Санкт-Петербургский государственный университет

Кафедра математической лингвистики

Направление: «Лингвистика»

Образовательная программа: «Прикладная и экспериментальная лингвистика»

Профиль: «Компьютерная лингвистика и интеллектуальные технологии»

**Параллельный конкорданс:**

**поиск и ранжирование переводных контекстов для иллюстрации переводов**

**в машинном словаре**

Выпускная квалификационная работа   
соискателя на степень магистра филологии

**Протопоповой Екатерины Владимировны**

Научный руководитель

к.ф.н., доц. Митрофанова О.А.

Рецензент: Тарелкин А.В.,

руководитель группы инструментов

оценки качества, «Яндекс»

Санкт-Петербург

2016

# Оглавление

Введение 4

1. Компьютерная двуязычная лексикография 7

1.1. Основные задачи и направления работы в компьютерной лексикографии 7

1.2. Задачи и методы их решения, общие для корпусной лингвистики, машинного перевода и компьютерной лексикографии 9

1.3. Компьютерная двуязычная лексикография 10

1.4. Электронные двуязычные словари 10

1.5. Словари систем машинного перевода 15

1.6. Автоматически составленные (машинные) двуязычные словари: методология составления и применение 18

1.7. Параллельные конкордансы 22

1.8. Выводы 26

2. Иллюстративный блок одно- и многоязычных словарей 28

2.1. Принципы формирования иллюстративного блока в словарях 28

2.2. Практическая реализация иллюстративного блока в двуязычных словарях 35

Русско-английский словарь под общим руководством проф. А.И. Смирницкого 35

Англо-русский словарь В.К. Мюллера 36

Большой англо-русский словарь под общим руководством И.Р. Гальперина 36

The Oxford Russian dictionary : Russian-English, English-Russian 37

Большой англо-русский словарь Abbyy Lingvo 37

Новый большой англо-русский словарь под руководством Ю.Д. Апресяна 38

2.3. Выводы 39

3. Устойчивые сочетания различных типов и опыт их лексикографирования 41

3.1. Основные концепции устойчивых сочетаний 41

3.2. Представление коллокаций в словарях сочетаний 50

Толково-комбинаторный словарь русского языка 51

A Dictionary of English Collocations 52

Oxford Collocations Dictionary 53

Collins Cobuild English Collocations 54

The BBI Combinatory Dictionary Of English 55

Англо-русский фразеологический словарь 56

Англо-русский словарь глагольных словосочетаний 57

Macmillan Collocations Dictionary 58

Устойчивые словосочетания русского языка 58

Русские глаголы и предикативы 59

Пособие по лексической сочетаемости слов русского языка 60

Словарь сочетаемости слов русского языка 61

3.3. Выводы 62

4. Поиск и ранжирование контекстов для англо-русского машинного словаря 63

4.1. Материал и инструменты исследования 64

4.2. Описание процедуры поиска контекстов 66

4.3. Подготовка обучающей и тестовой выборки 68

4.4. Анализ ошибок в контекстах-кандидатах 71

4.5. Общий алгоритм ранжирования параллельных контекстов 73

4.6. Факторы (признаки) ранжирования 74

Оценка по языковой модели (группа LM) 74

Относительная частота контекста (RelF) 76

Взаимная информация (MI) 76

Векторные модели (WV) 78

Семантическая близость (Sim) 80

4.7. Методы классификации 80

Случайный лес (Random forest) 80

Нейронная сеть прямого распространения 82

Оценка мультиклассификации 84

4.8. Оценка качества классификации 84

4.9. Результаты классификации 85

4.10. Сравнение ранжирования с простыми эвристическими методами 86

4.11. Значимость признаков классификации 87

4.12. Выводы 88

Заключение 90

Лексикографические источники 91

Литература 92

Приложение А. Сравнение и оценка ранжирования иллюстрирующих контекстов с помощью разных методов 101

Ранжирование контекстов на основе меры MI 101

Ранжирование контекстов на основе частоты перевода 104

Ранжирование 108

Ранжирование 111

Ранжирование 114

Ранжирование 118

# Введение

Работа посвящена поиску и ранжированию переводных контекстов для иллюстрации переводов в машинном словаре. Создание словарей автоматическими методами – сравнительно молодое направление компьютерной лексикографии, поэтому неудивительно, что исследования, посвящённые автоматическому составлению иллюстративного блока, очень редки. Можно предположить, что машинные словари в этом отношении следуют принципам классических «бумажных» словарей, но в действительности словарь, создаваемый «с нуля» или на основе бумажного издания и предназначенный для той или иной системы автоматического понимания текстов, – это реализация модели семантического, морфологического, синтаксического и других уровней естественного языка, основанной на идеях искусственного интеллекта [Караулов и др. 1982].

С другой стороны, та же проблема наблюдается и в лексикографии в целом и связана с разрывом между лингвистической теорией и лексикографической практикой. Лингвистическая описательная и аналитическая работа направлена на регистрацию широкого спектра явлений – нормы и отклонения от нормы, кода и узуса. Особенно это касается грамматики, то есть явлений морфологии и синтаксиса.

В лексикографической работе также проводится анализ наблюдаемых явлений языка и речи, но регистрации и систематизации подлежит не всё, а определённый класс явлений. В одних случаях описывается только норма, которая может быть в некоторых случаях результатом сознательного выбора специалистов, в других – параметры, характеризующие единицу языка с определённой точки зрения. Так, в словарях может описываться слово и его значение, восстанавливаемое на основе наблюдений за употреблением слова с текстах, слово и его происхождение, слово и его прагматические характеристики, например, сфера употребления.

Словари предназначаются обычно для широкой аудитории (носителей языка, изучающих язык) и при создании рассчитываются на долговременное использование. Лингвистические теории, в том числе касающиеся лексикографических проблем, разрабатываются специалистами для специалистов. Кроме того, они часто охватывают лишь конкретное языковое явление или класс случаев, тогда как словарь (в первую очередь, общего типа) должен охватить всё разнообразие одного или нескольких языков. Сложность разработки лексикографических принципов обусловлена и трудностью создания словаря вообще (некоторые толковые словари создаются десятилетиями, например, *Словарь современного русского литературного языка* издавался с 1948 по 1965 год, а работы по его созданию начались в 1937 году): словарь невозможно быстро переработать, чтобы принять во внимание новую трактовку какого-либо явления. Всё это приводит к тому, что составители словарей часто опираются либо на уже сложившуюся практику, либо на общепринятую теорию, а большинство теоретических разработок не используется вообще. Исключением является проект Толково-комбинаторного словаря, о котором подробнее будет сказано ниже. Следует лишь отметить, что в данном случае речь идёт о лексикографическом произведении, полностью основанном на лингвистической теории (модели «Смысл Текст»).

Лингвистические теории, касающиеся сочетаемости лексем, практически не употребляются в лексикографической практике. Говоря же об иллюстративных контекстах в двуязычном словаре, неизбежно приходится упоминать понятие сочетаемости. Поэтому в дальнейшем мы будем рассматривать и теоретические работы, возможно, отчасти отражённые в существующих словарях, и особенности лексикографической практики (то есть существующие словари).

***Целью*** работы является разработка алгоритма ранжирования иллюстрирующих контекстов для англо-русского машинного переводного словаря. Исходя из данной цели, а также учитывая указанные выше проблемы, формулируются следующие ***задачи***:

1. анализ принципов составления и особенностей существующих электронных словарей различных типов;
2. описание и анализ реализации иллюстративного блока в современной лексикографической практике;
3. описание подходов к сочетаемости и анализ их отражения в лексикографической практике;
4. выделение необходимых признаков для описания двуязычных контекстов, подходящих для иллюстрации статей в переводном словаре.

Предполагается также решение следующих ***практических задач***:

1. разметка потенциальных иллюстраций для создания обучающей выборки и дополнения набора признаков ранжирования;
2. реализация алгоритма ранжирования параллельных контекстов;
3. оценка качества ранжирования и релевантности выделенных признаков.

Для ранжирования параллельных контекстов, описанных с помощью набора признаков, используются такие ***методы*** машинного обучения, как метод случайного леса и нейронная сеть прямого распространения.

***Материалом*** для исследования реализации иллюстративного блока послужили существующие англо-русские словари и словари сочетаемости английского и русского языка. Параллельные контексты для машинного переводного словаря, которые используются при ранжировании, извлечены из параллельного англо-русского интернет-корпуса.

***Актуальность*** работы обусловлена тем, что, как сказано выше, проблема иллюстративного блока практически не разработана в машинной лексикографии вообще, а особенно – в двуязычной. С другой стороны, компьютерная лексикография практически не учитывает опыт «безмашинной» лексикографии и многочисленные лингвистические исследования. Наша работа направлена на восполнение этого пробела.

