**Уровень 1.**

**файл "5.2 Визуализация данных"**

Графики:

***Диаграммы-линии или графики:*** Для построения диаграмм-линий применяют прямоугольную систему координат. Обычно по оси абсцисс откладывается время (годы, месяцы и т. д.), а по оси ординат – размеры изображаемых явлений или процессов. На осях наносят масштабы.

***Столбчатые и линейные диаграммы (гистограммы):*** Построение столбчатой диаграммы заключается в изображении статистических данных в виде вертикальных прямоугольников или трёхмерных прямоугольных столбиков. Каждый столбик изображает величину уровня данного статистического ряда. Все сравниваемые показатели выражены одной единицей измерения, поэтому удаётся сравнить статистические показатели данного процесса.

***Круговые (секторные) диаграммы:*** Относительная величина каждого значения изображается в виде сектора круга, площадь которого соответствует вкладу этого значения в сумму значений.

***Радиальные (сетчатые) диаграммы:*** В отличие от линейных диаграмм, в радиальных или сетчатых диаграммах более двух осей. По каждой из них производится отсчёт от начала координат, находящегося в центре. Для каждого типа полученных значений создаётся своя собственная ось, которая исходит из центра диаграммы.

***Таблица:*** организация структуры данных, в которой отдельные элементы помещены в ячейки, каждой из которых сопоставлена пара значений – номер строки и номер колонки.



Wordcloud:

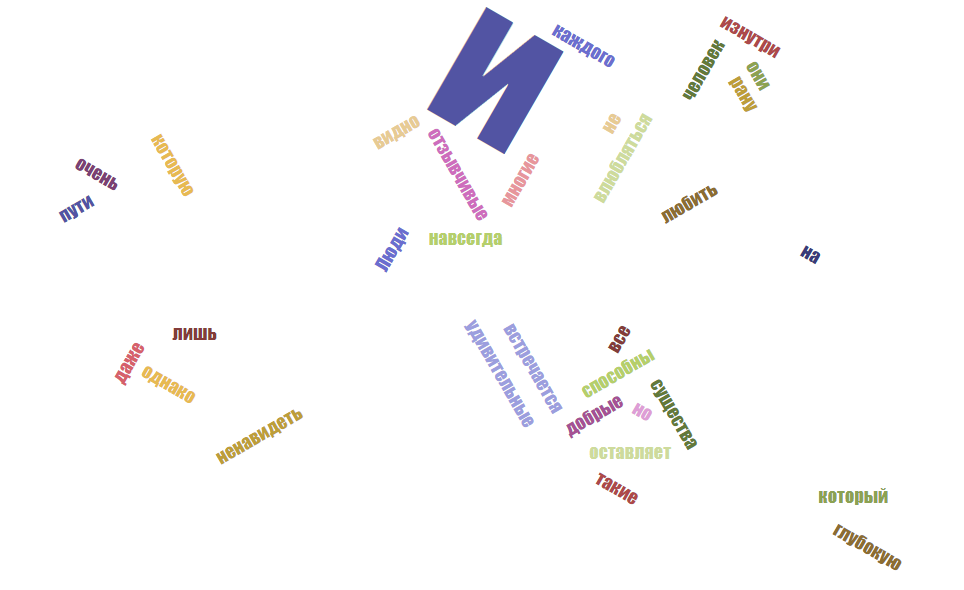


Wordle:  


Word it out:



Word cloud by Jason Davies:



Tagul:

