# Intrinsic Plagiarism Analysis with Meta Learning

Benno Stein and Sven Meyer zu Eissen

Bauhaus University Weimar
Web-Technology and Information Systems

# On Plagiarism Analysis

*"Plagiarism refers to the use of another's ideas, information, language, or writing, when done without proper acknowledgment of the original source."* [Wikipedia]

Fact: About 40% of the students admit to plagiarize from Internet documents (study on 50,000 students).

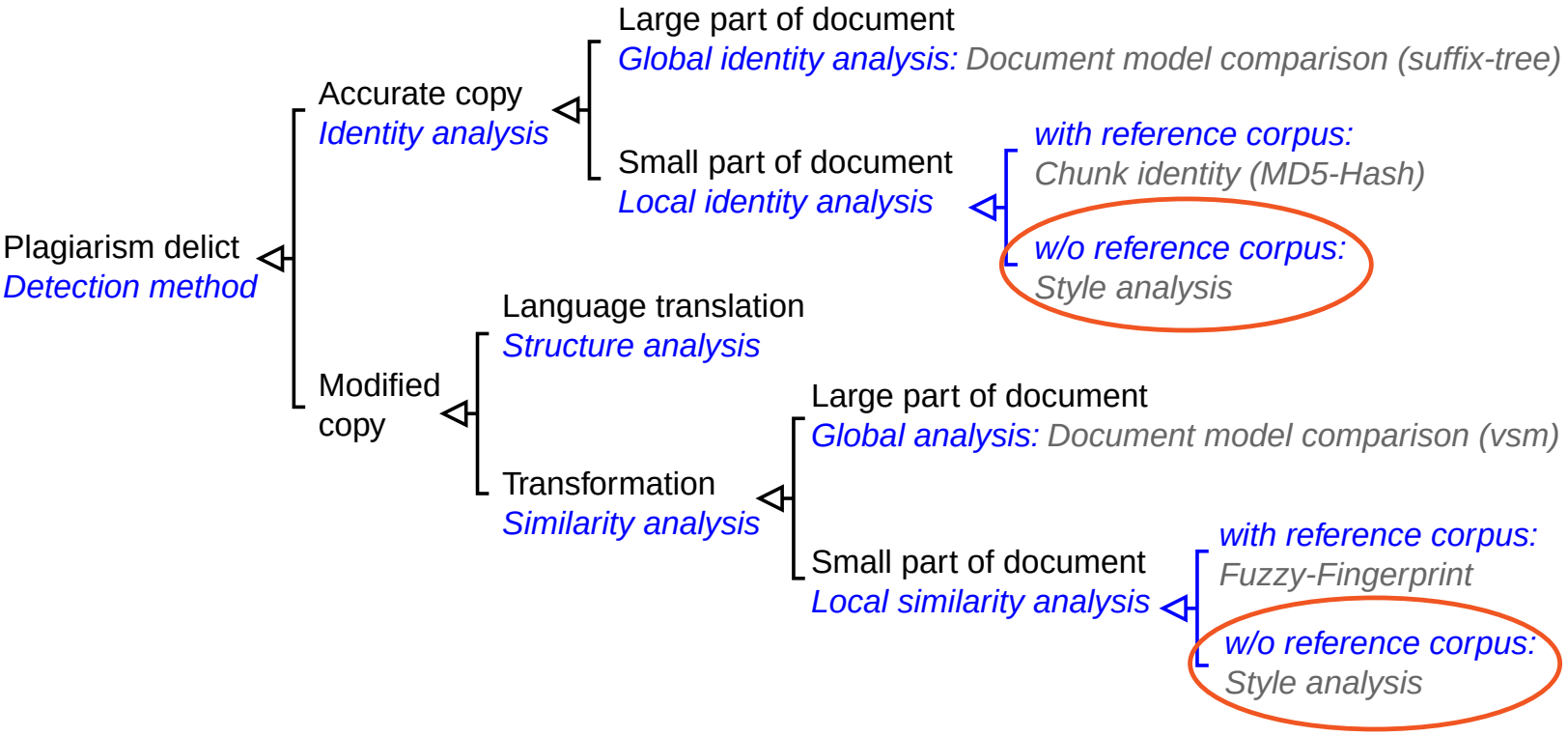
[McCabe 2005]

Plagiarism analysis:

*Given.* A suspicious document.
*Task.* Find copied parts
(and, if possible, provide references to original sources).

