Захаров В.П. , Хохлова М.В.

Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке

- Санкт-Петербургский государственный университет
- *vz1311@yandex.ru, khokhlova.marie@gmail.com*

Аннотация

- Коллокации как устойчивые сочетания.
- Их роль в лексикографии.
- В докладе описаны результаты исследования по выявлению устойчивых сочетаний в русском языке на основе статистических методов на базе корпусов текстов.
- Цель – изучить возможности автоматических методов извлечения коллокаций, сравнить наиболее популярные меры ассоциации.
- Рассматриваются требования к программному обеспечению.

Понятие коллокации в лингвистике

- В широком смысле это комбинация двух или более слов, имеющих тенденцию к совместной встречаемости. Под коллокациями понимаются характерные, часто встречающиеся сочетания слов, *«появление которых рядом друг с другом основывается на регулярном характере взаимного ожидания и задается не грамматическими, а чисто семантическими факторами»* [Firth 1957].

- Традиционные подходы к описанию понятия коллокации в целом можно свести к следующим:
  - подход, берущий начало в работах британских контекстуалистов (Firth 1957; Firth 1968);
  - семантико-синтаксический подход (Телия 1996; Cowie 1978; Hausmann 1979; Hausmann 1985 и др.);
  - подход в рамках теории «Смысл<–>Текст» (Мельчук 1974; Иорданская, Мельчук 2007).

Семантико-синтаксический подход

- Коллокации рассматриваются как подкласс более обширного класса несвободных словосочетаний, или фразем.
- Коллокацией называется словосочетание, в котором одно из слов является семантической доминантой, а второе выбирается в зависимости от него для передачи смысла всего выражения. Одним из ключевых свойств коллокаций является невозможность предсказания таких сочетаний **на основе значений входящих в них компонентов** (Телия, Мельчук, Борисова и др.).
- Коллокация – отношение между отдельными лексическими элементами **в пределах синтаксической единицы** (*The Concise Oxford Dictionary of Linguistics*).
- Терминологические словосочетания
- Прагматемы

Статистический подход

- Коллокация – это привычное, традиционное сочетание слов в речи, звучащее правильно, естественно для носителей языка.
- Характерные, часто встречающиеся сочетания слов, появление которых рядом друг с другом основывается на регулярном характере взаимного ожидания.
- Показатель: частота совместной встречаемости

Меры ассоциации

- Показатели силы синтагматической связи между элементами словосочетаний.
- Исходные данные: частота совместной встречаемости, частоты слов или словоформ (node – ключевое слово, collocate – слово, встречающееся слева или справа от ключевого, коллокат).
- Меры ассоциации: MI (mutual information), t-score, z-score, log-likelihood, Odds, Dice, X^2… (см. http://collocations.de)
- Корпуса как источники достоверных данных о частотах.

Мера MI (mutual information, взаимной информации)

$$MI = \log_2 \frac{f(n,c) \times N}{f(n) \times f(c)}$$

где MI = mutual information;

n – ключевое слово;

c – коллокат;

f(n,c) – частота встречаемости ключевого слова *n* в паре с коллокатом *c*;

f(n), f(c) – абсолютные (независимые) частоты ключевого слова *n* и слова *c* в корпусе (тексте);

N – общее число словоформ в корпусе (тексте).

Мера t-score

$$t-score=\frac{f(n,c)-\frac{f(n)\times f(c)}{N}}{\sqrt{f(n,c)}}$$

где

n – ключевое слово;

c – коллокат;

f(n,c) – частота встречаемости ключевого слова *n* в паре с коллокатом *c*;

f(n), f(c) – абсолютные (независимые) частоты ключевого слова *n* и слова *c* в корпусе (тексте);

N – общее число словоформ в корпусе (тексте).

Log-likelihood

$$\log\text{–}likelihood = 2\sum f(n,c)\times\log_2\frac{f(n,c)\times N}{f(n)\times f(c)}$$

где
n – ключевое слово;
c – коллокат;
f(n,c) – частота встречаемости ключевого слова *n* в паре с коллокатом *c*;
f(n), f(c) – абсолютные (независимые) частоты ключевого слова *n* и слова *c* в корпусе (тексте);
N – общее число словоформ в корпусе (тексте).