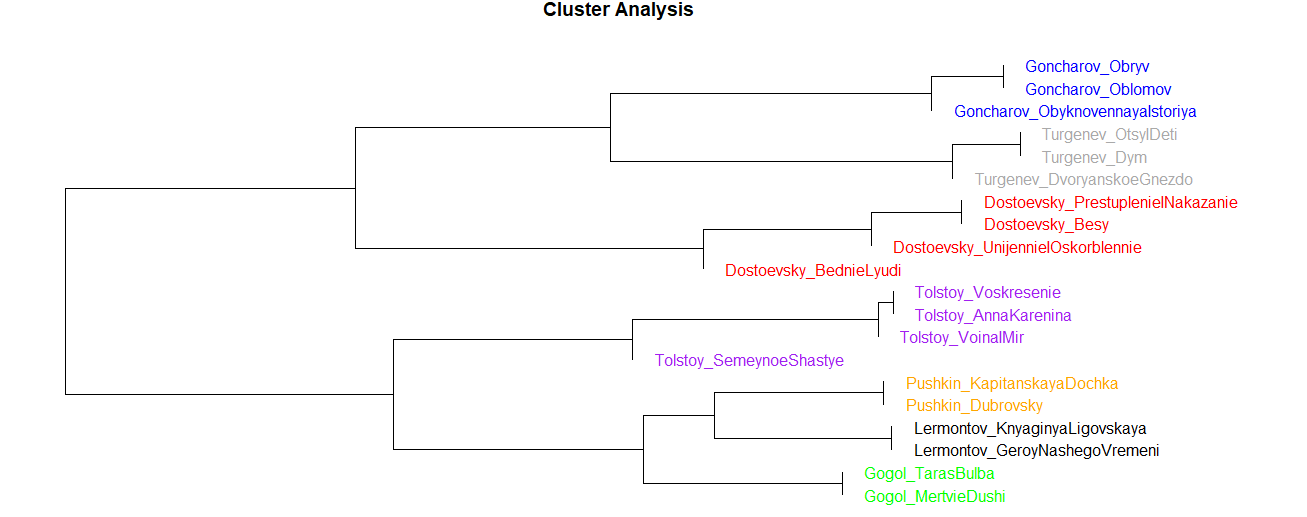


Аналоги:

1. библиотека WordCloud для python
2. <https://wordscloud.pythonanywhere.com/>
3. https://wordcloud.online/?ysclid=lve1mo0d71192325670

**файл "14. Стилеметрия и тематическое моделирование"**

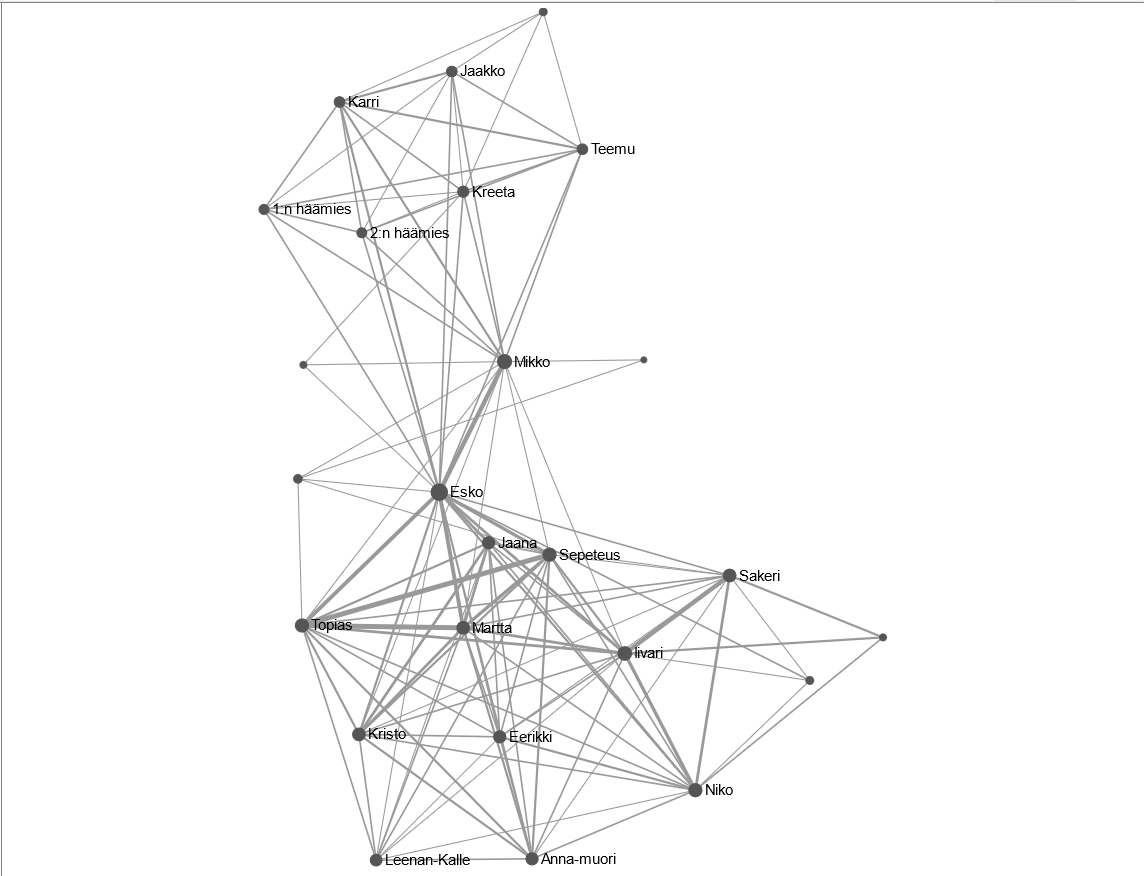
Для своего стилеметрического исследования я взял текста с гугл диска(<https://drive.google.com/drive/folders/1rc0CzioBpZpud7Re8e_XZHCvN9cUhDW5>) из файлов 1\_pushkin\_lermontov\_gogol.zip и 2\_fourteen\_russian\_novels.zip.



