Интеллектуальные информационные системы

Тема 9

# Стилистический анализ

Павел Исаакович Браславский
pb@imach.uran.ru
весенний семестр 2006

## План

- Что такое стиль?
- Области применения
- Определение авторства
- Инструменты MS Word, показатели удобочитаемости
- Стилистическая/жанровая классификация в контексте ИП
  - Эксперимент 1: функциональные стили
  - Эксперимент 2: жанры каталога Яндекс
- Стилистическое ранжирование

## Стиль (от лат. *stilus, stylus)*

- «Уровень» языка (нейтральный, высокий, низкий)
- Функциональный стиль
- Особенности конкретного речевого акта (ораторская речь, бытовой диалог, дружеское письмо и т.д.)
- Индивидуальный стиль
- Стиль эпохи



## Стиль как «уровень языка»

- Лексика:
  *очи – глаза – зенки*;
  *длань – рука – пакля*
- Произношение, формы слова
  *чё/что, пОртфель, саму себя/самоё себя, крепкий кофе/крепкое кофе*
- Фраза
  *Fen4uK, типа того, клиенты спрашивают, а на сайте у них них... нету*

# Функциональный стиль

- Язык ↔ Речь
- 5 функциональных стилей:
  - разговорный
  - художественный (?)
  - публицистический
  - научный
  - официально-деловой



# Несовпадение стиля/содержания

Сущность квантовой теории танца, представляющей собой своего рода компромисс между классической механикой условно периодических движений и классической эмоциодинамикой, заключается в следующем. Танцующие могут описывать определенные квантовые орбиты, не испуская и не поглощая при этом никаких эмоций. Последние испускаются и поглощаются прерывным образом при переходах с одной квантованной орбиты на другую. При этом в противоположность тому, что имеет место в случае электронных плясок в боровском атоме, эмоциональное излучение, как и поглощение, сопровождается переходом не на более низ-кий, а, наоборот, на более высокий уровень, т.е., другими словами, возбуждением. Таким образом, во время танца (особенно парного) возбуждение танцующих неизменно возрастает, пока не наступит релаксация, вызываемая истощением.

*Я.И.Френкель «Квантовая теория танца»*

## Индивидуальный стиль

- Кто написал «Тихий Дон»?
- Кто написал «Роман с кокаином»? (Агеев? Набоков?)

Идея: найти комбинацию параметров, которую сложно *сознательно* контролировать

Например: доля служебных слов

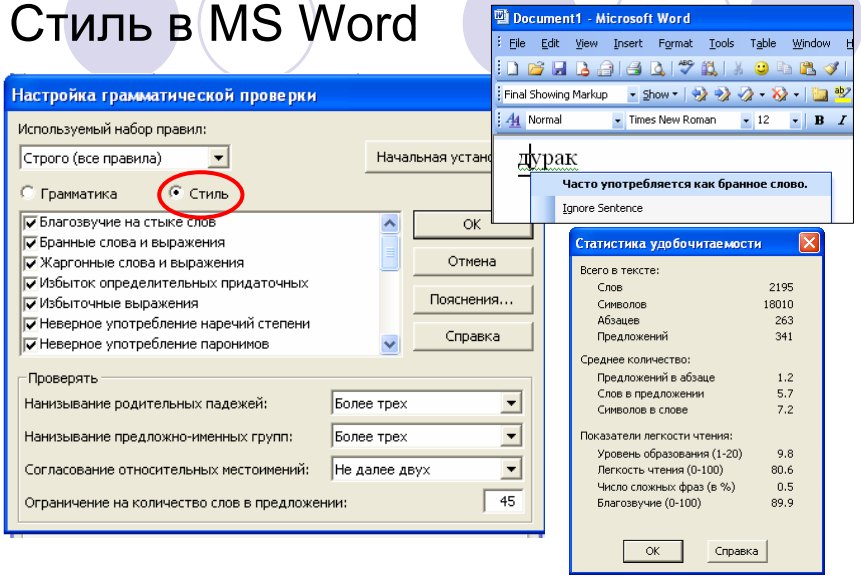


## Лингвоанализатор Дм. Хмелева

- → Последовательности пар букв (цепи Маркова)
- → Алгоритмы сжатия данных (!)

см. Лингвоанализатор и статью

# Стиль в MS Word






# Показатель легкости чтения

Flesch Reading Ease score, диапазон 0..100.
Стандартные документы: 60..70.

