Школа экономики
и менеджмента

Санкт-Петербург
2023

Использование Wordsmith

Куганова Дарья
Поторий Полина

WordSmith Tools — это интегрированный набор программ для наблюдения за тем, как слова ведут себя в текстах.

- **WordList**
- **Concord**
- **KeyWords**

Эти инструменты использовались издательством Oxford University Press для собственной лексикографической работы при подготовке словарей преподавателями языков и студентами, а также исследователями, изучающими языковые модели на множестве разных языков во многих странах мира.

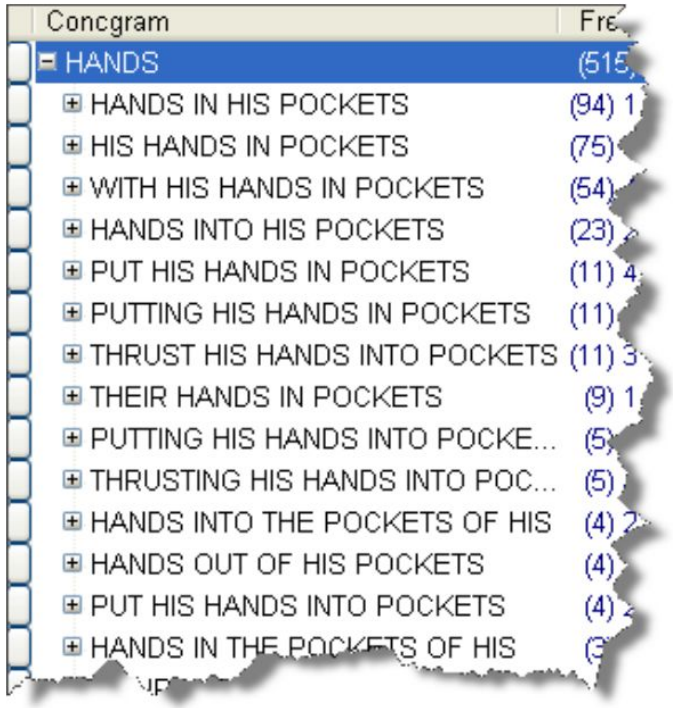
Определение “конграммы”

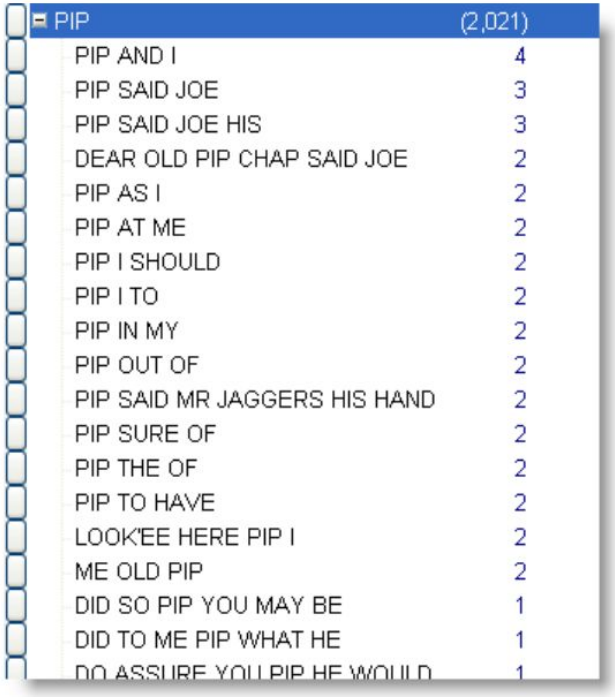
В течение многих лет было легко искать или идентифицировать последовательные кластеры (n-граммы), такие как ***AT THE END OF, MERRY CHRISTMAS*** or ***TERM TIME***

Также было возможно найти непоследовательные связи, такие как ***STRONG***, в пределах горизонтов ***TEA***, адаптировав поиск для поиска контекстных слов

WSConcGram

«Конграмма» — это все перестановки вариаций избирательного округа и позиционных вариаций, порожденные ассоциацией двух или более слов.