как видим авторы из первого файла располагаются рядом с собой, в то время как тексты одного автора из второго не всегда совпадают настолько сильно.

**файл “11 Сетевой анализ”**

Aleksis Kivi: Nummisuutarit (1864)



**Уровень 2.**

1. Ранг: Ранг - это порядковый номер слова по убыванию его частоты в тексте или корпусе. Он позволяет оценить важность слова в контексте анализа. Например, наиболее часто употребляемые слова будут иметь более низкий ранг.

2. Относительные частоты: Относительные частоты показывают долю конкретного слова или явления относительно общего количества слов или явлений в тексте. Они позволяют сравнивать частоту различных элементов и выявлять закономерности.

3. Показатель R (range): Показатель R, или диапазон, представляет разницу между наибольшим и наименьшим значениями в выборке. Он может использоваться для измерения вариации в частоте употребления слов или явлений.

4. Коэффициент Жуйяна: Коэффициент Жуйяна используется для оценки равномерности распределения частоты слов в тексте. Он вычисляется как отношение суммы квадратов частот слов к квадрату суммы всех частот.

5. ARF (Average Rank Frequency): ARF представляет собой средний ранг слова в тексте или корпусе. Эта метрика позволяет оценить общую значимость слова на основе его ранга.

6. Критерий логарифмического правдоподобия: Критерий логарифмического правдоподобия используется для оценки степени соответствия модели частотного распределения данным. Он позволяет оценить точность модели и ее способность предсказывать частоту слов.

7. Критерий Йенсена-Шеннона: Критерий Йенсена-Шеннона используется для измерения разнообразия текста на основе энтропии распределения частот слов. Он позволяет определить степень разнообразия лексического состава текста.

Глоттохронология - это научная дисциплина, занимающаяся реконструкцией исторических языковых связей на основе сравнительного анализа лексики и фонетики. Одним из методов глоттохронологии является метод Сводеша-Старостины, который использует количественные данные для оценки времени разделения языков.

