CorpusExplorer v2.0 - Seminartauglich in einem halben Tag

Rüdiger, Jan Oliver
jan.ruediger@uni-kassel.de
Universität Kassel, Deutschland

Grundzüge:

Die Fähigkeit zur Kritik setzt voraus, dass die notwendigen Fähigkeiten zur Durchführung vorhanden sind. Es bedarf jedoch des Mutes sich zu entschließen, durchzuhalten und nicht bequem zu werden und somit die eigenen Fähigkeiten/Methoden kontinuierlich zu verbessern. Methoden wie Sie in den Digital Humanities und speziell in der Korpuslinguistik zum Einsatz kommen, lassen sich nur verbessern, wenn man selbst tätig wird, hinterfragt, ausprobiert und gemeinsam diskutiert. Im Rahmen dieses Workshops wird der CorpusExplorer v2.0 vorgestellt (OpenSource), der unterschiedlichste Methoden aus dem Bereich der Forschung holt und diese für die universitäre Lehre bereitstellt. Studenten sollen mit dieser Software ermutigt werden, eigene kleine Forschungsprojekte zu realisieren (es wurden bereits Seminararbeiten, Bachlor-/Masterarbeiten sowie (laufende) Dissertationsprojekte mittels CorpusExplorer umgesetzt).

Dies ist nicht selbstverständlich, so weisen bereits (Bubenhofer 2011) „Oft bedingen korpuslinguistische Arbeiten einen großen Aufwand, sowohl für Lernende als auch die Betreuenden, der im Rahmen eines Studiums nicht geleistet werden kann.“ oder (Dipper 2011) „Bei der Arbeit mit ‚echten‘ Daten, [...] werden die Computerlinguistik-Studenten früh mit Problemen wie dem Daten-Encoding oder der Datengröße konfrontiert [...]“ auf elementare Probleme zu Seminar-/Projektstart hin. Außerdem ist es in der Regel notwendig, dass unterschiedliche Programme kombiniert werden, um ein (visuelles) Ergebnis zu erzielen.

Der CorpusExplorer v2.0 beseitigt viele dieser (Einstiegs-)Hürden. Unterschiedlichste Programme und Methoden werden unter einer benutzerfreundlichen Programmoberfläche kombiniert, die zudem vielfältige Visualisierung/Weiterverarbeitungsmöglichkeiten zur Verfügung stellt (wie auch in den Video-Tutorials von (Rüdiger 2017) bereits gezeigt wurde). Im Vergleich zu AntConc, TXM und anderen verbreiteten Tools wird schnell klar, wie stark sich der CorpusExplorer an Forschung -und- Lehre orientiert.

Einen Einblick in die Workshopinhalte:

Korpora erstellen

Der CorpusExplorer automatisiert den gesamten Erstellungsprozess. Die Möglichkeiten Korpusmaterial zu akquirieren sind vielfältig – PDF, eBooks, X/HML, Tweets, Blogs, uvm. – lassen sich einlesen, Text-/Metadaten werden getrennt, der Text bereinigt (z. B. von unerwünschten HTML/XML-Tags), abschließend erfolgt eine Annotation (z. B. durch den TreeTagger, Stanford POS oder TnT). Vom Rohtext zum analysefertigen Korpus ist die Nutzer*in nur wenige Mausklicks entfernt. Dies bietet verschiedene Möglichkeiten, Korpora werden entweder zentral durch Dozent*in/Tutor*in bereitgestellt oder selbst von den Student*innen aufgebaut. Außerdem besteht die Möglichkeit, Korpusmaterial im Seminarverlauf gemeinsam zu pflegen, zu erweitern und auszurollen.

Auswertungen

Im CorpusExplorer stehen über 45 Analysemodule zur Verfügung. Bei einigen davon überschneidet sich die Datengrundlage – es wird erprobbar, wie Visualisierungen in den Rezeptionsprozess eingreifen. Am Beispiel der Kookkurrenzanalyse wird dies besonders deutlich, optisch unattraktiv weil mächtig und funktional umfangreich – die Tabellen-Darstellung zeigt alle Daten auf einmal (filter-/gruppier-/sortierbar). Als WordCloud (auf Basis der TagCloud-Visualisierung von (Jänicke et al. 2015)) werden Bedeutungen/Schnittmengen einzelner Begriffe schnell sichtbar. In der Graphen-Darstellung lassen sich Zusammenhänge explorativ erkunden. Kombiniert mit der Auswertung von N-Grammen lassen sich Sprachgebrauchsmuster schnell identifizieren (welche Teile eines N-Gramms sind signifikante Kookkurrenzpartner?).

Nicht nur Text-Daten lassen sich auswerten, sondern auch Meta-Daten, diese werden während des Erstellprozesses automatisch mit erfasst oder lassen sich nachträglich manuell erweitern/ändern. Im Seminar können unterschiedliche Ansätze/Haltungen auf diese Weise erprobt werden – z. B. ob und wie sich die Korpuszusammensetzung auf das Analyseergebnis auswirkt (Bsp.: ausgewogenes Korpus, pragmatischer Ansatz oder „more data is better data“).

Das Konzept der Schnappschüsse erlaubt die Filterung/Zusammenstellung individueller Teilkorpora. Korpusmaterial und Analysen sind durch Schnappschüsse voneinander isoliert. Neu hinzukommendes Material verwirft keine bisherigen Ergebnisse. Aus vielen Analysen lassen sich Schnappschüsse erstellen, um konkreten Forschungsfragen nachzugehen – Es können individuelle Filter auf Text-, Meta- und Korpus-Ebene getroffen werden oder der CorpusExplorer kann anhand von Vorgaben mehrere Teilkorpora auf einmal erstellen (z. B. jeder Autor/

Verlag isoliert – Datums/Zeitabschnitte, etc.). Letztlich lassen sich Schnappschüsse durch Mengenoperatoren kombinieren.