Определение “конграммы”

По сути, то, что искали при разработке программы ***ConcGram***, было «поисковой машиной, которая помимо способности обрабатывать вариации избирательного округа

AB, ACB

также обрабатывает позиционные вариации

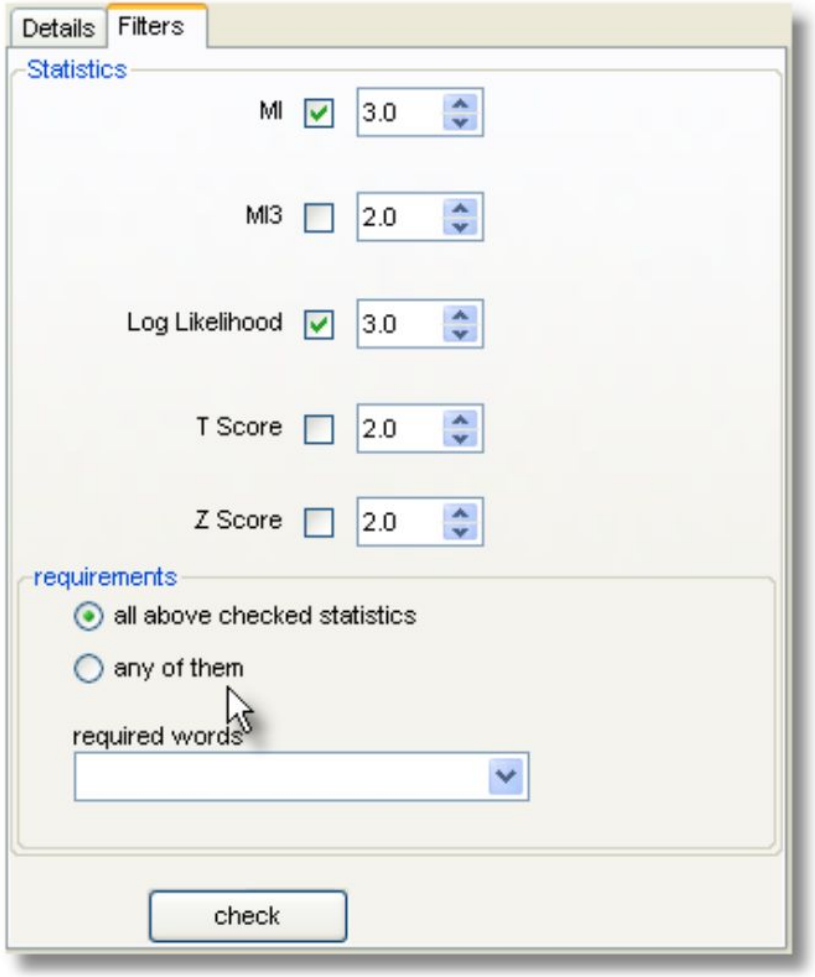
AB, BA

проводит полностью автоматический поиск и поиск словесных ассоциаций любого размера

WSConcGram разработан в честь этой идеи :)

Чтобы выбрать, какие элементы являются «связанными», нам нужны подходящие статистические процедуры.

Настройки фильтрации в контроллере позволяют указать, например, что вы хотите видеть только те, которые связаны с оценкой MI (взаимной информации) или оценкой логарифмического правдоподобия.



Есть ли важная разница между ключевым словом с keyness 50 и покателем keyness 500?

Предположим, вы обрабатываете текст о фермере, выращивающем 3 культуры (пшеницу, овес и нут) и он страдает от 3 проблем (дождь, ветер, засуха). Если каждая из этих культур одинаково важна в тексте, и каждая из трех задач требует объяснения по одному абзацу, читатель-человек может решить, что все три культуры одинаково важны, и все три проблемы одинаково важны. Но в английском языке эти три термина, обозначающего урожай, и термин, обозначающий погоду, сильно различаются по частоте (наименее часто встречаются нут и засуха).

Где ключевые слова распределены в тексте?

