




# Experimente zur automatischen Textstilerkennung

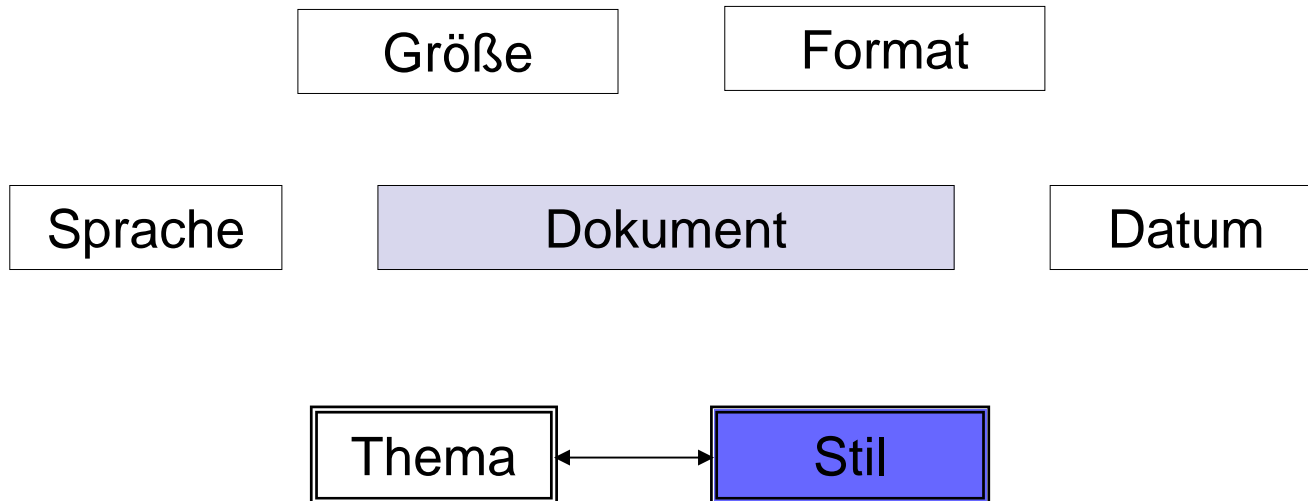


Pavel Braslavski  
29.01.2004

# Der Zweck der Stilerkennung

## Dokumentenstil als zusätzlicher Parameter bei der Internetsuche

(s. a. Karlgren, J. and Cutting, D. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In Proceedings of COLING 94, Kyoto.)





# Funktionale Stillehre

Sprache (=Mitteln)  $\leftrightarrow$  Reden (=Prozess)

Der Stil als eine Funktion der Gesellschaftsaktivität

Funktionalstile:

1. Offizieller Stil (Geschäftsstil)
2. Wissenschaftlicher Stil
3. Zeitungsstil (Publizistik)
4. Umgangssprache
5. Schöne Literatur (?)