# Plagiarism Forms

Plagiarism may happen in manyfold variants:



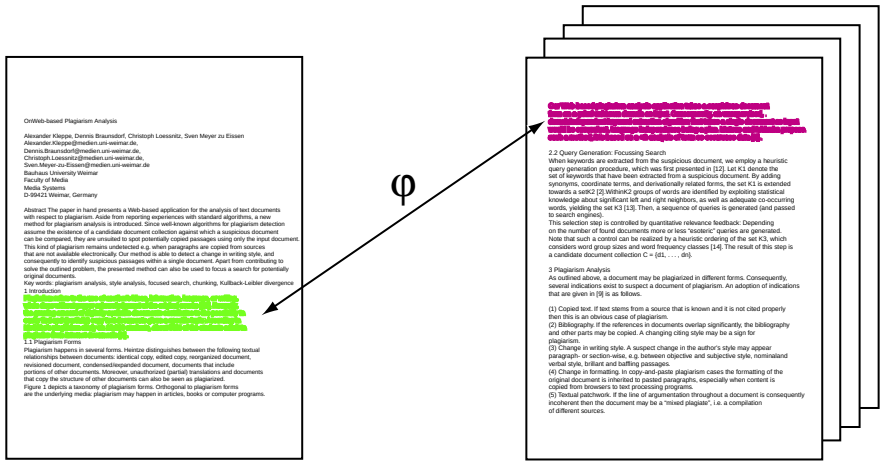
# Current Research on Plagiarism Analysis

Current research is mainly corpus-oriented.

e.g. [Stein et al. 2004-2006, Monostori et al. 2001-2004].

*Given.* A suspicious document $d$
and a corpus of original documents.

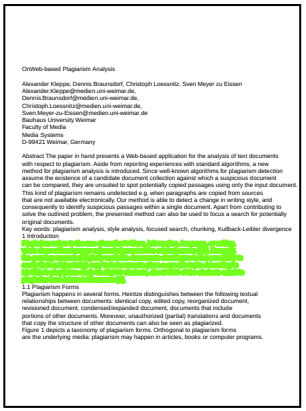
*Task.* Find potentially copied parts from $d$ in the corpus,
and provide references to original sources.



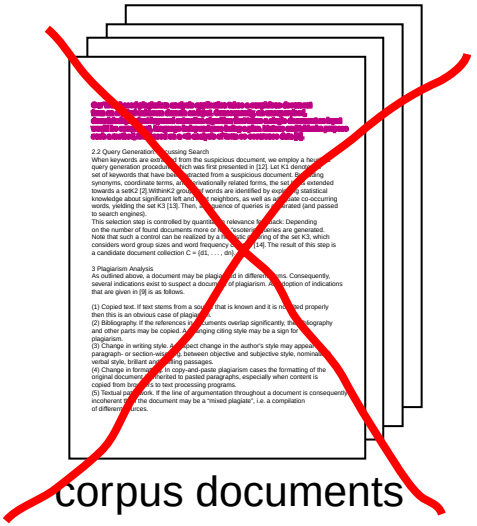

suspicious document corpus documents

# Intrinsic Plagiarism Analysis

What can be done if sources are *not available* in digital form?



suspicious document



corpus documents

Research focus:

*Given.* A suspicious document
~~and a corpus of original documents.~~

*Task.* Find potentially copied parts.

# Intrinsic Plagiarism Analysis

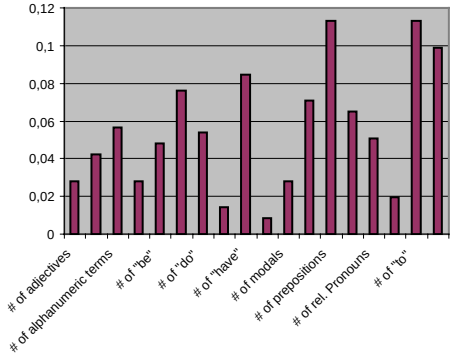
*Goal.* Model the human capabilities in detecting “somewhat different” sections.

