1.1 Purpose
1.2 Tasks
1.3 Forms of Data
1.4 Key Word in Context (KWIC)
1.5 Select Column in Dataset
1.6 Extract texts from tsv:
1.7 Variables
1.8 Better Tokenization
2.8 new, improved approach
3.8 old approach

- Tokenisierung = in Wörter splitten
- Interpunktion "klebt nicht mehr an Wörtern"
- -s = beliebig viele Zeichen
- -c = Komplement (also nicht diese Zeichen)
- angegebene Zeichen werden NICHT ersetzt

3.9 Batch Processing
3.10 Batch Renaming
4.10 since there are different versions, if this doesn't work try:
**5.10 rename 's/ /_/' *.txt
5.11 Imperfect Data: A Tail of Bias**

- fehlende, rauschende, selektive & verzerrte Daten
- sozialer Kontext
  - z.B. Budgetkürzung oder Neuausrichtung –> Wegfall von Thema
  - Sicht von weisen Männern auf Thema
- non-content elements
  - Metadaten, Kopfzeilen etc.

5.12 Outlook: NLP is on Fire

- die meisten haben ihr Schulwissen wieder vergessen, wieso kann das der Computer
- Intuition einfach, genaue technische Funktionsweise egal
  - Genauigkeit wichtig, aber noch zu advanced
  - best-practice
- viel genauer dank Embeddings (self-supervised)

5.13 Mind your Data
5.14 Anatomy of AI[@Crawford2018] illustrated by the Amazon Echo

- Es geht um mehr als nur Technologie
- Technologie ist eingewoben ins Soziale
- Soziotechnische Systeme

5.15 Grid Example