



Художественные и нехудожественные двусловные сочетания в русском языке XIX века

Борис Орехов

Проблема и исследовательский вопрос

- 
- Большие оцифрованные коллекции (НКРЯ) → системное изучение языка литературы
 - От интуиции исследователя — к количественным методам
 - Вопрос: какие двусловные сочетания специфичны для художественной сферы, а какие — для нехудожественной?
 - «Специфические» = ≥ 5 вхождений в одной сфере и 0 — в противоположной
 - Новизна: не авторские стили, а функциональные сферы; не слова, а биграммы



Материал

- 8 477 текстов НКРЯ, датированных XIX веком
- Метаданные: сфера, дата, автор, заголовок
- F (художественная): маркер «художественная»
- N (нехудожественная): маркер «нехудожественная» + публицистика, учебно-научная, бытовая, официально-деловая, реклама, церковно-богословская, производственно-техническая
- Тексты с неоднозначной сферой — исключены



Метрики анализа

4 уровня

Уровень 1 — формальный:
размеры, частоты

Уровень 2 — структурный: длины
токенов, цифры/латиница, POS-
матрица

Уровень 3 — семантический:
антропонимы, топонимы,
сокращения, метатекст,
предикативные, артефакты

Уровень 4 — контрастивный



Уровень 1: формальные характеристики

- F: 126 580 биграмм; N: 195 420 — в 1,54 раза больше при сопоставимом корпусе
- Средняя частота: 8,35 (F) vs 8,51 (N) — практически одинаково
- Медиана: 6 в обеих сферах; Q3: 8 (F) vs 9 (N); IQR: 3 vs 4
- Максимумы: 1819 («глафира семеновна») vs 1737 («вел кн»)
- Пороговый эффект: частота ровно 5 — 36,2 % (F) и 34,3 % (N); ≤ 10 — 85,3 % и 83,4 %
- Вывод: 85 % каждого списка — статистически «хрупкие» единицы; выводы делать по верхушке



Уровень 2: структурные паттерны

- Средние длины почти равны (11,24 vs 11,82), но форма распределений разная
- N — короткие сокращения (г, кн, вел) + длинные термины
- F — «средняя зона» 3–9 символов: имена, глаголы речи, служебные слова
- Цифры — сильнейший дискриминатор: 7,54 % (N) vs 0,06 % (F), разрыв в 136 раз
- Латиница: 2,47 % vs 0,63 % — но функция разная: транслитерация в F (gnadige frau, 380) vs термины в N (formula и, 362)
- POS: F — предикативные (VERB+NOUN 5,98 %, VERB+NPRO 2,36 %); N — атрибутивные (ADJF+NOUN 13,96 % vs 6,34 %)



Уровень 2: якорные токены

- F — «якоря разговора»: сказал (8915), спросил (4176), сказала (4047), отвечал (3520), проговорил (2011), спросила (2631); местоимения и имена персонажей
- N — «якоря формуляра»: г (15 711), въ (11 746), е (6744), цифры 2 (4470), 1 (3474), 3 (2851), кн (3637), вел (3254), стр (2515), месяцы (января, марта, февраля)
- Вывод: F и N различаются не лексикой, а «грамматическим каркасом»



Уровень 3: семантические классы

- Антропонимы всех типов: 7,82 % (F) vs 1,29 % (N) — разрыв в 6 раз
- «имя + отчество»: 1,30 vs 0,16 % (8×); «отчество + прочее»: 2,85 vs 0,33 % (9×)
- Предикативные: 8,80 % (F) vs 2,91 % (N) — втрое
- Сокращения: 1,30 % (F) vs 4,23 % (N) — в 3,3 раза
- Канцелярские предлоги (при, по, на, о, согласно, вследствие): 0,00 vs 0,07 %
- Метатекст: практически только в N — «лето правления» (482), «сего месяца» (397), «примечания на» (376)