Schnappschüsse erlauben es, das gegenwärtige Forschungsinteresse zu lenken und gezielt durch große Korpusmengen zu navigieren. Gegenwärtig stehen sich corpus-driven und corpus-based Ansätze sowie close- und distant-reading gegenüber. Da Ansätze beider Denkrichtungen in diesem Programm vereint sind, erlaubt der CorpusExplorer nicht nur einen schnellen Perspektivwechsel sondern schafft neue Möglichkeiten des Arbeitens.

Transparenz & Anbindung an andere Programme / Programmiersprachen

Ein wesentliches Teil des OpenSource-Gedankens im CorpusExplorers ist: Transparenz – Nicht nur, dass sich viele Formate verarbeiten lassen, auch der Prozess lässt sich ohne viel Aufwand verifizieren (Generator zur Erzeugung von Dummy-Korpora für die Prozessvalidierung) und alle Analyseergebnisse lassen sich exportieren (inkl. einer Konvertierung des CorpusExplorer-Formats in andere XML/JSON-Formate). Aus dieser, anfänglich als Möglichkeit gedachten Funktion, entstanden zwei Möglichkeiten für fortgeschrittene Nutzer*innen. Das HTML5-Labor erlaubt es, Daten/Analysen direkt im CorpusExplorer mittels HTML5, JavaScript und CSS zu visualisieren (Syntax-Editor inkl. HTML5 Rendering-Engine auf Basis von Chromium stehen bereit). Die zweite Möglichkeit erlaubt den kompletten Verzicht auf die Programmoberfläche. Mittels R, Kommandozeile oder anderer Programmiersprachen kann auf den CorpusExplorer zugegriffen werden. Der CorpusExplorer ist keine Ultima Ratio – er kann aber helfen, einen schnelleren Einstieg in die Korpuslinguistik zu finden und erleichtert selbst in fortgeschrittenen Szenarien die Arbeit.

Wie im Seminar einbinden

Ein didaktisches Konzept wird im Rahmen dieses Workshops nicht vorgestellt. Aus der Praxiserfahrung heraus werden einige Beispiele erfolgen, wie man diese Software bereits im Bachelor-Studium einsetzen kann. Zum Beispiel wie die von (Bubenhofer 2011) angesprochene Problematik „[…] die Studierenden überhaupt dazu zu motivieren, korpuslinguistisch, also empirisch zu arbeiten.“ durchbrochen werden kann. Dies gelingt im Wesentlichen dadurch, dass die Studenten bereits in der ersten Seminarsitzung aktiv arbeiten können und erkennen, welchen Mehrwert und Spaß empirisches Arbeiten bietet.

Workshopverlauf

Im Workshop wechseln sich vier Sozialformen in unterschiedlicher Reihung ab. In den Vortrags-/Frontal-Sequenzen werden die wesentlichen Funktionen und deren Hintergründe erklärt – z. B. wie werden N-Gramme ausgezählt? In Live-Demonstrationen wird die konkrete Anwendung gezeigt – Was muss eingestellt / ausgewählt / angeklickt werden, um ein Ergebnis zu erreichen? Dabei können die Teilnehmer*innen das Gezeigte an ihrem Rechner nachverfolgen. Außerdem erfolgt eine eigene Erprobungsphase – hier probieren die Teilnehmer*innen selbständig eigene Parameter aus und durchforsten das Korpusmaterial auf eigene Faust. Abschließend wird alles innerhalb der Workshopgruppe diskutiert. Für den Workshop werden unterschiedliche Korpustypen zur Verfügung gestellt, eigenes Korpusmaterial kann ggf. in der Erprobungsphase getestet werden.

Voraussetzungen

Im Idealfall würde ein PC-Poolraum mit Windows-Rechnern ab Windows 7 inkl. Beamer zur Verfügung gestellt, auf dem der CorpusExplorer bereits vorinstalliert ist. Ggf. könnten die Teilnehmer*innen die Software auch selbst installieren (für die Installation sind keine Administratoren-Rechte notwendig). Falls kein geeigneter Poolraum zur Verfügung stehen sollte, wäre ein Seminarraum mit Internetzugang, Beamer und ausreichenden Steckdosen für die Teilnehmer*innen notwendig. Die Teilnehmer*innen könnten ihren eigenen Windows-Rechner mitbringen. Teilnehmer*innen mit Linux/MacOS Notebook sollten sich zuvor melden, damit eine Virtualisierungslösung bereitgestellt werden kann.

Bibliographie

Bubenhofer, N. (2011): Korpuslinguistik in der linguistischen Lehre: Erfolg und Misserfolge. In: Journals for Language Technology and Computational Linguistics 26 (1), S. 141–156.

Dipper, S. (2011): Digitale Korpora in der Lehre: Anwendungsbeispiele aus der Theoretischen Linguistik und der Computerlinguistik. In: Journals for Language Technology and Computational Linguistics 26 (1), S. 81–95.

Jänicke, S.; Blumenstein, J.; Rücker, M.; Zeckzer D. & Scheuermann, G. (2015): Tagpies.

Rüdiger, J. (2017): Korpushermeneutische Analyse politischer Reden mittels CorpusExplorer. In: 10plus1 - Living Linguistics (3), S. 11–21.