А.Н.Савина, Е.В.Ягунова Исследование коллокаций с помощью экспериментов с информантами // Труды международной конференции «Корпусная лингвистика–2011». – СПб.: С.-Петербургский гос. университет, Филологический факультет, 2011

А.Н.Савина, Е.В.Ягунова

ИССЛЕДОВАНИЕ КОЛЛОКАЦИЙ С ПОМОЩЬЮ ЭКСПЕРИМЕНТОВ С ИНФОРМАНТАМИ

1. Введение

Доклад посвящен решению одного из вопросов широкой области изучения природы коллокаций и возможной их классификации. Этот доклад представляет работу в рамках общего проекта, посвященного этой теме. Под **коллокациями** понимается неслучайное сочетание двух и более лексических единиц, характерное для текстов определенного типа и выделяемое на основании статистических критериев. Мы идем от **реализации**, и максимально учитываем тип анализируемых текстов, собирая коллекции с учетом функционально стилистических и прочих характеристик текстов. Очевидно, что список, получаемый таким образом, не оказывается вполне однородным, требует дальнейшей классификации и теоретической интерпретации[1].

1 Ср. Ягунова Е. В., Пивоварова Л. М. От коллокаций к конструкциям // Русский язык: конструкционные и лексико семантические подходы / Отв. ред. С.С.Сай. (ACTA LIINGUIISTIICA PETROPOLIITANA. Труды Института лингвистических исследований РАН. Отв. редактор / Н. Н. Казанский) (в печати) СПб, 2011; Ягунова Е.В., Пивоварова Л.М. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов // Сб. НТИ, Сер.2, №6. М., 2010; Хохлова М.В. Исследование лексико-синтаксической сочетаемости в русском языке с

Широкие возможности для осмысления природы коллокаций – в рамках списков, полученных на основе тех или иных статистических мер – предоставляет обращение к эксперименту с носителем языка. Цель доклада – демонстрация таких возможностей.

2. Материал и методика

Мы хотим проиллюстрировать предлагаемую методику на примере монотематической коллекции материалов конференции «Корпусная лингвистика» 2004-2008 года, объемом около 220000 «токенов» – словоупотреблений и знаков препинания.

Для сравнения использовалась коллекция новостных текстов портала lenta.ru за 2009 год (объемом более 66000000 «токенов»). Через сравнение статистического обследования этих двух коллекций мы исследовали коллокации, характеризующие тексты определенного функционально стиля (типа, жанра).

Использовались две статистические меры. MI позволяет извлечь терминологические сочетания), а t-score выделяет научные клише и те терминологические сочетания, которые характеризуют все тексты коллекции (или большинство текстов)[1]. Нами рассматривалось два вида биграмм: с указанием целевого слова (для левого и для правого контекста) и без указания.

помощью статистических методов (на базе корпусов текстов). АКД СПб., 2011; Захаров В.П., Хохлова М.В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26-30 мая 2010 г.). Вып. 9 (16). - М.: Изд-во РГГУ, 2010; Пивоварова Л.М., Ягунова Е.В. Извлечение и классификация терминологических коллокаций на материале лингвистических научных текстов (предварительные наблюдения) // Терминология и знание. Материалы II Международного симпозиума (Москва, май 2010 г.). М., 2010.

1 См. подробнее в Пивоварова Л.М., Ягунова Е.В. Извлечение и классификация …. . М., 2010

А.Н.Савина написала удобный для исследовательских нужд программный код, который позволяет выделять списки биграмм с задаваемым пользователем целевым словом на основании указанных мер.

Для каждого типа биграмм рассматривались подсписки по 25 биграмм для каждой из мер (с максимальными значениями меры). В дальнейшем объединенные списки из 50 рандомизированных биграмм являются входными данными для проведения двух типов экспериментов с информантами[1].

Все информанты были студентами гуманитарных специальностей и имели представление о предметной области «корпусная лингвистика» хотя бы на уровне прослушанных лекций и семинарских занятий. В эксперименте 1 и 2 участвовали разные группы информантов.

Эксперимент 1 (25 информантов) предлагает информанту анкету, в которой от него требуется определить, к какому из трех классов – «правильные», «ожидаемые» и «остальные» – относится каждое из сочетаний предлагаемого списка. **Эксперимент 2** (22 информанта) позволяет оценить степень связности между словами – для тех же списков – в шкале от 0 до 5, где «0» соответствует минимальной, а «5» – максимальной степени связности с точки зрения информантов.

[1] В каждой из анкет были собраны сочетания, относящиеся только к одной коллекции. В инструкции информантам было сказано: «Перед Вами сочетания слов (биграммы) из научных текстов (материалов специализированной лингвистической конференции), выделенные на основании статистических критериев». Или – в другом варианте – было сказано, что в анкете «…сочетания слов из новостных текстов». Оценка информантами связанности предлагаемых в анкете коллокаций происходила на основании интуиции носителя языка (и его представления о предметной области корпусной лингвистики) без указания критериев оценки.

3. Результаты. Обсуждение результатов

По результатам эксперимента 1 была получена классификация в заданном наборе классов, каждый из классов подразделяется на ядро и периферию. Коллокация считалась ядерной, если более 65% информантов отнесли ее к данному классу и периферийной – если число информантов колебалось от 33% до 65% [1] Проиллюстрируем возможности методики на примере списков словоформных биграмм с ключевыми словами «корпус» и «слово». 42% коллокаций было отнесено к ядерным:

- ядро «правильных» содержит научные термины,
- ядро «ожидаемых» содержит составные слова (см. табл. 1),
- ядро «остальных» – сочетания, которые сложно интерпретировать[2].

