Slavica Helsingiensia 34
Инструментарий русистики: корпусные подходы, Хельсинки, 2008
Под ред.: А. Мустайоки, М.В. Копотева, Л.А. Бирюлина, Е.Ю. Протасовой

**М.В. Хохлова**
**(Санкт-Петербург)**

# Экспериментальная проверка методов выделения коллокаций

## 1. Понятие коллокации в лингвистике

Термин *коллокация* как синоним устойчивого сочетания, хотя и вошел в постоянное употребление недавно, по праву занимает одно из ключевых мест в современной лингвистике. В широком смысле это комбинация двух или более слов, имеющих тенденцию к совместной встречаемости, т.е. речь идет об осмысленных последовательностях слов в тексте.

Информация об устойчивых сочетаниях не всегда последовательно отражается в толковых словарях, и граница между ними и свободными сочетаниями слов определяется достаточно субъективно. Под сочетаемостью языковой единицы, как утверждает И.А. Мельчук (1960: 80), понимается «число других элементов, с каждым из которых данный элемент может вступать в определенное отношение (скажем, быть зависимым от него: например, сочетаемость прилагательного — это число существительных, к которым оно может быть определением, и т.д.)».

Мы придерживаемся подхода, постулированного И.А. Мельчуком в работе (Мельчук 1960), где устойчивые сочетания рассматриваются как подмножество несвободных сочетаний, их разновидность, и предлагаем использовать различные статистические меры, с помощью которых можно будет ранжировать словосочетания с точки зрения устойчивости.

Как нам кажется, бóльшему пониманию того, какая существует взаимосвязь между такими понятиями, как сочетаемость и устойчивость, может способствовать анализ коллокаций, полученных из корпусов на основании статистического метода.

Поскольку в российской лингвистике вместо термина «коллокация» часто используются различные синонимы (несвободные словосочетания, фразеологизмы, идиомы), рассмотрим это понятие в широком контексте.

Коллокации часто противопоставляются, с одной стороны, свободным словосочетаниям, а с другой стороны, идиомам. В свободных комбинациях слов элементы можно заменить синонимами или, если что-то опущено, можно догадаться о значении целого. В идиомах же, напротив, семантика целого не может быть выведена из значений отдельных

компонентов. В отличие от этих двух групп, в коллокациях значения отдельных слов могут привнести новое в значение целого.

Как мы полагаем, подходы к описанию понятия коллокации в целом можно свести к трем следующим:

1. подход, берущий начало в работах британских контекстуалистов (Firth 1957, 1968);
2. семантико-синтаксический подход (Cowie 1978; Hausmann, 1979, 1985);
3. подход теории «Смысл↔Текст» (Иорданская & Мельчук 2007).

Термин «коллокация» впервые был введен основоположником Лондонской школы структурной лингвистики, представителем *британского контекстуализма* Дж. Р. Фёрсом (Firth 1957: 94). В рамках этого направления значение рассматривается как сложное лингвистическое явление, требующее исследования на всех уровнях языковой структуры. Подчеркивается эмпирический характер лингвистического анализа, изучение языка непосредственно в его употреблении. Именно анализ употребления формы позволяет раскрыть значение. В анализе значения важнейшую роль играет контекстуализация, т.е. прием установления контекста применительно к каждому языковому уровню. На лексическом уровне — это коллокации, т.е. типичное и постоянное окружение слова, указание на его традиционную встречаемость (Firth 1968: 181; Ярцева 1990: 276). Таким образом, 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» (Сусов 1999: 153).

В рамках *семантико-синтаксического подхода* коллокации рассматриваются как семантико-синтаксические единицы или лексически определенные элементы грамматических структур. Они характеризуются семантической, синтаксической и дистрибутивной регулярностью, внутренне присущими свойствами сочетаний слов, а не их появлением в корпусах.

Одним из типов устойчивых сочетаний являются фразеологические единицы, изучением которых занимается фразеология. Однако в русскоязычной лингвистической традиции, помимо термина «фразеологизм», существует и термин «коллокация». Видимо, впервые он был введен О.С. Ахмановой (1966). Подобные речевые элементы в работах разных авторов называются по-разному: «устойчивые глагольно-именные сочетания» (Дерибас 1983), «аналитические лексические коллокации» (Телия 1996) и др. Первой работой в российской лингвистике, полностью посвященной исследованию понятия коллокации на материале русского языка, является монография Е.Г. Борисовой (1995а).

В *теории «Смысл↔Текст»* 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. Этому типу фразем соответствуют: англ. *land a job* (букв. «приземлиться на должность») — «найти работу», *stand a comparison* [with N] (букв. «выстаивать сравнение с N») — «выдерживать сравнение с N». Большинство полуфразем, или коллокаций, в теории «Смысл↔Текст» называется лексико-функциональными выражениями (Иорданская & Мельчук 2007: 239).