***Практические результаты*** работы используются в рамках сервиса «Яндекс. Переводчик» и обсуждались в статье [Protopopova и др. 2015], а также в докладе на конференции «Диалог’2015».

# Компьютерная двуязычная лексикография

## Основные задачи и направления работы в компьютерной лексикографии

Сфера компьютерной лексикографии довольно широка и включает в себя разнообразные методы составления различного рода словарей с использованием компьютерных технологий. Инструментами и материалом компьютерной лексикографии являются базы данных, компьютерные картотеки, инструменты анализа текста, которые позволяют в автоматическом режиме формировать словарные статьи, хранить словарную информацию и обрабатывать её. Множество различных компьютерных лексикографических программ можно разделить на две больших группы: программы поддержки лексикографических работ и автоматические словари различных типов, включающие лексикографические базы данных [Баранов 2001].

Развитие компьютерных методов в лексикографии началось в середине 1960 годов: с одной стороны, возникла необходимость перевести в машиночитаемый формат существующие словари, с другой стороны, стало возможным использование электронных корпусов при составлении новых словарей. Одним из первых словарей в машиночитаемом виде стал *Longman Dictionary of Contemporary English* [LDOCE; Proctor, 1978], который до сих пор является одним из наиболее популярных словарей английского языка. Поначалу электронные словари представляли собой копии бумажных (*Oxford English Dictionary*), но постепенно были разработаны специальные представления и форматы, позволяющие, например, хранить содержательную часть словаря отдельно от информации о графическом формате словарных статей.

Также следует упомянуть серию *Collins English Dictionary*, словари которой с 1979 года полностью создавались с использование электронных баз данных и корпусов.

Примерно в то же время (в 1985 году) по инициативе академика А.П. Ершова начинаются работы по созданию Машинного фонда русского языка[[1]](#footnote-1), одной из целей которого было развитие компьютерной лексикографии. Предполагалось решение следующих задач:

* Оборудование рабочих мест Института русского языка РАН компьютерами и объединение их в сеть;
* Сбор материала для теоретического и прикладного исследования русского языка и его преобразование в машинных формат; разработка средств хранения этого материала;
* Создание программных средств, для проведения теоретических и прикладных работ по исследованиям в области русского языка;
* Развитие прикладных направлений (лексикография, терминоведение, автоматическая обработка данных на естественном языке).

В дальнейшем были разработаны концепции компьютерного представления различных лингвистических данных, в том числе автоматические словари (*Автоматический Синтаксический словарь русского языка, Автоматический словарь синонимов русского языка, Автоматический вариант Словаря русского языка С.И. Ожегова, Автоматический словарь глагольного управления в русском языке* и др. (описания опубликованы в Бюллетене Машинного фонда русского языка, вып. 1–3)), система построения конкордансов и обеспечения лексикографической работы (UNILEX-T, UNILEX-D). В рамках проекта по созданию Машинного фонда русского языка была описана концепция компьютерного описания словосочетаний различной степени устойчивости [Борисова 1990].

Наши наблюдения, сделанные в ходе знакомства с направлениями работ в компьютерной лексикографии, позволяют сделать вывод о том, что автоматизация словарных проектов связана не столько с преобразованием существующих словарей в цифровой формат, а с созданием лексикографических комплексов, имеющих прямой выход к различным блокам лингвистического процессора, отвечающих за автоматический анализ текстов на уровне морфологии, синтаксиса, за автоматическое выделение устойчивых словосочетаний и ряд других задач.

## Задачи и методы их решения, общие для корпусной лингвистики, машинного перевода и компьютерной лексикографии

Прежде чем перейти к направлениям компьютерной двуязычной лексикографии, следует ввести группу понятий, относящихся скорее к области корпусной лингвистики и машинного перевода, которые в дальнейшем будут использоваться при описании различных словарей.

Двуязычный словарь предполагает описание лексики двух языков: *входного языка* – языка, с которого производится перевод, и *выходного* – языка перевода. Слово-заголовок статьи по отношению к статье обычно называют *заголовочным*. Если в словаре приводятся словосочетания, то в них выделяется *опорное слово* *–* слово, под которым это словосочетание приводится в словаре.

*Переводным эквивалентом* называют перевод, приписываемый лексеме входного языка, или пару «слово входного языка + слово выходного языка», которые являются переводами друг друга.

Для машинного создания словарей или в качестве вспомогательного материала используются параллельные и сопоставимые корпусы текстов. *Параллельный корпус –* многоязычный корпус, который состоит из текстов на одном языке и их переводов на другой (другие) язык (языки). *Сопоставимый (comparable) корпус –* набор текстов одной и той же тематической области, на двух или нескольких языках. Для удобства как машинного, так и ручного использования параллельных корпусов производится *выравнивание –* установление соответствия фрагментов исходного текста (слов, предложений) фрагментам перевода, выполняемое вручную или автоматически.

Лексикографы и лингвисты часто используют специальные инструменты для работы с корпусами – *корпус-менеджеры*. Для исследования удобен режим построения *конкорданса* – автоматическое извлечение набора контекстов для заданного явления (слово / словосочетание / грамматическая форма и др.).

## Компьютерная двуязычная лексикография

В области компьютерной двуязычной лексикографии постепенно выделилось несколько направлений:

* создание традиционных двуязычных словарей с помощью компьютерных технологий,
* разработка алгоритмов составления переводных эквивалентов для систем машинного перевода,
* автоматическое составление словарей, надстройки к параллельным и сопоставимым корпусам, параллельные конкордансы.

Рассмотрим каждое направление подробнее.

## Электронные двуязычные словари

Технология создания двуязычных компьютерных словарей наследует основные идеи работы над одноязычными лексикографическими ресурсами, поэтому начнём анализ существующих проектов с одноязычных компьютерных словарей. Зарождение компьютерной лексикографии в 60х годах 20 века связано именно с электронными версиями словарей. Так, *Webster’s Seventh New Collegiate Dictionary* (1967) и *New Merriam-Webster Pocket Dictionary* стали первыми словарями, электронная версия которых была подготовлена параллельно с бумажной. Как уже отмечалось, компьютерный набор тогда был ещё недостаточно развит, поэтому электронные словари создавались на основе оцифрованных бумажных препринтов.

В 70-е годы компьютерные технологии развивались по большей части в направлении автоматизации подготовки к изданию бумажных словарей. Компьютерный набор и некоторые другие средства подготовки статей использовались в *Longman Dictionary of Contemporary English* (1978) и первом издании *Collins English Dictionary* (1979). В то же время ведутся разработки по созданию так называемых карманных словарей и переводчиков (hand held dictionaries) – специализированных устройств, выполняющих поиск по словарю, представленному в машиночитаемом виде. Первые гаджеты появились в 1978 году (LK-3000, Craig M100), а технология оставалась популярной до первого десятилетия 21 века. Основными функциями карманных словарей был перевод заданного слова и озвучивание его произношения. Более продвинутые модели использовали распознавание речи и позволяли осуществлять поиск путём голосового ввода.

В 80-е годы, благодаря росту популярности и сокращению стоимости электронных носителей, появляются первые словари на дисках. Такой формат позволял мгновенно устанавливать словарь на любой компьютер и быстро распространять его. Среди многочисленных изданий словарей на CD-дисках нужно отметить издание *Oxford English Dictionary* (1988, использовалось издание 1928 года), *The American Heritage Dictionary of the English Language* (1992). Словари на электронных носителях улучшались с развитием технологий гипертекстовой разметки и вскоре ничем не уступали соответствующим бумажным изданиям. Кроме того, их пользователи могли задействовать такие функции, как озвучивание произношения, просмотр истории поиска, использование словаря непосредственно при чтении текста и т.п.

Развитие интернет-технологий привело к тому, что к 1998 году [Li 1998] в сети можно было найти около четырёхсот словарей английского языка, однако большинство были копиями, не защищёнными в соответствии с нормами авторского права. В 1999 году издательство Oxford University Press выложило в интернет с доступом по подписке *Oxford English Dictionary Online*, а вслед за ними многие другие издатели начали публиковать в интернете сокращённые версии словарей, предлагая купить подписку на расширенные версии. Затем стало понятно, что онлайн-словари могут приносить доход наряду с бумажными, поэтому сейчас многие популярные словари доступны без ограничений.

Создатели электронных словарей привлекают пользователей к дополнению и исправлению он-лайн версий или созданию новых словарных статей. Самый крупный полностью коллаборативный проект – это *Викисловарь*, который объединяет различные возможности лексикографического описания, в том числе, функционируя как многоязычный словарь. Описание слова включает фонетическую, этимологическую, синтаксическую, семантическую, сочетаемостную информацию о слове. В [Крижановский 2009] описывается разработка машиночитаемого словаря на основе данных русского Викисловаря. Русский Викисловарь является крупнейшим среди существующих викисловарей, его характерной особенностью является чёткая структура словарной статьи [Крижановский 2011].

