letzte Gruppe umfasst dann schließlich auch metaphysische Konstrukte, die personifiziert auftreten.

In Level 5 sollen diese dann genauer spezifiziert werden. Dabei stehen die Kategorien Tier, metaphysisches Wesen (z. B. Gottheit, Hexen und Magier) und Eigenschaft / Gefühl / Moral zur Auswahl.

Schließlich lässt sich noch der soziale Status bestimmen, sofern Berufsbezeichnungen, Adelstitel oder andere Indikatoren ausfindig zu machen sind.

Zwischen den einzelnen Levels gilt es die Eingaben anderer Spieler zu verifizieren oder auch zu falsifizieren. Diese Eingabe wirkt sich auf die eigenen Punkte immer positiv aus, die jeweils anonym bleibende begutachtete Spielerin wird bei Fehleingaben aber Punktabzüge bekommen. Da die Dramen immer zufällig gewählt werden und auch mehrfach erfasst werden, stehen damit verschiedene Qualitätskontrollen zur Auswahl, die auch kontinuierliche nicht sinnvolle Eingaben erkennen lassen. Die betreffenden Spielerinnen können weiterspielen, finden aber nur noch eine persönliche Highscoreliste vor, während sie aus den Highscorelisten anderer getilgt werden und ihre Eingaben auch nicht in den weiteren Forschungsprozess Einzug halten. Zudem stehen die Daten für 465 Dramen im LINA Zwischenformat (vgl. Trilcke et al. 2015) zur Verfügung, die mit den in Level 1 erfassten Eingaben übereinstimmen sollten. Zudem können alle hier getätigten Erhebungen ebenfalls in das Zwischenformat einfließen, womit sie dann für die Netzwerkanalysen des Projektes zur Verfüggun stünden. Bei Bedarf ließen sich die Ergebnisse sogar direkt in die Quelldokumente übernehmen.

Kritisch betrachtet stammen aus der Welt der Computerspiele die Levelstruktur und einzelne Elemente, wie Avatar und Highscoreliste. Tatsächlich ist das Angebot eines, das Social Editing auf einfachste Fragestellungen hin anwendet und jeder Spielerin die Möglichkeit bietet, aktiv an der Tiefenerschließung von literarischen Texten mitzuwirken. Außerdem gibt es einen didaktischen Aspekt, da komplexe Probleme im Hinblick auf Korpuserstellung implizit aufgezeigt werden.

Bibliographie

Prestopnik, Nathan R. (2011): *Citizen Science Case Study*. Galaxy Zoo / Zooniverse http://citsci.syr.edu/system/files/galaxyzoo.pdf [letzter Zugriff 15. Oktober 2015].

TEI Consortium (2015): *TEI P5*. Guidelines for Electronic Text Encoding and Interchange. Version 2.9.0. Updated on 9th October 2015. http://www.tei-c.org/Guidelines/P5/ [letzter Zugriff 15. Oktober 2015].

TextGrid Konsortium (2015): *Die Digitale Bibliothek bei TextGrid* https://textgrid.de/digitale-bibliothek [letzter Zugriff 15. Oktober 2015].

Trilcke, Peer / Fischer, Frank / Göbel, Mathias / Kampkaspar, Dario (2015): *Network Analysis of Dramatic Texts* https://dlina.github.io/about/ [letzter Zugriff 15. Oktober 2015].