1. выполняет согласование текста, находя все вхождения каждого ключевого слова;

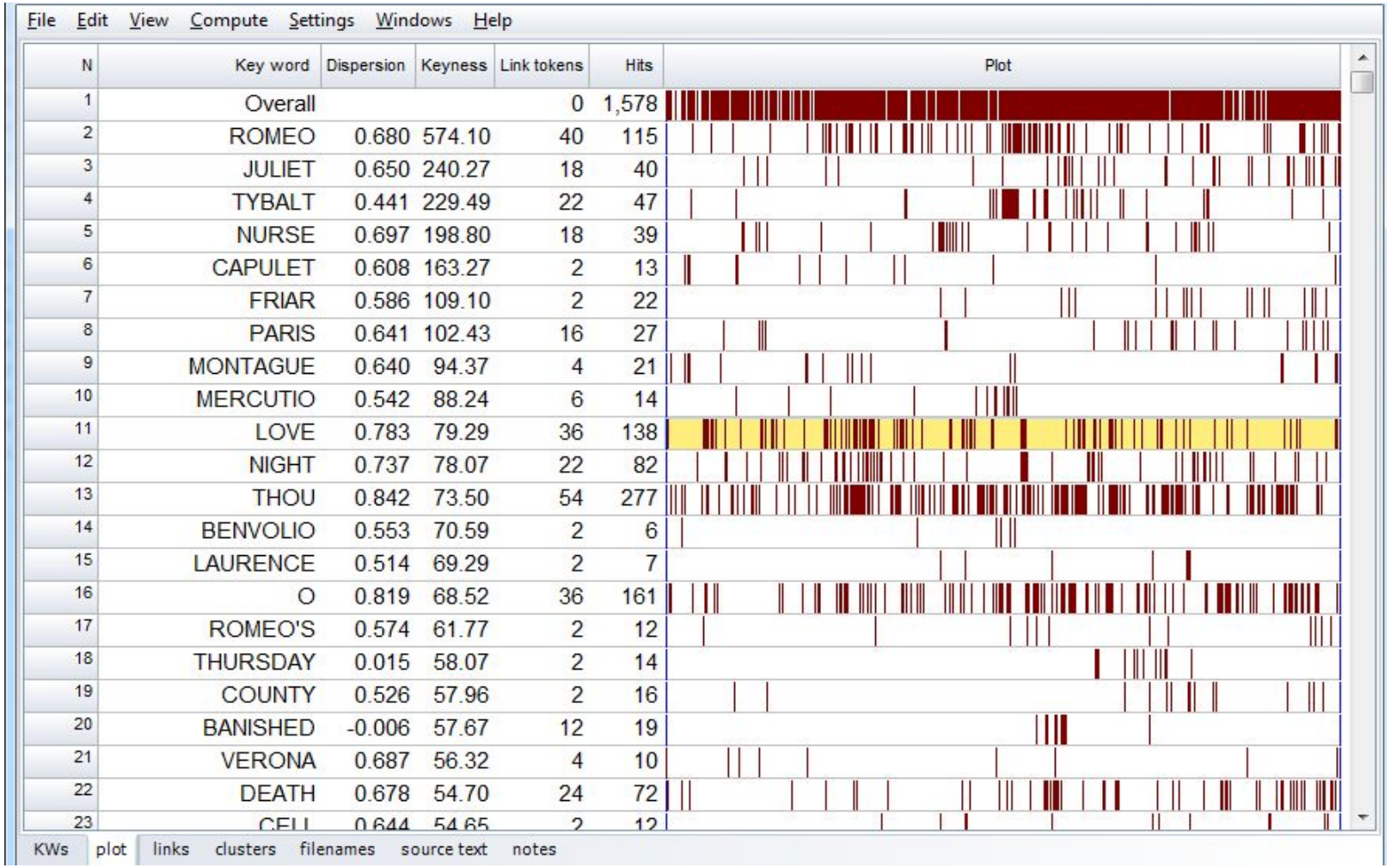
2. Затем он определяет, какие из остальных ключевых слов появляются в пределах горизонтов словосочетаний (установленных в настройках). Он использует больший из двух горизонтов.

3. Затем он отображает все слова, показывающие, где каждое вхождение встречается в исходном файле (с «линейкой», показывающей, сколько слов есть в каждой части файла).

4. вычисляет, сколько других ключевых слов встречается вместе с ним в пределах текущего коллокационного диапазона.