Мы предлагаем рассматривать термин «коллокация» как родовое понятие для обозначения определенных типов устойчивых словосочетаний, фразеологических единиц и фразем. Как нам кажется, наша трактовка понятия коллокации может быть весьма полезна и в лексикографической практике. Существующие словари устойчивых словосочетаний, во-первых, охватывают далеко не полный их перечень, во-вторых, часто делают это недостаточно последовательно. Есть потребность в словаре нового типа, который можно было бы назвать интегрированным словарем устойчивых словосочетаний, или словарем коллокаций.

## 2. Статистический метод

Как уже было отмечалось, мы полагаем, что словосочетания могут быть отнесены к высокоустойчивым (на практике такие сочетания называют просто устойчивыми) или низкоустойчивым (на практике их называют неустойчивыми) на основании значений каких-либо статистических мер.

В настоящее время в лингвистике существует несколько способов для вычисления степени связанности частей той или иной коллокации. В качестве таких статистических мер нами были выбраны меры ассоциации (*MI, t-score, log-likelihood*), которые чаще всего используются при вычислении степени близости между компонентами словосочетаний в корпусе.

Они основаны на данных о частоте из таблицы сопряженности 2x2 (в случае биграммы) рассматриваемых слов, где каждая из двух переменных может принимать одно из двух значений (истина — ложь (¬)), поэтому любая биграмма принадлежит одной из четырех комбинаций этих переменных. Конечно, статистическая связанность компонентов биграммы не всегда говорит о семантической или синтаксической связанности. Тем не менее, линейная близость может оказаться важной предпосылкой для нахождения устойчивых сочетаний, т.е. коллокаций и других типов словосочетаний в текстах.

Рассмотрим гипотетическое частотное распределение биграммы *принять решение* (X,Y), в которой случайные переменные имеют значения

(X = *принять*, Y = *решение*). Ниже приведена таблица сопряженности для биграммы.

**Табл. 1. Сопряженность для биграмм**

| | Y | ¬ Y | |
|---|---|---|---|
| X | $O_{11}$ | $O_{12}$ | $N_{1P}$ |
| ¬ X | $O_{21}$ | $O_{22}$ | $N_{2P}$ |
| | $N_{P1}$ | $N_{P2}$ | $N_{PP}$ |

Таким образом, в таблице представлены наблюдаемые частоты (Oij). $O_{11}$ — частота данной биграммы, т.е. количество случаев, когда оба компонента X и Y сочетания встречаются вместе. $O_{12}$ — частота биграмм, в которых встретилось слово X встретилось, а слово Y нет. $O_{21}$ — количество биграмм, в которых встретилось слово Y и не встретилось слово X. $O_{22}$ — частота биграмм, в которых не присутствуют ни слово X, ни слово Y. $N_{1P}$, $N_{P1}$, $N_{2P}$ и $N_{P2}$ представляют собой маргинальные частоты, которые содержат информацию, встречается ли слово X или Y в данной биграмме. $N_{PP}$ — общее количество биграмм в корпусе.

Ожидаемые частоты (Eij) вычисляются по таблице 2 на основании маргинальных частот.

**Табл. 2. Вычисление ожидаемых частот слов по таблице сопряженности**

| | Y | ¬ Y |
|---|---|---|
| X | $E_{11} = N_{1P} * N_{P1} / N_{PP}$ | $E_{12} = N_{1P} * N_{P2} / N_{PP}$ |
| ¬ X | $E_{21} = N_{P1} * N_{2P} / N_{2P}$ | $E_{22} = N_{P2} * N_{2P} / N_{PP}$ |

### *2.1. MI*

В работе (Church & Hanks 1990) был введен коэффициент ассоциации (*association ratio*) для вычисления степени связанности между словами. В его основе лежало понятие взаимной информации (*mutual information*), заимствованное из теории информации и впервые примененное в работе (Fano 1961: 28), определяемое как:

$$I(x,y) \equiv log_2 \frac{P(x,y)}{P(x) \times P(y)}, \text{ где}$$

x, y — слова;

P(x,y) — вероятность сочетания x и y;

P(x), P(y) — вероятности слов x и y соответственно.