# Experiment I (1999-2000)

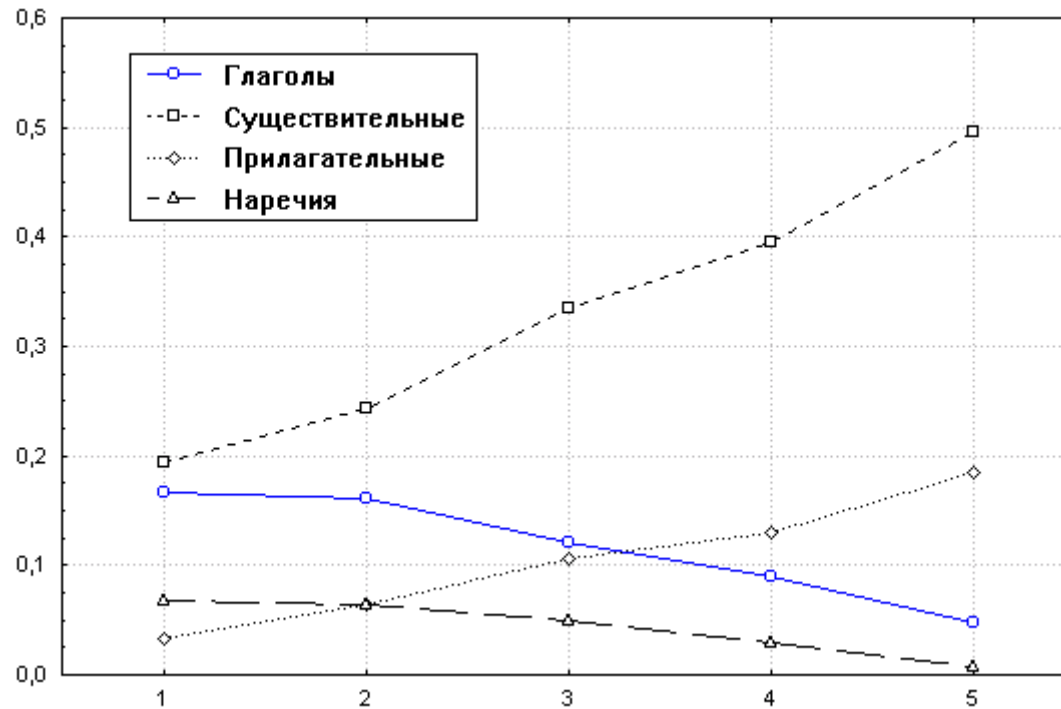
- Methode: Lineare Diskriminanzanalyse
- Klassen: 5 Funktionalstile
- Lernmenge: 305 Dokumente
- „Einfache“ Textparameter
- „Überflüssige“ Startparametermenge



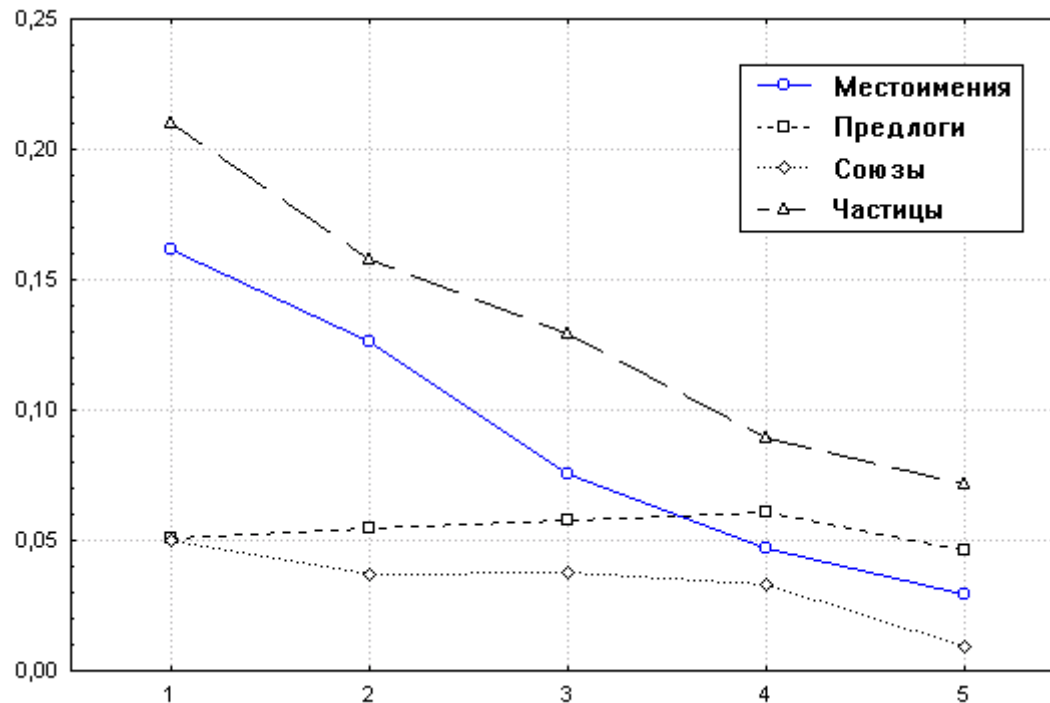
# Die Parameterebenen

- **Wortbildung** (wissenschaftliche Präfixe wie *aqua-*, *aero-* usw.)
- **Morphologie** (Wortarten, Neutra, Reflexiva usw.)
- **Lexik** (Wortlisten)
- **Syntax** (Genitivketten, bestimmte Konjunktionen usw.)
- **Formale Parameter** (Satz- und Wortlänge, Zeichensetzung usw.)