Формула для английского:

206.835 – (1.015 x ASL) – (84.6 x ASW)

ASL = average sentence length, средняя длина предложения в словах

ASW = average number of syllables per word, средняя длина слова в слогах

# Показатель уровня образования

Flesch-Kincaid Grade Level score, соответствует классам американской образовательной системы (Значение 8.0 – текст должен без труда понимать восьмиклассник). Для большинства документов целевой диапазон - 7.0..8.0.

Формула показателя уровня образования по Флешу/Кинсайду:

(0.39 x ASL) + (11.8 x ASW) – 15.59

где:

ASL = average sentence length, средняя длина предложения в словах

ASW = average number of syllables per word, средняя длина слова в слогах



# Стилистическая категоризация

- Задача: автоматически отнести документ к одному из заранее определенных стилей (ср.: тематическая категоризация)
- Стиль ↔ Тема: часто независимы
- Применение: Веб-поиск

**Космическая связь**

До сих пор любому желающему сделать звонок, скажем, из пустыни Сахара, приходилось нести с собой тяжелую и объемную аппаратуру, включающую передающее устройство и антенну.

«Известия»

…в шнурках бегунов будет расположено маленькое передающее устройство, которое зафиксируют антенны, расположенные на расстоянии пяти километров друг от друга.

**Парламентская газета, 5/09/ 2000**

**Приёмо-передающее устройство (трансивер) ПТ-100.**

Организация симплексных и полудуплексных каналов телефонной и телеграфной связи в КВ диапазоне. Одноконтурный перестраиваемый преселектор. Цифровая обработка и формирование сигналов. Частотная адаптация в телефонном канале.

передающие устройства, антенна

**ИНСТРУКЦИЯ о порядке приема и рассмотрения заявок на выявление помех радиоприему**

Зона уверенного приема определяется расчетным путем при решении задачи обеспечения электромагнитной совместимости данного передающего устройства с действующими и планируемыми к установке РЭС различного назначения в этой зоне и прилегающих к ней.

Для уменьшения помех работающим в эфире радиостанциям при налаживании передающих устройств применяют эквивалент антенны. Его нетрудно превратить в измеритель выходной мощности передатчика.

В.Скрыпник, Радио, №9'79

**Происхождение негосударственного телевидения в Закавказье**

…Единственным исключением является передающее устройство, оставленное отступавшей армией Звиада Гамсахурдия в 1992 году.

**Законодательство и практика средств массовой информации**



# Эксперимент I (1999-2000)

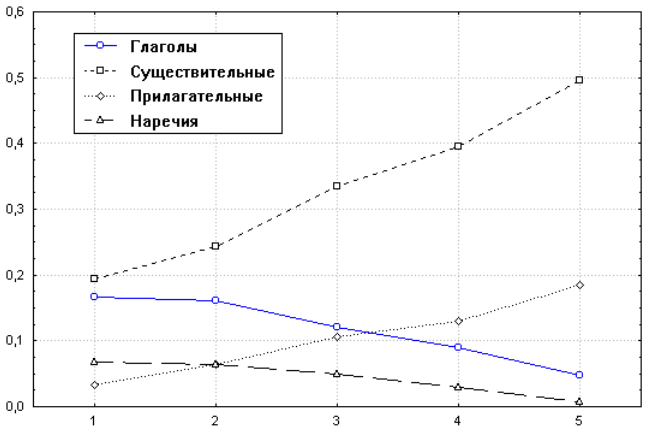
- Категории – 5 функциональных стилей
- Обучающая выборка – 305 документов
  - Разг: чаты и ICQ
  - ХудЛит: рассказы-участники конкурса сетевой литературы
  - Публ: сетевые СМИ
  - Науч: он-лайн версии естественно-научных статей
  - ОфДел: федеральные законы
- Признаки легко вычисляются (первичный набор – 31)
- Метод построения классификатора – линейные классифицирующие функции (~многомерная линейная регрессия) STATISTICA, модуль Discriminant Analysis

# Первичный набор признаков

| Уровень | Признаки | |
|---|---|---|
| | формальные | формально-семантические |
| графика | формулы | smiles (“улыбки”) |
| слово-образование | нет | научные приставки |
| лексика | средняя длина слова | • общенаучная лексика<br>• названия официальных документов<br>• слова организации логики повествования |
| морфология | • распределение по частям речи<br>• существительные среднего рода<br>• возвратные глаголы<br>• аббревиатуры | • личные местоимения 1-го и 2-го лица: *я, ты, мы, вы*<br>• частицы *бы*<br>• частицы *ну, вот, ведь* |
| синтаксис | • цепочки имен существительных в родительном падеже<br>• средняя длина предложения в словах<br>• доля предложений с экспрессивной пунктуацией | доля предложений с подчинительными союзами |

