«Зачем», «что» и «как» в исследовании коллокаций. Вопросы и возможные ответы

Размышления на тему

Елены Ягуновой & Co

iagounova.elena@gmail.com

место доклада в миниконференции

В рамках мини-конференции **«Коллокации и сочетаемостные особенности: методы исследования»** мой доклад взаимосвязан с докладом Л.М. Пивоваровой «Подводные камни статистических мер»:

- определяет цели, задачи, гипотезы работы;
- задает критерии выбора и описание материала (новостных и научных коллекций);
- задает критерии выбора статистических мер;
- предлагает обсуждение полученных результатов;
- т.е. подготавливает к обсуждению «подводных камней статистических мер» в контексте конктретного цикла работ

Что -1 (у других)

- чаще всего – коллокации как несвободные сочетания, не относящиеся к идиомам:
  - ключевое слово этих сочетаний может появляться в контексте разных языковых единиц,
  - эти единицы (т.е. контекст ключевого слова) можно перечислить в виде закрытого списка

Что -2 (у нас)

- Коллокации: неслучайное сочетание двух и более лексических единиц, характерное как для языка в целом (текстов любого типа), так и определенного типа текстов (или даже (под)выборки текстов).

Зачем???

Исследование

- характеристик единиц языка,

и/или

- характеристик текстов и их структурных составляющих

Что-1? Зачем-1

- рассматриваются большие массивы текстов
  - изучаются характеристики языка,
  - исследуемые единицы можно перечислить в виде закрытого списка,
  - напр., работы, которые ведутся на базе НКРЯ

Что-1? Зачем-1 (примеры)

- Корпусной словарь неоднословных лексических единиц (оборотов) http://ruscorpora.ru/obgrams.html
- При каждом обороте указано количество употреблений в НКРЯ (по данным на сентябрь 2008 г.).
- Словарь составлен на основе базы данных частотных коллокаций НКРЯ, с дополнениями из словарей Р.П. Рогожниковой (Толковый словарь сочетаний, эквивалентных слову, М., 2003) и МАС (Словарь русского языка в 4-х томах под ред. А.П.Евгеньевой, М., 1999).
- Обороты в функции предлога
- Наречные и предикативные обороты
- Вводные обороты
- Обороты в функции союза и союзного слова
- Обороты в функции частиц

Корпусной словарь неоднословных лексических единиц (оборотов). Плюсы и минусы

- Есть закрытый список коллокаций (по словарям),
- требуется оценить количество – в абсолютных единицах! – соответствующих коллокаций в корпусе,
- нет стат. оценки степени связанности коллокаций,
- возможен выход на контексты (на запрос в НКРЯ),
- но неоднозначность не снимается (напр., *может быть, в качестве*)
  - автоматически снять неоднозначность свободное сочетание vs. неоднословная лексическая единица практически невозможно
  - Вы́явленная осо́бенность **мо́жет быть** ва́жной при прогнози́ровании ис хо́да заболева́ния. (пример свободного сочетания из НКРЯ)

Что-1? Зачем-1 (примеры)

на http://dict.ruslang.ru/

- **Г. И. Кустова** СЛОВАРЬ РУССКОЙ ИДИОМАТИКИ (выход на запрос в НКРЯ)
- Сочетания слов со значением высокой степени
- Алфавитный список всех сочетаний
- Алфавитный общий список степенных слов
- Алфавитный список прилагательных
- Алфавитный список наречий и наречных выражений

Степенное слово: *Характеризуемое слово:*

ЧАСТЬ РЕЧИ	ЧАСТЬ РЕЧИ

Пример алфавитного списка всех сочетаний слов со значением высокой степени

- абсолютная анархия
  абсолютная бездарность
  абсолютная безопасность
  абсолютная безысходность
  абсолютная бесперспективность
  абсолютная беспечность
  абсолютная беспомощность
  абсолютная беспринципность
  абсолютная беспристрастность
  абсолютная бессмыслица
  абсолютная бесспорность
  абсолютная бесцеремонность

Что-1? Зачем-1 (примеры)

- **О. Л. Бирюк, В. Ю. Гусев, Е. Ю. Калинина** СЛОВАРЬ ГЛАГОЛЬНОЙ СОЧЕТАЕМОСТИ НЕПРЕДМЕТНЫХ ИМЕН РУССКОГО ЯЗЫКА
- Выбор параметров:
- существительное фазовое значение
- прилагательное оценка
- глагол количество
- абстрактное значение отрицание
- конкретное значение порядок слов
- синтаксические отношения