# Die morphologischen Parameter I



# Die morphologischen Parameter II



# Die Klassifizierungsparameter

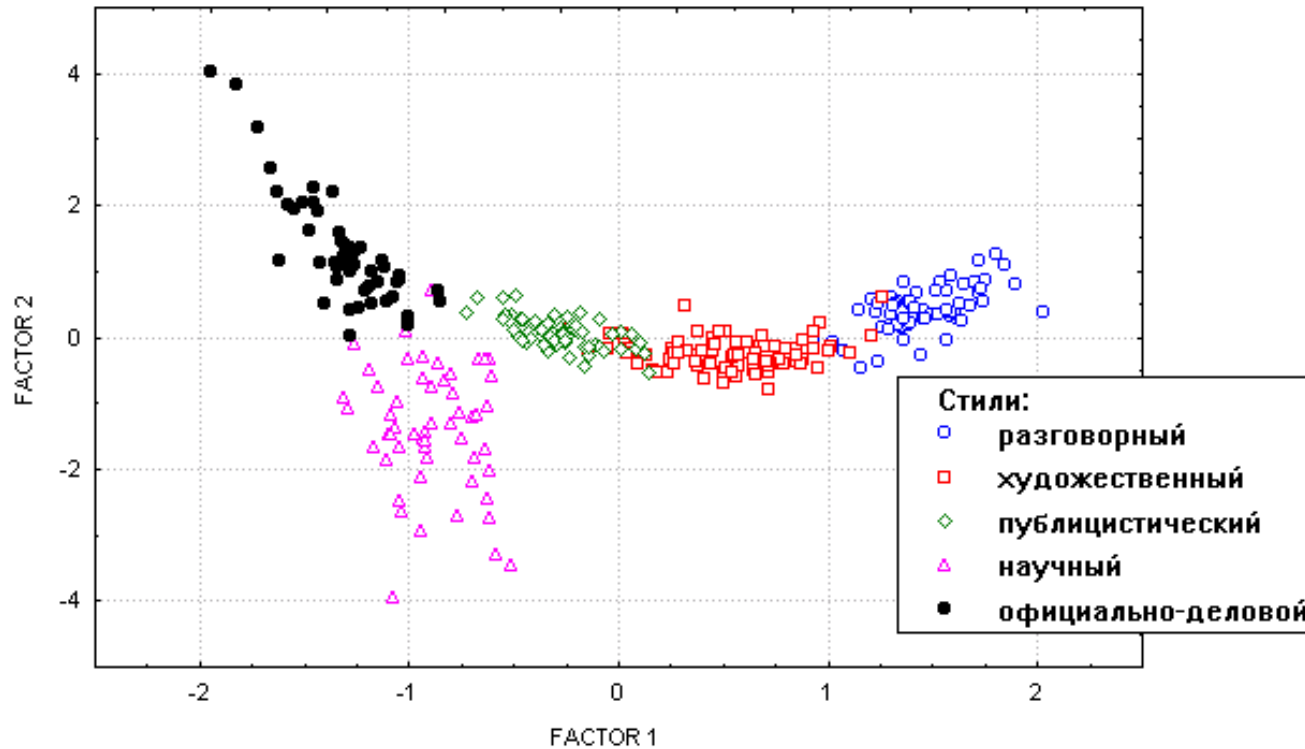
- Anteil der Verben
- Anteil der Adverbien
- Durchschnittswortlänge
- Durchschnittssatzlänge
- Anteil der Wörter der *Science* Liste
- Anteil der Wörter mit wissenschaftlichen Präfixen
- Anteil der Wörter der *OfficialDoc* Liste

# Das Klassifizierungsbeispiel

Stil	UmgSpr.	SchLit.	Publ.	Wiss.	Offizieller Stil	Recall
Umgangssprache	0	1	0	0	0	0,0 %
Schöne Literatur	0	1	0	0	0	100,00 %
Publizistik	0	2	40	0	2	90,91 %
wissenschaftliche Texte	0	0	5	20	0	80,00 %
Precision	0%	25,0%	88,89%	100,0%	2	85,92 %

