CL 2 Übung

Korpora & Annotationen

Dr. des. Johannes Hellrich

https://julielab.de

18.6.2019

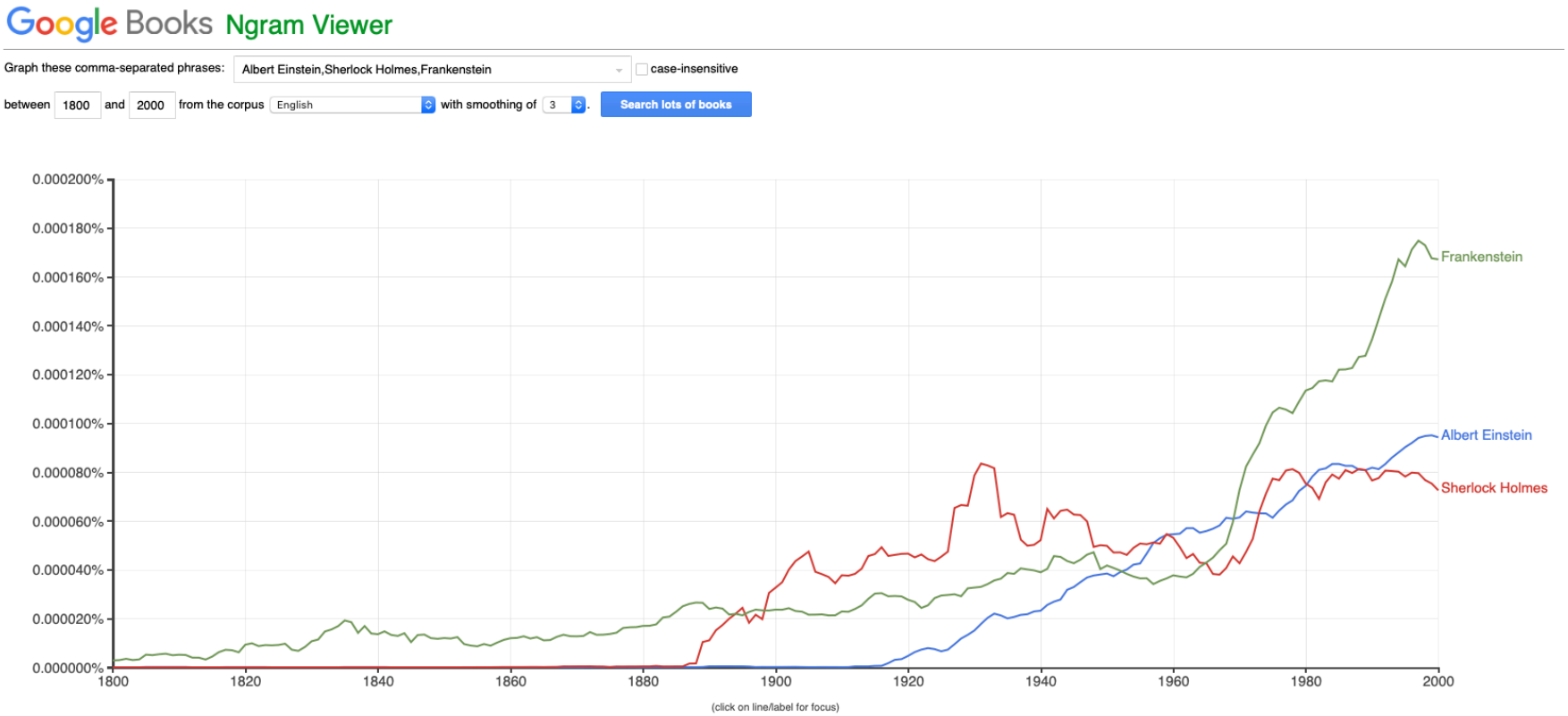
Jena University Language & Information Engineering Lab