*Method.* Quantify changes in writing style.

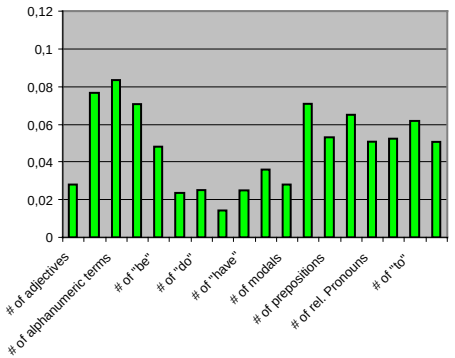
[Meyer zu Eissen and Stein 2006]

*Operationalization.*

style markers for the entire document (global)




style markers for a single paragraph (local)



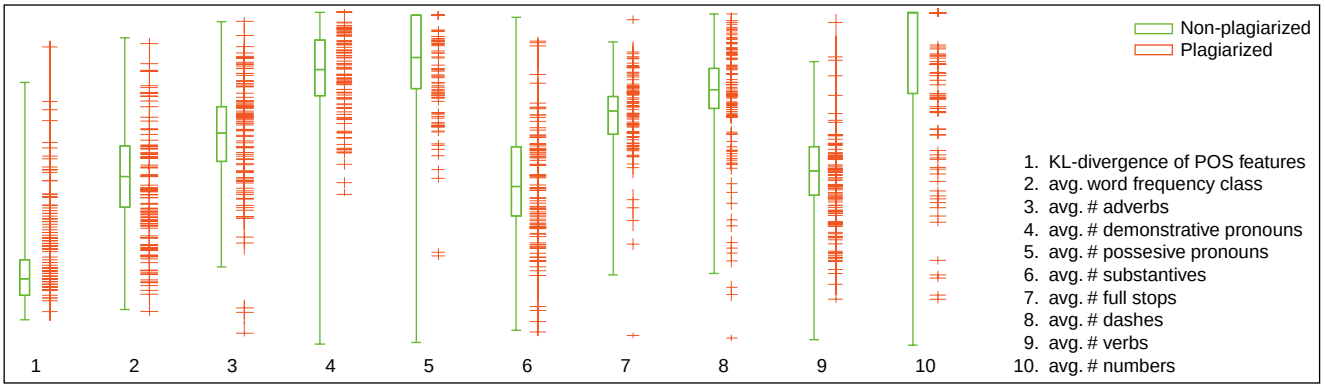
# Intrinsic Plagiarism Analysis

Algorithm for intrinsic analysis:

1. Let $\sigma_1, \ldots, \sigma_m$ denote style markers.
2. For each section $s \subseteq d$:
3. compute style model $\mathbf{s} = \begin{pmatrix} \sigma_1(s) \\ \vdots \\ \sigma_m(s) \end{pmatrix} \in \mathbf{R}^m$
4. compute relative deviations $\mathbf{s}_\Delta = \begin{pmatrix} \frac{\sigma_1(s)-\sigma_1(d)}{\sigma_1(d)} \\ \vdots \\ \frac{\sigma_m(s)-\sigma_m(d)}{\sigma_m(d)} \end{pmatrix} \in \mathbf{R}^m$
5. use instances of $\mathbf{s}_\Delta$ for an outlier analysis.

# Intrinsic Plagiarism Analysis

Distribution of 10 style markers:
16,000 non-plagiarized sections (green)
1,500 plagiarized sections (red)



# Intrinsic Plagiarism Analysis

Success using a discriminant analysis on the $\mathbf{s}_\Delta$ on a hand-made corpus:

About 70% in precision, 80% in recall.

Improvement if the fraction $\theta$ of plagiarized passages is known.