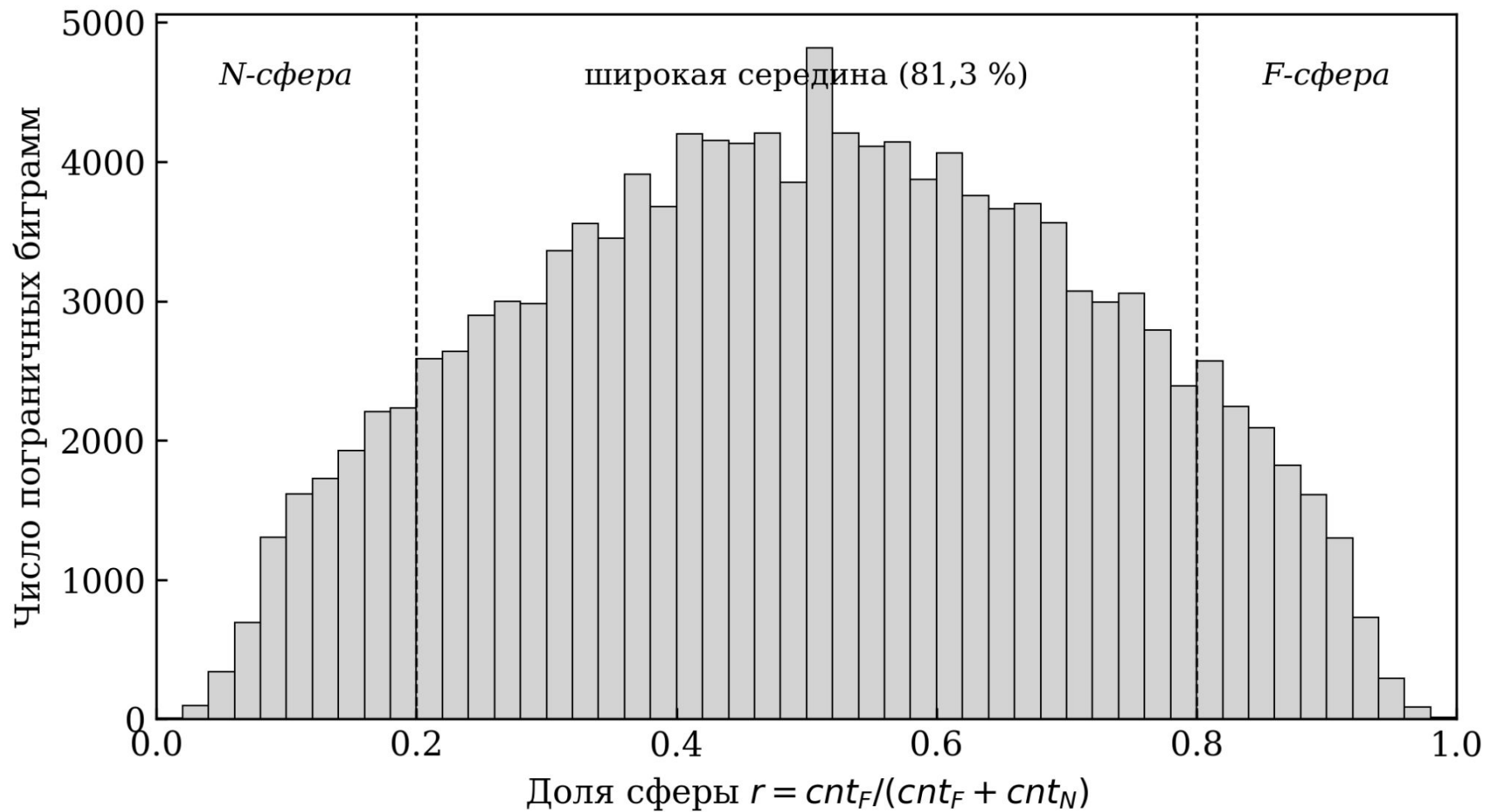
Уровень 3: примеры-контрасты

- F: «глафира семеновна» (1819), «патап максимыч» (1718), «марко данылыч» (1088), «егор егорыч» (1003), «василий борисыч» (818) — дилогия Мельникова-Печерского
- F: «сентъ джонъ» (938), «мистрисъ дарлингъ» (432); «сказал патап» (248), «сказал андрей» (207), «сказала анюта» (181)
- N: «вел кн» (1737), «вел князь» (832), «ваш п» (492), «п анненков» (331), «г успенский» (248)
- N: «лето правления» (482), «сего месеца» (397), «приписано на» (187); «тургеневъ князю» (502), «князь вяземск» (476)