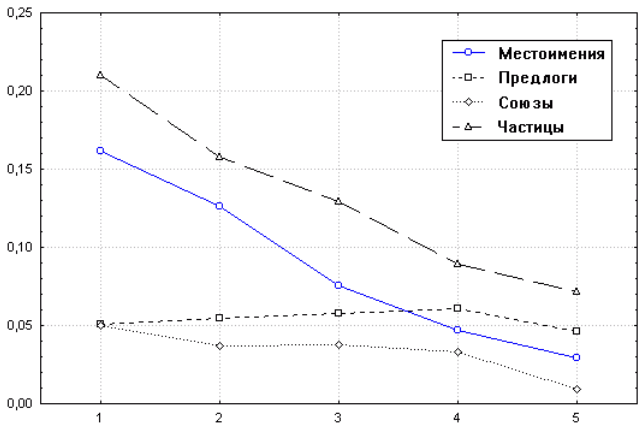


# Морфология/стили




**1 – разговорный; 2 – художественный; 3 – публицистический; 4 – научный; 5 – официально-деловой.**

# Морфология/стили – 2






# Классифицирующая функция

$$s = Ax + b$$

$$\mathbf{A} = \begin{pmatrix} 458{,}75 & 318{,}61 & 37{,}47 & 0{,}09 & 8{,}95 & -252{,}01 & -238{,}23 \\ 471{,}56 & 313{,}31 & 40{,}60 & 0{,}25 & 23{,}62 & -292{,}64 & -244{,}43 \\ 408{,}92 & 275{,}27 & 44{,}98 & 0{,}29 & 67{,}34 & -393{,}37 & -197{,}99 \\ 367{,}12 & 173{,}79 & 50{,}59 & 0{,}11 & 436{,}48 & 157{,}17 & -201{,}34 \\ 287{,}77 & 122{,}34 & 48{,}41 & 0{,}40 & 190{,}42 & -423{,}40 & 160{,}86 \end{pmatrix} \quad b = \begin{pmatrix} -139{,}79 \\ -158{,}25 \\ -173{,}39 \\ -211{,}02 \\ -192{,}68 \end{pmatrix}$$

NB: сокращение размерности: 31→ 7 признаков, см. слайд 21

Стили:

$s_1$ – разговорный; $s_2$ – художественный; $s_3$ – публицистический; $s_4$ – научный; $s_5$ – официально-деловой

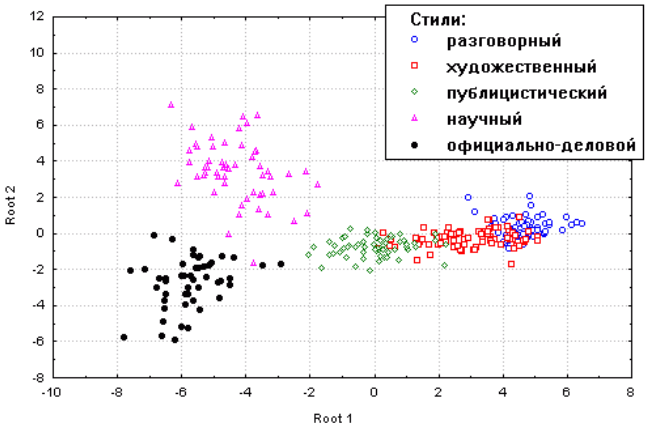
# Оценка: пример

| | 1 | 2 | 3 | 4 | 5 | Всего | Recall |
|---|---|---|---|---|---|---|---|
| 1. Разг | - | - | - | - | - | - | - |
| 2. ХудЛит | - | 2 | 1 | - | - | 3 | 0.67 |
| 3. Публ | - | 2 | 44 | 8 | - | 54 | 0.81 |
| 4. Науч | - | - | 10 | 17 | - | 27 | 0.63 |
| 5. ОфДел | - | - | - | - | 7 | 7 | 1.0 |
| Undefined | 1 | 1 | 7 | 3 | - | 12 | - |
| Всего | 1 | 5 | 62 | 28 | 7 | 103 | |
| Precision | 0.0 | 0.2 | 0.71 | 0.61 | 1.0 | | 0.74 |

Запрос: ‘небесные тела’



# Структура стилей



# Первое каноническое направление

$$R_1 = 18{,}44\ x_1 + 22{,}35\ x_2 - 1{,}36\ x_3 - 0{,}01\ x_4 - 37{,}74\ x_5 - \\ - 15{,}41\ x_6 - 31{,}07\ x_7 + 5{,}73$$

$x_1$ – доля глаголов;

$x_2$ – доля наречий;

$x_3$ – средняя длина слова;

$x_4$ – средняя длина предложения;

$x_5$ – доля слов научной лексики;

$x_6$ – доля слов с научными префксами;

$x_7$ – доля слов-названий официальных документов.