Как ожидается, если слова действительно связаны, тогда наблюдаемая совместная вероятность их встречи P(x,y) будет много больше, чем случайная P(x)*P(y), следовательно, I(x,y) >> 0. Если между словами нет

особой зависимости, тогда P(x,y) ≈ P(x)*P(y), следовательно, I(x,y) ≈ 0. Если слова находятся в отношении дополнительной дистрибуции, тогда P(x,y) будет много меньше P(x)*P(y), отсюда I(x,y) << 0. Высказывается гипотеза, согласно которой значение I(x,y) > 3 выявляет интересные для рассмотрения случаи, в то время как величина меньше 3 — нет (Church & Hanks 1990: 78). Как указывается, эта мера позволяет выделять не только коллокации, но и семантические классы.

*MI* (коэффициент взаимной зависимости, объем информации) сравнивает зависимые контекстно-связанные частоты с независимыми, как если бы слова появлялись в тексте совершенно случайно:

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

MI — объем информации;
n — ключевое слово;
c — коллокат;
f(n, c) — частота встречаемости ключевого слова *n* в паре с коллокатом *c*;
f(n), f(c) — абсолютные частоты ключевого слова *n* и слова *c* в корпусе;
N — общее число словоформ в корпусе.

Если значение MI (n,c) больше 1, тогда данное сочетание слов считается статистически значимым. В случае если MI (n,c) примерно равно 0, сочетание слов является менее статистически значимым, слова появляются в паре крайне редко. MI (n,c) меньше 0 означает, что n и c находятся в отношении дополнительной дистрибуции. Вопрос о том, какие значения MI следует считать пороговыми, остается открытым. Так, в работе (Biber 1998: 266) говорится, что значения, намного превышающие 0, свидетельствуют, что слова встречаются не случайно.

Мера MI позволяет выделить устойчивые словосочетания, имена собственные, а также низкочастотные специальные термины. Это особенно важно в задачах информационного поиска, поскольку, если будет предоставлена информация об их сочетаемости в разных областях знаний, это позволит более эффективно сравнивать документы, релевантные запросу. Слова, у которых MI-score принимает наибольшую величину, менее частотны и обладают ограниченной сочетаемостью.

### *2.2. T-score*

Мера *t-score* также учитывает частоту совместной встречаемости ключевого слова и его коллоката, отвечая на вопрос, насколько не случайной является сила ассоциации (связанности) между коллокатами:

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

К недостаткам использования этой меры можно отнести то, что она в первую очередь выделяет коллокации с очень частотными словами, в частности, со служебными словами. Слова с наибольшим значением t-score оказываются частотными и могут сочетаться с множеством единиц (подробнее см. Stubbs 1995). Поэтому для t-score необходимо задавать stop list, чтобы «отбросить» самые частотные слова, сочетания с которыми неизменно окажутся в самом верху таблицы: предлоги, например, местоимения или союзы.

### *2.3. Log-Likelihood*

Также широко применяется формула, известная под названием *log-likelihood*, известная как *логарифмическая функция правдоподобия*. В ней используется «отношение функций правдоподобия, соответствующих двум гипотезам — о случайной и неслучайной природе двусловия» (Браславский & Соколов 2006: 89).

$$\log - likelihood = 2\sum_{ij} Oij \times \log \frac{Oij}{Eij},$$

где Oij, Eij — наблюдаемая и ожидаемая частоты, вычисляемые по вышеприведенным таблицам (Evert 2004: 83).

## 3. Выделение коллокаций на основе статистического метода

### *3.1. Постановка задачи*

Исследование сочетаемости методами корпусной лингвистики базируется на вероятностно-статистических характеристиках лексических единиц. Как указывает Н.Д. Андреев (1967), статистические характеристики языковых единиц в речи оказываются весьма важным фактором при описании языкового материала.

Задачи нашего исследования заключаются в следующем: провести эксперименты, чтобы найти пригодную меру ассоциации, позволяющую обнаруживать устойчивые словосочетания; определить возможности статистических методов в целом и отдельных мер в частности.

Для решения данных задач была проведена серия экспериментов с целью оценки эффективности статистических методов. В ходе эксперимента проверялось, насколько предложенные методы применимы для русского языка и выделяются ли коллокации на основании данных мер.

### *3.2. Материал и инструменты исследования*

Материалом для нашего исследования послужили коллокации 19 существительных, которые были отобраны по следующему принципу. Первоначально были отобраны существительные, входящие в первую 1000 самых частотных слов, из электронного частотного словаря русского языка С.А. Шарова (2002). Далее по Малому академическому словарю (МАС 1981-1984) проверялось, имеют ли данные слова омонимы, которые могли бы искажать их частоту (например, сущ. *брак* в значениях ‘супружество’ и ‘изъян’; мест. *друг друга*, в котором оба элемента при лемматизации возводятся к одной лексеме). Следовательно, такие слова исключались из списка и не рассматривались в эксперименте. Затем список оставшихся существительных сверялся с данными в словаре коллокаций русского языка Е.Г. Борисовой (1995b). В случае отсутствия словарных статей для данного слова или ограниченной информации о его сочетаемости, представленной в словаре, такие слова тоже исключались из списка. Таким образом, был получен список из 19 существительных: *власть*, *внимание*, *возможность*, *война*, *вопрос*, *дождь*, *жизнь*, *закон*, *любовь*, *место*, *мнение*, *мысль*, *ночь*, *ответ*, *помощь*, *радость*, *слово*, *случай*, *смысл*.