Пример списка (параметры не выбраны), выход на запрос в НКРЯ

(не) ведать стыда действие
(не) видеть логики знание понимание
(не) видеть надобности знание понимание
(не) видеть оснований знание понимание
(не) видеть причины знание понимание
(не) видеть разницы Neg знание понимание
(не) внушать доверия действие каузация
(не) возникает сомнения действие субъект начало
(не) встретить сопротивления действие получатель
(не) встречать сопротивления действие получатель
(не) выдержать напряжения объект оценка соответствие
(не) выдержать характера прерывание демонстрация
(не) выдерживать критики действие объект мало соответствие

особенности этого подхода

- Заданность списка анализируемых коллокаций (частичная или по параметрам)
- Отношение к текстовым коллекциям
- работает
  - с материалом репрезентативного корпуса (что это такое?)
  - относится безразлично к типу текстов, входящих в корпус

Что-2? Зачем-2

- рассматриваются большие массивы текстов
  - тексты разных функциональных стилей и предметных областей,
- список потенциальных коллокаций для них принципиально не задан,
  - этот список является отражением тех характеристик, которые заложены в анализируемых текстах.

разные ФС текстов и различие списков коллокаций

http://corpus.leeds.ac.uk/ruscorpora.html

A query to Russian corpora

Выбор:

- Russian National Corpus (2009 version)
- Russian Fiction (disambiguated)
- Russian Newspapers
- Russian Internet Corpus RNC+NEWS-RU+I-RU (for rare words)
- Russian Business Internet Corpus

разные ФС текстов, разные стат. меры и различие списков коллокаций

A query to Russian corpora

Collocation scores:

- Mutual Information
- T-score
- Loglikelihood score

Context:

- ? words on the left ? words on the right

Но

- нет порогов отсечения,
- практически нет возможности работать со словоформными биграммами,
- очень грязная морфологическая разметка

Зачем-2 и Что-2 и Как-2?

Если коллокации не заданы списком,

если коллокации не заданы правилами, то что такое «коллокация»?

Какова природа коллокации?

Как понимать: неслучайное сочетание двух и более лексических единиц, характерное

- для языка в целом (текстов любого типа)?
- для определенного типа текстов (или даже (под)выборки текстов)?

Текст и коллокации

- текст есть структурированная последовательность единиц разных уровней,
- Коллокации как сложносоставные подструктуры текста – важный объект при исследовании процедур анализа (и синтеза) текста.

- Выделяя и исследуя коллокации мы исследуем текст:
  - структурные единицы текста разных языковых – и текстовых – уровней
  - их роль в процедурах анализа и синтеза речи (текстов).

Текстовые коллекции и коллокации

- Мы не привязаны к заданной коллекции или Корпусу
- На коллекциях **разных** текстов мы можем изучать характеристики наиболее связанных структурных составляющих, и через них выходить на структуру **разных** текстов
  - Прежде всего, текстов разных функциональных стилей (новостные, научные, деловые, художественные)

Что мы можем получить, на разных коллекциях-корпусах?

Варьируя коллекции, мы можем организовать систему вложенных друг в друга корпусов:

- тексты определенного функционального стиля,
  - тексты определенного источника,
    - тексты определенной предметной области,
      - однородная выборка текстов определенных источников и предметной области,
        - и т.д.

Что мы можем получить, на разных коллокциях-корпусах?

Например, вложенные друг в друга:

- научные тексты,
  - лингвистические научные тексты,
    - научные тексты предметной области «Теоретическая и прикладная лингвистика» (материалы конференции «Диалог»),
      - научные тексты предметной области «Корпусная лингвистика».

