Mein Computer sagt mir, welche Quellen ich lesen soll! – Qualitative und quantitative Textanalysen verbinden

@historytoby

Datum/Uhrzeit: 27.11.2015, 16:15 Uhr
Protokoll: Johannes Mirus, Bundesstadt.com

Ausgangslage

Masterarbeit zum Thema Homosexualität in Qualitätszeitungen der USA. 40.000 Texte müssen verarbeitet und bewertet werden.

Inhalt

- Sprache wird im Zusammenhang von gesellschaftlicher Macht und sozialer Ungleichheit untersucht → qualitative Diskursanalyse
- Korpuslinguistik (große Mengen echte Sprachdaten untersuchen) → quantitative Analyse
- Methodik: CADS (Corpus Assisted Discourse Study), hybride Methok qual/quant
- Vorgehensweise:
  - Datenbank Nexis (Washington Post, New York Times) → nach Suchbegriffen suchen => fast 20.000 Artikel mit fast 20 Millionen Wörtern, nach Einschränkung 2.864 Artikel mit 2,7 Millionen Wörtern (endgültiger Korpus muss noch besser werden, AIDS/HIV haben zu viel Rauschen verursacht)
  - Sortierung der Treffer nach gefundenen Begriffen, welche kommen sehr oft gemeinsam vor (Resultat z.B.: Homosexualität und Alkoholismus kommen oft gemeinsam vor)
  - Statistische Untersuchung dieser Paare
  - Da hört die Korpusanalyse normalerweise auf. Jetzt kommt die qualitative Zweitanalyse.

  - Die ergibt z.B., das fast alle Alkoholismus-Homosexualität-Verbindungen auf nur einem “politischen Skandal” beruhen. Quantitative Analyse hat einen Zusammenhang also nur vorgegaukelt.
- Der Computer sagt also, was man lesen soll: Häufige Funde werden näher untersucht. Der Computer hilft, eine Masse von Texten vorzuqualifizieren.
- Am längsten dauerte erst einmal das Zusammensuchen der Zeitungsartikel. Die quantitative Analyse kostete gerade einmal einen Vormittag.

Diskussion

- Wie kannst du sicher sein, dass woanders nicht auch Homosexualität diskutiert wurde, ohne das Wort zu nennen? – *Kann ich nicht. Man muss sich eine Art Wörterbuch aufbauen, aber sichergehen kann man nicht.*
- Die Transformation von Print zu Digital mittels OCR ist fehlerhaft! – *Ja, es gibt Fehler, aber eine stichprobenhafte Untersuchung zeigte, dass die Fehlerquote sehr gering ist.* – OCR-Genauigkeit sollte über 99 Prozent betragen, sonst ist ein Text nicht mehr leserlich. – Ist sowieso ein Grundproblem der Geschichtswissenschaft, ältere Dokumente wie Urkundenabschriften aus dem Mittelalter sind noch viel ungenauer!
- Wie viele von den 40.000 Artikel hast du dann wirklich gelesen? – *Es bricht runter auf bis zu nur noch zehn Artikel, die man wirklich aufmerksam durcharbeiten muss. Weitere fünf bis zehn überfliege ich.*
- Wie wählst du die Quellen aus? Homosexualität wurde nicht nur in Zeitungen diskutiert! – *Ja, ich wollte aus dem Ungefähren raus. Ich musste selektieren.* – Aber dann musst du das begründen! – *Tue ich.* – Vollständig erfassen klappt sowieso nicht.
- Wie verträgt sich dein Ansatz von Objektivität in Verbindung mit dem Thema “Queer”? Das ist ja sehr subjektiv für die Linguistik. – *Ich bewerte nicht, ich werte nur statistisch aus.*
- Die Auswahl der Zeitungen ist schon nicht mehr objektiv. Menschen an der Ostküste sehen das Thema bestimmt anders als die in der Mitte der USA! – *Ja. Es ist ja auch nicht DIE Auswertung über die Stimmung der USA, sondern aus den beiden Zeitungen.*