Challenge:
Find style markers that are reliable for short texts.

| style marker $\sigma_i$ | unit of measure | reliability level |
|---|---|---|
| avg. paragraph length | paragraph | document |
| Flesch index | document | document |
| avg. sentence length | sentence | paragraph? |
| avg. word length | word | paragraph |
| avg. word frequency class | word | paragraph |

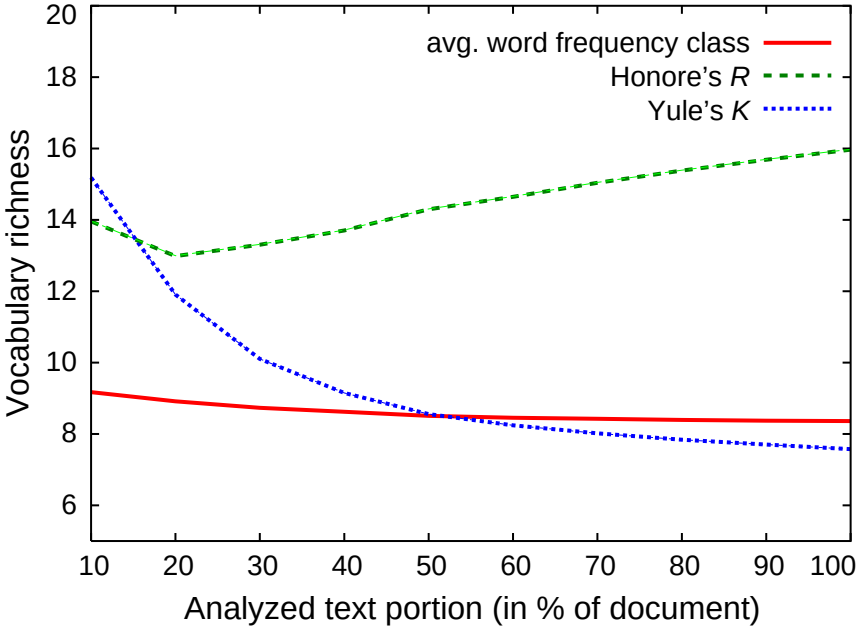
# Intrinsic Plagiarism Analysis

Success using a discriminant analysis on the $\mathbf{s}_\Delta$ on a hand-made corpus:

About 70% in precision, 80% in recall.

Improvement if the fraction $\theta$ of plagiarized passages is known.

Challenge:
Find style markers that are reliable for short texts.



# Intrinsic Plagiarism Analysis

An intrinsic analysis (as shown)

- is very useful for preselecting suspicious sections (for human inspection, for Web search)
- is ambitious from the modeling perspective.

An intrinsic analysis can be used to answer the following question (with high probability):

*Is a given document $d$ written by a single author?*

# Meta Learning

Meta Learning: Method for authorship verification.

[Koppel and Schler 2004]

Authorship verification:

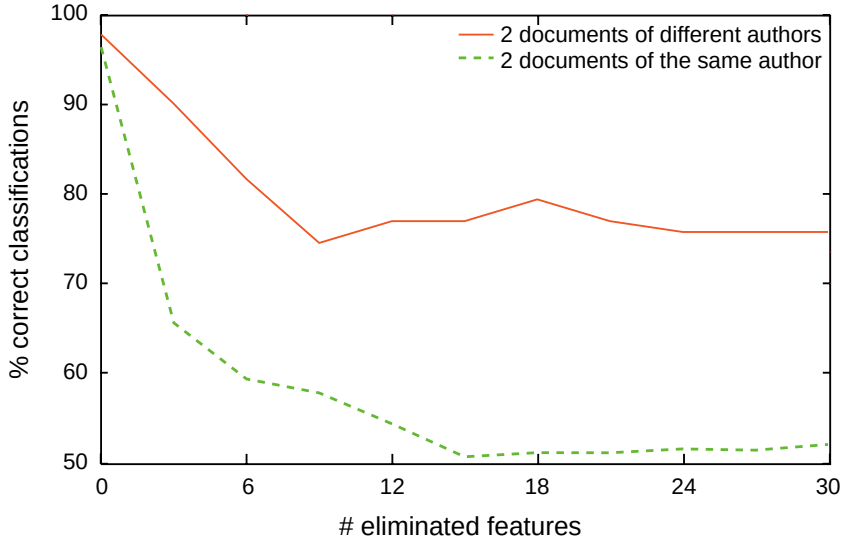
*Given.* $d_1, d_2$.
*Task.* Decide whether $d_1, d_2$ are written by the same author.