Исследование проводилось на базе корпусов русских текстов, созданных в университете Лидса (Великобритания) под руководством С.А. Шарова, с помощью корпус-менеджера CQP (http://corpus1.leeds.ac.uk/ruscorpora.html). В случае работы в режиме поиска коллокаций предоставляется возможность выбора статистических мер (одной или нескольких сразу: MI, t-score, log-likelihood), установления требуемого диапазона (в словах), а также можно задавать часть речи для коллоката.

### *3.3. Методика исследования*

Нами был проанализированы коллокации на базе газетного корпуса. Этот корпус включает в себя около 78 млн. слов из разных видов газет («Известия», «Труд» и Strana.ru), его морфологическая разметка была выполнена при помощи программы Mystem.

Следует сразу сказать, что для каждого слова в качестве коллокаций нами были рассмотрены только биграммы, т.е. сочетания заданного слова со словом, находящимся справа или слева от него, и поэтому «разрывные» коллокации оказываются не учтенными в данном исследовании.

Результат выдачи по запросу представлен списком коллокаций, организованным в виде одной, двух или трех таблиц (в зависимости от количества выбранных мер) с шестью столбцами данных (см. Табл.3). Результаты запроса для каждого существительного были сведены нами в одну таблицу, а потом мы сравнили их со словарными статьями, приве-

денными для этих существительных в словаре коллокаций (Борисова 1995), в толковых словарях русского языка: БАС-17 (1948-1965), БАС-25 (2004-2006), МАС (1981-1984) и в Словаре синонимов и сходных по смыслу выражений (2006).

### *3.4. Выявление коллокаций для слова война*

Ниже в таблице приведены данные для первых 30 коллокаций с опорным словом *война*, отсортированные по значению меры MI.

**Табл. 3. Значения мер ассоциации для слова *война* (левый контекст)**

| Collocation | Joint | Freq1 | LL score | MI | T-score |
|---|---|---|---|---|---|
| необъявленный война | 9 | 76 | 30,19 | 11,03 | 3,00 |
| междоусобный война | 4 | 54 | 12,43 | 10,35 | 2,00 |
| партизанский война | 45 | 728 | 135,77 | 10,09 | 6,70 |
| рельсовый война | 6 | 100 | 18,00 | 10,05 | 2,45 |
| победоносный война | 9 | 174 | 26,31 | 9,84 | 3,00 |
| вялотекущий война | 6 | 142 | 16,92 | 9,54 | 2,45 |
| позиционный война | 5 | 128 | 13,90 | 9,43 | 2,23 |
| холодный война | 171 | 4747 | 469,90 | 9,31 | 13,06 |
| грянуть война | 14 | 457 | 37,19 | 9,08 | 3,73 |
| финляндский война | 4 | 148 | 10,37 | 8,90 | 2,00 |
| минный война | 8 | 332 | 20,28 | 8,73 | 2,82 |
| кровопролитный война | 6 | 251 | 15,18 | 8,72 | 2,44 |
| полномасштабный война | 11 | 491 | 27,48 | 8,63 | 3,31 |
| затяжной война | 12 | 571 | 29,59 | 8,54 | 3,45 |
| священный война | 31 | 1526 | 75,95 | 8,49 | 5,55 |
| гражданский война | 194 | 12469 | 451,11 | 8,10 | 13,88 |
| ценовый война | 13 | 1062 | 28,53 | 7,76 | 3,59 |
| разражаться война | 9 | 881 | 18,94 | 7,50 | 2,98 |
| кончаться война | 13 | 1285 | 27,29 | 7,48 | 3,59 |
| стальной война | 10 | 1021 | 20,83 | 7,44 | 3,14 |
| локальный война | 8 | 860 | 16,46 | 7,36 | 2,81 |
| тотальный война | 8 | 978 | 15,94 | 7,18 | 2,81 |
| начинаться война | 138 | 19820 | 264,97 | 6,94 | 11,65 |
| превентивный война | 4 | 584 | 7,62 | 6,92 | 1,98 |
| крымский война | 6 | 904 | 11,33 | 6,87 | 2,43 |
| разгораться война | 7 | 1071 | 13,17 | 6,85 | 2,62 |
| продлиться война | 9 | 1425 | 16,78 | 6,80 | 2,97 |

| Collocation | Joint | Freq1 | LL score | MI | T-score |
|---|---|---|---|---|---|
| мировой война | 154 | 25171 | 285,92 | 6,76 | 12,29 |
| чеченский война | 83 | 13558 | 153,79 | 6,76 | 9,03 |
| затягиваться война | 8 | 1316 | 14,76 | 6,75 | 2,80 |