5. вычисляет значение дисперсии графика

KW пьесы «Ромео и Джульетта», показывающие, где встречается каждый термин



File Edit View Compute Settings Windows Help

N	Key word	Dispersion	Keyness	Link tokens	Hits	Plot
1	Overall			0	1,578	
2	ROMEO	0.680	574.10	40	115	
3	JULIET	0.650	240.27	18	40	
4	TYBALT	0.441	229.49	22	47	
5	NURSE	0.697	198.80	18	39	
6	CAPULET	0.608	163.27	2	13	
7	FRIAR	0.586	109.10	2	22	
8	PARIS	0.641	102.43	16	27	
9	MONTAGUE	0.640	94.37	4	21	
10	MERCUTIO	0.542	88.24	6	14	
11	LOVE	0.783	79.29	36	138	
12	NIGHT	0.737	78.07	22	82	
13	THOU	0.842	73.50	54	277	
14	BENVOLIO	0.553	70.59	2	6	
15	LAURENCE	0.514	69.29	2	7	
16	O	0.819	68.52	36	161	
17	ROMEO'S	0.574	61.77	2	12	
18	THURSDAY	0.015	58.07	2	14	
19	COUNTY	0.526	57.96	2	16	
20	BANISHED	-0.006	57.67	12	19	
21	VERONA	0.687	56.32	4	10	
22	DEATH	0.678	54.70	24	72	
23	CELL	0.644	54.65	2	12	

KWs plot links clusters filenames source text notes

Text dispersion

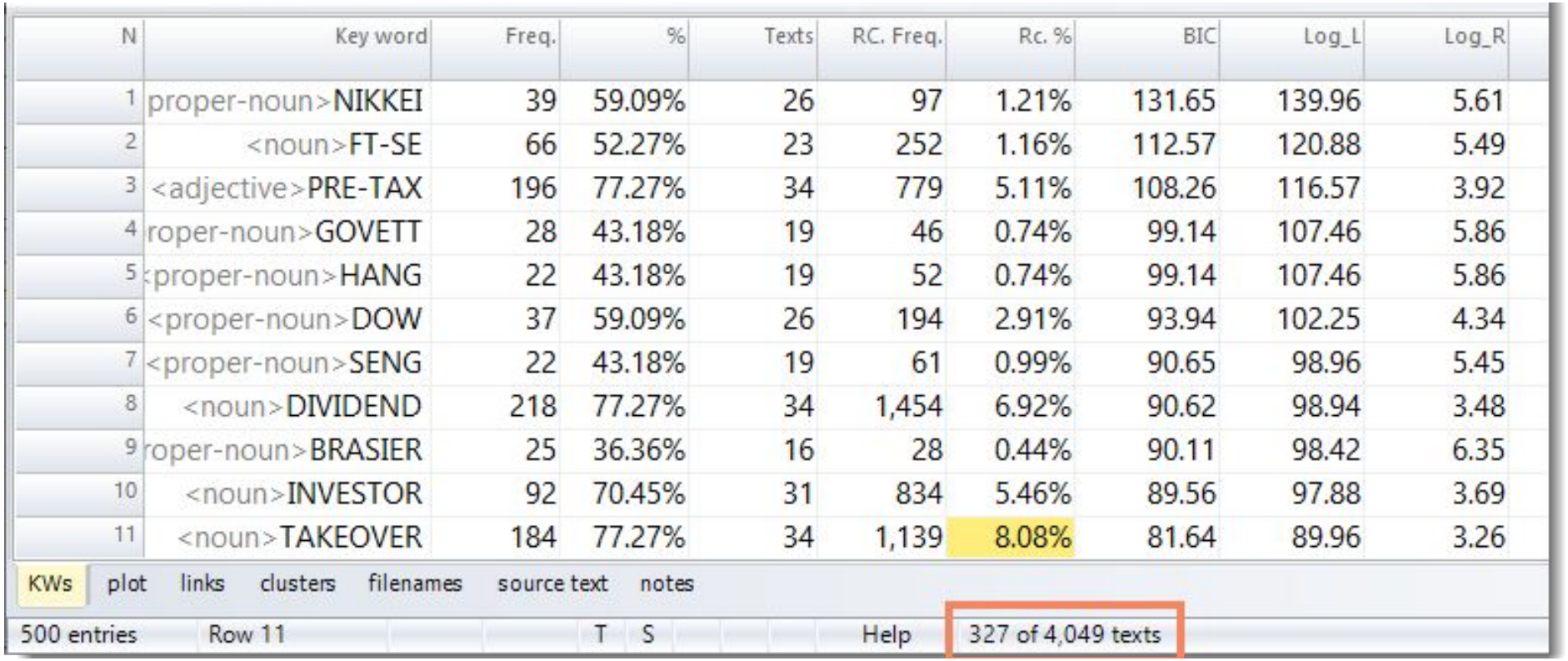
Было рассчитано с использованием 24 письменных академических текстов BNC по медицине. Справочным корпусом была вся версия BNC XML.