Уровень 4: пограничная зона

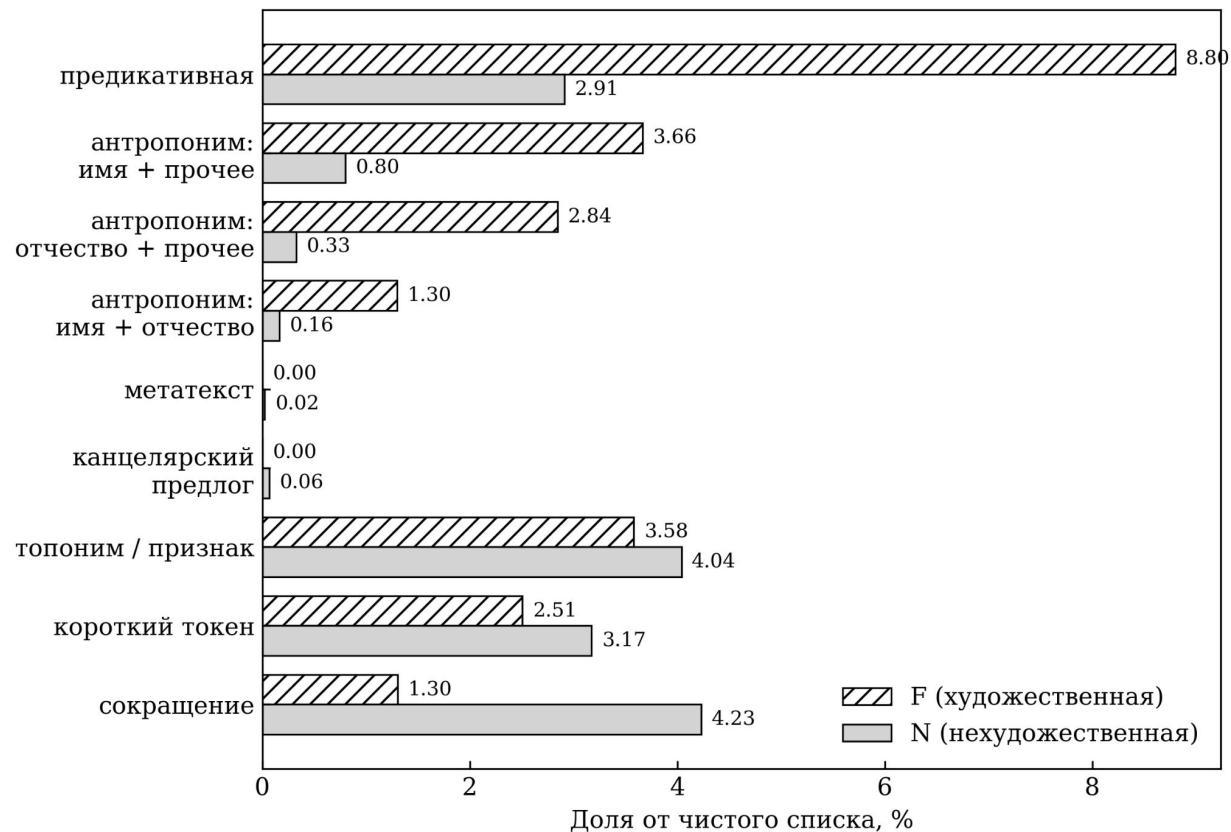
- Всего пограничных: 459 143 → устойчивых (≥ 50) 131 693
- Распределение r : 0,19 % ($r < 0,05$), 9,04 % $[0,05-0,2)$, 39,93 % $[0,2-0,5)$, 41,15 % $[0,5-0,8)$, 9,51 % $[0,8-0,95)$, 0,18 % ($> 0,95$)
- 81,3 % — «широкая середина» ($r \in [0,2, 0,8]$), но она из низкочастотных биграмм
- Устойчивые: 231 с $r > 0,95$ и 247 с $r < 0,05$ — резкая бимодальность





Уровень 4: примеры

- Тяготеют к F: «донъ кихоть» (826/6, $r = 0,993$), «катерина александровна» (689/6, $r = 0,991$), «николай ивановичъ» (2158/31, $r = 0,986$), «господин голядкин» (487/9, $r = 0,982$)
- Тяготеют к F (предикативные): «заговорила она» (496/6, $r = 0,988$), «проговорила она» (1496/25, $r = 0,984$), «подумал он» (1194/23, $r = 0,981$)
- Тяготеют к N: «князю вяземскому» (6/547, $r = 0,011$), «князь кутузов» (6/330, $r = 0,018$), «граф толстой» (8/372, $r = 0,021$)
- Тяготеют к N (канцелярские): «государственном совете» (7/405, $r = 0,017$), «востоку от» (5/363, $r = 0,014$), «прим чан» (10/552, $r = 0,018$)
- Вывод: перетекание — 20:1 ... 140:1, редкое и семантически мотивированное (цитаты, критика, сцены с документами)



Распределение устойчивых пограничных биграмм по доле сферы r



Итоги

- F — коммуникативный режим «персонаж и речь»: антропонимы + предикативные с глаголами речи
- N — режим «документ, факт, метатекст»: сокращения, канцелярские предлоги, даты, метатекст
- Сферная специфика — на лексическом, структурно-грамматическом и формульном уровнях