Что касается англо-русских словарей, необходимо прежде всего упомянуть электронные словари *ABBYY Lingvo*. Первая версия, которая тогда называлась *LINGuist Volume, v1.0* была выпущена в 1990 году и содержала 35 тысяч словарных статей.

С 2008 года под названием ABBYY Lingvo выпускаются версии x3, x5, x6, содержащие одно- и двуязычные словари, общее количество статей достигает 1,5 миллионов. Некоторые из них составлены лексикографами компании ABBYY, другие представляют собой электронные версии бумажных словарей (*Большой толковый словарь русского языка* Т.Ф. Ефремовой, *Англо-русский словарь* под редакцией В.К. Мюллера и т.п.). Возможности электронных словарей Lingvo включают:

* поиск слова по заголовкам словарных статей, поиск по неначальной форме слова, поиск по всем доступным языкам;
* полнотекстовый поиск, в том числе, по примерам употребления;
* озвучивание транскрипции;
* перевод по наведению курсора в браузере, текстовых документах;
* сохранение и просмотр избранных словарных статей.

На данный момент доступен также онлайн-сервис *LingvoLive* (https://lingvolive.ru), где пользователи могут не только осуществлять поиск по словарям, но и комментировать словарные статьи, добавлять свои примеры употребления и сохранять историю поиска.

Хранение информации электронном словаре осуществляется в базах данных или в специально разработанных для этой цели форматах представления, основанных на XML. Например, рекомендации Text Encoding Initiative (TEI) содержат отдельную главу, посвященную хранению лексикографических данных (<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html>). Элемент *entry* позволяет хранить разнообразную информацию, содержащуюся в словарной статье, причём возможно сохранение словарной статьи в различных видах:

1. *типографическое представление* – позволяет восстановить статью в том виде, в котором она должна быть напечатана;
2. *«редакторское» (editorial) представление* – хранит текст без специфической мета-информации;
3. *лексическое представление* – содержит структурированную информацию из словарной статьи безотносительно особенностей типографии, однако сохраняет лингвистические категории (поле грамматических помет, особенностей употребления и т.п.)

Статья *entry* может содержать отдельные разделы для омографов (hom), внутри каждого из которых возможны описания форм слова (form), грамматических помет (gramGrp), толкования (def), примеры употребления с отсылкой к источнику (cit), информацию об особенностях употребления (usage), перекрёстные ссылки на другие статьи (xr, re), этимологическую информацию (etym). Возможности хранения лексикографической информации в формате TEI рассматриваются, например, в работах [Захаров и др. 2011; Захаров 2013, Буторова и др. 2016].

Другой вариант представления – Lexical markup formal (LMF) – использует только последний тип (лексическое представление), поскольку он предназначен в первую очередь для хранения всевозможных словарей систем обработки текстов. Этот детально разработанный формат используется в сложных системах хранения лингвистических данных, например, CLARIN (http://www.clarin-d.de/en/).

Вторая, не менее важная составляющая электронного словаря – программа, отвечающая за отображение словарных статей, поиск, редактирование. Как уже было сказано, многие электронные словари предоставляют веб-интерфейс, в котором взаимодействие между пользователем и словарём осуществляется обычно с помощью запросов к базе данных. Программы, устанавливаемые на компьютер или портативное устройство, часто имеют возможность подключения дополнительных словарей и расширений. Среди программ, содержащих англо-русские словари, наиболее популярны *ABBYY Lingvo*, *GoldenDict*, *AtomicDict* и другие.

Обычно главными элементами интерфейса являются окно поиска и окно отображения словарной статьи (см. Рис. 1):

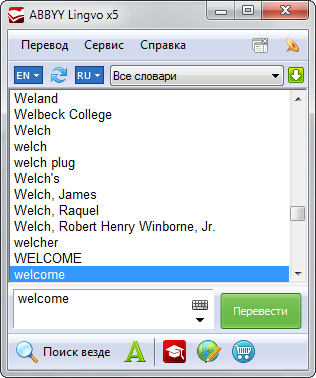


Рисунок 1. Окно поиска ABBYY Lingvo x5.

Электронные двуязычные словари также используются при создании широкомасштабных лексических баз данных. Например, в проекте многоязычного словаря *PanDictionary* [Mausam и др. 2009; Mausam и др. 2010] было объединено около 600 онлайн-словарей, в частности из Викисловарей, причём переводы из разных словарей были автоматически сгруппированы по смыслу с помощью графового алгоритма. В статье приводится объём полученной базы – 10 миллионов слов для тысячи языков, однако эти результаты не были опубликованы в открытом доступе. Проект *Linguee* (<http://www.linguee.ru>, рис. 2) также объединяет материалы различных словарей в рамках одного онлайн-ресурса, причём эти словари пополняются на основе данных, собранных по большому параллельному корпусу. Этот же корпус служит источником для построения параллельного конкорданса (см. ниже).

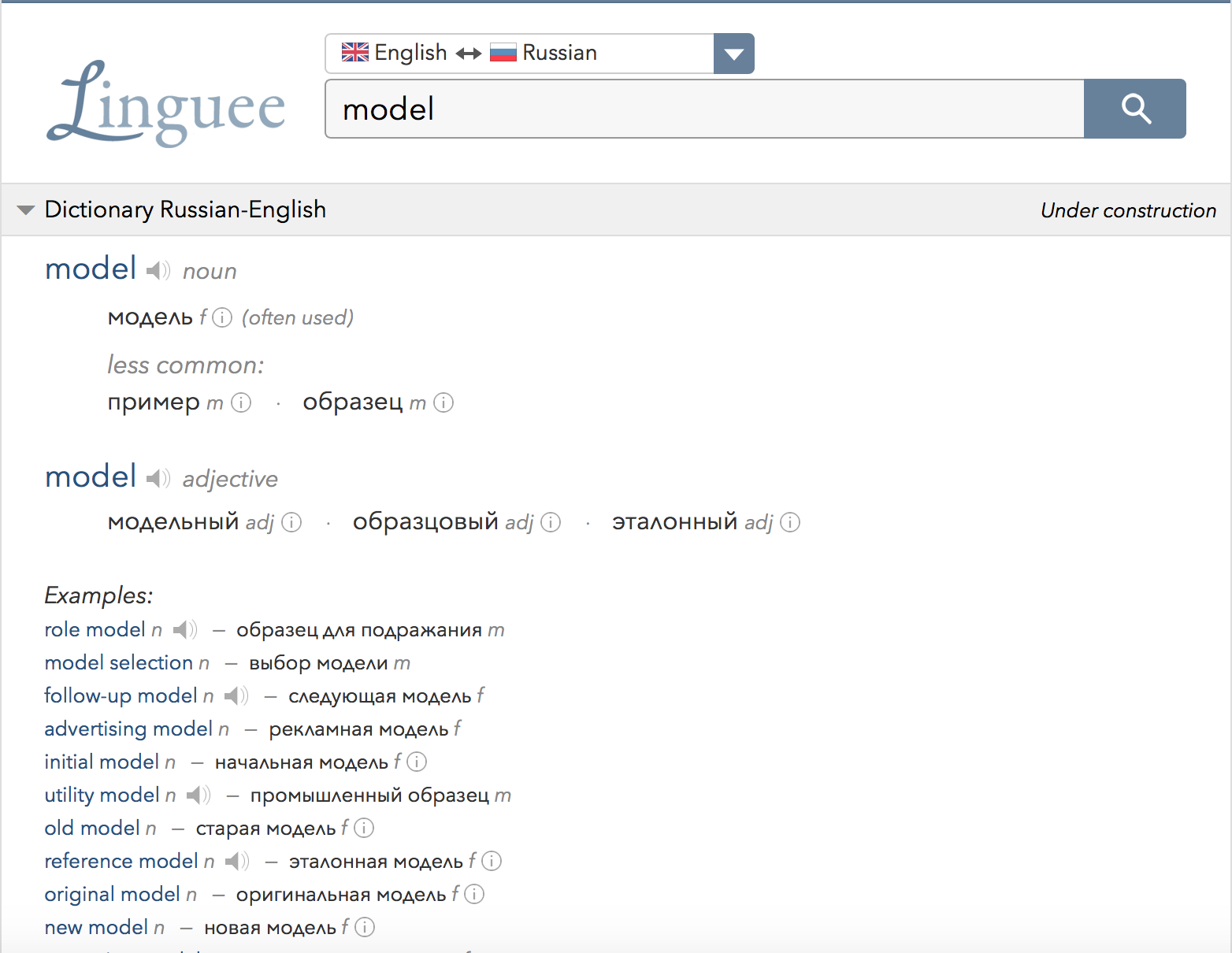


Рисунок 2. Пример статьи из электронного словаря системы Linguee.

