Delta vs. N-Gram Tracing: Evaluating the Robustness of Authorship Attribution Methods

Thomas Proisl[*] Stefan Evert[*] Fotis Jannidis[†]
Christof Schöch[‡] Leonard Konle[†] Steffen Pielström[†]

[*]Friedrich-Alexander-Universität Erlangen-Nürnberg
[†]Julius-Maximilians-Universität Würzburg
[‡]Universität Trier

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Authorship attribution

- Goal: Identify true author of text of unknown or disputed authorship (Juola 2006; Koppel et al. 2009; Stamatatos 2009)
  - based on quantitatively measured linguistic evidence
- Assumption: Authors' idiosyncratic habits of language use lead to stylistic similarities between their texts
- Typical approach: Similarity between feature vectors
  - relative frequencies of function words, vocabulary richness, syntactic complexity, ...
- Important for real-world applications: Reliability and robustness of methods
  - length of disputed text
  - size of comparison corpus
  - composition of comparison corpus

Authorship attribution

- Goal: Identify true author of text of unknown or disputed authorship (Juola 2006; Koppel et al. 2009; Stamatatos 2009)
  - based on quantitatively measured linguistic evidence
- Assumption: Authors' idiosyncratic habits of language use lead to stylistic similarities between their texts
- Typical approach: Similarity between feature vectors
  - relative frequencies of function words, vocabulary richness, syntactic complexity, ...
- Important for real-world applications: Reliability and robustness of methods
  - length of disputed text
  - size of comparison corpus
  - composition of comparison corpus

Delta measures

- Delta measures (Burrows 2002; Argamon 2008) are popular in literary stylistics
  - Treat texts as bags of words
  - Use n most frequent words (nMFW) from corpus
  - Standardize relative frequencies to z-scores
  - Optional: normalize feature vectors
  - Quantify similarity with some metric, e.g. Manhattan distance
  - Optional: hierarchical clustering of distance matrix and dendrogram
  - Assign disputed text to author of most similar text or to most frequent author in cluster
- Cosine Delta usually superior to other variants of Delta (Jannidis et al. 2015)
  - also robust to choice of nMFW
- We use Cosine Delta with 3000 MFW

N-gram tracing

- N-gram tracing: Novel method from forensic linguistics (Grieve et al. submitted)
  - Designed for short disputed texts
  - Extract all word or character n-gram types of certain length(s)
  - Determine percentage of overlap with each candidate author in corpus
  - Frequency is ignored!
  - Combination of different n-gram lengths via majority voting
- We use majority vote of word 1-to-3-grams and of character 4-to-10-grams (following Grieve et al. submitted)

Shortening experiments

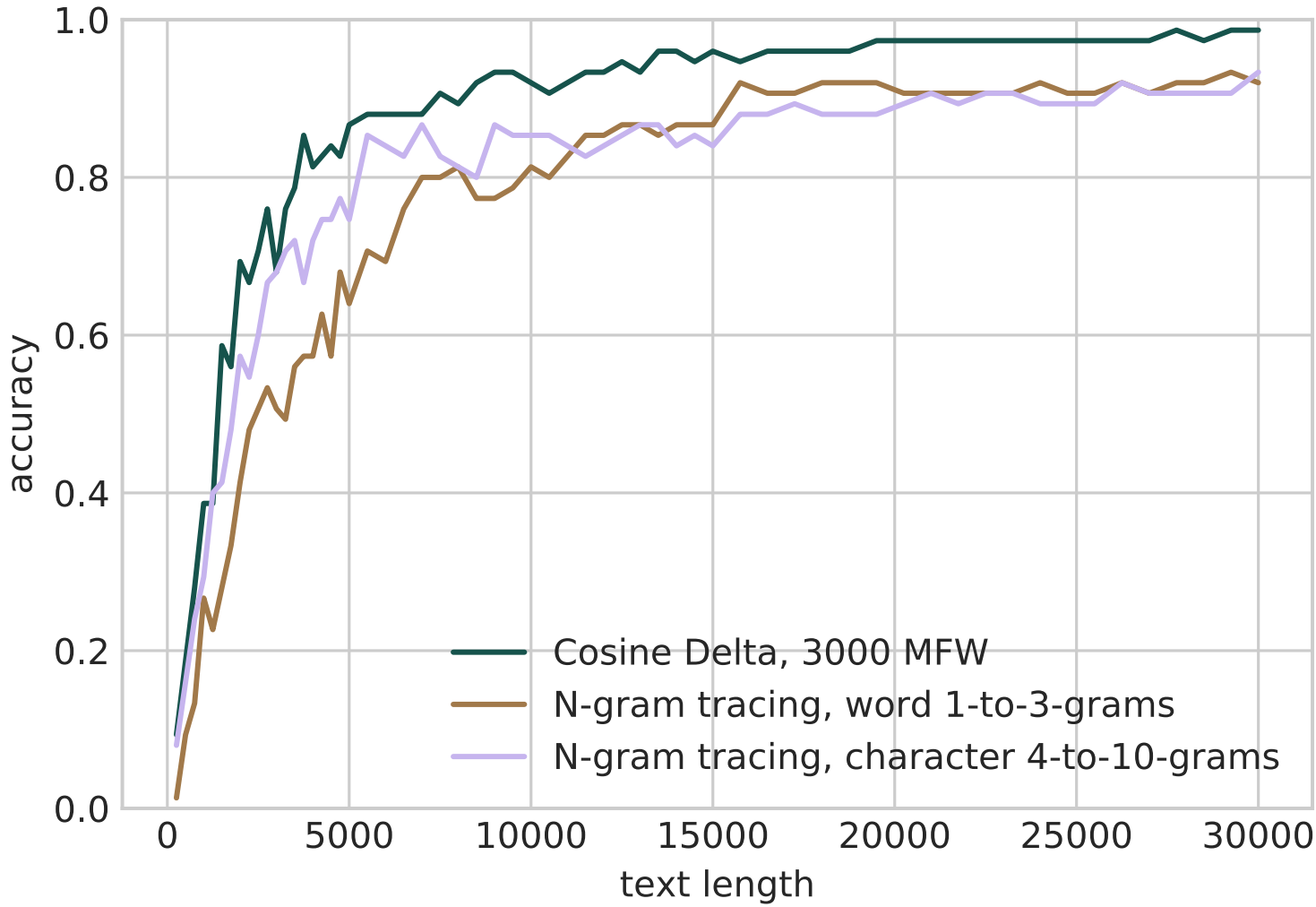
- Three corpora of German, English and French novels[1] (Jannidis et al. 2015; Evert et al. 2017)
  - 75 novels per corpus (25 authors with 3 novels each)
- Stratified three-fold cross-validation
  - 25 test texts per fold (one per author)

Experiment 1a: Shorten **all texts** (test and comparison) to same number of tokens (250–30,000 tokens)