Joint — абсолютная частота данной коллокации в корпусе;
Freq1 — абсолютная частота первого слова биграммы, т.е. коллоката для слова *война*;
LL score, MI, T-score — значения мер log-likelihood, MI и t-score для данной коллокации.

Как можно увидеть, в список попали сочетания, которые, с одной стороны, являются устойчивыми, с другой, обладают довольно высокими показателями меры MI.

В нижеследующей таблице приведен список коллокаций из найденных на слово *война*, которые отражены в словарных статьях словаря коллокаций Е.Г. Борисовой:

**Табл. 4. Меры ассоциации для коллокаций со словом *война* по словарю 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 |

Будем называть коллокации, приведенные в словаре Е.Г. Борисовой и входящие в выданные таблицы, «правильными».
Ниже приведены словосочетания, найденные на слово *война* в БАС-17.

**Табл. 5. Меры ассоциации для коллокаций со словом война по БАС-17**

| Collocation | Joint | Freq1 | LL score | MI | T-score |
|---|---|---|---|---|---|
| гражданский война | 194 | 12469 | 451,11 | 8,10 | 13,88 |
| идеологический война | 4 | 1678 | 5,53 | 5,40 | 1,95 |
| мировой война | 154 | 25171 | 285,92 | 6,76 | 12,29 |
| партизанский война | 45 | 728 | 135,77 | 10,09 | 6,70 |
| холодный война | 171 | 4747 | 469,90 | 9,31 | 13,06 |

Ниже приводится график, на котором изображены значения меры log-likelihood по оси ординат и ранги биграмм по оси абсцисс. Темным цветом обозначены «правильные» коллокации и коллокации, найденные в БАС-17.

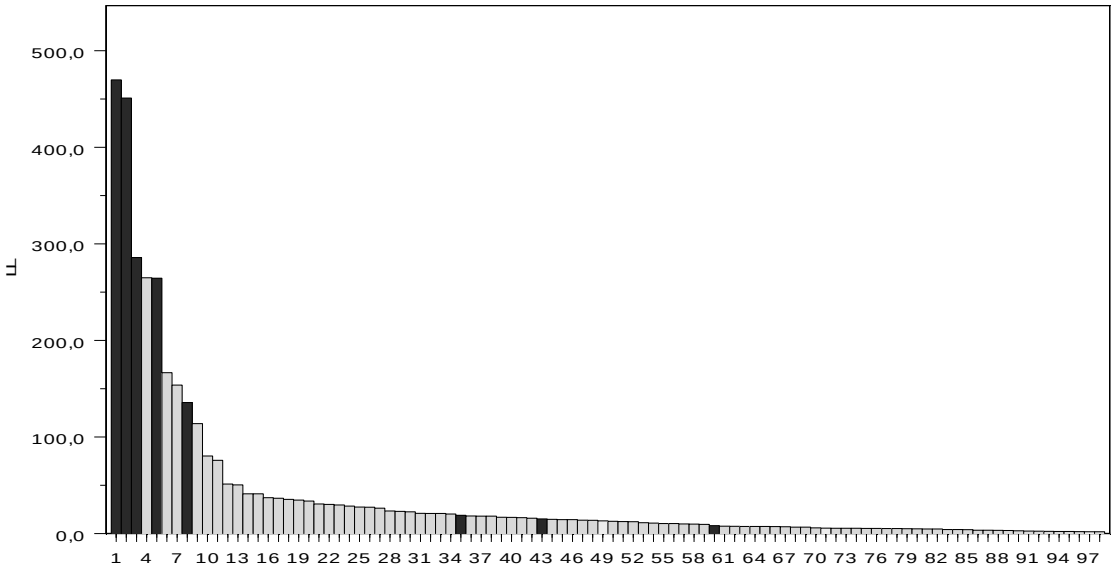


*Рис. 1. Значения меры log-likelihood для коллокаций со словом война*

Соответственно, ранги «правильных» коллокаций равны 5, 35, 43 и 60. На графике видно, что наибольшие значения LL соответствуют словосочетаниям, приведенным в БАС-17.

На графике, который приводится ниже, по оси ординат изображены значения меры MI, а по оси абсцисс ранги сочетаний.

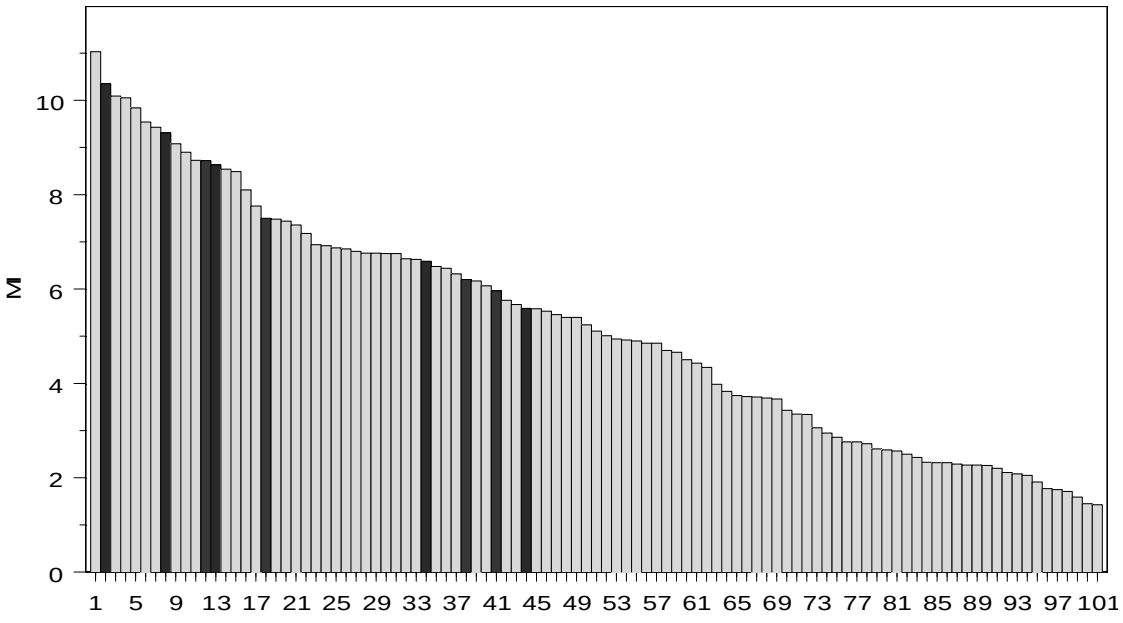


*Рис. 2. Значения меры MI для коллокаций со словом война*

Ранги «правильных» коллокаций равны 12, 18, 38 и 41. Устойчивые словосочетания, приведенные в БАС, распределены в первой половине шкале. Тем не менее, нельзя сказать о том, как тесно связан их ранг со значением меры.

Для меры t-score результаты приведены на графике ниже.