N	Key word	Freq.	%	Texts	RC. Freq.	Rc. %	BIC	Log_L	Log_R
1	E	273	91.67%	22	4	0.05%	203.84	212.16	10.86
2	<adjective>RANDOMISED	164	87.50%	21	181	0.69%	140.73	149.04	6.98
3	<adjective>MYOCARDIAL	212	83.33%	20	238	0.64%	134.37	142.69	7.02
4	<adjective>HISTOLOGICAL	378	83.33%	20	399	0.69%	132.18	140.49	6.91
5	<noun>INFARCTION	237	79.17%	19	269	0.62%	126.90	135.21	7.00
6	MMOL	350	79.17%	19	370	0.67%	124.73	133.04	6.89
7	<adjective>RENAL	419	91.67%	22	478	1.46%	123.54	131.85	5.98
8	<adjective>POSTOPERATIVE	82	79.17%	19	101	0.74%	121.70	130.01	6.74
9	<noun>MORBIDITY	173	91.67%	22	365	1.63%	119.40	127.71	5.81
10	<noun>OUTPATIENT	197	91.67%	22	287	1.75%	116.67	124.98	5.71
11	<adjective>ASYMPTOMATIC	109	75.00%	18	129	0.67%	116.26	124.58	6.81
12	<adjective>ISCHAEMIC	63	70.83%	17	79	0.52%	114.24	122.55	7.09

Text dispersion

Приведенные ниже результаты были получены из 44 коммерческих текстов:

TAKEOVER был обнаружен в 34 из 44 коммерческих текстов (77,27%), а 8,08% составляют 327 из 4049 справочных корпусов, как показано в status bar.



N	Key word	Freq.	%	Texts	RC. Freq.	Rc. %	BIC	Log_L	Log_R
1	proper-noun>NIKKEI	39	59.09%	26	97	1.21%	131.65	139.96	5.61
2	<noun>FT-SE	66	52.27%	23	252	1.16%	112.57	120.88	5.49
3	<adjective>PRE-TAX	196	77.27%	34	779	5.11%	108.26	116.57	3.92
4	roper-noun>GOVETT	28	43.18%	19	46	0.74%	99.14	107.46	5.86
5	<proper-noun>HANG	22	43.18%	19	52	0.74%	99.14	107.46	5.86
6	<proper-noun>DOW	37	59.09%	26	194	2.91%	93.94	102.25	4.34
7	<proper-noun>SENG	22	43.18%	19	61	0.99%	90.65	98.96	5.45
8	<noun>DIVIDEND	218	77.27%	34	1,454	6.92%	90.62	98.94	3.48
9	roper-noun>BRASIER	25	36.36%	16	28	0.44%	90.11	98.42	6.35
10	<noun>INVESTOR	92	70.45%	31	834	5.46%	89.56	97.88	3.69
11	<noun>TAKEOVER	184	77.27%	34	1,139	8.08%	81.64	89.96	3.26

KWs | plot | links | clusters | filenames | source text | notes

500 entries | Row 11 | T | S | Help | 327 of 4,049 texts

Возможности работы со стилем

1. Шаблоны
2. Команды форматирования
3. Автоформатирование
4. Функция стилизации
5. Редактирование стилей

Некоторые функции WordSmith

1. Формат
2. Стили
3. Шрифты
4. Абзацы
5. Таблицы стилей
6. Темы оформления


ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ •