

---

## Projects involved

- Archive of General Reference Corpora of Contemporary Written German
- Methods of Corpus Analysis and Corpus Mining


## Researchers involved

- Dipl.-Ing. Cyril Belica
- Dr. Marc Kupietz
- Dipl.-Inf. Rainer Perkuhn
- Dr. Andreas Witt

## Contact

Dr. Marc Kupietz  
Corpus Linguistics Programme Area  
IDS – Institute for the German Language  
Postfach 10 16 21  
D-68016 Mannheim  
Germany  
Phone: +49-621-1581-0  
Fax: +49-621-1581-200  
Email: [corpuslinguistics@ids-mannheim.de](mailto:corpuslinguistics@ids-mannheim.de)



 INSTITUT FÜR  
DEUTSCHE SPRACHE  
R 5, 6-13  
D-68161 Mannheim  
[www.ids-mannheim.de](http://www.ids-mannheim.de)

Mitglied der  
  
Leibniz-Gemeinschaft

*The Institute for the German Language (IDS) is the central institution for the study and documentation of the contemporary usage and recent history of the German language. Together with 85 other non-university research institutes and service facilities, it belongs to the Leibniz Association, one of the four major research organisations in Germany.*

---

 INSTITUT FÜR  
DEUTSCHE SPRACHE

Corpus Linguistics Programme Area

# (Near) Duplicate Detection in the IDS Corpora

We used an elaborate algorithm to compute complete similarity matrices, e.g., for newspaper corpora as a whole and for similarities across different newspaper volumes and agency articles of one year. The algorithm ...

- is based on *shingling*
- uses normalised token pentagrams as basic units of comparison
- employs an improved, intuitive similarity metric, particularly well suited for short texts: the *shared shingle coverage ratio*
- computes similarities in, e.g., 20 years of *die tageszeitung* in less than four hours
- has linear space complexity

Stand: 04/12

T03JUL36384 die tageszeitung, 25.07.2003, S. 28, Ressort: tazplan-Programm;  
Diese Woche frisch

**Neu im Kino:**

## Diese Woche frisch

**Brandzeichen – Momente der Rebellion:** Doku über den Kampf gegen die Neoliberalisierung in Argentinien **Das verordnete Geschlecht:** Interviews mit Hermaphroditen **Die Blume des Bösen:** Claude Chabrol sezziert wieder die französische Provinzbourgeoisie und deren kellerlichen Früchte der Liebe: ein schwuler Pianistengott sein jugendlicher Liebhaber und dessen Mutter bilden ein Dreieck **Natürlich blond 2:** Lustig gemeinter Blondinenfilm **Planet der Kannibalen:** Düstere Science Fiction, schwarzweiß mit einem kleinen Lichtstreif am Horizont **Raumputzrolle Orion – Rücksturz ins Kino:** Das Weltraumbauunternehmen unserer Eltern jetzt endlich im **Kino Sindbad – Herr der 7 Meere:** Der Held aus 1001 Nacht als cooler Slacker **The Gathering:** Horror mit Christina Ricci **Vampire Hunter D:** Zeichentrickfassung des beliebten japanischen Vampiromics, ganz ohne Bisse

T03JUL37208 die tageszeitung, 30.07.2003, S. 28, Ressort: tazplan-Programm;  
Diese Woche frisch

**Neu im Kino:**

## Diese Woche frisch

**Brandzeichen – Momente der Rebellion:** Doku über den Kampf gegen die Neoliberalisierung in Argentinien **Das verordnete Geschlecht:** Interviews mit Hermaphroditen **Die Blume des Bösen:** Claude Chabrol sezziert die französische Provinzbourgeoisie und deren kellerlichen Früchte der Liebe: Ein schwuler Pianistengott, sein jugendlicher Liebhaber und dessen Mutter bilden ein Dreieck **Natürlich blond 2:** Lustig gemeinter Blondinenfilm **Planet der Kannibalen:** Düstere Sciencefiction, schwarzweiß mit einem kleinen Lichtstreif am Horizont **Raumputzrolle Orion – Rücksturz ins Kino:** Das Weltraumbauunternehmen unserer Eltern jetzt endlich im **Kino Sindbad – Herr der 7 Meere:** Der Held aus 1001 Nacht als cooler Slacker **The Gathering:** Horror mit Christina Ricci **Vampire Hunter D:** Zeichentrickfassung d  
beliebten japanischen Vampiromics, ganz ohne Bisse