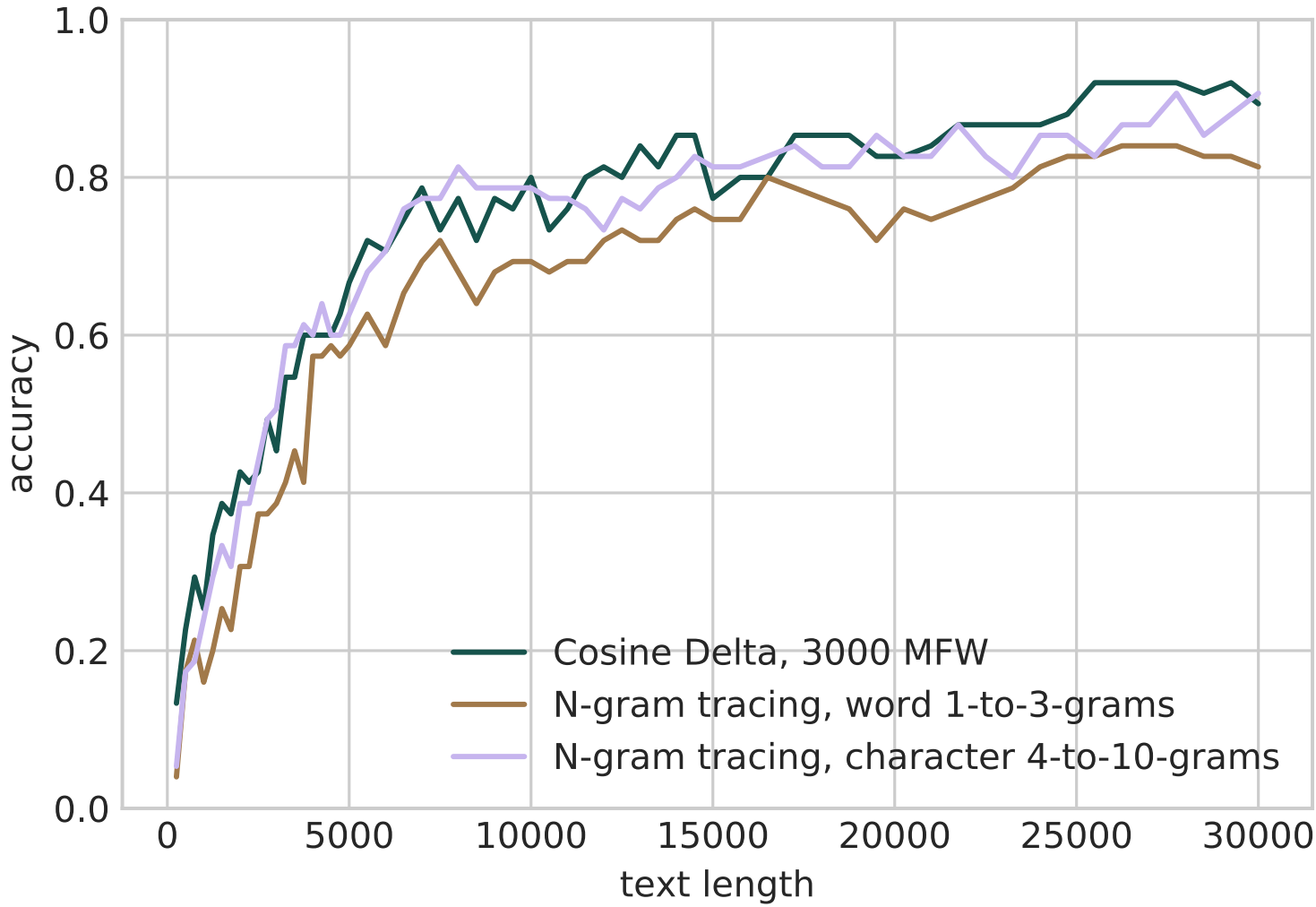
Experiment 1b: Shorten **only test texts** (250–30,000 tokens), length of comparison texts capped at 30,000 tokens

[1] `https://github.com/cophi-wue/refcor`

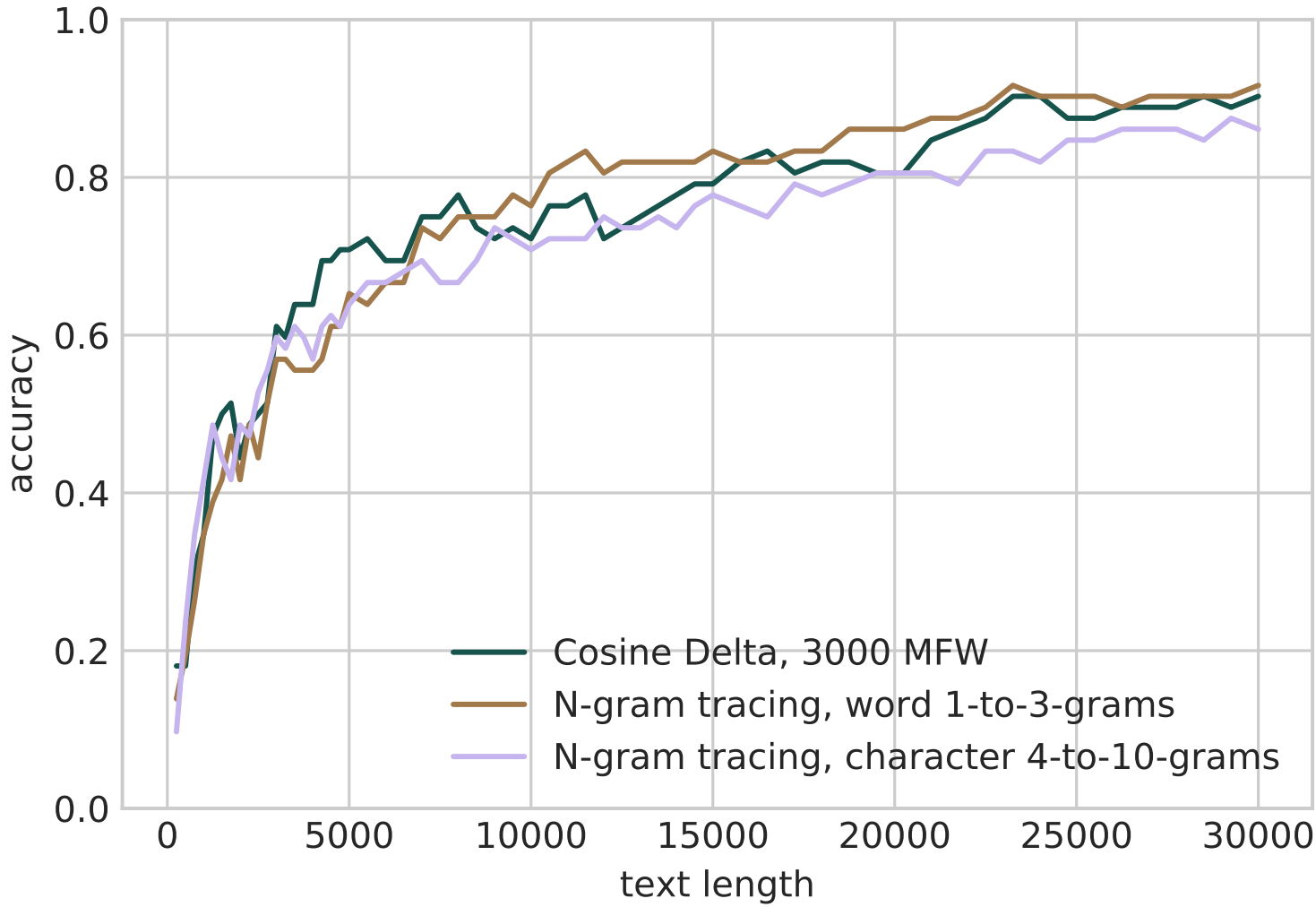
Experiment 1a (shorten all texts): German