Таблица 1. Классы научных биграмм с ключевыми словами «корпус» и «слово»

ядро «правильных»	**ядро «предсказуемых»**
аннотированный корпус	корпус является
национальный корпус	корпус содержит
параллельный корпус	корпус представляет
международный корпус	корпус позволяет
представительный корпус	данный корпус
размеченный корпус	большой корпус
электронный корпус	второе слово
служебное слово	первое слово
составное слово	данное слово
	слово встретилось

Результаты эксперимента 2, во-первых, подтверждают данные эксперимента 1 (все биграммы класса «ядро правильных» соответствуют усредненной оценке связности в 4 балла). Во-

1 Часть периферийных коллокаций было отнесено к пересечению классов (если указанное требование соблюдалось по отношению к двум классам).

2 Такого типа распределение характерно для всех коллокаций научного текста.

вторых, данные эксперимента 2 позволяет говорить о высокой связанности еще трех терминологических биграмм *«главное слово», «зависимое слово», «отдельное слово»,* которые по результатам эксперимента 1 относились к пересечению классов «правильные» и «ожидаемые» (тем самым повышают статус этих терминов на дополнительной шкале «связаности»).

Выделение в качестве «ядра правильных» именно терминов является ярой чертой научных текстов, в случае новостных текстов в этот класс попадают, прежде всего, сложные номинации (персоны, организации, географические наименования), единицы терминологического характера занимают вспомогательное место (*ценные бумаги*, *тротиловый эквивалент*) (см. табл. 2).

Таблица 2. Пример биграмм класса «ядро правильных» для научных и новостных текстов

научные»	**новостные**
математической лингвистики	РИА Новости
художественной литературы	Саудовская Аравия
русского языка	Нижнем Новгороде
корпусная лингвистика	Бараком Обамой
имена собственные	Соединенных Штатов
словарной статьи	Млечного Пути
машинного перевода	встречную полосу
корпусной лингвистике	Палестинская автономия
корпусной лингвистики	сообщает РИА
речевой деятельности	ценным бумагам
XX века	Бритни Спирс
корпуса текстов	тротиловом эквиваленте

Термин *«контекстная предсказуемость»* является хорошим примером различия между результатами только вычислительного эксперимента и экспериментального подхода, учитывающего данные от информантов. У этой коллокации максимальное

значение меры MI, однако, по оценке информантов она отнесена к пересечению «правильных» и «ожидаемых» биграмм.

Данные эксперимента 2 подтверждают классификацию по результатам эксперимента 1. Биграммы, связность которых оценивается группой информантов не менее чем 4 баллами (среднее по группе), соответствуют ядру «правильных», эти биграммы выделяются на основании MI (наиболее частотные и t-score), а те, связность которых меньше 2,8, – ядру «остальных» (на основании t-score).

Экспериментальная классификация производных служебных слов, дискурсивных слов и наречных образований оказывается не менее интересной. Анализируемые коллокации допускают отнесение более чем к одному типу, а предложенная классификация позволяет выделить зоны ядра и периферии в трех заданных классах. Наличие пересечений говорит о наличии нестабильности в отношении ряда коллокаций (что проявляется как в эксперименте с информантами, так и в функционировании названных коллокаций в текстах и языке). Напр.,: научные, «ядро ожидаемых» – *в качестве*, *за счет*, *(в) свою очередь*, *(в) том числе*; «пересечение ожидаемых и других» – *в виде*, *(по) крайней мере*, *не только* и т.д.

Пример нестабильности для научных текстов: биграммы «*с помощью*» и «*в частности*» характеризуются большей целостностью, чем «*в качестве*», «*на основе*», «*за счет*», т.к. первые попадают в пересечение “правильных” и “ожидаемых”, а вторые – в ядро «ожидаемых». Наши данные демонстрируют также зависимость от типа коллекции, напр., «*в частности*» и «*(в) том числе*» характеризуется большей целостностью для научных текстов, чем для новостных[1].

[1]Для научных текстов «*в частности*» попадает в «пересечение правильных и ожидаемых», а для новостных – в ядро «ожидаемых»; для научных «*(в) том числе*» оказывается в «ядре ожидаемых», для новостных – в «пересечении ожидаемых и других»

Данные экспериментов с информантами позволяют установить дополнительные шкалы, определяемые уже не только значениями статистических мер, но и связностью (целостностью), ощущаемой носителями языка и эксплицируемой в ходе экспериментов.

Anna Savina, Elena Yagunova

COLLOCATIONS: EXPERIMENTAL INVESTIGATION (CORPUS STUDY AND EXPERIMENTS WITH INFORMANTS)

Our research is devoted to solving one of the most important problems of collocation study: about the nature of collocations and their possible classification. The report presents the result of the first stage of work within the overall project on this topic. We understand a **collocation** as a non-random combination of two or more lexical items that characterizes a certain text type. Ample opportunities for understanding the nature of the collocations – within the lists received on the basis of statistical measures – are given by the reference to experiments with the native speaker informants. The advantages of this combined approach – the corpus based and informant-used methods – are in classifying of multiword units with multi-scales.