Die ersten 71 Yandex-Antworte auf die Anfrage *'радикал отношение'* (März 2000)

# Die Hauptkomponenten der Lernmenge



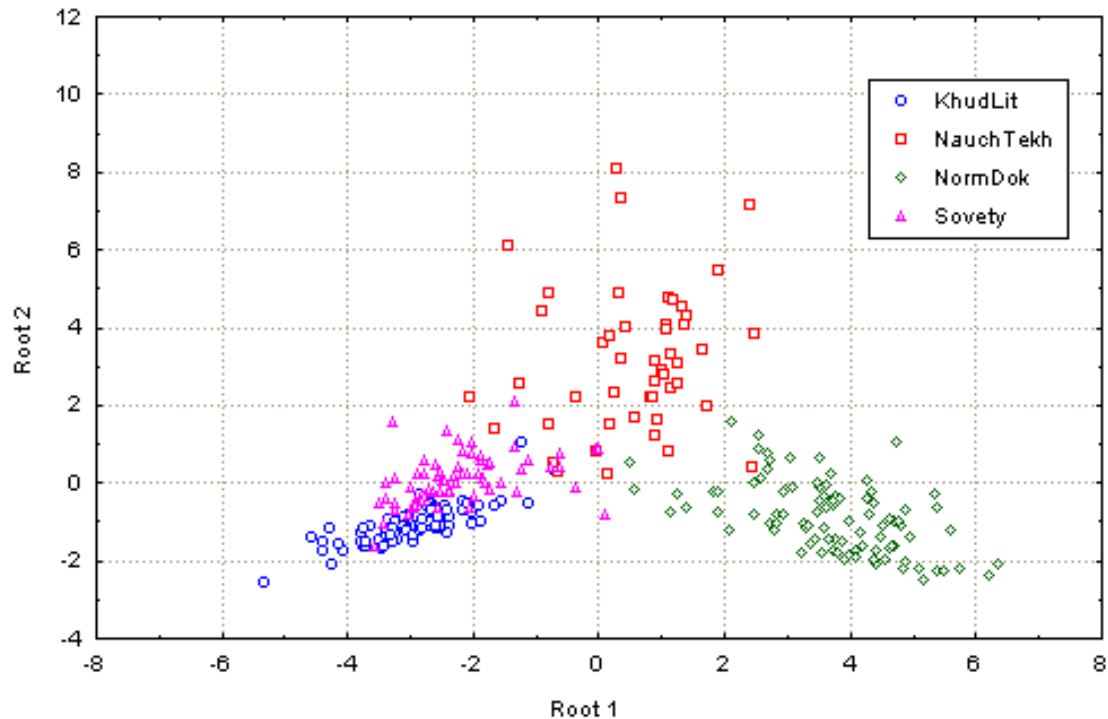
# Experiment II (2001-2002)

- Klassen: 4 Facetten des Yandex-Katalogs ([www.yaca.yandex.ru](http://www.yaca.yandex.ru))
- Lernmenge: 285 Dokumente
- Methode: Lineare Diskriminanzanalyse
- Distanzbasierten Abschätzungstechniken
- „Einfache“ Textparameter
- „Überflüssige“ Startparametermenge