Experiment 1a (shorten all texts): English



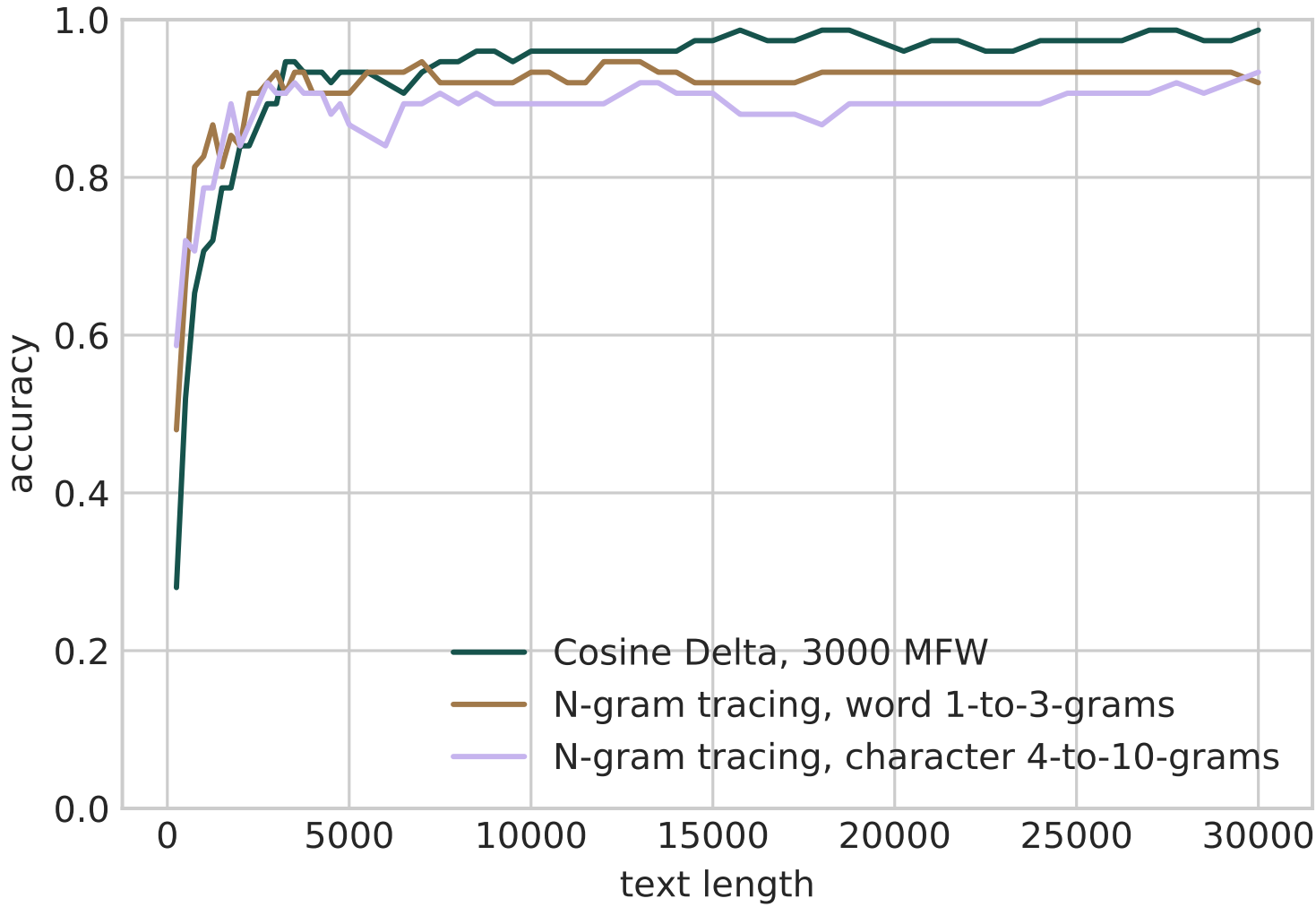
Experiment 1a (shorten all texts): French



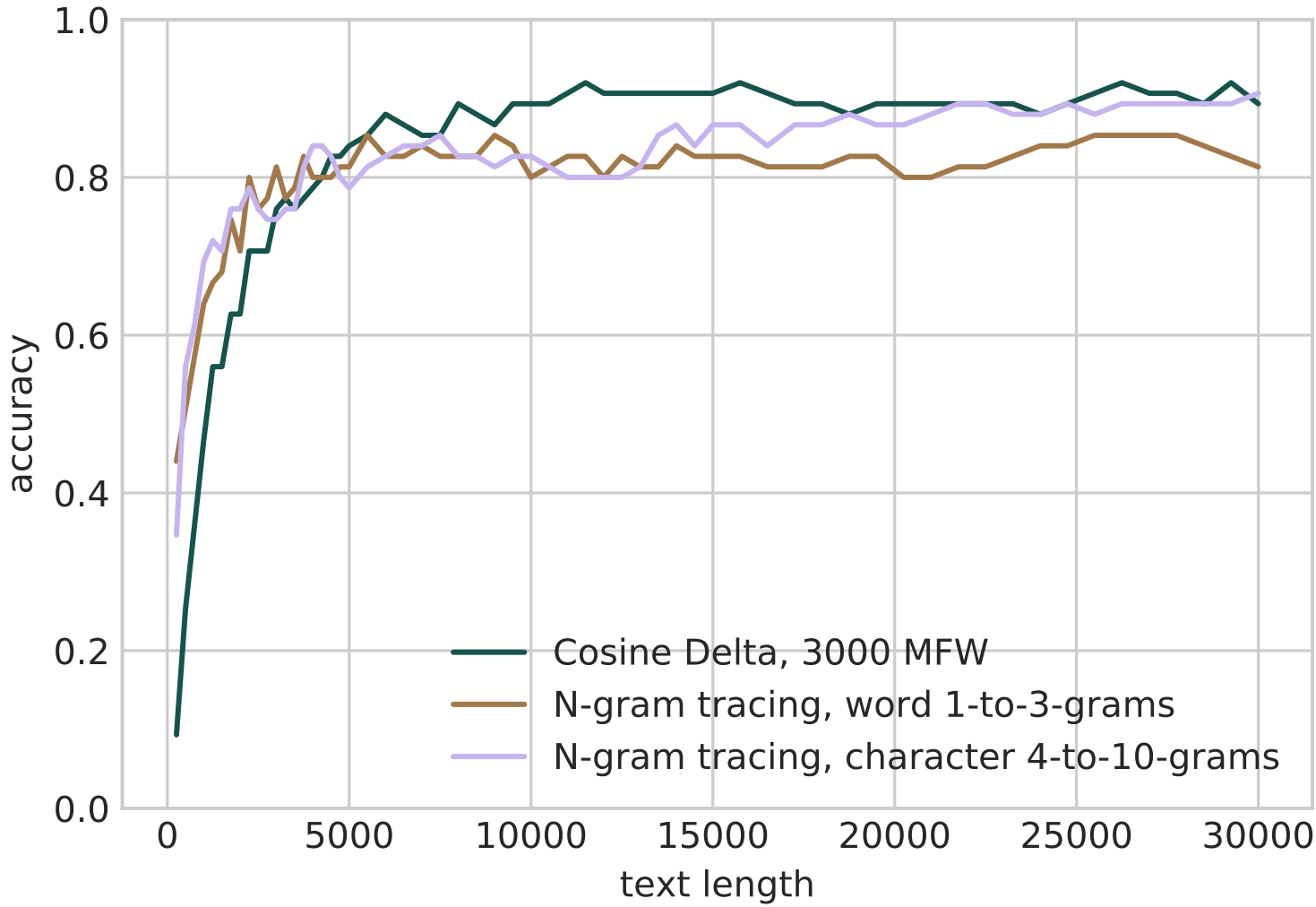
Experiment 1a: Summary

- Accuracy of all three methods improves with larger text sizes
- All methods perform rather poorly for very short texts
  - Extreme case: attribute 250 word fragment to one of 25 possible authors with only 500 words comparison text per author
- Delta usually as good as or better than N-Gram Tracing
- Not clear if word or character n-grams perform better for N-Gram Tracing
- Performance on English and French corpora notably worse than on German corpus