## Словари систем машинного перевода

*Словарями систем машинного перевода* (*фразовыми таблицами, моделями перевода, phrase table*) обычно называют автоматически составленные таблицы переводов последовательностей слов с весами, которые затем учитываются при машинном переводе текста. Переводы извлекаются из параллельного корпуса на основании выравнивания предложений. Для построения модели перевода были предложены алгоритмы, известные как модели IBM [Brown и др. 1993]. Они оценивают вероятности перевода с помощью EM-алгоритма, который позволяет итеративно оценить модель на неполных данных. Он состоит из двух шагов, которые обычно повторяются до сходимости:

* E-шаг (expectation), на котором имеющаяся модель применяется к данным;
* M-шаг (maximization), на котором из данных выводится новая модель (корректируется старая).

Итак, IBM model 1 действует следующим образом. Пусть есть корпус, состоящий из трёх параллельных фраз (см. Рисунок 3):

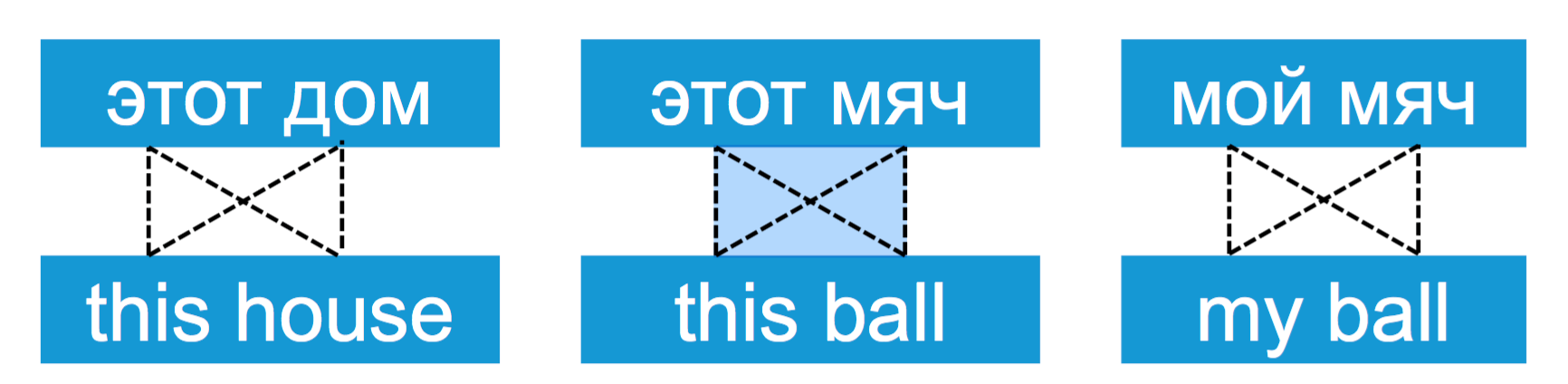


Рисунок 3. Пример параллельного корпуса для обучения модели IBM.

Изначально в корпусе нет информации о пословном выравнивании, веса задаются равномерно (0.25 для каждого перевода). Для оценки вероятности того или иного пословного выравнивания и используется ЕМ-алгоритм. На каждом М-шаге для каждого перевода (пары) суммируются все веса. На первом шаге сначала инициализируются вероятности всех возможных выравниваний, а затем по имеющимся данным для каждого слова в каждом входном предложении суммируются веса различных вариантов его выравнивания. Так, например, по первому предложению вероятность выравнивания и вес буду равны 0.5. Затем для каждого входного слова *e* (в данном случае на русском языке) и всех вариантов его перевода *f* (на английском) вычисляется итоговый вес , равный отношению вероятности выравнивания к весу слова на языке перевода. На первом шаге это значение 0.5.

Приведём псевдокод алгоритма оценки весов модели IBM-1 по [Koehn 2010]:

Вход: набор пар предложений

Выход: вероятность перевода

Задать равномерно

Пока не сходится

// инициализация

для всех

для всех

для всех предложений

// вычислим нормализацию

для всех слов *е* в *е*

для всех слов в

// суммируем

для всех слов в

для всех слов *f* в *f*

// оценим вероятности

для всех слов

для всех слов

Следующие модели являются усовершенствованиями первой. Основные дополнения следующие:

* *IBM Model 2* определяет вероятность всего выравнивания;
* *IBM Model 3* добавляет так называемую *fertility model* – как много переводов возможно для каждого входного слова;
* *IBM Model 4* добавляет модель относительного выравнивания;
* *IBM Model 5* регулирует заполнение только свободных позиций при выравнивании.

Подробное сравнение моделей IBM можно найти в [Och, Ney 2003].

Эти модели могу применяться как на уровне пословного перевода (*word-based translation*), так и на уровне перевода фраз (*phrase-based translation*). С этим связано и основное отличие словарей систем машинного перевода от классических словарей – помимо слов и многословных выражений (обычно устойчивых) в них включаются словосочетания, а также последовательности слов (n-граммы), которые синтаксически не связаны между собой. Соответственно, в таких словарях отсутствует понятие нормальной формы: для каждой формы слова существуют свои варианты перевода.

За прошедшие 20 лет были предложены различные улучшения и новые способы подбора параметров для моделей IBM [Turkato 1998; Och, Ney 2000; Tsunakawa и др. 2008; Huck и др. 2011; Luo, Lepage 2015]. Кроме того, одним из актуальных направлений остаётся фильтрация машинных лексиконов, полученных путём автоматического выравнивания [Melamed 1996]. Улучшенные модели анализируются и сравниваются в [Huck и др. 2011].

Морфологическая и синтаксическая информация также может быть использована при построении моделей перевода, в частности, для фильтрации шумных переводов. Некоторые авторы предлагают извлечение переводных эквивалентов по структурам зависимостей [Yamamoto, Matsumoto 2000].

## Автоматически составленные (машинные) двуязычные словари: методология составления и применение

С 1990 годов начинаются многочисленные исследования, посвящённые автоматическому извлечению переводных эквивалентов из параллельных корпусов.

Полученные словари обозначали как *machine-readable dictionaries (MRD)* или *bilingual lexicons*. Однако, поначалу такие словари использовались в системах машинного перевода (см. выше) или в качестве вспомогательных источников для лексикографов. Подобные ресурсы обычно содержат только информацию о возможных переводных эквивалентах, вероятности перевода (или просто частоте конкретного перевода в корпусе); иногда входные слова и словосочетания снабжаются морфологической информацией.

В противоположность традиционным бумажным словарям, которым присущи «универсальность, всеобъемлемость и мультиприложимость» [Марчук 1976: 2], машинный словарь строго ориентирован на заданную совокупность текстов. Отбор материала для машинного словаря происходит на протяжении всего времени его функционирования в системе автоматической обработки текстов. Принципы формирования словника машинных словарей основываются в большей степени на точных методах, и в меньшей – на логико-интуитивных методах, которыми руководствуются разработчики традиционных словарей. Можно сказать, что традиционный бумажный словарь представляет собой статическое описание, существующее в диахронии, тогда как машинный словарь – динамическое описание, существующее в строгой синхронии.

Развитие интернет-технологий, увеличение объёма доступных параллельных корпусов текстов сделали возможным построение машинных словарей, сопоставимых по размеру с традиционными бумажными или электронными словарями. Кроме того, существует большое количество инструментов автоматической обработки текста, которые позволяют преобразовать информацию из машинного словаря, предназначенного, например, для системы машинного перевода, в электронный словарь, ориентированный на пользователя.

Как уже было сказано, технология извлечения переводных эквивалентов из корпуса разрабатывается довольно давно, и одним из наиболее значимых результатов многолетних исследований явился класс разнообразных методов фильтрации шумных переводов из фразовых таблиц. Таким образом, главной задачей при создании автоматических переводных словарей является включение в них вспомогательной информации: группировка переводов в синонимические ряды, иллюстрация особенностей употребления, грамматические и стилистические пометы.

Первая задача – группировка переводов из фразовой таблицы – несмотря на кажущуюся очевидность её трактовки, имеет лишь несколько известных прикладных решений. В [Bansal 2012] описывается реализация алгоритма группировки переводных эквивалентов, основанная на кластеризации методом K-средних. В качестве факторов кластеризации используются информация об обратном переводе и контекстные представления переводных эквивалентов. Полученные «кластеры смыслов» сравниваются со смыслами из WordNet. Небольшая часть статьи посвящена иллюстрации различных вариантов перевода, оказавшихся в одной группе.