**NEU KINO WOCHE FRISCH BRANDZEICHEN MOMENTE REBELLION DOKU KAMPE NEOLIBERALISIERUNG ARGENTINIEN VERORDNETE GESCHLECHT INTERVIEWS HERMAPHRODITEN BLUME BSEN CLAUDE CHABROL SEZERT FRANZÖSISCHE PROVINZBOURGEOISIE DEREN KELLERLICHEN OTE LIEBE SCHWULER PIANISTENGOTT SEIN JUGENDLICHER LIEBHABER DESSEN MUTTER BILDEN DREIECK NATÜRLICH BLOND LUSTIG GEMEINTER BLONDINENFILM PLANET KANNIBALEN DÜSTER SCIENCE FICTION SCHWARZWEIß KLEINEN LICHTSTREIF HORIZONT RAUMPUTZROLLE ORION RUCKSTURZ INS KINO WELTRAUMBÄUUNTERNEHMENS UNSERER ELTERN JETZT ENDLICH KINO SINDBAD HERR MEERE HELD NACHT COOLER SLACKER THE GATHERING HORROR CHRISTINA RICCI VAMPIRE HUNTER D ZEICHENTRICKFASSUNG BELIEBTEN JAPANISCHEN VAMPIROMICS GANZ OHNE BISSE**

**NEU KINO WOCHE FRISCH BRANDZEICHEN MOMENTE REBELLION DOKU NEOLIBERALISIERUNG ARGENTINIEN VERORDNETE GESCHLECHT INTERVIEWS HERMAPHRODITEN BLUME BSEN CLAUDE CHABROL SEZERT FRANZÖSISCHE PROVINZBOURGEOISIE DEREN KELLERLICHEN OTE LIEBE SCHWULER PIANISTENGOTT SEIN JUGENDLICHER LIEBHABER DESSEN MUTTER DREIECK NATÜRLICH BLOND LUSTIG GEMEINTER BLONDINENFILM PLANET KANNIBALEN DÜSTER SCIENCEFICTION SCHWARZWEIß KLEINEN LICHTSTREIF RAUMPUTZROLLE ORION RUCKSTURZ INS KINO WELTRAUMBÄUUNTERNEHMENS UNSERER ELTERN JETZT ENDLICH KINO SINDBAD HERR MEERE HELD NACHT COOLER SLACKER THE GATHERING HORROR CHRISTINA RICCI VAMPIRE HUNTER D ZEICHENTRICKFASSUNG BELIEBTEN JAPANISCHEN VAMPIROMICS GANZ OHNE BISSE**

**Neu im Kino:** Diese Woche frisch Brandzeichen Momente der Rebellion: Doku über den Kampf gegen die Neoliberalisierung in Argentinien **Das verordnete Geschlecht:** Interviews mit Hermaphroditen **Die Blume des Bösen:** Claude Chabrol sezziert wieder die französische Provinzbourgeoisie und deren kellerlichen Früchte der Liebe: ein schwuler Pianistengott sein jugendlicher Liebhaber und dessen Mutter bilden ein Dreieck **Natürlich blond 2:** Lustig gemeinter Blondinenfilm **Planet der Kannibalen:** Düstere Science Fiction schwarzweiß mit einem kleinen Lichtstreif am Horizont **Raumputzrolle Orion Rücksturz ins Kino:** Das Weltraumbauunternehmen unserer Eltern jetzt endlich im **Kino Sindbad Herr der 7 Meere:** Der Held aus 1001 Nacht als cooler Slacker **The Gathering:** Horror mit Christina Ricci **Vampire Hunter D:** Zeichentrickfassung des beliebigen japanischen Vampiromics ganz ohne Bisse

**Neu im Kino:** Diese Woche frisch Brandzeichen Momente der Rebellion: Doku über den Kampf gegen die Neoliberalisierung in Argentinien **Das verordnete Geschlecht:** Interviews mit Hermaphroditen **Die Blume des Bösen:** Claude Chabrol sezziert die französische Provinzbourgeoisie und deren kellerlichen Früchte der Liebe: Ein schwuler Pianistengott sein jugendlicher Liebhaber und dessen Mutter bilden ein Dreieck **Natürlich blond 2:** Lustig gemeinter Blondinenfilm **Planet der Kannibalen:** Düstere Sciencefiction schwarzweiß mit einem kleinen Lichtstreif am Horizont **Raumputzrolle Orion Rücksturz ins Kino:** Das Weltraumbauunternehmen unserer Eltern jetzt endlich im **Kino Sindbad Herr der 7 Meere:** Der Held aus 1001 Nacht als cooler Slacker **The Gathering:** Horror mit Christina Ricci **Vampire Hunter D:** Zeichentrickfassung des beliebigen japanischen Vampiromics ganz ohne Bisse