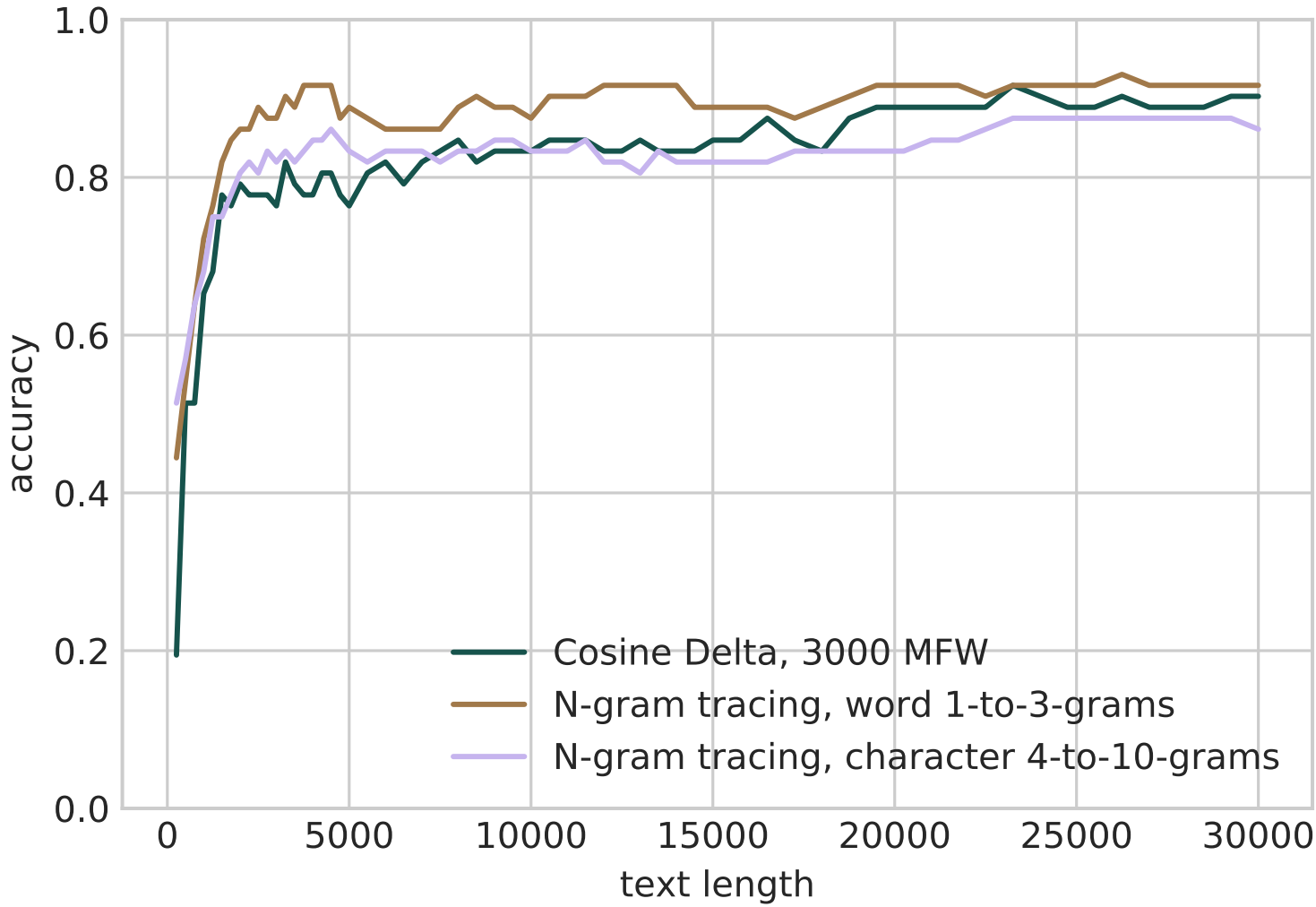
Experiment 1b (shorten test texts): German



Experiment 1b (shorten test texts): English



Experiment 1b (shorten test texts): French



Experiment 1b: Summary

- Results for shorter text lengths much better than in experiment 1a
  - Much larger comparison corpus
- N-Gram Tracing outperforms Delta on very short texts by large margin
  - $\approx$ 50% accuracy on 250-word fragments
- Not clear if word or character n-grams perform better for N-Gram Tracing
- 1,000–5,000 words sufficient for 80% accuracy
- Performance on English and French corpora notably worse than on German corpus

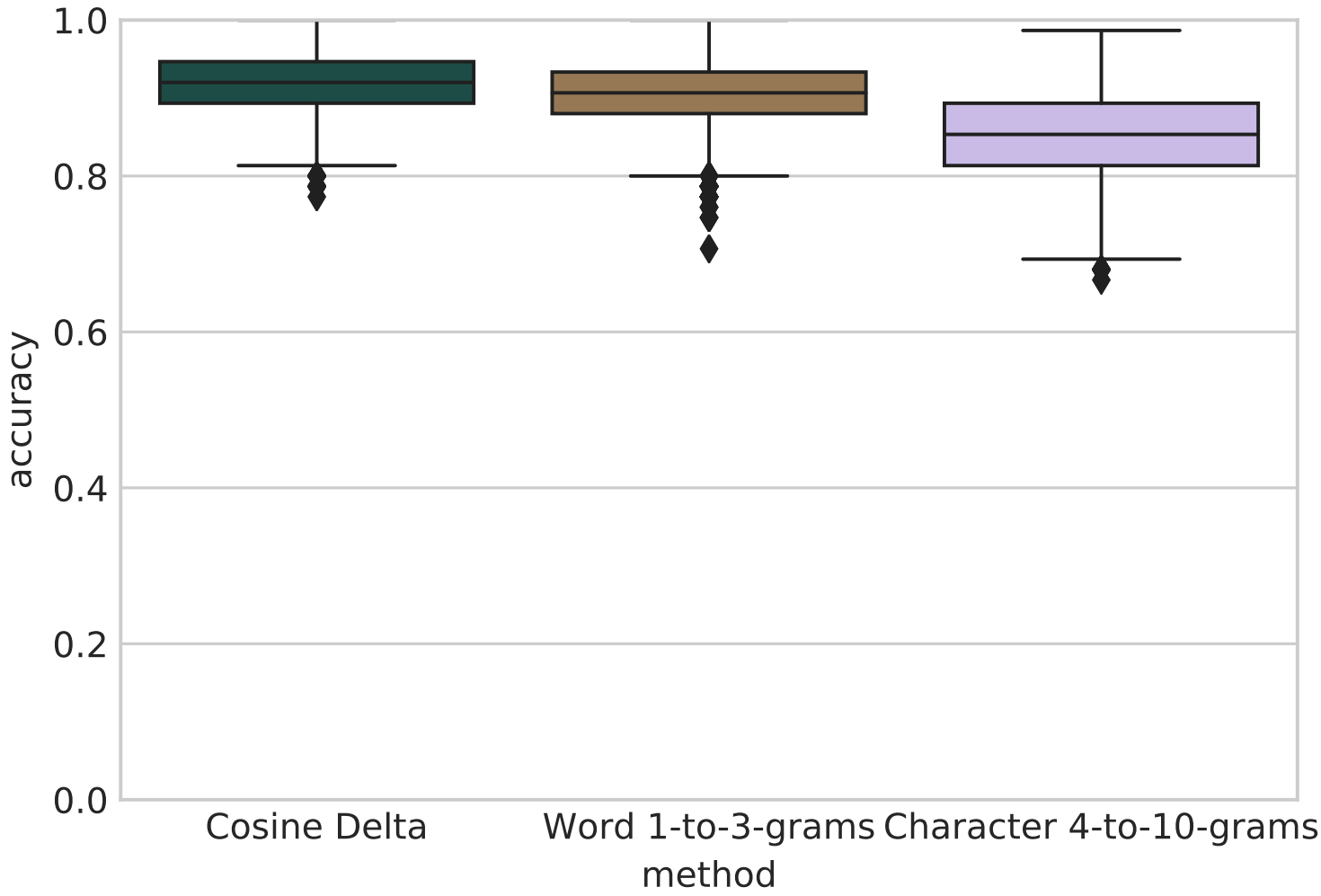
Sampling experiments

- 973 German novels by 131 authors
  - At least three novels from each author
  - All authors native speakers
  - No translations
  - Novels written 1789–1914
- Draw samples of 75 novels (25 authors with 3 novels each)
- For each sample: Stratified three-fold cross-validation
  - 25 test texts per fold (one per author)