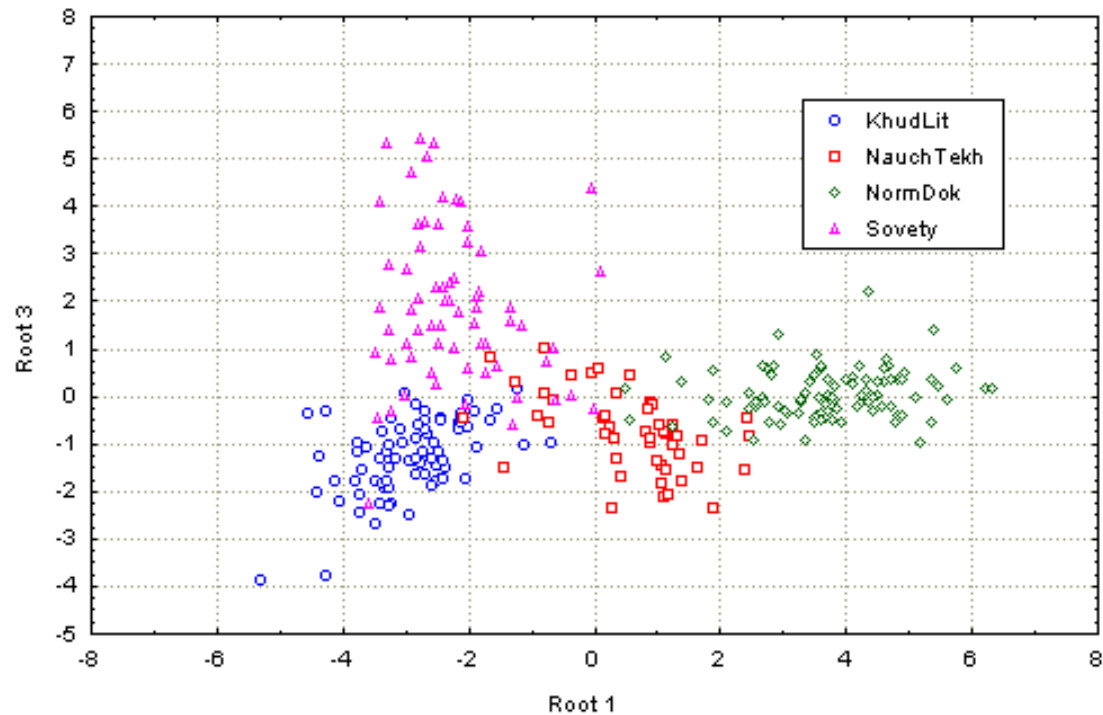
# Die Facettenklassifikation von Yandex

- Thema (600)
- Region (230)
- Informationsquelle (4)
- Informationsempfänger (4)
- Wirtschaftsbranche (3)
- **Genre (11/6)**
  - *Schöne Literatur (Fiction)*
  - *Wissenschaft, Technik (SciTech)*
  - *Populärwissenschaft (PopSci)*
  - *Offizielle Dokumente (NormDoc)*
  - *Ratschläge (Tip)*
  - *Publizistik (Publ)*

# Kanonische Analyse der Lernmenge I



# Kanonische Analyse der Lernmenge II



# Klassifizierungsparameter

- Anteil der Verben
- Anteil der Adverbien
- Anteil der Wörter der *SciTech* Liste
- Anteil der Wörter der *NormDoc* Liste
- Anteil der Wörter der *Tip* Liste
- Durchschnittswortlänge
- Anteil der Sätze mit dem Muster ‘{*МОЖНО|НУЖНО*} + *Infinitiv*’

# Ergebnisse

	Test 1		Test 2		Test 3		Test 4	
	P	R	P	R	P	R	P	R
Fiction	0.857	0.913	0.729	0.913	0.506	0.913	0.709	0.848
Sci (Tech)	0.912	0.912	0.933	0.724	0.553	0.724	0.682	0.517
NormDoc	0.950	0.864	0.950	0.864	0.452	0.864	0.842	0.727
Tip	0.750	0.727	0.585	0,727	0.267	0.727	0.423	0.333
NoGenre	-	-	-	-	-	-	0.633	0.705
Total	0.859		0.799		0.436		0.649	



# Zukunftspläne

Experimente zum stilistischen Rangieren (ranking) mit RIRES/ROMIP Daten.

- 7+ Gb
- 600 000+ HTML Seiten
- 20 000+ Websites
- 54 bewertete Anfragen

RIRES (Russian Information Retrieval Evaluation Seminar)

<http://romip.narod.ru>



# Das Fazit

- Der Textstil kann mit ausreichender für die praktischen Anwendungen Qualität aufgrund der formalen Parametern erkannt werden.
- Statistische Methoden der Textverarbeitung brauchen eine gute konzeptuelle Begründung.
- Stilistische Textcharakteristiken müssen an die bestimmte Anwendungen angepasst werden (Rangieren, Filterung, Klassifikation...).