База исследования

- Данные:
  существительные: *власть, внимание, возможность, война, вопрос, дождь, жизнь, закон, любовь, место, мнение, мысль, ночь, ответ, помощь, радость, слово, случай, смысл;*
  глаголы: *быть, сказать, мочь, говорить, знать, стать, есть, хотеть, видеть, идти*
- Инструменты: сервис на сайте университета г. Лидс (автор С.А. Шаров), включающий различные корпуса русского языка и различные меры ассоциации (http://corpus.leeds.ac.uk/ruscorpora.html),
  сервис биграмм на сайте АОТ (http://aot.ru/demo/bigrams.html)
- Словари русского языка

Результаты для глагола «говорить» (левый контекст) (модель Adv+V), отсортированных по мере MI

Collocation	Joint	Freq1	Rank MI	MI score (7,08-2,14)	Rank LL	LL score (1064,06-2,96)	Rank T-score	T- score (22,79-1,96)
честно говорить	527	2339	**1**	7,08	**1**	1064,06	**101**	1,96
постоянно говорить	62	4158	**2**	7,04	**14**	40,59	**85**	2,26
условно говорить	90	585	**3**	6,53	**8**	162,73	**91**	2,11
обиженно говорить	5	208	**4**	6,46	**77**	4,37	**16**	6,52
грубо говорить	130	988	**5**	6,30	**6**	224,23	**93**	2,10
умело говорить	23	2034	**6**	6,20	**33**	12,26	**70**	2,40
откровенно говорить	139	1203	**7**	6,12	**5**	230,24	**94**	2,09
собственно говорить	333	3114	**8**	6,00	**2**	538,32	**99**	1,97

Часть таблицы результатов для глагола «говорить» (левый контекст) (модель Adv+N), отсортированная по мере t-score

Collocation	Joint	Freq1	Rank MI	MI score (7,08-2,14)	Rank LL	LL score (1064,06-2,96)	Rank T-score	T- score (22,79-1,96)
умоляюще говорить	4	85	**15**	4,82	**65**	4,80	**1**	22,79
Примири-тельно говорить	4	111	**23**	4,43	**80**	4,27	**2**	17,96
скупо говорить	4	138	**31**	4,12	**86**	3,85	**4**	15,46
…	…	…	**…**	…	**…**	…	**…**	…
восхищенно говорить	9	149	**69**	2,91	**34**	11,91	**39**	3,24
убедительно говорить	10	502	**17**	4,76	**47**	7,87	**44**	3,01
неуверенно говорить	10	615	**53**	3,25	**49**	6,94	**45**	3,01
смело говорить	10	782	**58**	3,09	**56**	5,87	**46**	2,99
собственно говорить	333	3114	**8**	6,00	**2**	538,32	**99**	1,97

Частотные данные и меры ассоциации для глагола «*говорить*» (первое значение для леммы /*второе значение курсивом* для формы деепричастия).

Collocation	MI score	LL score	T score
искренне говоря	2,94/***4.92***	4,49/***6.11***	2,74/***2.16***
точно говоря	2,64/***5.29***	21,09/***55.31***	2,21/***6.24***
просто говоря	2,19/***5.60***	79,38/***209.98***	2,02/***11.75***
откровенно говоря	6,12/***9.67***	230,24/***299.54***	2,09/***10.19***
честно говоря	7,08/***10.98***	1064,06/***1690.55***	1,96/***22.33***
объективно говоря	4,24/***6.82***	4,37/***11.22***	4,16/***2.43***
образно говоря	3,00/***10.80***	102,07/***145.01***	2,32/***6.63***
строго говоря	4,55/***8.34***	184,16/***351.80***	2,08/***12.05***

Сравнение рангов коллокаций для глагола «*говорить*» (модель Adv+N), полученных по мере MI и вычисленных на основе двух корпусов русского языка (НКРЯ (117 млн. словоупотреблений) и газетный корпус (70 млн.))