Формула, используемая в методе Сводеша-Старостины, выглядит следующим образом:

T = log(2)/R

Где:

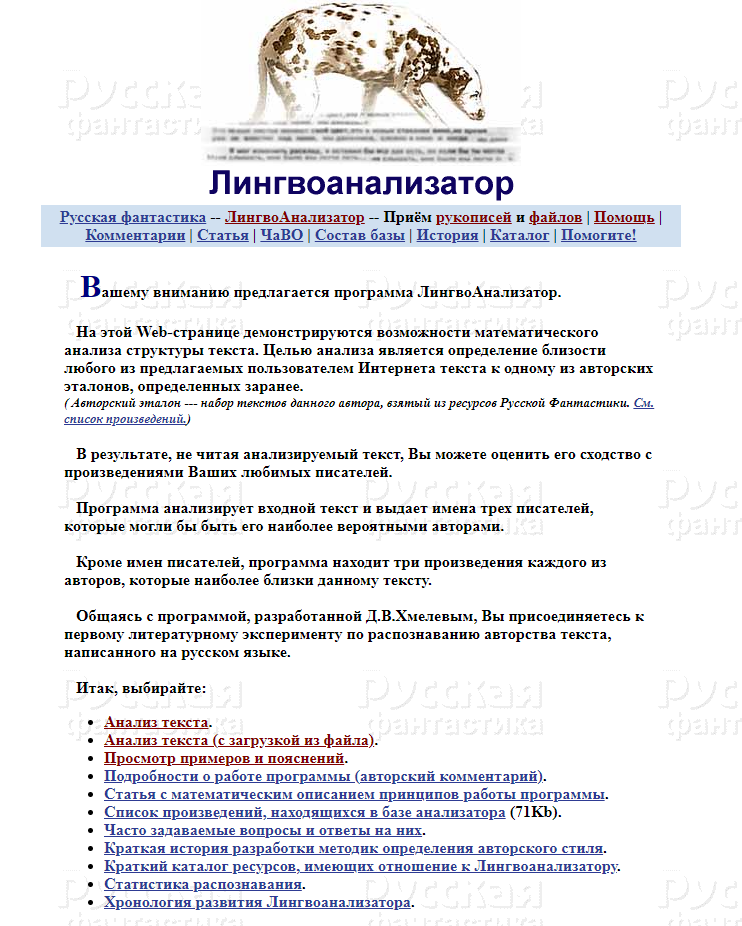
- T - время разделения двух языков (в тысячелетиях);

- R - средняя доля лексики, сохраняющейся в двух языках за определенное время.

Эта формула позволяет оценить примерное время разделения двух языков на основе степени сходства их лексики. Чем выше значение R, тем более близки языки по лексике и тем менее времени прошло с момента их разделения.

Метод Сводеша-Старостины и соответствующие формулы используются для реконструкции исторических языковых связей и определения времени разделения языковых групп.