- Weitere Korpora
- Inter-Annotator Agreement

Google Books N-gram Korpus
(Michel et al., Science 2011)

- Frei verfügbar
- Extrem groß, ca. 6% aller Bücher bis 2009
- 1-gram bis 5-gram wegen Copyright/Speicherplatz
- Mehrere Sprachen (en, de, es, fr, ...)
- Für Englisch mehrere Varietäten (amerikanisch, britisch, fiction)
- Basiert auf OCR Scans (v.a. Bibliotheksbestände)
- Nachteile: unbalanciert & opak

Google Books N-gram Viewer



https://books.google.com/ngrams

COHA: Corpus of Historical American English
(Davies, Corpora 2012)

- Online frei verfügbar, offline kostenpflichtig
- Balanciert
- Diachron
- Schwärzt 10% des Texts wg. Copyright
- https://www.english-corpora.org/coha/

OPUS: open parallel corpus

- Sammlung paralleler Korpora in verschiedenen Sprachen
- Nutzung v.a. als Trainingsmaterial für maschinelle Übersetzung
- Beispielsinhalte:
  - EU Parlamentsdebatten
  - Wikipedia
  - Online News
- http://opus.nlpl.eu

- Weitere Korpora
- Inter-Annotator Agreement

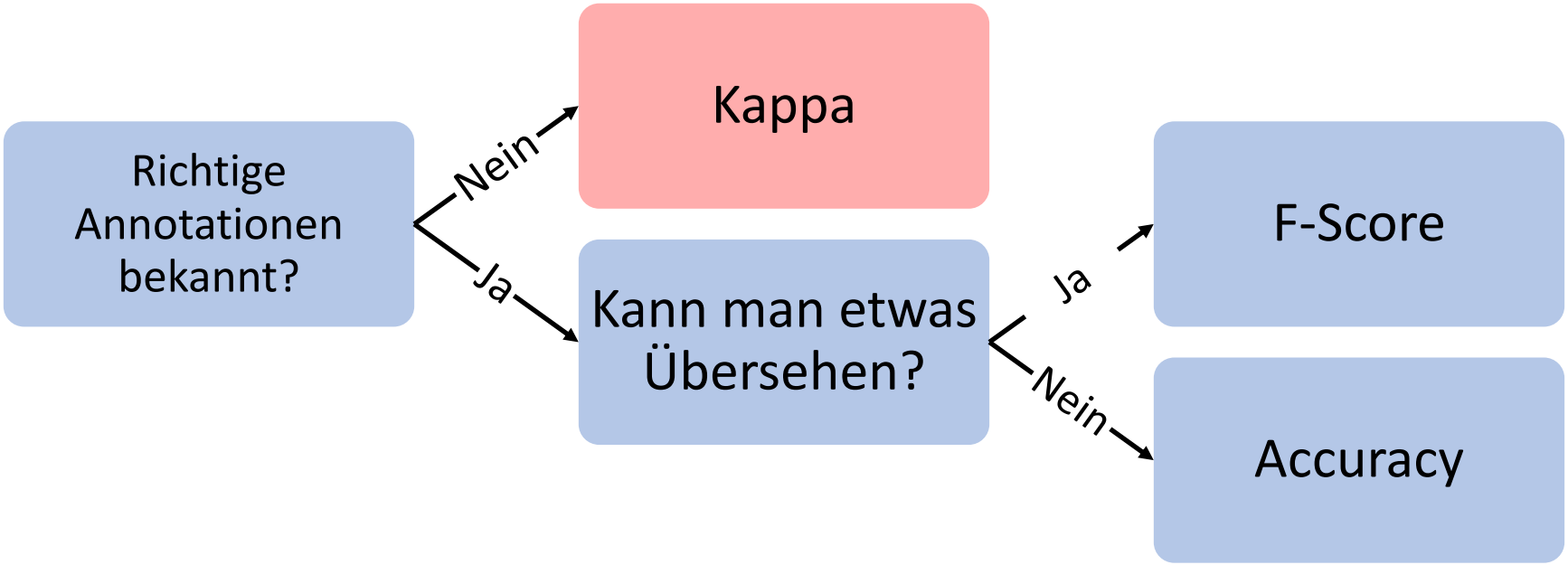
Wir wollen wissen...