Другой возможный подход описан в [Antonova, Misyurev 2014]: группировка смыслов опирается на предварительно собранный словарь синонимов. Этот словарь генерируется автоматически на основании дистрибутивной векторной модели [см. также Antonova и др. 2016] и сходстве переводов слова. В этой статье [Antonova, Misyurev 2014] описывается создание машинного словаря сервиса Яндекс.Переводчик (http://translate.yandex.ru), который будет использован в практической части данной работы. Рассмотрим на этом примере технологию создания машинного словаря в общем виде. Автоматическое построение переводного словаря предполагает следующие этапы:

1. Предобработка параллельного корпуса (извлечение фразовой таблицы, морфосинтаксический анализ, лемматизация, фильтрация «шумных» переводных эквивалентов).
2. Объединение переводных эквивалентов для каждого слова в ряды синонимов.
3. Добавление грамматических помет.
4. Генерация обратных переводов, то есть синонимов входного слова. В данном случае этот этап производится на основании полученных прямых переводов.
5. Иллюстрация каждой группы переводов примерами употребления из параллельного корпуса.

В результате формируется словарная статья, пример который представлен на рис. 4.

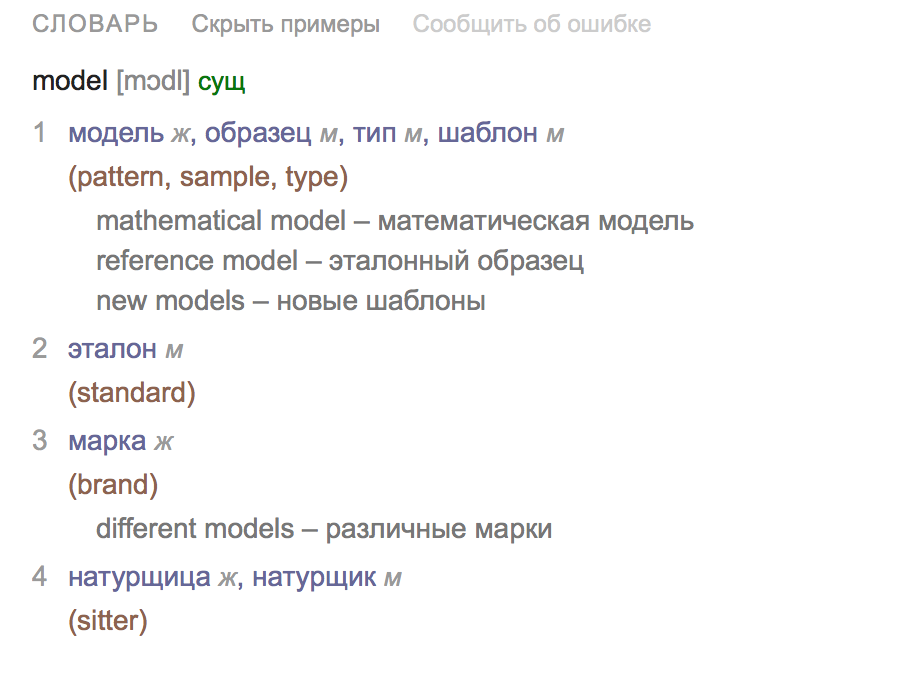


Рисунок 4. Пример словарной статьи из машинного словаря сервиса Яндекс.Переводчик.

Качество параллельного корпуса представляет определённую проблему в развитии автоматической переводной лексикографии. Поэтому в некоторых работах по автоматическому построению двуязычных словарей в качестве параллельных корпусов используются специфические источники. Так, в [Tyers, Pienaar 2008] описывается метод создания словаря на основе заголовков статей Википедии, которым авторы предлагают пользоваться в частности для языков, для которых нет корпусов достаточного объёма – так называемых *under-resourced languages*.

Последним направлением, активно развивающимся в последнее десятилетие, является построение переводного словаря по данным сопоставимого (*comparable*) корпуса. С одной стороны, задача оказывается более актуальной, поскольку для многих пар языков параллельные корпуса доступны в ограниченном объёме (если вообще доступны). С другой стороны, извлечение переводных эквивалентов из похожих корпусов требует более сложных методов. Все эти методы используют начальный словарь, состоящий из небольшого набора переводных эквивалентов для заданной языковой пары, а затем осуществляется поиск «похожих» слов входного языка и их переводов в сопоставимом корпусе. Часто применяются графовые модели: в [Sadat и др. 2002; Laws и др. 2010; Yu, Tsujii 2009; Fung 2000; Chatterjee идр. 2010] используется сходство синтаксических деревьев для предложений на разных языках, при условии, что имеется исходный словарь небольшого размера. Возможны разные подходы в зависимости от используемой метрики сходства, способа построения контекстных векторов и так далее. Авторы [Tamura и др. 2012] предлагают улучшения для методов, основанных на сходстве, которое заключается в использовании непрямых отношений между словами (когда вершины графа связаны более чем через одно ребро). Для близких языков предлагаются методы, которые используют не специальный начальный словарь, а, например, слова, общие для обоих языков [Fiser, Ljubesic 2011]. В [Ismail, Manandhar 2010] сначала извлекаются наиболее близкие и относящиеся к одной теме слова, для которых перевод будет наиболее надёжным. Улучшение словаря также может достигаться за счёт добавления параллельного корпуса [Morin, Prochasson 2011] или использования методов разрешения лексической неоднозначности [Bouamor 2013]. Отдельные работы посвящаются проблеме состава начального словаря [Hazem, Morin 2012].

## Параллельные конкордансы

Задача иллюстраций словарных статей контекстами употребления заголовочных слов частично совпадает с задачей построения конкордансов, то есть поиска контекстов заданных слов или языковых явлений в корпусах текстов.

Системы, строящие параллельные конкордансы, как правило, находят параллельные предложения или даже целые абзацы, а не короткие примеры употребления. Самые простые системы осуществляют поиск только по одной части параллельного корпуса (только входному / выходному языку) и не ранжируют выдачу [Kjaersgaard 1987, Langlois 1996].

Некоторые из систем отображают сведения о частоте данного слова и его переводов – например, в [Barlow 2004] описана система поиска по параллельному корпусу *ParaConc*, которая выводит информацию о частоте отдельных слов и их коллокатов. Системы, которые ранжируют контексты, используют статистику встречаемости перевода. Так, в [Wu и др. 2003] авторы располагают предложения и их переводы в соответствии с частотой перевода запроса, а в [Bai и др. 2012] наиболее релевантные контексты находятся на основании статистики Дайса [Dice 1945].

Такие системы обычно предназначены для лексикографов и профессиональных переводчиков. Рассмотрим две системы с доступным веб-интерфейсом, которые осуществляют поиск по параллельным англо-китайским корпусам: *DOMCAT* [Bai и др. 2012] и *TotalRecall* [Wu и др. 2003].

Примеры выдач на поисковые запросы в системах *DOMCAT* и *TotalRecall* (собственно конкордансы) представлены на рисунках 5 и 6. Система DOMCAT, как описывается в [Bai и др. 2012], находит наиболее релевантные контексты с помощью статистического критерия Дайса (коэффициент Сёренсена-Дайса):

где и - частоты слова и переводного эквивалента соответственно, а - частота совместного появления этих слов.

Однако, в веб-интерфейсе не содержится никакой информации о ранжировании и принципах работы системы.