Что мы можем получить,

используя разные

- статистические меры (напр., MI, t-score, LL),
- а может где-то и абсолютные частоты коллокаций?
- пороги отсечения,
- разные единицы (коллокации из словоформ и/или лексем),
- … расстояния между коллокатами

используя разные параметры,

Мы получаем разные типы коллокаций = типы структурных составляющих текста:

- неоднословных номинаций
  - в новостном тексте – наименования персон (*Бенедикт XVI, Бритни Спирс, президент Венесуэллы Уго Чавес*), организации (*РИА Новости, Арбат Престиж*), географические наименования (Саудовская Аравия, *Соединенные Штаты, Нижнем Новгороде*),
  - в новостном тексте – наименования событий или ?? (*умышленное причинение тяжкого вреда здоровью, защищать принадлежащий ему титул чемпиона*),
  - в научном тексте – термины (*корпусная лингвистика, часть речи, машинный перевод*);

используя разные параметры, (продолжение)

Мы получаем еще другие типы коллокаций = типы структурных составляющих текста:

- составные слова (*в качестве, в связи, в результате*),
- газетные клише (*по словам, сообщает РИА, как сообщает или сообщает Интерфакс со ссылкой на*),
- конструкции с управлением глаголов (*зависит от, состоит в,* а также – *имеет место, обращать внимание*), и т.д.

статистические меры (напр., MI vs. t-score)-1

Новостные тексты (напр., на материале lenta.ru за 2009)

- мера MI (порог 40): определение наименования объектов, терминов, сложных номинаций, отражающих предметную область (– как?) ,
- мера t-score (порог 40) – выделение:
  - «общеязыковых устойчивых сочетаний» (производных служебных слов, дискурсивных слов)
  - «устойчивых конструкций», где и те, и другие характеризуют стилистические особенности новостных текстов

статистические меры (напр., MI vs. t-score)-2

Научные тексты (напр., на материале «Диалог 2003-2009» и «Корпусная лингвистика» (2004, 2006, 2008))

- мера MI: «ключевые» неоднословные термины, которые характеризуют предметную область коллекции;
- t-score:
  - «общеязыковых устойчивых сочетаний» (производных служебных слов, дискурсивных слов),
  - «устойчивых конструкций», где и те, и другие характеризуют стилистические особенности научных текстов,
  - коллокации, общие для ***всех*** (или ***подавляющего большинства***) текстов коллекции

Степень тематической однородности коллекции научных текстов соотносится с однородностью множества выделяемых коллокаций

Таблица 1. Биграммы (MI-score), выделяющиеся и для лексем, и для словоформ. Материал конференции «Диалог» (из доклада на симпозиуме "Терминология и знание" -- Пивоварова, Ягунова 2010)

п.п.	биграммы		п.п.	биграммы	
1	ударном	слоге	30	корпусная	лингвистика
2	концептуальных	графов	33	отглагольных	существительных
4	внешним	посессором	37	знаки	препинания
5	оперативной	памяти	38	педагогической	коммуникации
8	вокального	жеста	42	основного	тона
14	*крайней*	*мере*	46	машинного	перевода
16	XIX	века	61	устойчивых	словосочетаний
17	лингвистического	процессора	63	точки	зрения
21	положение	дел	*70*	*меньшей*	*мере*
22	*первую*	*очередь*	*72*	*вряд*	*ли*
25	картине	мира	73	предметной	области
26	множественного	числа	*85*	*вплоть*	*до*
28	интеллектуальные	технологии			

Биграммы (MI-score), выделяющиеся и для лексем, и для словоформ. Табл. 1 и 2а. Пояснения

- **Пороги для коллекций «Корпусная лингвистика» и «Диалог»: 16 и 40**

- **Курсивом** в таблице выделены сочетания, которые были удалены на этапе выделения терминологических коллокаций с использованием морфологического фильтра.
- **Подчеркиванием** выделены те сочетания, которые на основании формальных критериев должны были быть ошибочно отнесены к терминологическим.

Таблица 2а. Терминологические биграммы (MI-score), выделяющиеся и для лексем, и для словоформ. Материал конференции «Корпусная лингвистика» (из доклада на симпозиуме "Терминология и знание" -- Пивоварова, Ягунова 2010)

п.п	биграммы		п.п.	биграммы	
4	речевой	деятельности	40	разрешения	неоднозначности
5	художественной	литературы	41	английский	язык
9	общим	объемом	47	Национальный	корпус
11	корпусная	лингвистика	48	грамматических	категорий
13	имена	собственные	52	устная	речь
15	математической	лингвистики	54	база	данных
16	словарной	статьи	61	лексических	единиц
18	предметной	области	65	русский	язык
19	машинного	перевода	67	корпусные	данные
26	большое	количество	79	частей	речи
35	семантических	состояний	86	морфологической	разметки

Биграммы (MI-score), выделяющиеся и для лексем, и для словоформ. Почему мы выбрали этот список?