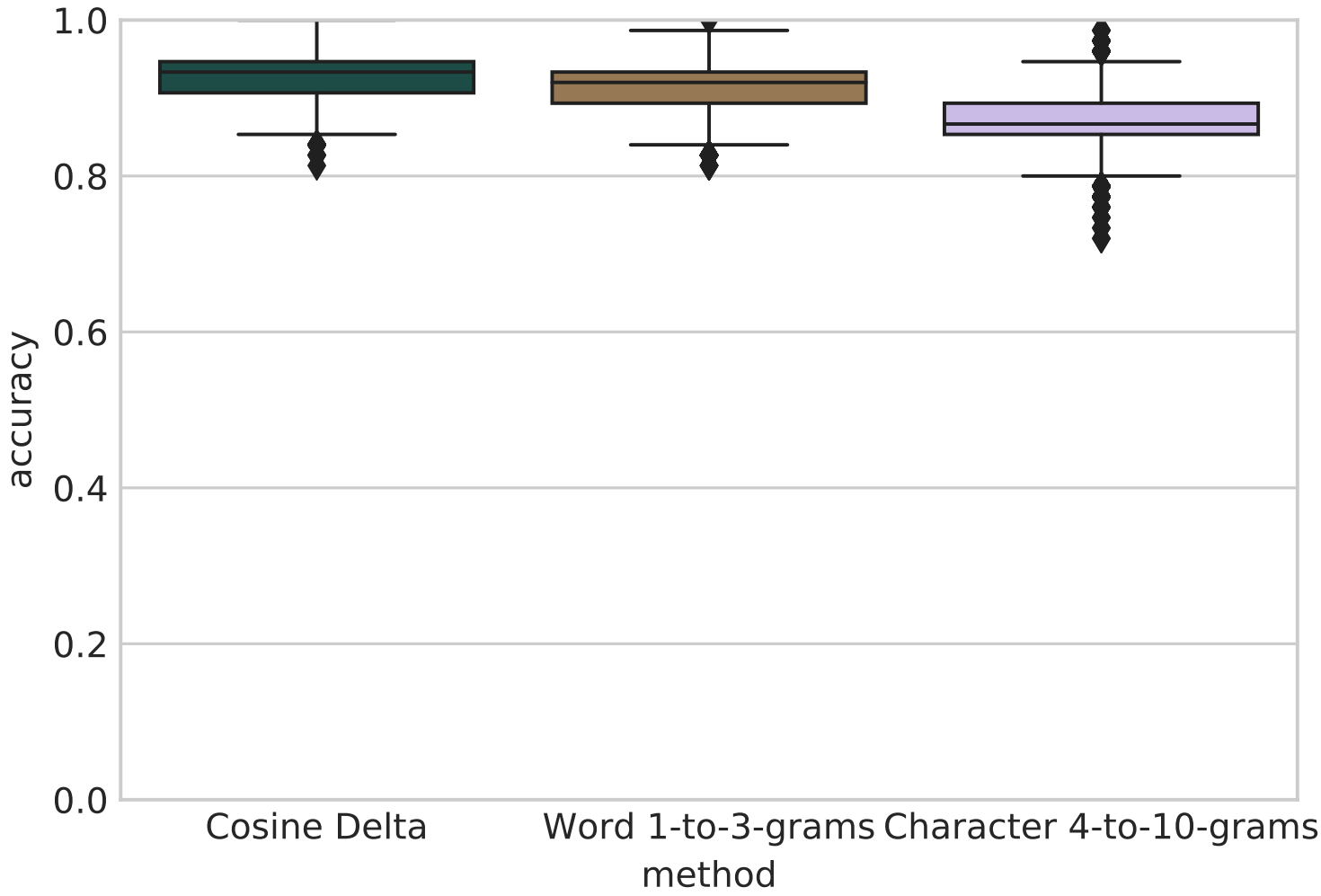
Experiment 2a: 5,000 random samples, each text shortened to 30,000 tokens

Experiment 2b: 5,000 random samples **from 25 authors with most texts**, each text shortened to 30,000 tokens

Experiment 2a: Samples from all authors



Experiment 2b: Samples from fixed authors



Sampling experiments: Summary

- Central 50% of samples lie in fairly narrow range around median
  - ± 5 points in experiment 2a, even less in 2b
- Considerably larger range for remaining 50%
  - Accuracies between 70% and 100% in experiment 2a
  - Accuracies between 80% and 100% in experiment 2b
- Delta usually a little bit better than N-Gram Tracing
- Accuracies can easily fluctuate by 15 points even with fixed set of comparison authors

Conclusion & future work

- Conclusion
  - Short texts and little material in comparison corpus: Both methods unreliable
  - Short texts and much material in comparison corpus: N-Gram Tracing better than Delta
    - ★ N-Gram Tracing requires at least 1,000–3,000 words and large enough comparison corpus for 80% accuracy
  - Longer texts ($>$ 5,000 words) and much material in comparison corpus: Delta better than N-Gram Tracing
  - Composition of comparison corpus has large and unpredictable impact on accuracy of authorship attribution
- Future work
  - Run shortening experiments on large number of samples drawn from large collections of texts in many languages

References

Shlomo Argamon. Interpreting Burrows's Delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2):131–147, 2008.

John Burrows. 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 2002.

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32(suppl_2):ii4–ii16, 2017.

Jack Grieve, Emily Carmody, Isobelle Clarke, Hannah Gideon, Annina Heini, Andrea Nini, and Emily Waibel. Attributing the Bixby Letter using n-gram tracing. *Digital Scholarship in the Humanities*, submitted. Submitted on May 26, 2017.

Fotis Jannidis, Stefen Pielström, Christof Schöch, and Thorsten Vitt. Improving Burrows' Delta – an empirical evaluation of text distance measures. In *Digital Humanities 2015: Conference Abstracts*, 2015.

Patrick Juola. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334, 2006.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26, 2009.