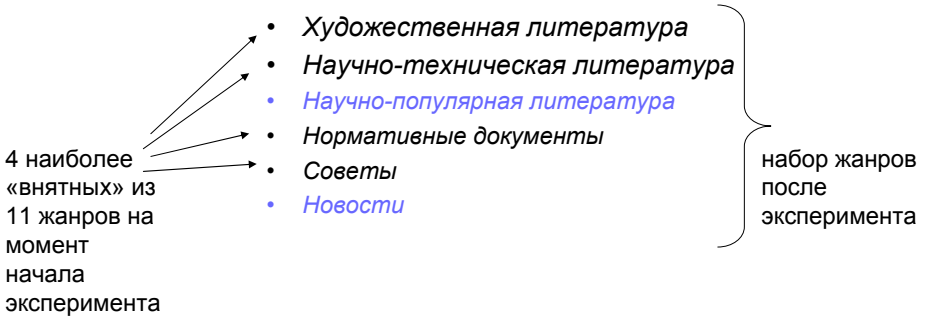
Можно использовать в качестве непрерывного стилистического показателя?
См. *стилистическое ранжирование*



# Эксперимент II (2001-2002)

- Категории: 4 жанра каталога Яндекс (www.yaca.yandex.ru)
- Обучающая выборка: 285 док-тов
- Метод: линейные классифицирующие функции (тот же)
- Критерий отбрасывания результата классификации на основе расстояния Мехаланобиса
- «Простые» признаки
- Избыточный первичный набор признаков (→ снижение размерности)

# Жанры каталога Яндекс




NB: жанры сайтов, а не документов



# Выбранные признаки

- доля глаголов в личной форме
- доля наречий
- доля слов из списка «НаучТех»
- доля слов из списка «НормДок»
- доля слов из списка «Советы»
- средняя длина русского слова
- доля предложений с конструкцией ‘*{можно|нужно} + инфинитив*’

# Оценка

| | Тест 1 | | Тест 2 | | Тест 3 | |
|---|---|---|---|---|---|---|
| | P | R | P | R | P | R |
| ХудЛит | 0.857 | 0.913 | 0.506 | 0.913 | 0.709 | 0.848 |
| Науч | 0.912 | 0.912 | 0.553 | 0.724 | 0.682 | 0.517 |
| НормДок | 0.950 | 0.864 | 0.452 | 0.864 | 0.842 | 0.727 |
| Советы | 0.750 | 0.727 | 0.267 | 0.727 | 0.423 | 0.333 |
| Прочее | - | - | - | - | 0.633 | 0.705 |
| Всего | 0.859 | | 0.436 | | 0.649 | |

Тестовая выборка – 260 документов, классифицированных вручную в соответствии с новым набором жанров (6)



# Стилистическое ранжирование

- Идея: включить в схему ранжирования поисковой выдачи стилистические показатели
- Когда может пригодиться: поиск «серьезных» документов (наука, аналитика, право)
- Эксперимент на данных РОМИП (см. Тему 1), многие расширенные описания информационной потребности предполагают «серьезность» релевантного документа
- Подробнее см. [Braslavski, 2005]

# Коллекция РОМИП

7+ Gb – подмножество домена narod.ru
600 000+ HTML страниц на русском
20 000+ Веб-сайтов
54 оцененных запросов

Пример запроса:
rb4095: *Ель обыкновенная*
Описание: Релевантная страница должна содержать информацию об ели обыкновенной – например, основные характеристики этой породы деревьев, места произрастания, применения в народном хозяйстве и т.п.



# Данные для эксперимента

Результаты ИПС Кодекс на РОМИП-2003 :
51 набор документов, соответствующий оцененным запросам
2608 страниц (вкл. 388 релевантных)

+ случайная выборка (500 документов)

Обработка:
HTML → plain text
1824 документа, длиннее 50 предложений (68%)

