FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Delta vs. N-Gram-Tracing: Wie robust ist die Autorschaftsattribuierung?

Thomas Proisl · Stefan Evert
Professur für Korpuslinguistik, FAU Erlangen-Nürnberg

Einleitung

Untersuchte Einflussfaktoren auf die Autorschaftsattribuierung

- Abhängigkeit von der Länge des untersuchten Texts und der Größe des Vergleichskorpus.
- Robustheit in Bezug auf Zusammensetzung des Vergleichskorpus (Auswahl der Autoren und Texte).

Deltamaße (Burrows, 2002; Argamon, 2008; Smith/Aldridge 2011)

- Bag-of-Words-Modell der *n* häufigsten Wörter im Korpus.
- Standardisierung der relativen Häufigkeiten zu *z*-Scores.
- Abstandsmaß (z. B. Kosinus), anschließend (hierarchisches) Clustering oder Nearest-Neighbor-Klassifikation.

N-Gram-Tracing (Grieve et al., eingereicht bei DSH)

- Buchstaben- und Wort-N-Gramm-Typen.
- Für jeden Autor: Wie viele N-Gramm-Typen aus dem untersuchten Text kommen im Vergleichskorpus vor; Nearest-Neighbor-Klassifikation.
- Kombination der Ergebnisse für unterschiedliche N-Gramm-Längen über Mehrheitsentscheid möglich.

Samplingexperimente

Datenbasis

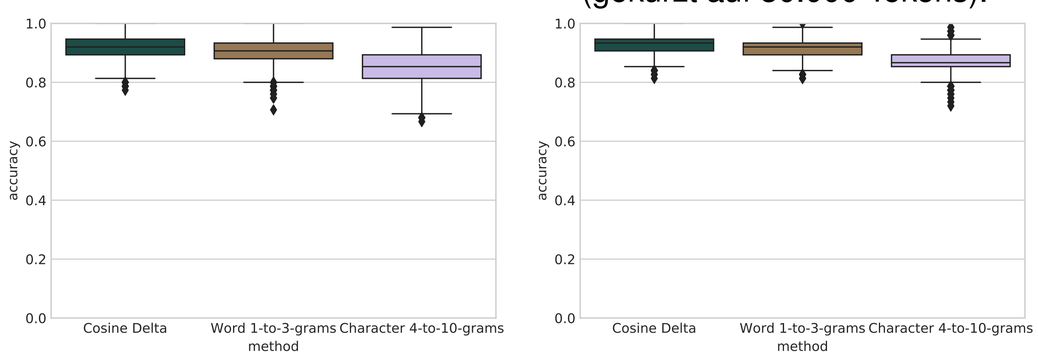
973 deutsche Romane von 131 Autoren (mindestens 3 Texte pro Autor) von Projekt Gutenberg.

Samplingexperiment 1 (S1)

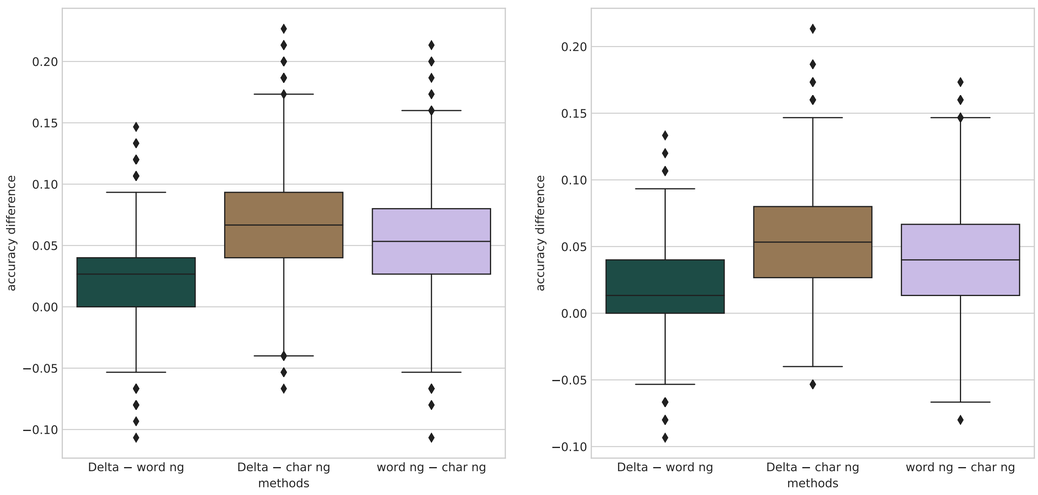
5.000 Samples von 25 × 3 Romanen über alle Autoren (gekürzt auf 30.000 Tokens).

Samplingexperiment 2 (S2)

5.000 Samples von 25 × 3 Romanen über die 25 am stärksten vertretenen Autoren (gekürzt auf 30.000 Tokens).



Boxplots der Accuracies



Paarweise Differenzen

Fazit

- Erhebliche Zufallsschwankungen bei S1 und S2.