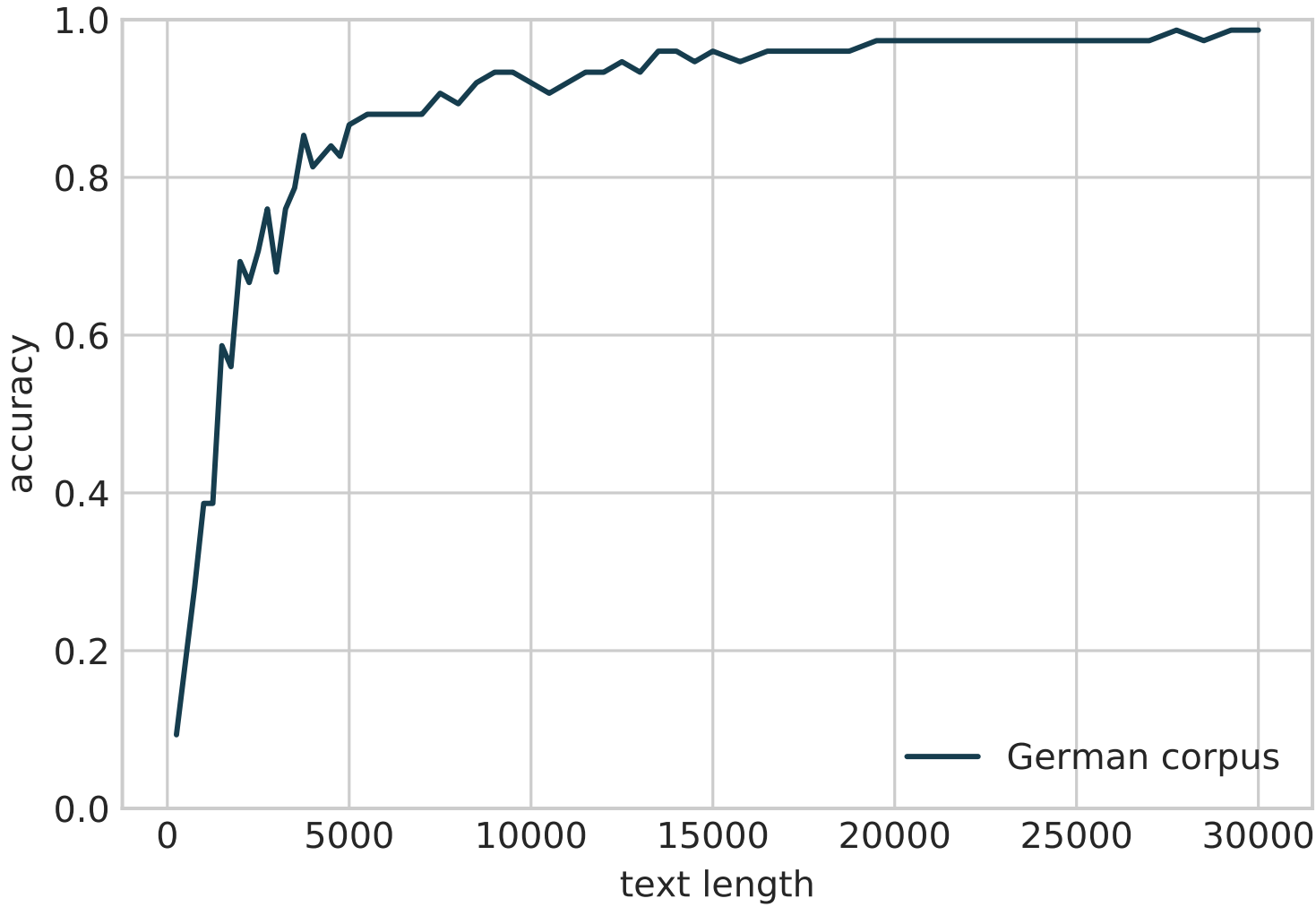
Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

Discussion

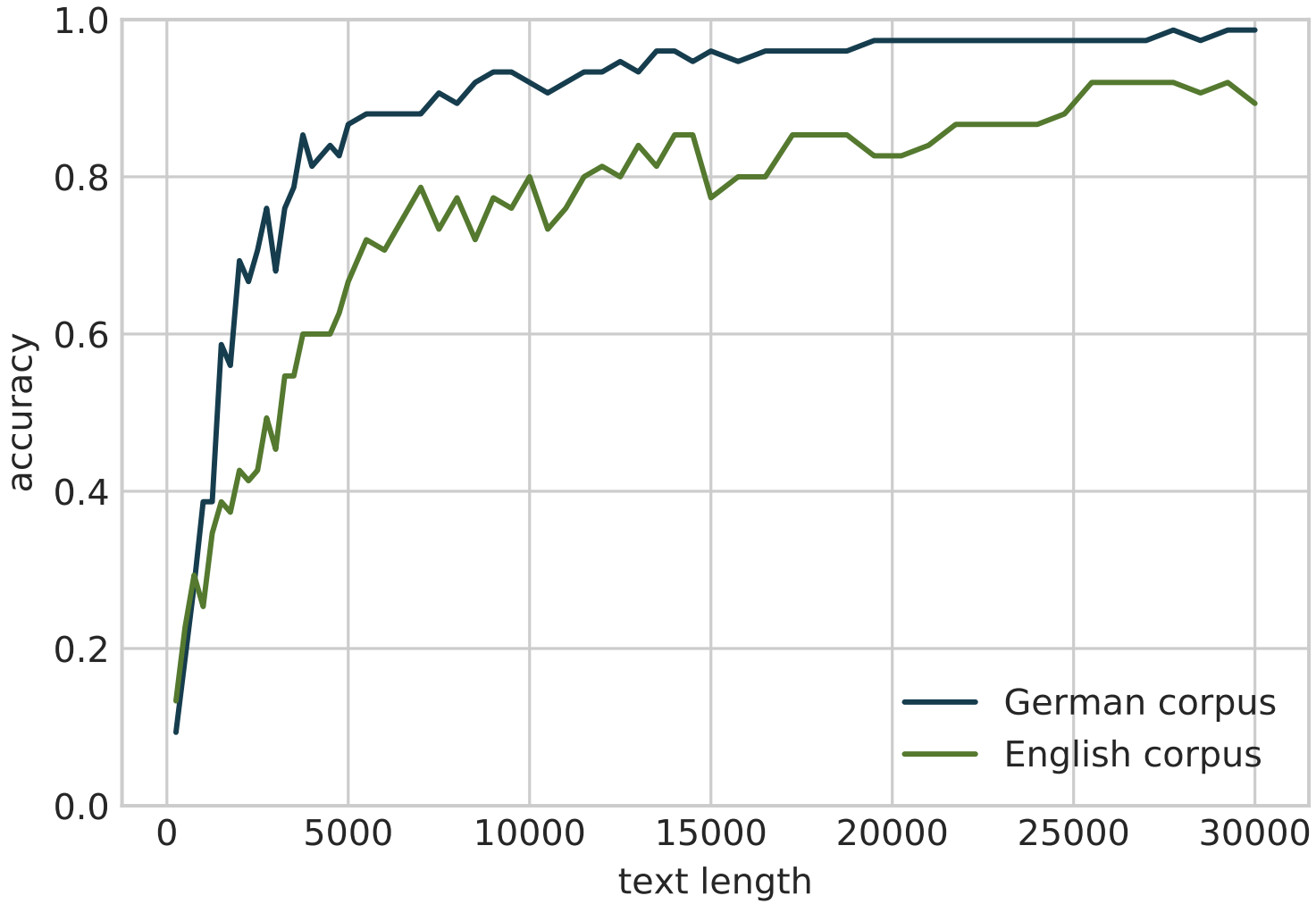
Thank you!

Time for questions!