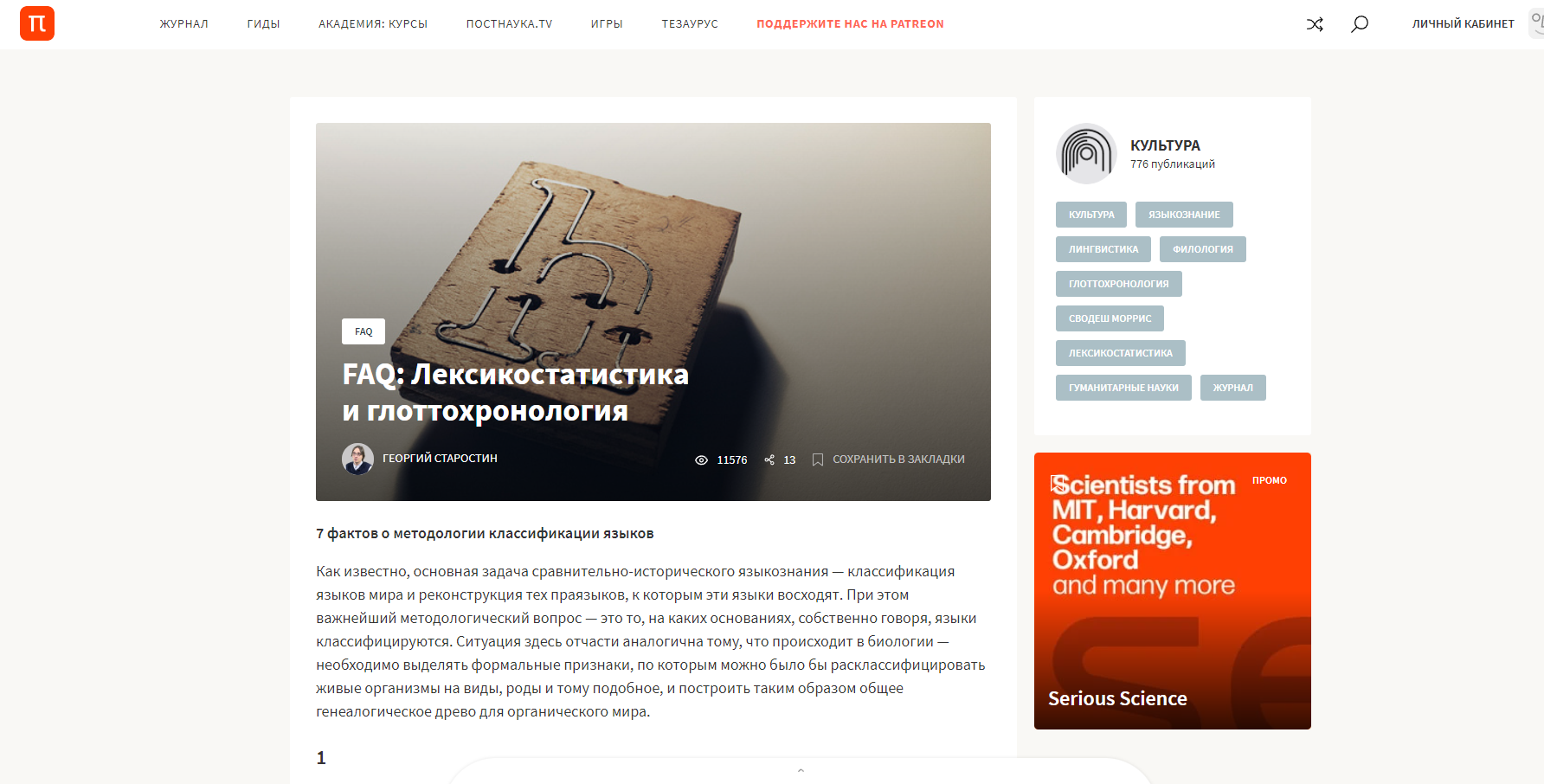
http://www.rusf.ru/books/analysis/ - Этот ресурс представляет собой сайт, посвященный вопросу определения авторства текстов. Автор Д. В. Хмелев разработал лингвоанализатор для определения авторства анонимных или псевдонимных произведений.