- В список 1 попадают составные номинации, характеризуемые максимальной свободой (максимальным разнообразием, минимальной ограниченностью) набора выполняемых ими в предложении семантико-синтаксических ролей.
- Примеры: 9 *винительный падеж*, 17 *именительный падеж*, 24 *актуальный членение*, 29 *инструментальный среда*.
- Биграммы списка 2 – номинации в определенной синтаксической позиции.
- Примеры: 10 *речевой акт*, 50 *речевых актов*, 19 *именная группа*, 65 *именных групп*, 27 *коммуникативного акта*, 62 *коммуникативных актов*, 77 *просодических характеристик*, 78 *прошедшего времени*, 74 *речевого сигнала*. Кроме того, биграммы этого подкласса могут относиться к части целостной номинации, напр., сочетание *речевых актов* часто является частью триграммы «теории речевых актов».
- У биграмм списка 3 (см.табл.1 и 2а) наиболее простая структура: нет ни закрепленности, ни противоречий между смысловыми, лексическими и синтаксическими связями. Биграммы этого класса занимают в текущем словарном составе некое **промежуточное место** между биграммами класса «1» и биграммами класса «2».
- Анализ разных списков показал, что список 3 является наиболее адекватным при решении задачи определения ключевых тем (неоднословных терминов), характерных для рассматриваемых коллекций.

Статистические меры (напр., MI vs. t-score)-3. Дельта. Порог

Новостные тексты (напр., на материале lenta.ru), в которых представлена коллекция за год и подколлекции за каждый месяц (дельта за месяц)

Дельты за месяц имеют гораздо большую однородность тем!

- MI (порог 3): в списках коллокаций за разные месяцы – небольшое число пересечений,
  - ок. 50% биграмм появляется только в одном списке, менее 50% процентов из первой сотни годового списка попали в первую сотню какого-либо из месячных списков,
  - мера лучше отражает тематику текстов, а темы новостных текстов непрерывно меняются.
- t-score (порог 3): в списках коллокаций за разные месяцы – большое число пересечений,
  - первые сто биграмм из «года» повторяются в нескольких месячных списках (часто во всех двенадцати списках),
  - мера лучше отражает стратегию выбора тем (?) и стилистку текстов, а они в рамках одного и того же СМИ меняется сравнительно медленнее

Выделении основных тем новостной коллекции. Мера. Дельта. Порог

Гипотеза об иерархии используемых мер (с учетом дельт (списков по месяцам) и разных порогов) для новостных коллекций:

См. еще раз слайд 26 на материале научных коллекций.

1. **традиционно –** использование t-score для выделения основных тем новостных коллекций гораздо хуже MI,
   - НО **пересечения** списков коллокаций, полученных для разных месяцев (тематически более однородных выборок) с помощью t-score (**Δt-score**) --
   - дают представление **о ведущих темах**
     - более, чем списки, традиционно полученные с помощью меры MI;
2. MI **с высоким порогом отсечения** – при прочих равных -- более информативна для определения тематики коллекции, чем Δt-score.
3. Пересечение списков, полученных для разных месяцев с использованием меры MI (ΔMI), – почти пустое

Дополнительная проверка гипотезы. Дельта. Порог

- Еще раз про гипотезу: t-score < MI < Δt-score < MI^T (подробнее про стат. обоснование в докладе Л.М.Пивоваровой)
- Дельта нужда для увеличения тематической однородности выборки. КАК лучше определять дельту?
- Порог нужен для отсечения редких для коллекции коллокаций. Он зависит от объема коллекции и степени тематической однородности. КАК определять порог в каждом конкретном случае?

зачем? что? как?

- Сейчас мы не ставим перед собой задачу практически востребованного метода
  - напр., извлечения ***всех*** терминов или тестирования разных методик (см., напр., [Браславский, Соколов 2006]).

- Задача – изучение возможности выделения формальных признаков, необходимых для определения предметной области коллекций текстов и ключевых слов, описывающих рассматриваемые коллекции;
- формирование наборов информационно значимых для коллекции коллокаций и выделение общих для текстов коллекции коллокаций.