Procedure:

1. *Chunking.* Decompose $d_1, d_2$ into sets of chunks $D_1, D_2$.
2. *Model fitting.* Build a VSM for each chunk in $D_1, D_2$.
   The VSM includes only the 250 most frequent words.
   Learn a function that discriminates between $D_1$ and $D_2$.
3. *Impairing.* Drop the 3 most discriminating features from the VSMs.
4. Goto Step 2 until feature space is sufficiently reduced.
5. *Meta Learning.* Analyze the degradation in the quality of model fitting.

# Meta Learning

Expected outcome:




Rationale:

- A large fraction of the 250 words are function/stop words.
- Only some of the words are related to topic.
- Only some words do the discrimination job (e.g. these topic words).
- Different authors can be distinguished by their use of function words.

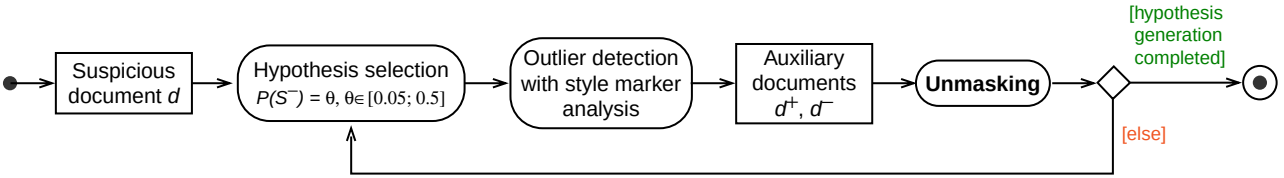
# Meta Learning

Problem: Länge der Texte unklar.

Meta learning cannot be applied directly

(there is a combinatorial problem)

The proposed process:



# Case study

Setting:

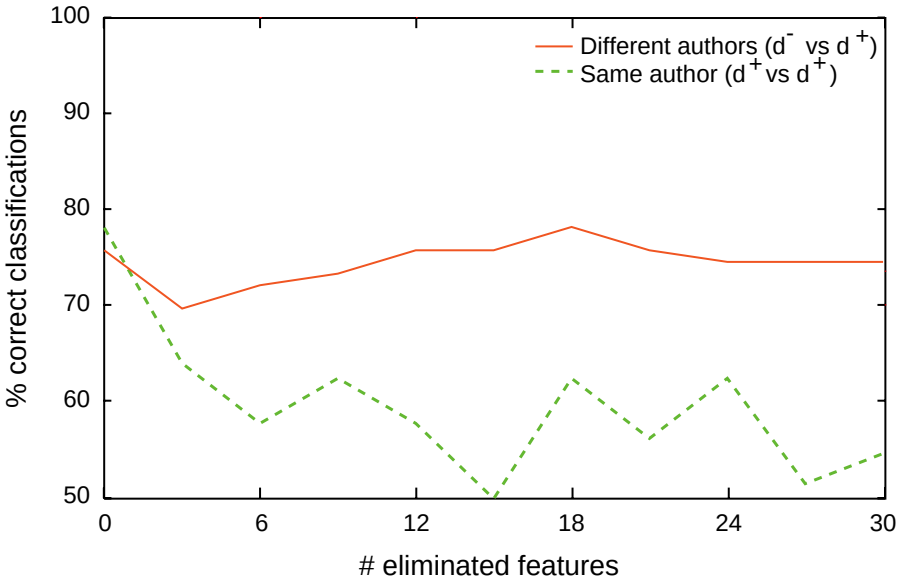
- Given: A German habilitation thesis from the 1980s.
- The habilitation was suspected to be plagiarized.
- Related books are not available in electronic form.

Procedure:

- The thesis was scanned.
- It was converted to plain text using OCR technology.
- It was decomposed into 138 natural sections.
- 13 suspicious sections were identified as $d^-$ (using intrinsic plagiarism analysis).
- (Three of them are confirmed to be plagiarized)
- Meta learning was applied: $d^-$ versus randomly drawn sections, $d^+$, from the remainder.

# Case study

Results of the meta learning approach:




➜ Clear indication that $d^-$ contains plagiarized passages.

# Thank You!

Questions?