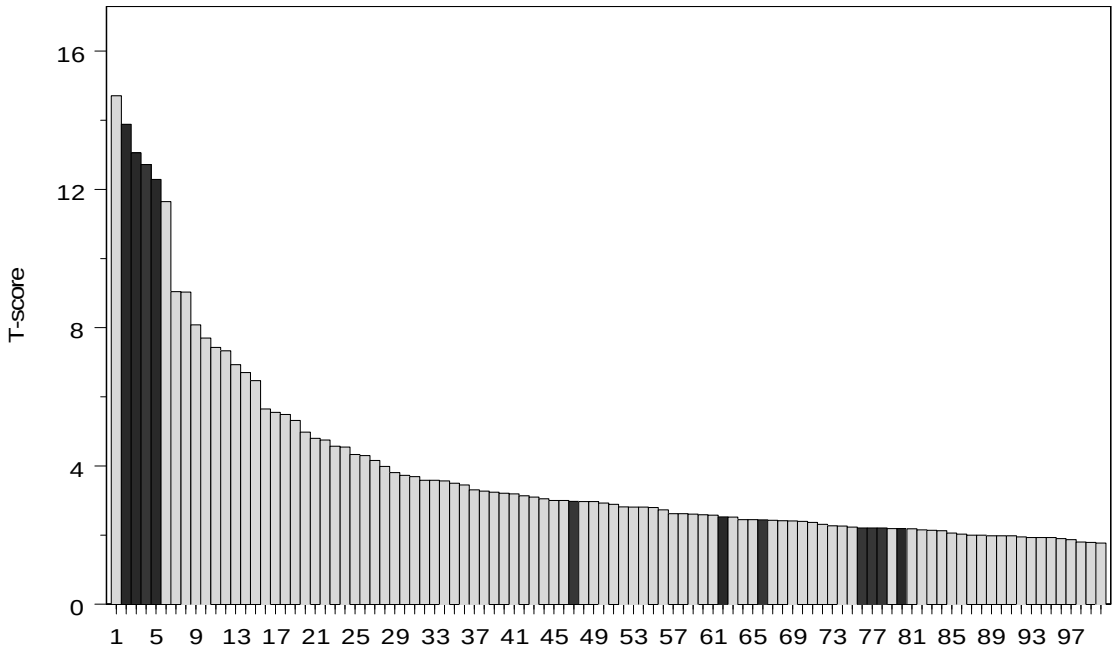


*Рис. 3. Значения меры t-score для коллокаций со словом война*

Ранги «правильных» коллокаций соответствуют 4, 47, 66 и 76-78.

Опять же, как и в случае с мерой log-likelihood, мы наблюдаем «скученность» словосочетаний, найденных в БАС, в самом верху графика.

Оказалось, что «правильных» коллокаций было получено мало, но это объясняется тем, что словарь коллокаций Е.Г. Борисовой, на который мы опирались, слишком мал, и его следует расширить. Можно сказать, что требуется новый словарь коллокаций, который на самом деле будет содержать разные типы устойчивых словосочетаний (фраземы, например, или словосочетания, представленные в толковых словарях за ромбом). Всего было обнаружено 106 биграмм для левого контекста слова *война*.

### *3.5. Анализ эффективности мер log-likelihood, MI, t-score для выявления коллокаций*

При анализе коллокаций для всех 19 существительных были получены следующие результаты.

Результаты для меры log-likelihood

Всего было найдено 1763 биграммы. Из них:
47 присутствуют в двух или более словарях;
79 только в словаре Е.Г. Борисовой;
48 только в словаре МАС;
20 только в словаре синонимов;

11 только в словаре БАС-25;
6 только в словаре БАС-17;
15 сочетаний со знаками пунктуации.

Значения меры log-likelihood оказались наибольшими для коллокаций, найденных в двух или более словарях. Возможно, это свидетельствует о том, что признаваемые устойчивыми словосочетания «весят» больше согласно мере log-likelihood.

Результаты для меры MI

Всего было найдено 1755. Из них:
68 присутствуют в двух или более словарях;
73 только в словаре Е.Г. Борисовой;
27 только в словаре МАС;
13 только в словаре синонимов;
9 только в словаре БАС-25;
25 только в словаре БАС-17;
11 сочетаний со знаками пунктуации.

Значения меры MI оказались наибольшими для коллокаций, найденных только в МАС, а также найденных в двух или более словарях. Также оказалось, что в диапазоне значений от 0 до 1 не были найдены словосочетания, которые можно было бы причислить к устойчивым. Это позволяет сделать вывод, что сочетания, значение меры ассоциации MI которых попадает в данный интервал, оказываются статистически незначимыми.

Результаты для меры t-score

Всего было найдено 1755. Из них:
71 присутствуют в двух или более словарях;
73 только в словаре Е.Г. Борисовой;
22 только в словаре МАС;
14 только в словаре синонимов;
8 только в словаре БАС-25;
23 только в словаре БАС-17;
20 сочетаний со знаками пунктуации.

Сочетания, обладающие большим значением t-score, оказались весьма частотными, поскольку, в отличие от предыдущих мер, одним из элементов являются предлоги, местоимения. А также было выделено бóльшее число (по сравнению с другими мерами) биграмм, в которых компонентом являются знаки препинания.

## 4. Заключение

Проведенное нами исследование показало возможность применения статистического аппарата для выделения устойчивых сочетаний на базе корпусов русского языка.

Для всех полученных сочетаний наблюдается одинаковая тенденция: чем меньше значение меры, тем больше вероятность, что эти словосочетания не зафиксированы как устойчивые в словарях русского языка. Таким образом, можно сказать, что данные о сочетаемости, приведенные в словарях, совпадают с данными, полученными на основе мер ассоциации. Большинство коллокаций (фразем), зафиксированных в словарях, оказывается в верхней части списка, составленного на основе одной из мер ассоциации. Это говорит о том, что данные коллокации имеют высокие показатели связанности.

Крайне важным представляется тот факт, что в результате эксперимента были выделены сочетания, не зафиксированные ни в одном из словарей. Анализ подобных сочетаний показал, что биграммы, находящие на самом верху списка (отсортированного по убыванию по одной из мер), с некоторой долей вероятности оказываются устойчивым и, следовательно, могут быть внесены в словарь. В нижней части списка в подавляющем большинстве случаев оказываются свободные сочетания. Отметим, что списки словосочетаний, приведенные в толковых словарях за ромбом, не могут считаться полными. Тем не менее, важно, что помещаемые туда единицы, обладают некоторой степенью устойчивости. Как следствие этого, результаты нашего эксперимента, с одной стороны, говорят о применимости описанных статистических мер в лексикографической практике, и, с другой стороны, указывают на известную неполноту существующих словарей.