... wie zuverlässig arbeiten Mitarbeiter?

... wie gut funktionieren Tools?

... wie einig kann man sich über Kategorien sein?

Auswahl des richtigen Verfahrens



Was ist Kappa?

- Kommt aus der Psychologie (sind sich 2 Ärzte einig woran der Patient leidet?)
- Annahme: Annotatoren sind gleich (un-)fähig, die Qualität ihrer Urteile wird klar, wenn man ihre Annotationen vergleicht
- Berechnung der Übereinstimmung bei der Vergabe von **nominalen** Urteilen unter Ausschluss der **zufällig zu erwartenden Übereinstimmung**

Nominale Urteile am Beispiel POS-Tags

- Schließen einander aus (je Wort nur eine Wortart)
- Haben keine Reihenfolge/Wertigkeit (jede Wortart ist gleich wichtig)
- Decken alles ab (es gibt nur die definierten Wortarten und jedes Wort bekommt eine)
- Wurden für den Vergleich unabhängig voneinander erhoben

Beobachtete Übereinstimmung

$$A_o = \frac{\sum_{i=0}^{n} Urteil_{ii}}{\sum_{i=0}^{n} \sum_{j=0}^{n} Urteil_{ij}}$$

- Urteil ist eine Tag x Tag Matrix mit den Häufigkeiten für das Vorkommen dieser Tagkombination
- A_O, die beobachtete Übereinstimmung, wird berechnet, indem man die Anzahl der Fälle in denen das gleiche Tag vergeben wurde (Diagonale) durch die Summer aller vergeben Tags teilt

Wozu Ausschluß der zufällig zu erwartenden Übereinstimmung?

- Beim Münzwurf gibt es eine beobachtete Übereinstimmung von 50%…
- Effekt besonders stark bei wenigen Kategorien oder sehr ungleicher Kategorienhäufigkeit

Annotator 1	Annotator 2		
	Kopf	Zahl	Summe
Kopf	0,25	0,25	0,50
Zahl	0,25	0,25	0,50
Summe	0,50	0,50	1,00

Kontingenztafel / Kreuztabelle zum Münzwurfbeispiel, enthält bereits Anteile

Kappa berechnen

$$\kappa = \frac{A_o - A_e^\kappa}{1 - A_e^\kappa}$$

- Der Wert bewegt sich typischerweise zwischen 0 und 1
- Ein Wert von 0 entspricht Übereinstimmung nur durch Zufall
- Bei einem Wert von 1 ist die Übereinstimmung völlig zufallsunabhängig
- Was als guter Wert gilt ist vom Problemfeld abhängig...
  - CL-Faustregel: 0.8 ist gut, 0.7 akzeptabel
  - Bessere Werte für POS oder Tokenisierung
  - Schlechtere Werte etwa für Sentiment Analysis (Emotionen beurteilen)

Berechnung der zufällig zu erwartenden Übereinstimmung

$$A_e = \frac{1}{N^2} \sum_{i=0}^{u} \left(\sum_{j=0}^{u} U_{ji} \cdot \sum_{j=0}^{u} U_{ij} \right)$$

- Verschiedene Verfahren:
  - Hier „original“ nach Cohen, nimmt Bias der Annotierenden an
  - Variante von Siegel & Castellan arbeitet ohne Bias
  - Ergebnisse ähnlich

- Die erwartete Übereinstimmung basiert auf Häufigkeit, mit der jedes Label von jedem Annotierendem vergeben wurde. Die Produkte dieser Häufigkeiten werden summiert und das Ergebnis normalisiert.