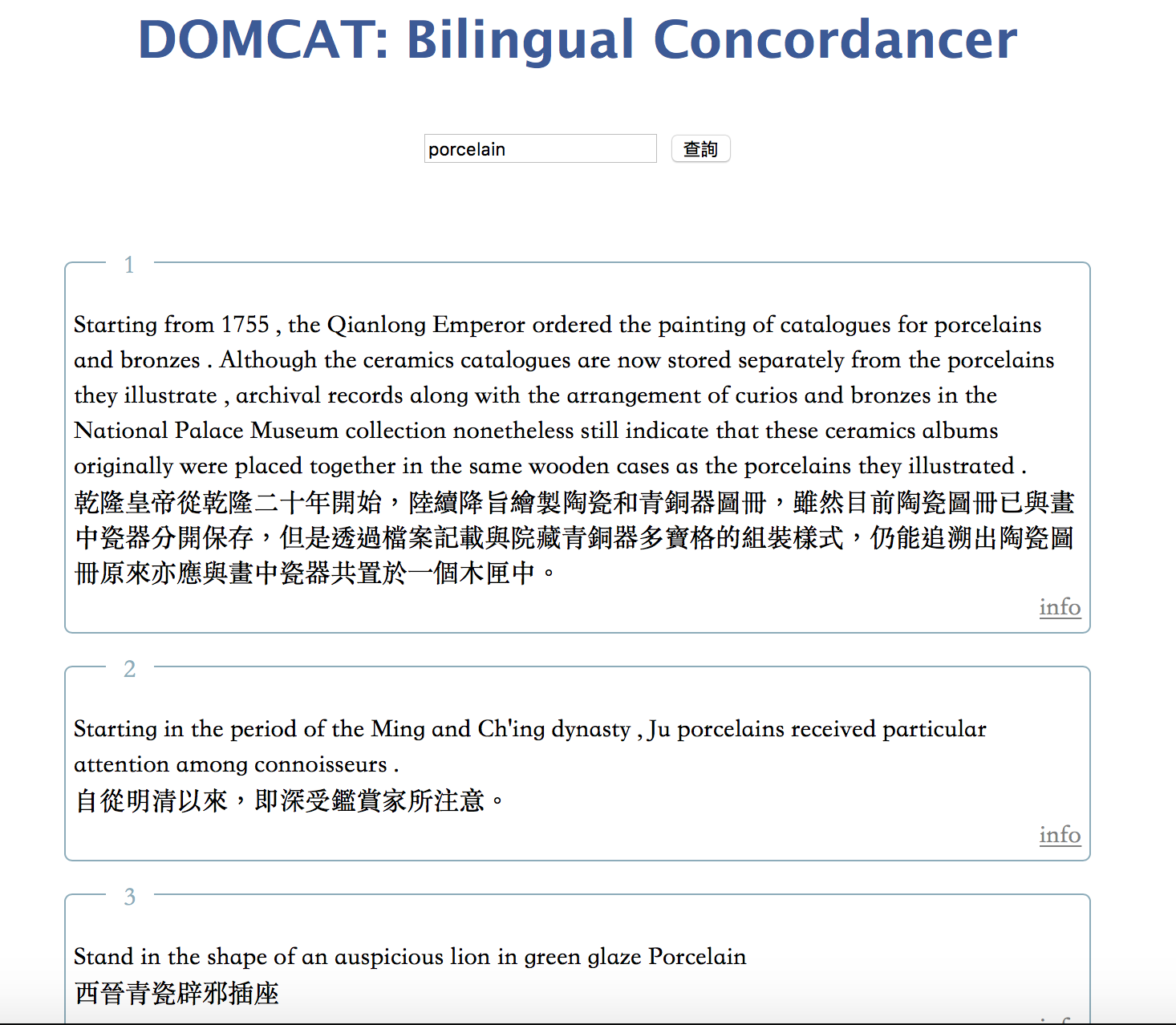


Рисунок 5. Результаты поиска по запросу "porcelain" в системе DOMCAT.

Вторая система – TotalRecall – предоставляет две возможности – ранжирование по частоте и по логарифму правдоподобия (log-likelihood ratio, LLR). Применительно к связи между двумя явлениями (например, встречаемостью слова и его перевода) последняя мера вычисляется как [Dunning 1993].

где – совместная частота слова и перевода, а , – частоты *i* и *j* соответственно.

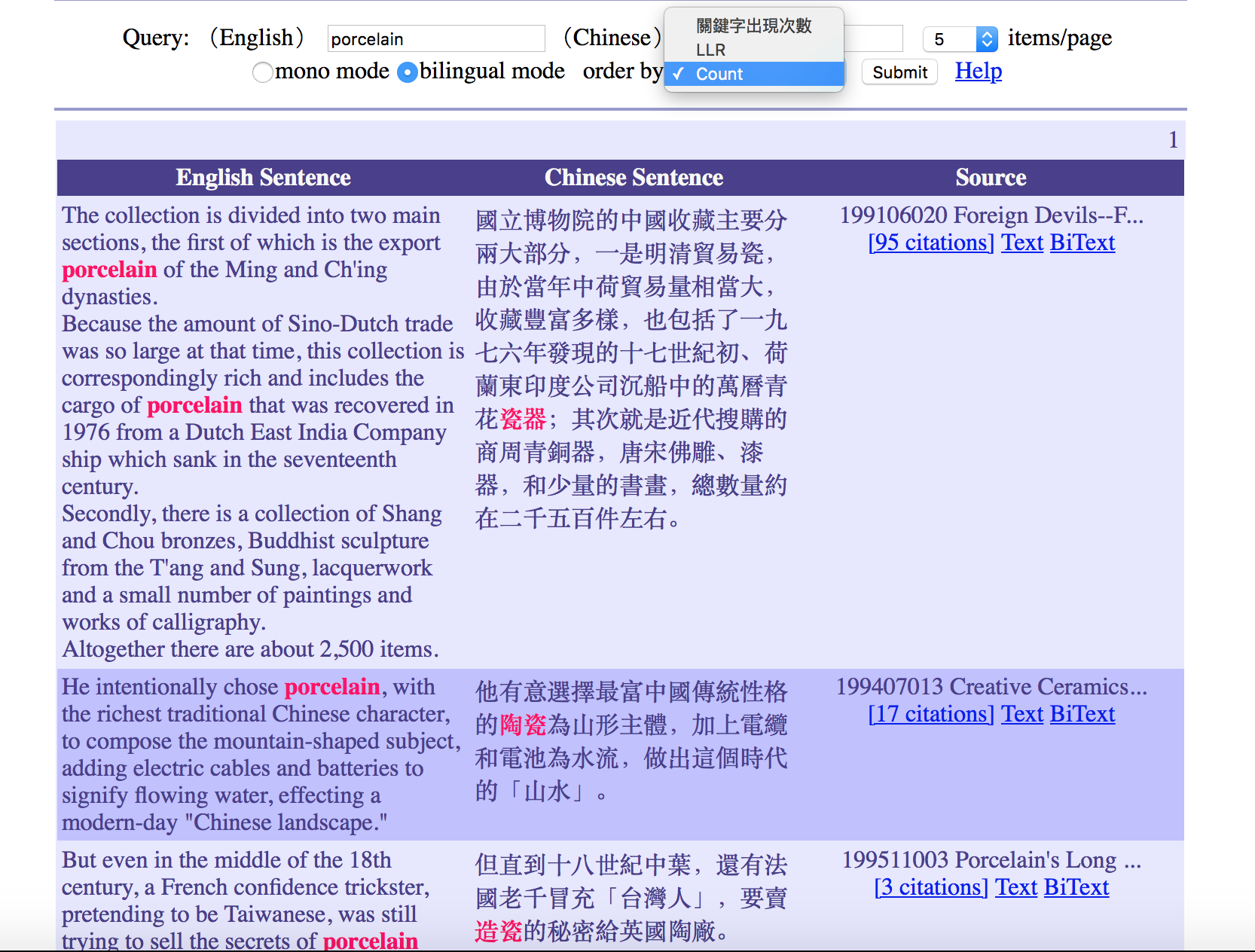


Рисунок 6. Результаты поиска по запросу "porcelain" в системе TotalRecall.