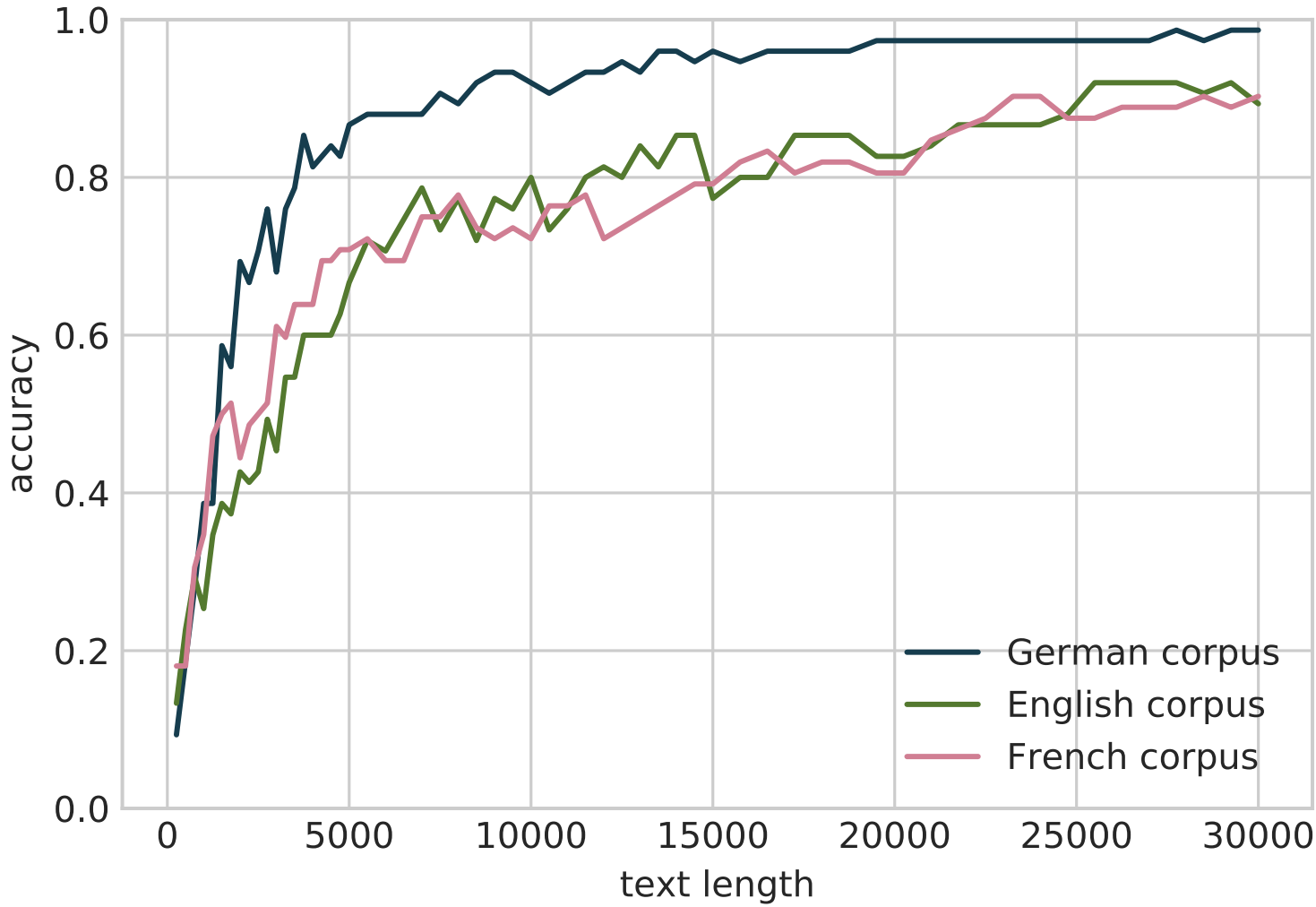
Cosine Delta in 1a and 2a



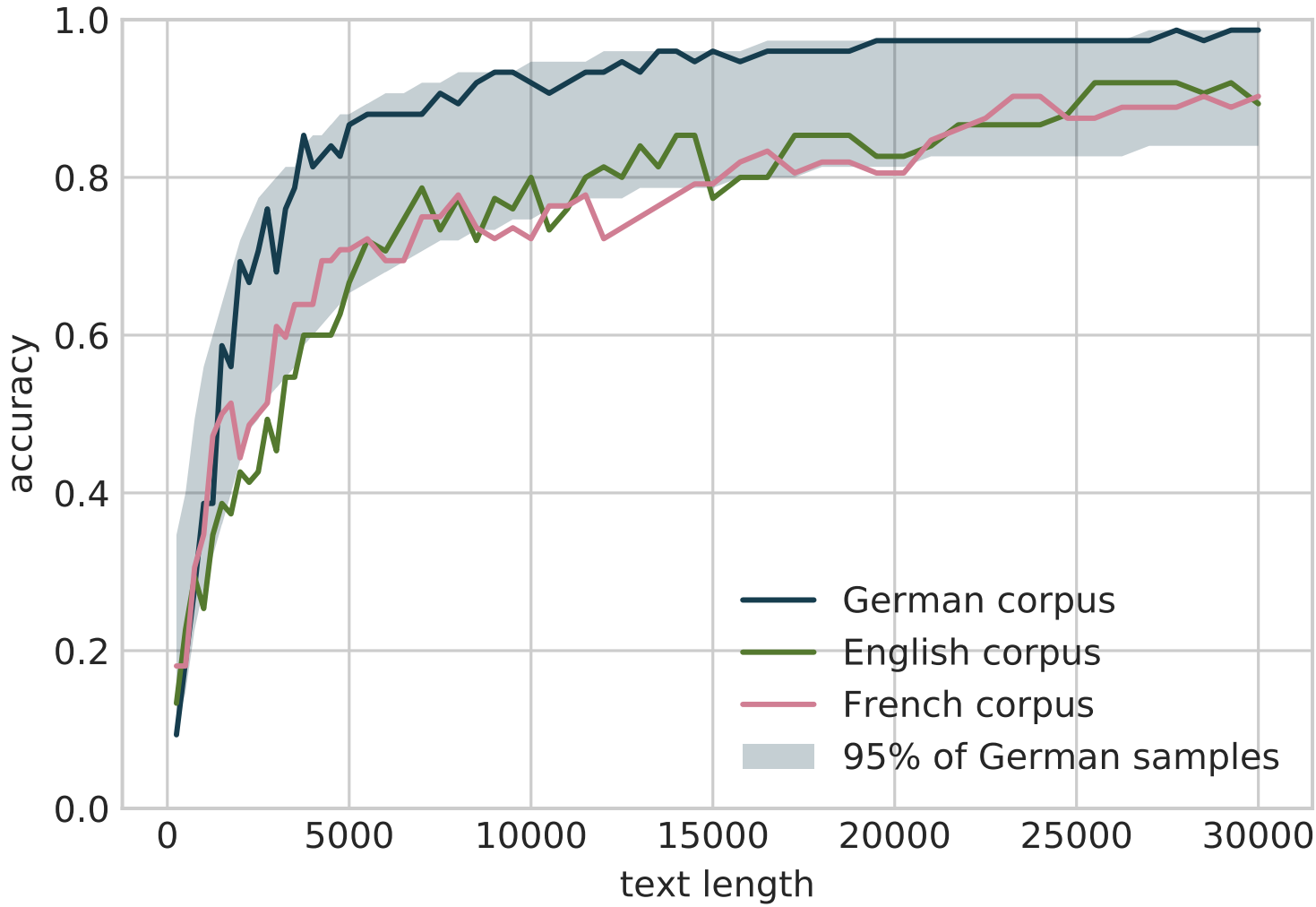
Cosine Delta in 1a and 2a



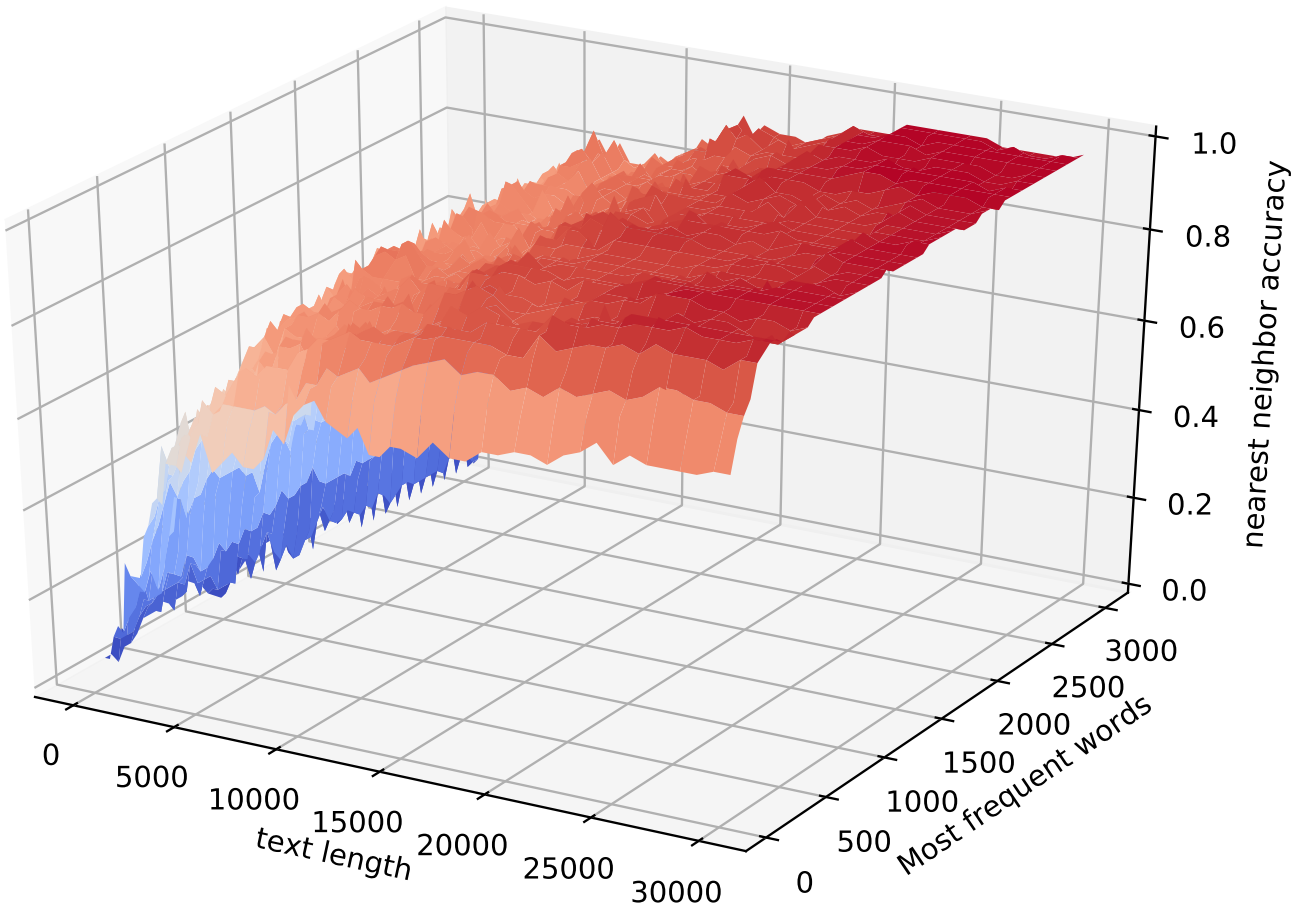
Cosine Delta in 1a and 2a