# Извлечение факторов

Анализ случайной выборки

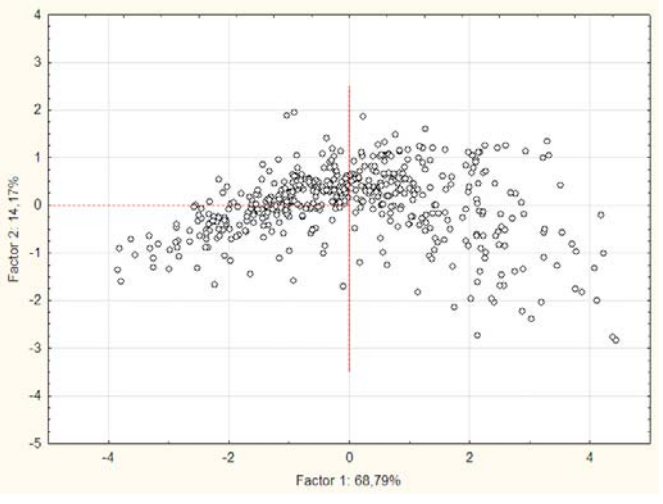
$$S_G = -0{,}32 \cdot x_{1S} + 0{,}31 \cdot x_{2S} - 0{,}30 \cdot x_{3S} + 0{,}28 \cdot x_{4S}$$

$x_1$ – *длина слова;*
$x_2$ – *доля глаголов в личной форме;*
$x_3$ – *доля прилагательных;*
$x_4$ – *доля местоимений первого лица.*

Объясняет 68,8% дисперсии выборки.



# Проекция на плоскость факторов



# Корреляция

| | *GL* | *RE* | $S_G$ | $S_L$ |
|---|---|---|---|---|
| *GL* | 1,0 | -0,91 | -0,73 | -0,50 |
| *RE* | -0,91 | 1,0 | 0,57 | 0,38 |
| $S_G$ | -0,73 | 0,57 | 1,0 | 0,81 |
| $S_L$ | -0,50 | 0,38 | 0,81 | 1,0 |

GE – Grade Level; RE – Reading Ease;
*SG* – фактор, выделенный на основе анализа случайно выборки,
SL – фактор, выделенный на основании локального анализа каждого из набора документов, соответствующих запросу

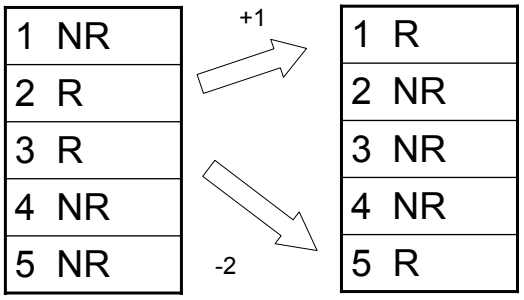


# Ранговая корреляция

| | $R_K$ | $R_{GL}$ | $R_{RE}$ | $R_{SG}$ | $R_{SL}$ |
|---|---|---|---|---|---|
| $R_K$ | 1,0 | 0,11 | 0,09 | 0,18 | 0,18 |
| $R_{GL}$ | 0,11 | 1,0 | 0,97 | 0,73 | 0,73 |
| $R_{RE}$ | 0,09 | 0,97 | 1,0 | 0,67 | 0,68 |
| $R_{SG}$ | 0,18 | 0,73 | 0,67 | 1,0 | 0,998 |
| $R_{SL}$ | 0,18 | 0,73 | 0,68 | 0,998 | 1,0 |

$R_K$ – ранг ИПС Кодекс (релевантность)

# Мера для сравнивания списков




$\Sigma = -1$



# Агрегированные ранги: оценка

| | $D_R$ | $AD_R$ | $AD_N$ | $\oplus$ | $\varnothing$ | $\otimes$ |
|---|---|---|---|---|---|---|
| $R_{SG}$ | -1377 | -3,55 | 0,62 | 16 | 0 | 35 |
| $R_K + R_{SG}$ | -95 | -0,24 | 0,04 | 21 | 1 | 29 |
| $R_K + 0{,}5 * R_{SG}$ | 73 | 0,19 | -0,03 | 22 | 0 | 29 |
| $R_K + 0{,}25 * R_{SG}$ | 57 | 0,15 | -0,03 | 22 | 6 | 23 |
| $R_K + 0{,}125 * R_{SG}$ | 54 | 0,14 | -0,02 | 22 | 11 | 18 |

# Агрегированные ранги – 2

| | $D_R$ | $AD_R$ | $AD_N$ | ⊕ | ∅ | ⊗ |
|---|---|---|---|---|---|---|
| $R_1$ | -1263 | -3,26 | 0,57 | 17 | 1 | 33 |
| $R_K + R_1$ | 146 | 0,38 | -0,066 | 27 | 2 | 22 |
| $R_K + 0,5 * R_1$ | 164 | 0,42 | -0,074 | 23 | 7 | 21 |
| $R_K + 0,25 * R_1$ | 119 | 0,31 | -0,05 | 26 | 5 | 20 |
| $R_K + 0,125 * R_1$ | 73 | 0,19 | -0,03 | 23 | 13 | 15 |

$R_1$ – см. слайды 20, 21.