- Анализ коллокаций, полученных на этих двух корпусах, показывает, что грубо их можно разбить на две части: присутствующие в обоих корпусах (часто с близкими рангами) и присутствующие только в одном из них. Видимо, это говорит о принадлежности коллокатов, в данном случае, наречий, выданных только по одному из корпусов, к определенному жанру. И действительно, анализ контекстов употребления наречий *пространно, модно, полусерьзно, фигурально* в корпусе показывает преобладание художественных текстов.

- Но еще более разительную картину дает сравнений коллокаций, полученных на основе НКРЯ (117 млн. словоупотреблений) [1] и Интернет-корпус (188 млн.), где из 13 первых коллокаций из НКРЯ, отсортированных по мере MI, в последнем присутствует только одна.

Анализ

- Ранги коллокаций, полученных на основе разных мер, не совпадают.
- Иногда статистические меры для поиска коллокаций следует применять к словоформам, а не к леммам.
- Зависимость состава и ранжирования списков коллокаций от типа корпуса; для разных жанров, возможно, следует применять разные меры.
- Разные меры по-разному реагируют на частоту слов, образующих коллокацию, и на частоту совместной встречаемости:
  MI – низкочастотные слова;
  T-score – высокочастотные сочетания.

Сравнение коллокаций, полученных автоматически на основе разных мер ассоциации, с данными различных словарей

- Материалом послужили коллокации 19 вышеперечисленных существительных
- Исследование проводилось на базе газетного корпуса на сайте Ун-та Лидс (78 млн. слов).
- Результаты запроса для каждого существительного сравнивались со словарными статьями, приведенными для этих существительных в Словаре коллокаций (Борисова 1995а), в толковых словарях русского языка: БАС-17 (Словарь современного русского языка 1948-1965) и МАС (Словарь русского языка 1981-1984) – и в Словаре синонимов и сходных по смыслу выражений (Абрамов 2006).
  Приведем некоторые результаты для слова *война*.

Значения мер ассоциации для коллокаций со словом *«война»*, совпавших с коллокациями словаря Е.Г. Борисовой

Collocation	Joint	Freq1	LL score	MI	T-score
вспыхивать война	5	1201	***8,29***	***6,20***	***2,21***
идти война	167	47464	***264,43***	***5,96***	***12,72***
кровопролитный война	6	251	***15,18***	***8,72***	***2,44***
разражаться война	9	881	***18,94***	***7,50***	***2,98***