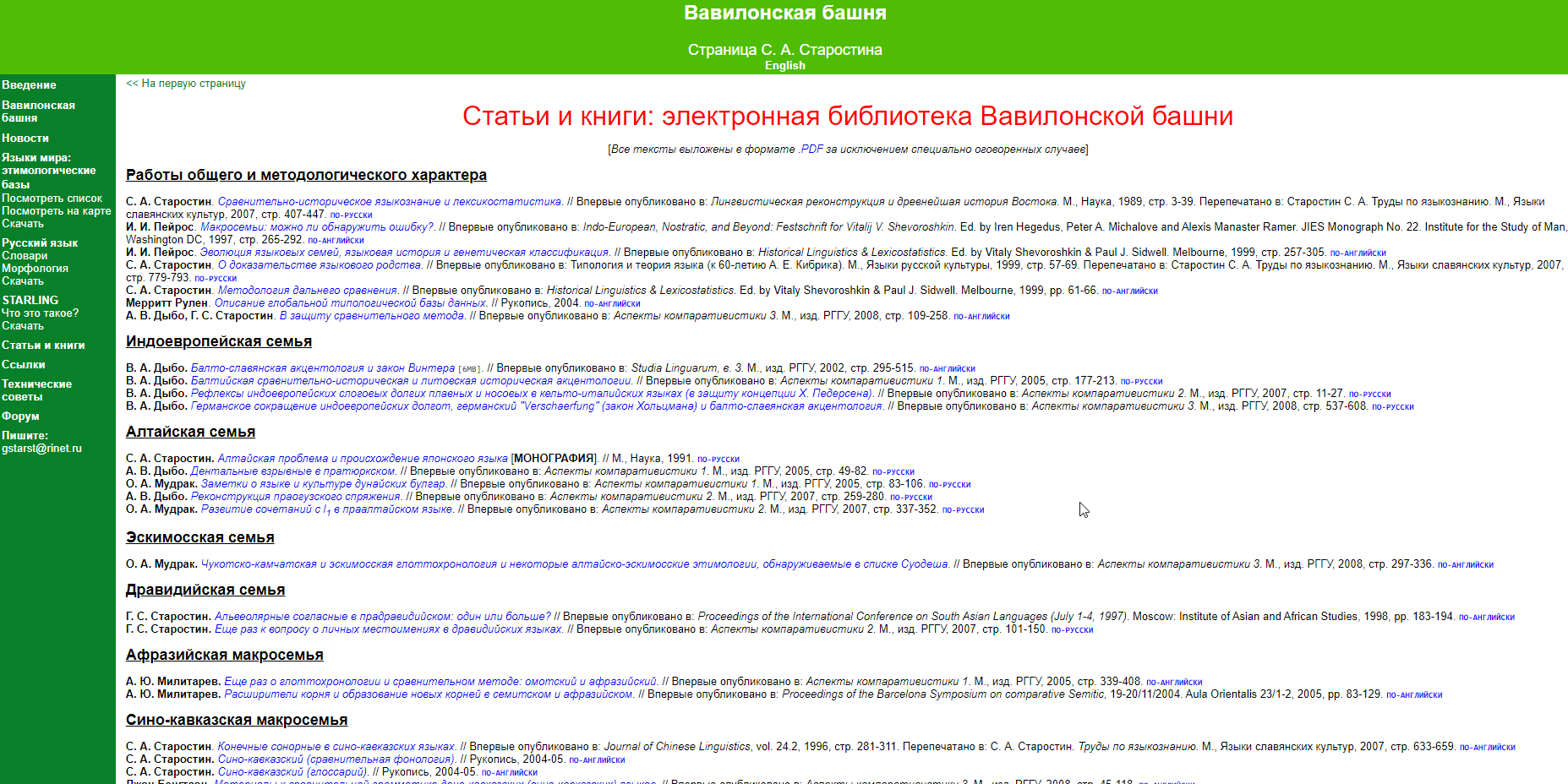
http://attribution.corneille-moliere.com/?p=page19&m=main&l=rus - Этот информационный ресурс также посвящен проблеме определения авторства анонимных и псевдонимных произведений. Он может содержать методы и инструменты для анализа текстов и определения их авторства.