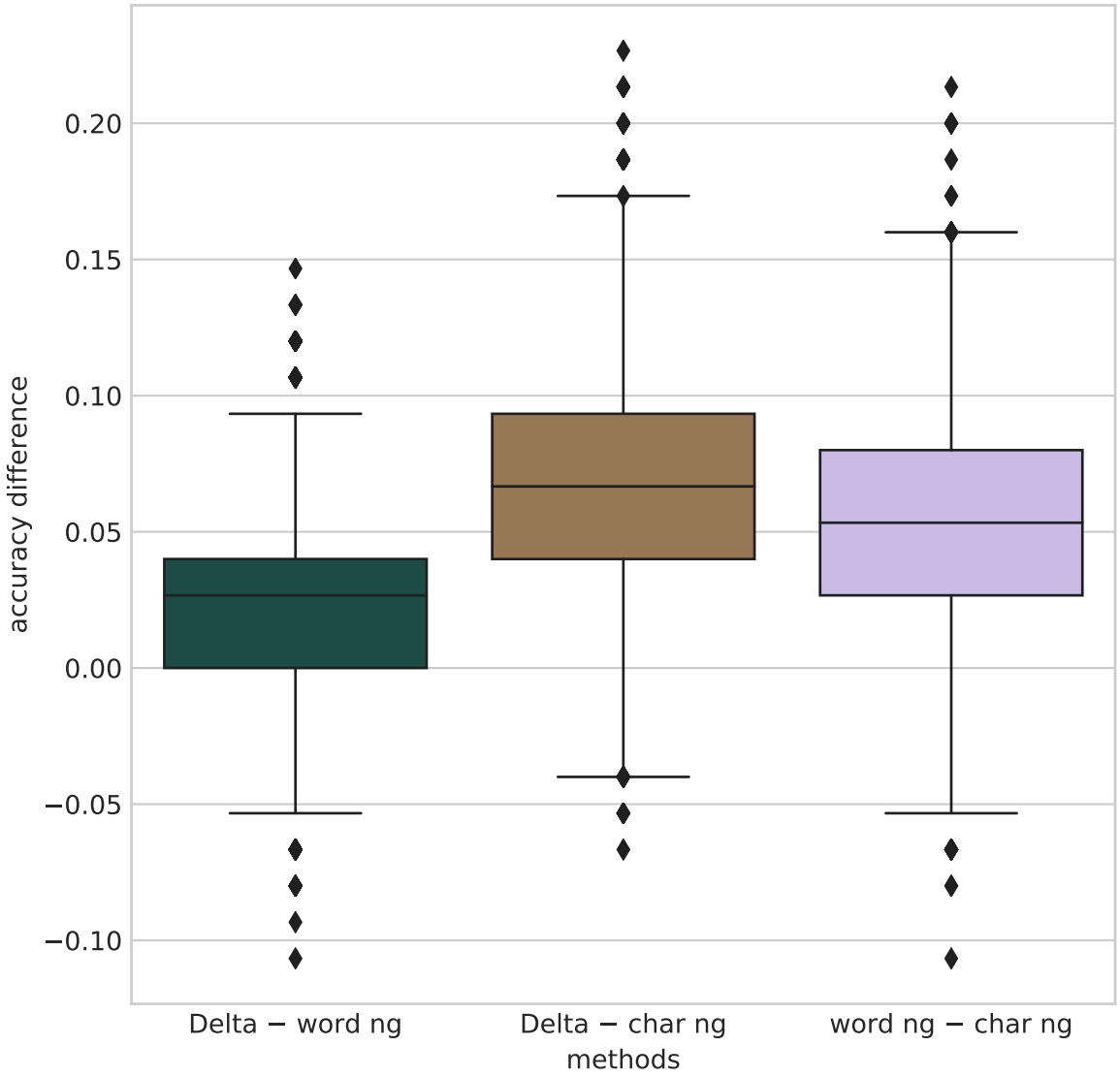
Cosine Delta in 1a and 2a



Text length vs. nMFW for Cosine Delta (German)



Experiment 2a: Pairwise accuracy diffs between methods



Experiment 2b: Pairwise accuracy diffs between methods