Мера MI

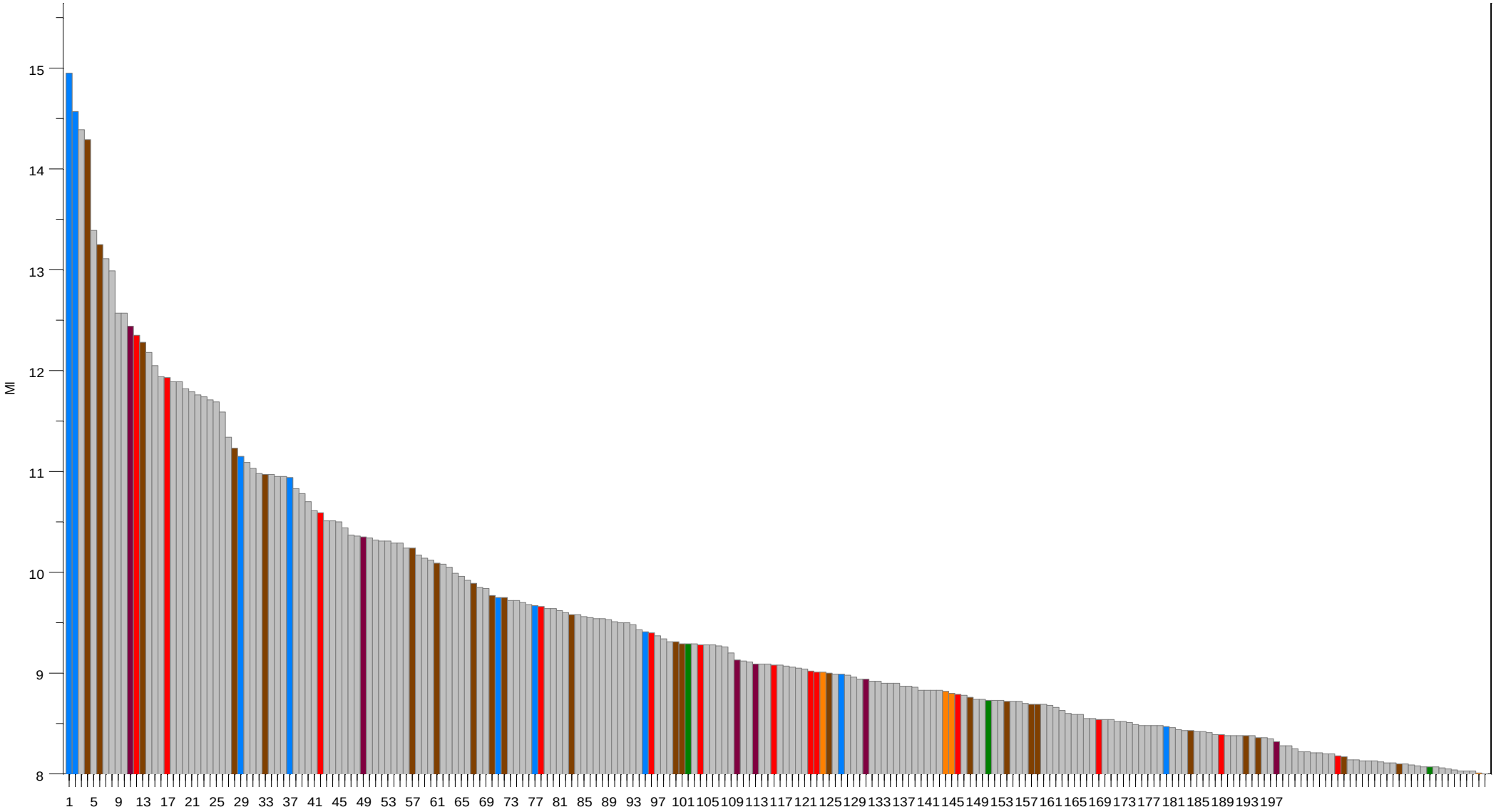
- Всего было найдено 1755.
  Из них:
- 68 присутствуют в двух или более словарях;
- 73 только в словаре Борисовой;
- 27 только в словаре МАС;
- 13 только в словаре синонимов;
- 9 только в словаре БАС (нов.);
- 25 только в словаре БАС (ст.);

- Значения меры MI оказались наибольшими для коллокаций, найденных только в МАС, а также найденных в двух или более словарях.