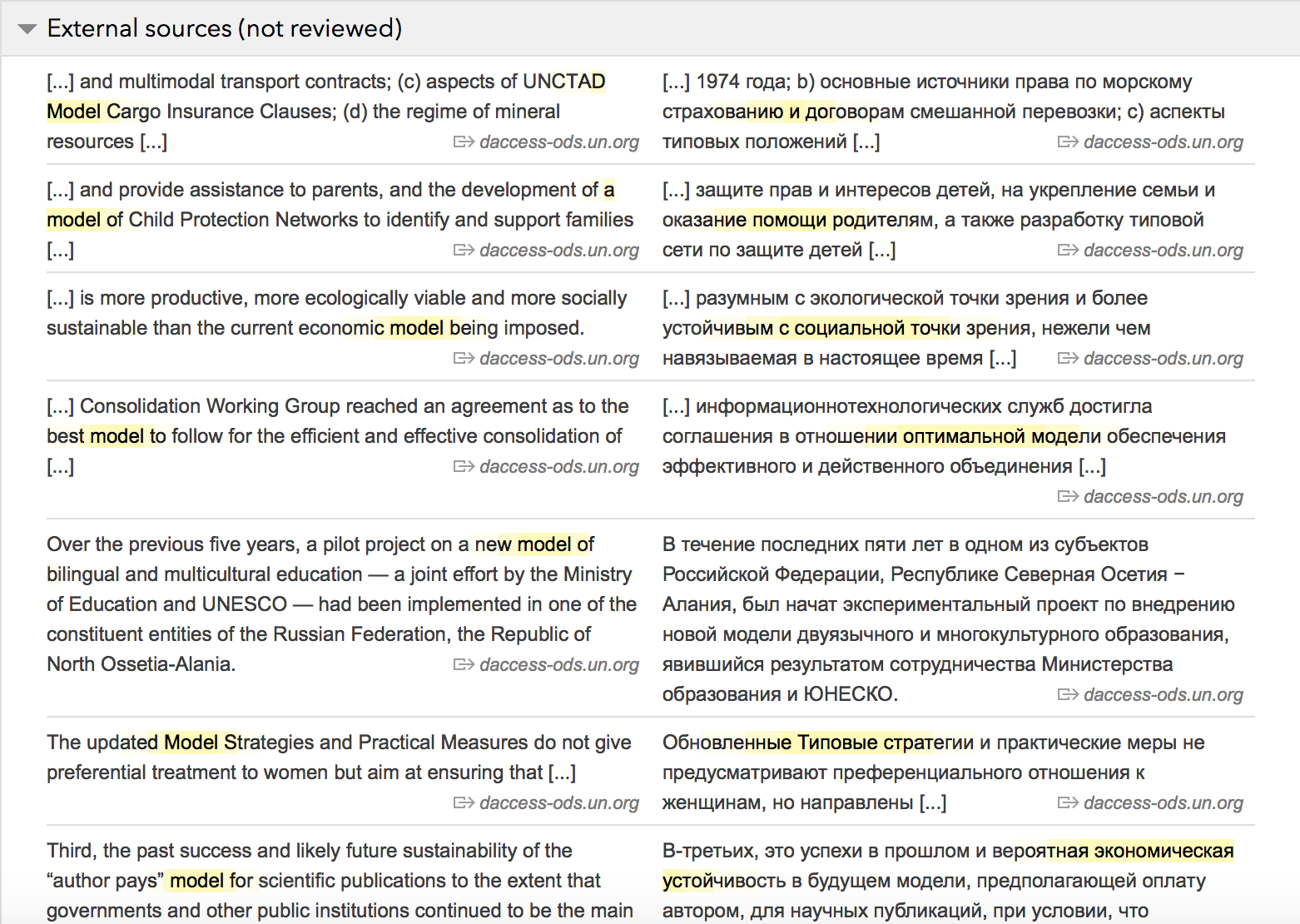


Рисунок 7. Пример конкорданса системы Linguee.

Параллельный конкорданс в качестве дополнения к электронному словарю строит многоязычная система Linguee (http://www.linguee.ru). Она предназначена как для изучающих язык, так и для профессионального перевода. Конкорданс для запроса на входном языке строится по большому корпусу текстов различных жанров; особое внимание уделяется технической и специальной (например, юридической и медицинской) литературе. Параллельные тексты собраны из интернета и не проверялись. Весь многоязычный корпус содержит, по описанию разработчиков, около 10 миллионов примеров, при этом в документации не уточняется, идёт речь о коротких словосочетаниях или о параллельных предложениях.

Следует отметить одну из особенностей системы – поиск может осуществляться не только по запросу из одного слова, но и по фразе. Поскольку корпус содержит информацию о пословном выравнивании, при отображении результатов поиска выделяется как текст запроса, так и его вероятный перевод в тексте на языке перевода, однако не всегда это соответствие корректно (см. Рисунок 7).

## Выводы

В первой главе были рассмотрены основные направления компьютерной двуязычной лексикографии: технологии создания электронных словарей традиционными методами, построение словарей систем машинного перевода, автоматическое построение электронных словарей и параллельных конкордансов.

Автоматический поиск иллюстрирующих контекстов в компьютерной лексикографии в основном сводится к задаче построения параллельного конкорданса. Существующие решения, касающиеся поиска коротких примеров употребления и перевода, практически не описываются подробно, а чаще всего составители электронных словарей обращаются к иллюстрациям, подобранным вручную. Таким образом, поскольку эта задача не имеет полного и универсального решения в компьютерной лексикографии, нам представляется необходимым рассмотреть теоретические предпосылки и практические методы отбора иллюстративных примеров в традиционных и компьютерных словарях. Этому посвящена Глава 2.

# Иллюстративный блок одно- и многоязычных словарей

## Принципы формирования иллюстративного блока в словарях

В одноязычных словарях примеры предназначены для иллюстрации основного содержания словарных статей, а метод их подбора практически полностью определяется функциями словаря. Например, в толковых словарях это иллюстрация значения: примеры дополняют толкование (причём могут даже содержать информацию энциклопедического характера), иллюстрируют грамматическую и лексическую сочетаемость, стилистические особенности употребления, коннотации и ассоциации. Количество и характер иллюстраций при заголовочном слове даже указывает на частоту употребления и может отражать данные живой речи. Поскольку данная работа посвящена двуязычным словарям, мы рассмотрим лишь некоторые характеристики иллюстраций одноязычных словарей – в основном, те проекты и идеи, которые касаются компьютерной обработки лингвистических данных.

Двуязычные словари выполняют другие функции – они предназначаются для перевода с родного языка на неродной и обратно, как при обучении, так и при, например, профессиональном переводе. Из одноязычных словарей по цели наиболее близки к двуязычным словари устойчивых сочетаний различных типов и учебные словари.

«К словарям, тем более двуязычным, не принято писать теоретических введений», пишет Ю.Д. Апресян во введении к *Новому большому англо-русскому словарю*, составленному под его руководством [Апресян 2000: 12]. Выбор той или иной стратегии для описания конкретного явления в словаре часто явно не выражается в словаре. Исключения составляют теоретические опыты, например, *Толково-комбинаторный словарь* (Мельчук и др. 1984), о котором речь пойдёт в следующей главе. Что касается отбора иллюстрирующих контекстов, то в одноязычной лексикографии они изначально служат материалом для составления толкований. В этом плане интересным представляется замечание В.П. Беркова в его книге «Двуязычная лексикография»: «Проблема отбора словосочетаний – одна из важнейших в двуязычной лексикографии. Именно отбором словосочетаний переводные словари с одним и тем же входным языком чаще всего отличаются один от другого, именно отбор словосочетаний представляет собой наиболее спорную и уязвимую часть практически каждого двуязычного словаря» (Берков 2004: 61). Словосочетания, по его мнению, вообще составляют часть словника двуязычного словаря (действительно, во многих словарях указывается как количество статей, так и количество включённых словосочетаний). В роли иллюстраций, по мнению В.П. Беркова, выступают свободные сочетания: они дополняют семантизацию, иллюстрируют сходную структуру значений, показывают нюансы конкретного значения. Примеры должны

1. быть информативными, сообщать некую информацию, дополнительную по отношению к переводу входного слова;
2. располагаться в порядке следования соответствующих переводных эквивалентов;
3. группироваться по значению.

Несвободные, фразеологизированные сочетания, очевидно, по мнению В.П. Беркова, включаются в словарь независимо от того, является ли перевод пословным.

Наиболее чётко принципы отбора иллюстраций для двуязычных словарей (как пассивного, так и активного типа) сформулированы в пособии по лексикографии «The Oxford Guide to Practical Lexicography» [Atkins, Rundell 2008]. Цель иллюстрирующих примеров – помочь пользователям словаря выбрать необходимый переводной эквивалент и правильно его использовать. В связи с этим нужно:

* указать, какое значение заголовочного слова иллюстрируется;
* предупредить пользователя о некорректных употреблениях;
* указать значения переводных эквивалентов в случае полисемии.

Поскольку в пособии [Atkins, Rundell 2008] речь идёт о традиционном двуязычном словаре, авторы замечают, что было бы слишком трудоёмко отбирать примеры переводов из корпуса. Вместо этого они рекомендуют использовать имеющиеся списки коллокаций, таким образом, иллюстрировать переводы в контексте этих коллокаций.

Функциональное описание иллюстраций в двуязычных словарях даётся в монографии [Butina-Koller 2005]: «Стандартные ситуации, в которых используется двуязычный словарь (восприятие текста на иностранном языке, перевод текста с иностранного языка на родной, перевод с родного языка на иностранный или более или менее свободное текстопроизводство на иностранном языке), и обусловленные этими ситуациями различные функции словарей должны учитываться при принятии конкретных решений в лексикографической практике». В работе далее подробно анализируются существующие решения в русско-французской лексикографии по четырём основным словарям – *Новому французско-русскому словарю* В.Г. Гака и К.А. Ганшиной, *Dictionnaire Francais-Russe* В.Г. Гака и Ж. Триомфа, *Русско-французскому словарю* Л.В. Щербы и М.И. Матусевич, *Русско-французскому словарю* Л.В. Щербы, М.И. Матусевич и Д.В. Сеземана.

В литературе обычно выделяются два критерия отбора коллокаций в качестве материала для словаря: частотность и типичность. В случае с двуязычными словарями важен контрастивный аспект – является ли то или иное сочетание понятным (transparent) для пользователя словаря. Понятные (transparent) словосочетания в рамках двуязычной лексикографии – те, значение которых выводится из значения составляющих и внутренняя форма сочетания на родном языке эквивалентна форме сочетания на иностранном. В противном случае трудности у пользователей словаря обычно возникают не при переводе на родной язык, а при порождении текста на иностранном языке. В формальной семантике такие выражения называются некомпозициональными, и именно эти случаи должны учитываться в двуязычном словаре.

В зарубежной лексикографии особое внимание уделяется так называемым *типовым контекстам*. Эта идея разрабатывается П. Хэнксом [Hanks 2012] в рамках методологии Corpus Pattern Analysis (CPA), на основе которой автор руководит составлением словаря английских глаголов *The Pattern Dictionary of English Verbs* (http://pdev.org.uk).