Пока трудно сказать, что какой-то определенный тип устойчивых словосочетаний может быть выделен конкретной мерой ассоциации.

В рамках корпусных исследований, следует уточнять вероятностно-статистические методы и развивать программный инструментарий.

Следует также принимать во внимание структурные формулы, которые лежат в основе коллокаций. Их комбинация со статистическими подходами, по нашему мнению, может дать неплохой результат. При этом свое место должны занять программные способы исключения из речевого материала (или просто учета) так называемых стоп-слов и знаков препинания.

Очевидно, что автоматический анализ текста (например, с помощью описанного выше статистического аппарата) — это только первоначальный этап для выявления коллокаций. Затем требуется ручная обра-

ботка полученных результатов, в том числе с привлечением данных из словарей (в первую очередь, толковых и словарей сочетаемости).

**Литература**

Андреев, Н.Д.: 1967, *Статистико-комбинаторные методы в теоретическом и прикладном языковедении*, Ленинград.

Ахманова, О.С.: 1966, *Словарь лингвистических терминов*, Москва.

БАС-17: 1948-1965, *Словарь современного русского литературного языка: В 17 т.*, Москва, Ленинград.

БАС-25: 2004-2006, *Большой академический словарь*, Т. 1-6, Москва, Санкт-Петербург.

Борисова, Е.Г.: 1995a, *Коллокации. Что это такое и как их изучать*, Москва.

Борисова, Е.Г.: 1995b, *Слово в тексте. Словарь коллокаций (устойчивых словосочетаний) русского языка с англо-русским словарем ключевых слов*, Москва.

Браславский П., Соколов Е.: 2006, 'Сравнение четырех методов автоматического извлечения двухсловных терминов из текста', *Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая-4 июня 2006 г.)*, Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея (ред.), Москва, 88-94.

Дерибас, В.М.: 1983, *Устойчивые глагольно-именные словосочетания русского языка*, Москва.

Иорданская, Л.Н. & Мельчук, И.А.: 2007, *Смысл и сочетаемость в словаре*, Москва.

МАС: 1981-1984, *Словарь русского языка в 4-х тт.*, под ред. А.П. Евгеньевой, Москва.

Мельчук, И.А.: 1960, 'О терминах «устойчивость» и «идиоматичность»', *Вопросы языкознания*, Москва, 4, 73-80.

Словарь русских синонимов и сходных по смыслу выражений: 2006, Москва.

Сусов, И.П.: 1999, *История языкознания: Учебное пособие для студентов старших курсов и аспирантов*, Тверь.

Телия, В.Н.: 1996, *Русская фразеология: семантический, прагматический и лингвокультурологический аспекты*, Москва.

Шаров, С.А.: 2002, *Частотный словарь русского языка*, [электронный ресурс], http://www.artint.ru/projects/frqlist.asp

Ярцева, В.Н. (ред.): 1990, *Лингвистический энциклопедический словарь*, Москва.

Church K., Hanks, P.: 1990, 'Word association norms, mutual information, and lexicography', *Computational Linguistics*, 16(1), 22-29.

Cowie A.P.: 1978, 'The place of illustrative material and collocations in the design of a learner's dictionary', Strevens P (ed), *Honour of A S Hornby*, Oxford.

Evert, S.: 2004, *The Statistics of Word Cooccurences Word Pairs and Collocations*, PhD thesis, Universität Stuttgart.

Fano R.: 1961, *Transmission of Information*, Cambridge (MA).

Firth, J.R.: 1957, *Papers in Linguistics 1934-1951*, London.

Firth, J.R.: 1968, *Selected Papers of J.R. Firth, 1952-1959*, London.

Hausmann F.J.: 1979, ‘Un dictionnaire de collocations est-il possible?’, *Travaux de Linguistique et de Litterature XVII(1)*, Centre de philologie et de littérature romanes de l'université de Strasbourg, 187-195

Hausmann F.J.: 1985, ‘Kollokationen im deutschen Wörterbuch: ein Beitrag zur Theorie des lexicographischen Beispiels’, Bergenholtz, H. and Mugdon, J. (eds.), *Lexicographie und Grammatik*, Tübingen.

Stubbs, M.: 1995, ‘Collocations and semantic profiles: On the cause of the trouble with quantitative studies’, *Functions of Language*, 1.