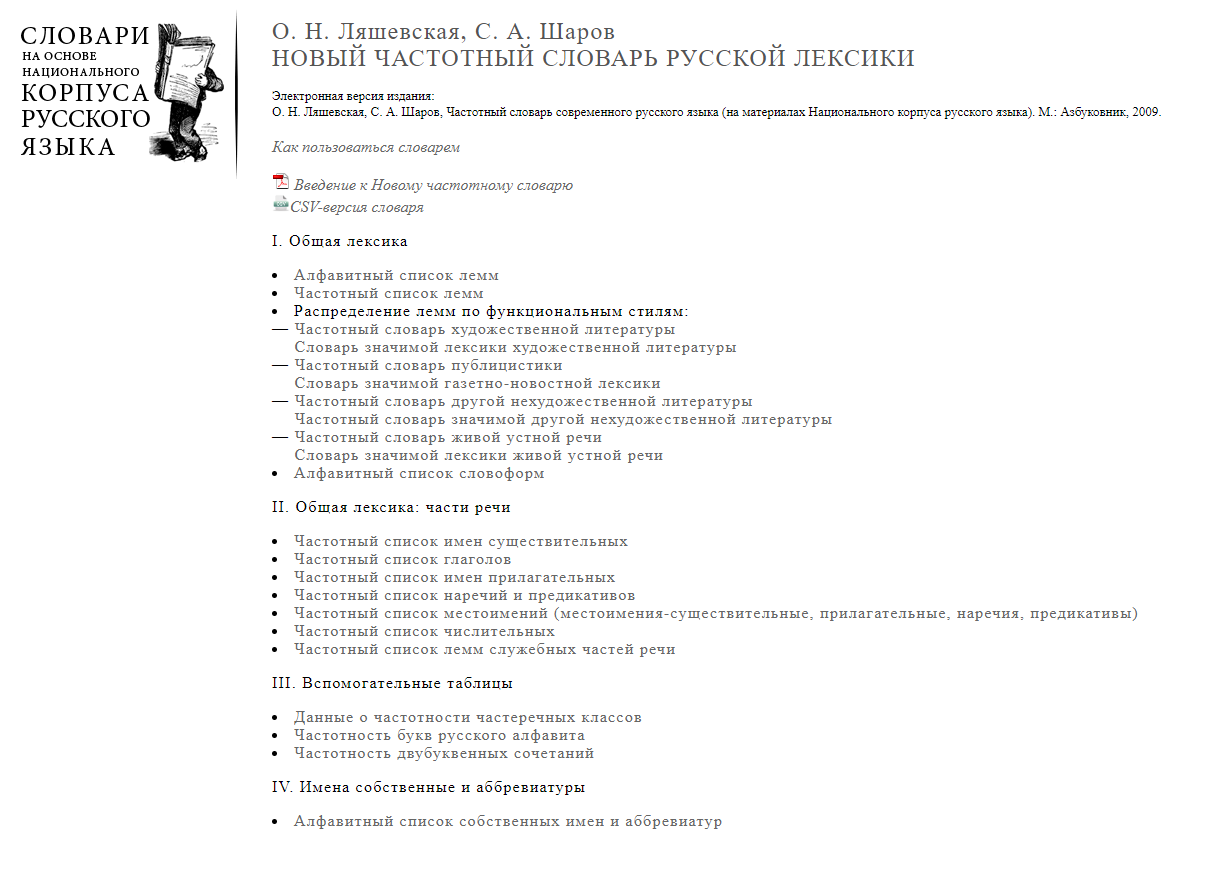
http://postnauka.ru/faq/11053 - Ссылка на статью или материал на портале "ПостНаука" по вопросам авторства или другим лингвистическим темам.



http://starling.rinet.ru/Texts/textsr.htm - Этот ресурс, возможно, содержит тексты на различных языках для анализа и сравнения.



http://dict.ruslang.ru/freq.php - Новейший академический частотный словарь русского языка, который вероятно содержит информацию о частотности употребления слов в русском языке.



http://www.artint.ru/projects/frqlist.php - не открылся  
http://morpher.ru/Russian/Stats.aspx - Частотный словарь, созданный программистом для сравнения с предыдущими двумя. Этот ресурс, вероятно, содержит статистические данные о частотности слов в русском языке и может использоваться для сравнительного анализа.