Зачем-2 и Что-2 и Как-2? продолжение… на будущее

- **что задано для списка потенциальных коллокаций ??**

1. не заданы даже ключевые слова,
2. ключевые слова заданы, варьируют коллокаты,
3. задан морфолого-синтаксический шаблон (в комбинации с п.1. или 2),
4. заданы ключевые слова, вместо слова-коллоката
5. и т.д.

Литература

- Бирюк О. Л., Гусев В. Ю., Калинина Е. Ю. Словарь глагольной сочетаемости непредметных имен русского языка М., 2008 http://dict.ruslang.ru/abstr_noun.php
- Браславский П., Соколов Е. Сравнение четырех методов автоматического извлечения двухсловных терминов из текса // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.) / Под ред. Н.И. Лауфер, А. С. Нариньяни, В. П. Селегея. – М.: Изд-во РГГУ, 2006.
- Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" - RCDL2003, Санкт-Петербург, 2003
- Иорданская Л. Н., Мельчук И. А.. Смысл и сочетаемость в словаре. М.: Языки славянских культур, 2007
- Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю. Поверхностные фильтры для разрешения семантической омонимии в текстовом корпусе // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог'2005" (Звенигород, 1-6 июня, 2005 г.)/ Под ред. И.М. Кобозевой, А.С. Нариньяни, В.П. Селегея. - М.: Наука, 2005.
- Кустова Г. И. Словарь русской идиоматики. Сочетания слов со значением высокой степени М., 2008 http://dict.ruslang.ru/magn.php
- Ляшевская О. Н., Шаров С. А. Новый частотный словарь русской лексики 2008 http://dict.ruslang.ru/freq.php

Литература (продолжение)

- Пивоварова Л.М., Ягунова Е.В. Извлечение и классификация терминологических коллокаций на материале лингвистических научных текстов. Предварительные наблюдения // Материалы второго Международного симпозиума “Терминология и знание” М., 2010 (в печати)
- Шайкевич А.Я., Андрющенко В.М., Ребецкая Н.А. Статистический словарь русской газеты (1990 гг.) М., 1998
- Хохлова М.В. Экспериментальная проверка методов выделения коллокаций // Slavica Helsingiensia 34. Инструментарий русистики: Корпусные подходы. Под ред. А. Мустайоки, М.В. Копотева, Л.А. Бирюлина, Е.Ю. Протасовой. Хельсинки, 2008. С.343–357
- Ягунова Е.В. Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей). Пермь, 2008.
- Ягунова Е.В. Формальные и неформальные критерии вычленения ключевых слов из научных и новостных текстов // Материалы IV Международного конгресса исследователей русского языка «Русский язык: исторические судьбы и современность». М., 2010
- Ягунова Е.В., Пивоварова Л.М. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов – Сб. НТИ, Сер.2, №5. М., 2010 (в печати)

Литература (продолжение)

- Degand L., Bestgen Y.Towards automatic retrieval of idioms in French newspaper corpora // Literary and Linguistic Computing, 18, 2003, 249-259
- Iordanskaja, L., Paperno, S.: A Russian-English Collocational Dictionary of the Human Body, Columbus/Ohio 1996
- Khokhlova M. Extracting Collocations in Russian: Statistics vs. Dictionary // JADT 2008: actes des 9es Journées Internationales d'Analyse Statistique des Données Textuelles, Lyon, 12-14 mars 2008 : Proceedings of 9th International Conference on Textual Data statistical Analysis, Lyon, March 12-14, 2008 (editors : Serge Heiden, Bénédicte Pincemin). P. 613–624.
- Petrovic S., Snajder J., Basic B.D., Kolar M. Comparison of collocation extraction for document indexing // Journal of Computing and information technology – CIT 14, 2006, 4, 321-327
- Stubbs M. Collocations and semantic profiles: om the case of the trouble with quantitative studies.Functions of language 2:11, 23-55, Benjamins, 1995.
- Manning C., Schutze H. Collocations // Manning C., Schutze H. Foundations of Statictical Natural Language Processing, 2002, pp.151-189
- Rayson, Paul & Roger Garside (2000). Comparing corpora using frequency profiling // Proceedings of the Comparing Corpora Workshop at ACL 2000. Hong Kong, 2000. P. 1-6.