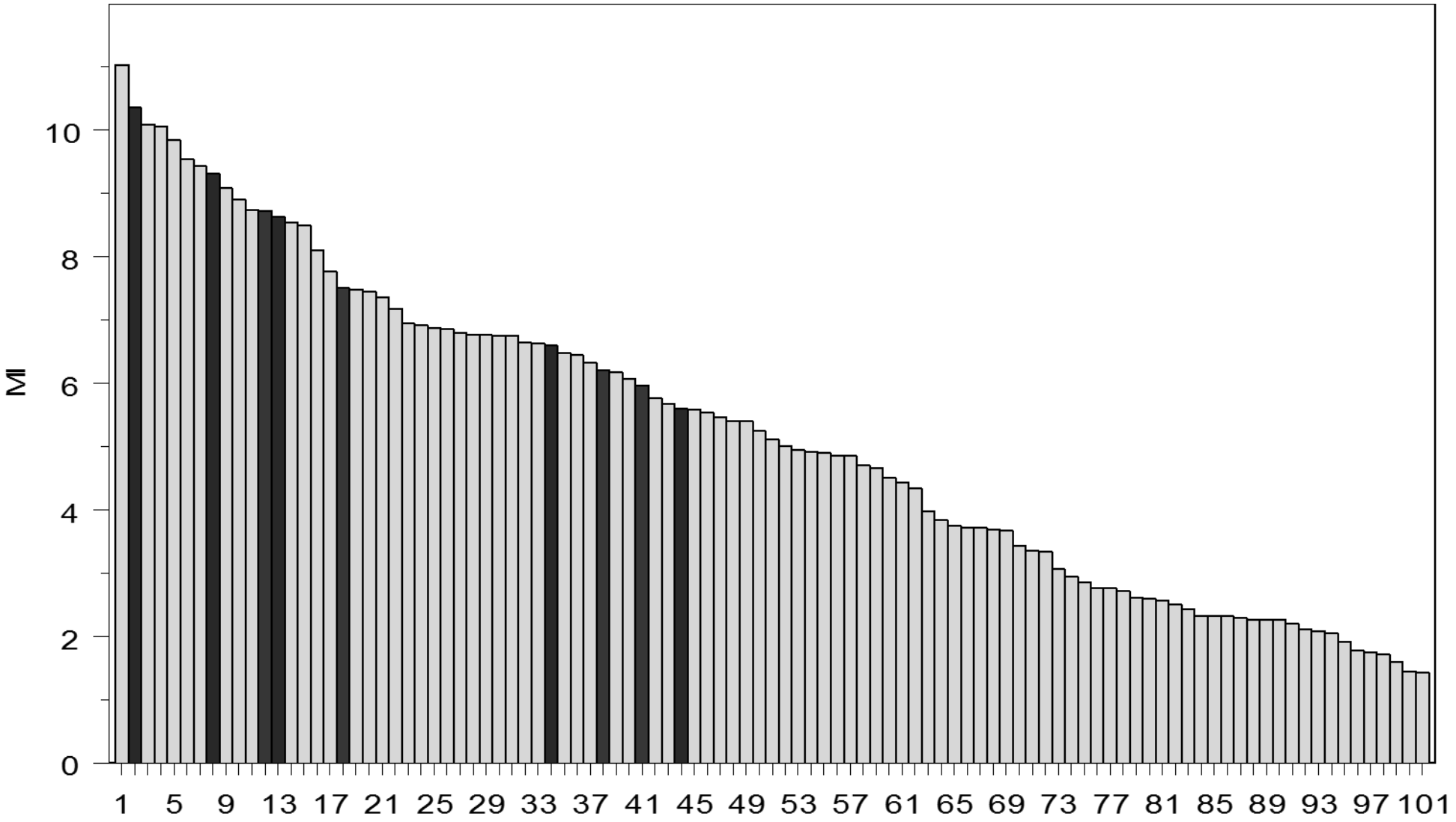
Графики

- На представленном ниже графике следующими цветами обозначены:
- красным – сочетания, зафиксированные только в словаре коллокаций Е.Г. Борисовой;
- синим – сочетания, зафиксированные только в МАС;
- зеленым – сочетания, зафиксированные только в БАС (нов. изд.);
- оранжевым – сочетания, зафиксированные только в БАС (ст. изд.);
- темно-фиолетовым – сочетания, зафиксированные только в словаре синонимов;
- коричневым – сочетания, зафиксированные по крайней мере в двух из вышеперечисленных словарей;
- серым – прочие сочетания, выделенные согласно рассматриваемой мере.
- Ось ординат – значения рассматриваемой меры, ось абсцисс – ранги выделенных согласно этой мере сочетаний (коллокаций).

Mepa MI


MI
15
14
13
12
11
10
9
8
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93 97 101 105 109 113 117 121 125 129 133 137 141 145 149 153 157 161 165 169 173 177 181 185 189 193 197

MI: БАС-17 (ранги 2, 8, 13, 34, 44), Борисова (12, 18, 38, 41)



Коллигации

- Коллигации – это коллокации с учетом грамматических отношений между элементами коллокаций

- ***Colligatio****n* is a type of collocation, **but where a lexical item is linked to a grammatical one.** *Surprising*, *amazing* and *astonishing* are nearly synonymous. We can say *it is astonishing/surprising/amazing*, but we tend to say *it is not surprising* and not the others- *surprising* colligates with the negative.