Kürzungsexperimente

Datenbasis

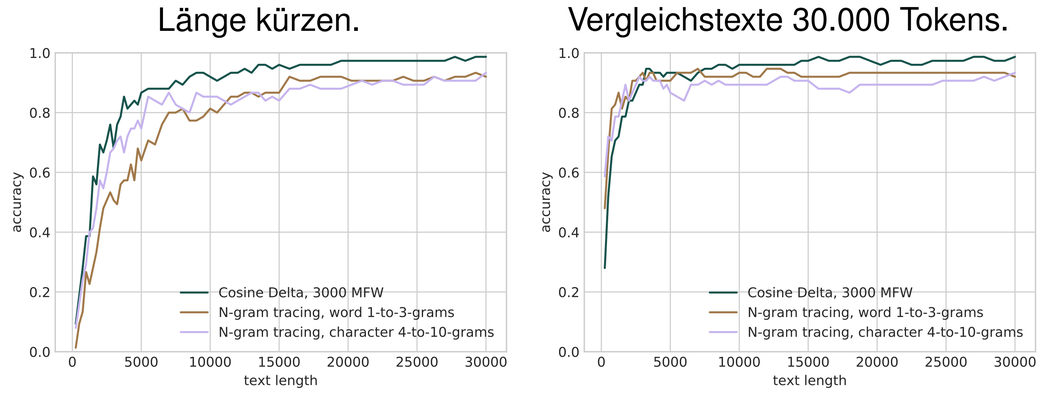
3 Romankorpora für Deutsch, Englisch, Französisch; je 75 Texten von 25 Autoren (3 Texte pro Autor) → https://github.com/cophi-wue/refcor.

Kürzungsexperiment 1 (K1)

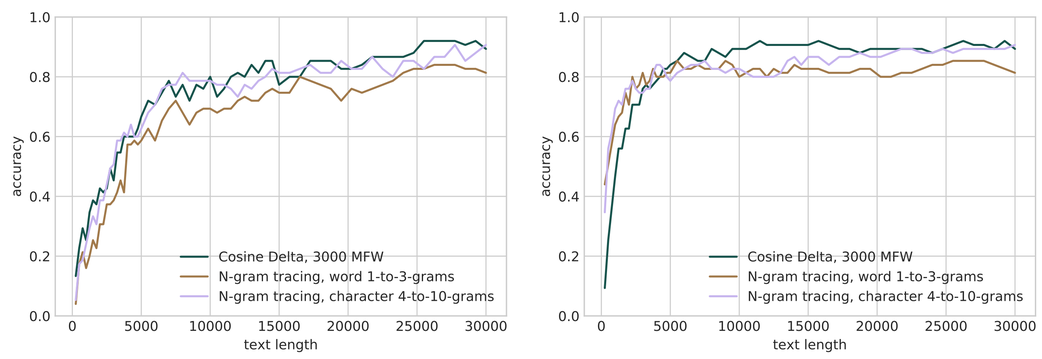
Alle Texte im Korpus auf gleiche Länge kürzen.

Kürzungsexperiment 2 (K2)

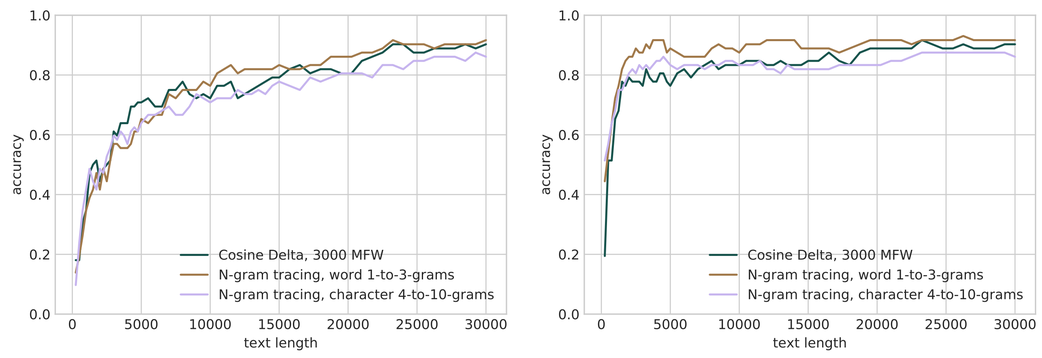
Nur Testtexte kürzen, Vergleichstexte 30.000 Tokens.



Deutsch



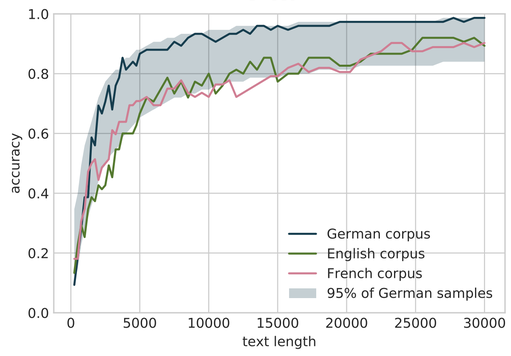
Englisch



Französisch

Kürzen vs. Sampling

Kombination von S1 und K1 für Delta



Fazit

- N-Gram-Tracing nur bei K2 im Vorteil.
- Deutliche Unterschiede zwischen Sprachen.
- Unterschiede allerdings im Bereich von Zufallsschwankungen.

Future Work

Untersuchung des Spracheinflusses

Systematische Kombination von Sampling- und Kürzungsexperimenten auf großen Korpora in mehreren Sprachen.

Proisl et al. (2018): „Delta vs. N-Gram Tracing: Evaluating the Robustness of Authorship Attribution Methods.“ In *Proceedings of LREC 2018*.