Statistik (Token)	
Mindest-Sequenzlänge	5 Token
Token-S-Gramm-Überlappung:	164 Token (97.62%)
Rest:	4 Token
Shingles	90
Common shingles:	70
Common Single Ratio:	77.78%
Längendifferenz:	2 Token (1.19%)
Rest-Überlappung (fett):	1 Token (25.00%)
Unmatched:	3 Token (1.79%)
Sequenzen:	3
Transpositionen:	0

Statistik (Wörter)	
Mindest-Sequenzlänge	5 Wörter
Token-S-Gramm-Überlappung:	220 Wörter (96.07%)
Rest:	9 Wörter
Shingles	131
Common shingles:	90
Common Single Ratio:	68.70%
Längendifferenz:	3 Wörter (1.31%)
Rest-Überlappung (fett):	0 Wörter (0.00%)
Unmatched:	9 Wörter (3.93%)
Sequenzen:	5
Transpositionen:	0

Snetliges	
[t(a)-(t)b]	5
[dow(a)-dow(b)]	5
Absatz-Anteil	0.036
Cluster-Größe	
p(Versions):	0.269
p(Varian):	0.731

typical portion of (near)  
duplicates in newspaper  
cms dumps

run time T (in seconds)

100

90

80

70

60

50

40

30

20

10

(a) Two short newspaper texts and their normalisations. Tokens in the normalised texts covered by common 5-shingles (see below) are displayed in bold:

T02/NOV.53095 die tageszeitung,  
01.11.2002, S. 12, Ressort: Meinung und  
Diskussion; Betr.: Dieter Rulff

Dieter Rulff ist freier Journalist in Berlin. Nach langen Jahren bei der taz war er zuletzt leitender Redakteur der Wochenzeitung „Die Woche“. Sein Interesse gilt seit langem der Entwicklung der deutschen Innen- und Parteipolitik.

**DIETER RULFF FREIER JOURNALIST  
BERLIN LANGEN JAHREN TAZ ZULETZT  
LEITENDER REDAKTEUR WOCHENZEITUNG  
WOCHE INTERESSE GILT SEIT LANGEM  
ENTWICKLUNG DEUTSCHEN INNEN  
PARTEIPOLITIK**

T03/JUL.31966 die tageszeitung,  
01.07.2003, S. 12, Ressort: Meinung und  
Diskussion; Betr.: Dieter Rulff

Dieter Rufft ist freier Journalist in Berlin. Nach vielen Jahren bei der taz war er zuletzt leitender Redakteur der Zeitung „Die Woche“. Sein Interesse gilt seit langem der Entwicklung der deutschen Innen- und Parteipolitik.

**DIETER RULFF FREIER JOURNALIST  
BERLIN VIELEN JAHREN TAZ ZULETZT  
LEITENDER REDAKTEUR ZEITUNG WOCHEN  
INTERESSE GILT SEIT LANGEM  
ENTWICKLUNG DEUTSCHEN INNEN  
PARTEIPOLITIK**

**(b) Shared 5-shingles of the normalised texts:**

$$S_A \cap S_B = \{(\text{DIETER, RULFF, FREIER, JOURNALIST, BERLIN}),$$
$$(\text{JAHREN, TAZ, ZULETZT, LEITENDER, REDAKTEUR}),$$
$$(\text{WOCHE, INTERESSE, GILT, SEIT, LANGEM}),$$
$$(\text{INTERESSE, GILT, SEIT, LANGEM, ENTWICKLUNG}),$$
$$(\text{GILT, SEIT, LANGEM, ENTWICKLUNG, DEUTSCHEN}),$$
$$(\text{SEIT, LANGEM, ENTWICKLUNG, DEUTSCHEN, INNEN}),$$
$$(\text{LANGEM, ENTWICKLUNG, DEUTSCHEN, INNEN, UND}),$$
$$(\text{ENTWICKLUNG, DEUTSCHEN, INNEN, UND, PARTEIPOLITIK}) \}$$

(c) Similarity according to the *shared shingle ratio* (*ssr*) and the *shared shingle coverage ratio* (*sscr*) metric, respectively:

$$r_{\text{ssr}}(A, B) = \frac{|S_A \cap S_B|}{|S_A \cup S_B|} = \frac{8}{28} \approx \mathbf{0.2857}$$
  

$$r_{\text{sscr}}(A, B) = \frac{\text{marked tokens}}{\text{total tokens}} = \frac{40}{44} \approx \mathbf{0.9091}$$