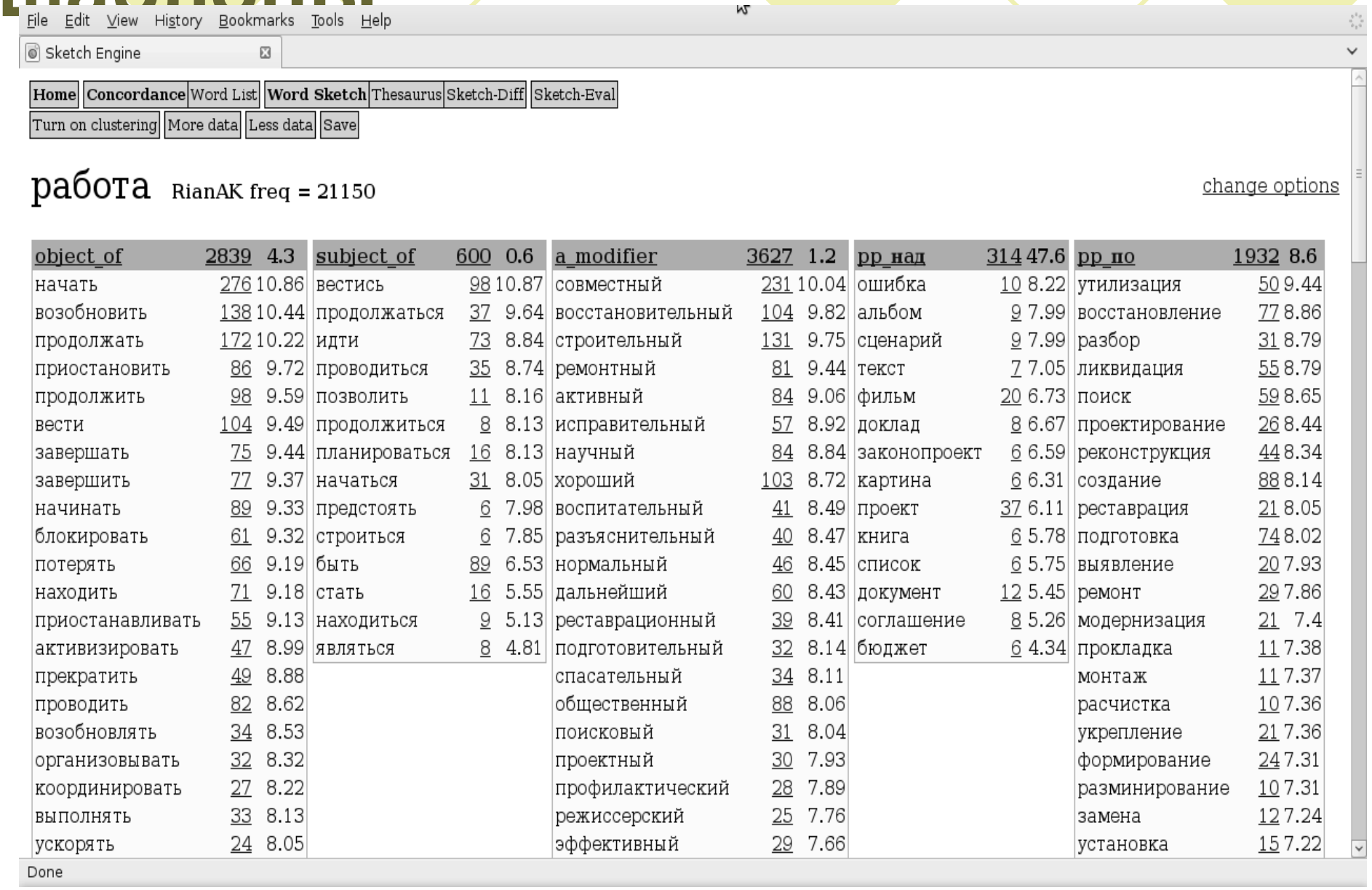
- Примеры формул:
  V + Adv
  V + N
  Adv + V
  V + V

Sketch Engine

- Типичные словосочетания:
- синтаксис, накладывающий ограничение на сочетаемость слов в языке;
- вероятностные закономерности.

- Лексикографическая система:
- Oxford University Press;
- Cambridge University Press;
- Collins, Macmillan.

Лексико-синтаксические шаблоны



File Edit View History Bookmarks Tools Help

Sketch Engine

Home | Concordance | Word List | Word Sketch | Thesaurus | Sketch-Diff | Sketch-Eval

Turn on clustering | More data | Less data | Save

работа RianAK freq = 21150

change options

object_of	2839	4.3
начать	276	10.86
возобновить	138	10.44
продолжать	172	10.22
приостановить	86	9.72
продолжить	98	9.59
вести	104	9.49
завершать	75	9.44
завершить	77	9.37
начинать	89	9.33
блокировать	61	9.32
потерять	66	9.19
находить	71	9.18
приостанавливать	55	9.13
активизировать	47	8.99
прекратить	49	8.88
проводить	82	8.62
возобновлять	34	8.53
организовывать	32	8.32
координировать	27	8.22
выполнять	33	8.13
ускорять	24	8.05

subject_of	600	0.6
вестись	98	10.87
продолжаться	37	9.64
идти	73	8.84
проводиться	35	8.74
позволить	11	8.16
продолжиться	8	8.13
планироваться	16	8.13
начаться	31	8.05
предстоять	6	7.98
строиться	6	7.85
быть	89	6.53
стать	16	5.55
находиться	9	5.13
являться	8	4.81

a_modifier	3627	1.2
совместный	231	10.04
восстановительный	104	9.82
строительный	131	9.75
ремонтный	81	9.44
активный	84	9.06
исправительный	57	8.92
научный	84	8.84
хороший	103	8.72
воспитательный	41	8.49
разъяснительный	40	8.47
нормальный	46	8.45
дальнейший	60	8.43
реставрационный	39	8.41
подготовительный	32	8.14
спасательный	34	8.11
общественный	88	8.06
поисковый	31	8.04
проектный	30	7.93
профилактический	28	7.89
режиссерский	25	7.76
эффективный	29	7.66

pp_над	314	47.6
ошибка	10	8.22
альбом	9	7.99
сценарий	9	7.99
текст	7	7.05
фильм	20	6.73
доклад	8	6.67
законопроект	6	6.59
картина	6	6.31
проект	37	6.11
книга	6	5.78
список	6	5.75
документ	12	5.45
соглашение	8	5.26
бюджет	6	4.34

pp_по	1932	8.6
утилизация	50	9.44
восстановление	77	8.86
разбор	31	8.79
ликвидация	55	8.79
поиск	59	8.65
проектирование	26	8.44
реконструкция	44	8.34
создание	88	8.14
реставрация	21	8.05
подготовка	74	8.02
выявление	20	7.93
ремонт	29	7.86
модернизация	21	7.4
прокладка	11	7.38
монтаж	11	7.37
расчистка	10	7.36
укрепление	21	7.36
формирование	24	7.31
разминирование	10	7.31
замена	12	7.24
установка	15	7.22

Done

Выводы (1)

- Сравнении со словарями:
  тенденция: чем меньше значение меры, тем больше вероятность, что эти словосочетания не зафиксированы как устойчивые. Т.е., статистические меры ассоциации достаточно хорошо выявляют реально существующие семантико-синтагматические связи.
- Высокоранговые коллокации являются кандидатами на включение.
- Анализ различных мер ассоциации: мера MI, возможно, дает наилучшие усредненные результаты.
- Необходимо задавать список стоп-слов, чтобы «отбросить» самые частотные слова, сочетания с которыми неизменно оказываются вверху таблицы: предлоги, местоимения или союзы.

Выводы (2)

- Дополнительные возможности:

- Возможность объединения разных мер, например, ввести величину, равную сумме их рангов.
- Учитывать в статистических мерах при поиске коллокаций леммы или словоформы?
- Разрывные коллокации
- Меры ассоциации для 3-грамм
- Следует принимать во внимание структурные синтаксические формулы

Требования к программному инструментарию

- уметь находить разрывные коллокации со свободным порядком;
- искать коллокаты не только по леммам, но и по словоформам;
- искать коллокаты для гнезда опорных однокоренных слов;
- уметь варьировать размер окна;
- искать коллокации n-граммы;
- обработка знаков препинания и служебных слов, имен собственных и т.п.;
- поиск и выдача коллигаций;
- гибкие выходные интерфейсы
- ……

Тоже коллокация, она же прагматема ☺

- ***Спасибо за внимание!***