Методика базируется на теории нормы и употребления (Theory of Norm and Exploitation, TNE), а та в свою очередь вдохновлена работами М. Хэллидея [Halliday 1966] и Дж. Синклера [Sinclair 1966] по лексической семантике. Эти работы примечательны тем, что описанный в них подход лёг в основу методики создания словарей серии COBUILD и проекта Hector [Atkins 1993].

Основное внимание уделяется прототипическим синтаксическим шаблонам (patterns), фактически – рамкам валентности, с которыми ассоциируются эти слова при использовании их в речи. Эти шаблоны для существительных и глаголов существенно отличаются. Для существительных они строятся на основе корпусных данных и состоят обычно из нескольких групп, в которые объединяются значимые коллокаты. Для глаголов шаблонов включают не только базовую аргументную структуру, но и характеристики актантов, например, наличие определителя в подчинённой именной группе. Значение лексемы определяется не в изоляции, а в её прототипическом контексте. Поэтому первый этап CPA – группировка строк конкорданса по найденным синтаксическим шаблонам, а затем приписывание соответствующему значению толкования. Для практических приложений шаблону может соответствовать не толкование, а синонимический ряд или переводной эквивалент.

В немецком проекте ***elexiko*** [Storjohann 2005, URL: http://www1.ids-mannheim.de/lexik/elexiko.html] этот подход используется при разработке блока типовых контекстов (*typische Verwendungen*) в электронном словаре. Авторы ссылаются на подход П. Хэнкса в рамках проекта COBUILD [Hanks 1987] и формулируют своё неформальное определение типового контекста: это прототипические образцы с конкретными, выбранными по корпусным данным, лексическими элементами; многоуровневые лексико-синтаксические конструкции, элементы которых часто встречаются вместе. Типовые контексты в первую очередь показывают синтагматические ассоциации заголовочного слова – как особенности употребления конкретных грамматических форм, так и лексическую сочетаемость: см. Рисунок 8.

Более разработанное представление типовых контекстов описано в (Storjohann, Mohrs 2007). Было замечено, что типовые контексты представляют чрезвычайно ценный материал для изучающих язык. Поэтому важной задачей становится группировка и ранжирование этих фраз в электронном словаре. На рисунке 8 показаны типовые контексты для слова *unbegrenzt* (безграничный) в первой версии словаря и с группировкой. Можно заметить, что в первой версии отсутствовала возможность объединения контекстов по их синтаксическому типу, а также пример реализации для слота (обозначается как […]), где возможна широкая вариативность. Это было связано как с ограниченностью функционала графического отображения словарных статей, так и с неразработанностью методики группировки и ранжирования.

Разработчики отмечают, что иллюстрации должны быть упорядочены в соответствии с частью речи опорного слова: так, для прилагательного *unbegrenzt* в примере на рисунке выделяются синтаксические конструкции, в которых оно выступает в различных функциях – атрибутивной, предикативной и т.д. Затем для каждой группы вырабатывается свой принцип сортировки, иногда довольно специфический: например, в атрибутивном значении контексты располагаются в алфавитном порядке следования второстепенных слов конструкции (предлогов, глагольных форм).

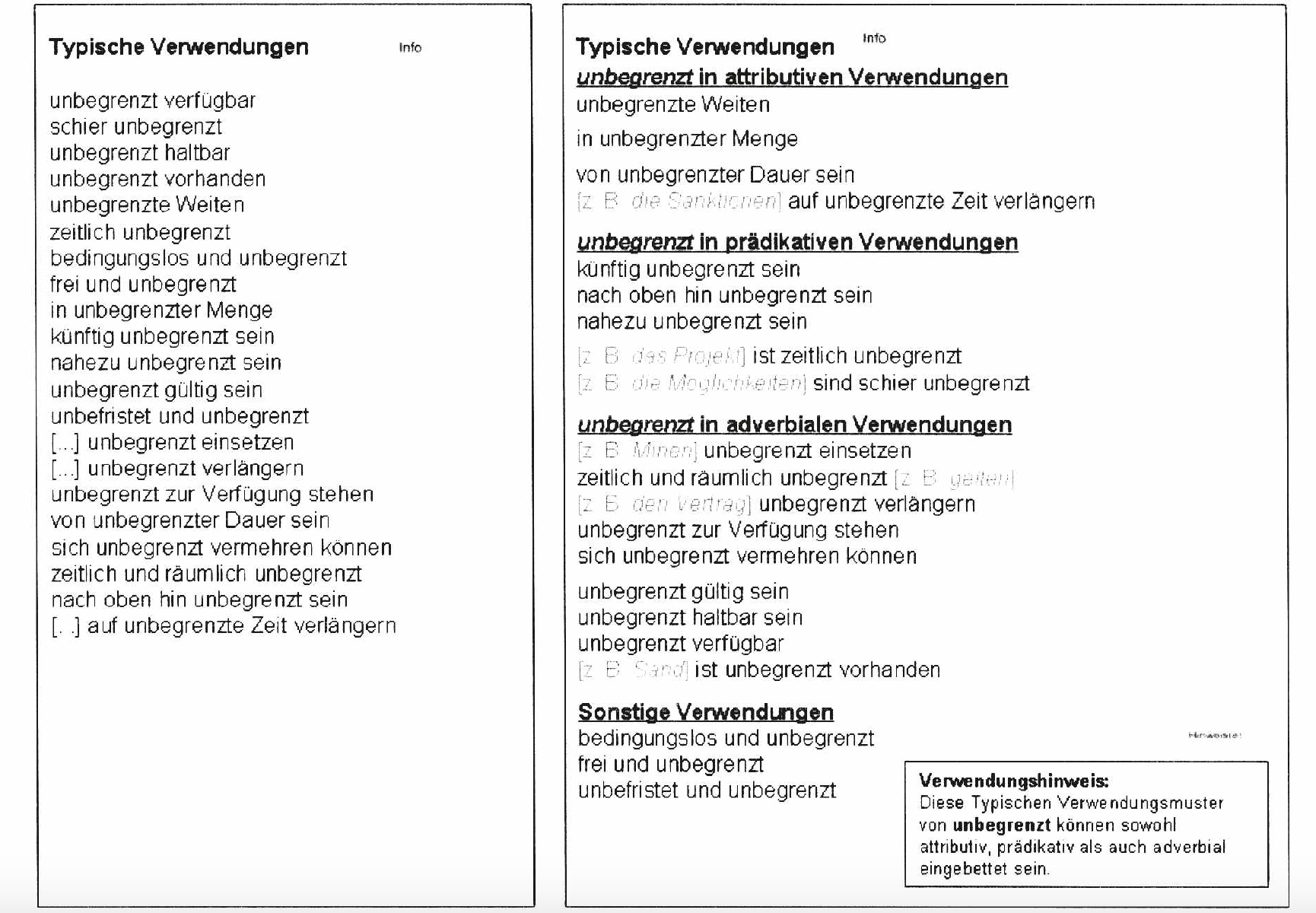


Рисунок 8. Стандартное и упорядоченное представление типовых контекстов в elexiko.

Интересен также подход лексикографов elexiko к иллюстрации реализаций широко вариативных слотов: в примере это, скажем, *[…] unbegrenzt verlängern* и *[z.B. den Vertrag] unbegrenzt verlängern*. Очевидно, что часто возможность реализации слота ограничена конкретной лексической группой (например, одушевлённые лица), однако часто требуется конкретизация. Поэтому на основании корпусных данных были выбраны наиболее регулярные реализации, которые затем использовались в качестве примера заполнения слота.

В отечественной лингвистике тоже употребляется понятие типового контекста (обычно без определения). Например, в книге «Лингвистическая семантика» И.М. Кобозевой [Кобозева 2000] акцентируется необходимость типового контекста для описания значения лексических единиц: «Рассматривая проблему описания значения слова, мы пришли к выводу о том, что в общем случае невозможно правильно, адекватно описать значение слова, изъяв его из типового синтаксического контекста. Таким образом, описание значения (парадигматического свойства слова) требует учёта его синтагматических свойств» [Кобозева 2000: 148].

Типовые контексты, представленные в словарях, и реальные употребления, зафиксированные в корпусе, обсуждаются в статье [Крылов, Митрофанова 2006]. Авторы не формулируют своего определения типового контекста, но принимают как данность, что контексты, иллюстрирующие словарные статьи, считаются типовыми. В статье сравниваются контексты из корпуса Бокрёнок [Азарова, Синопальникова 2004], иллюстрации из словаря Ожегова [Ожегов 1989] и типовая сочетаемость по лексической базе СО-Starling [Крылов, Старостин 2005]. В результате исследования, проведённого на материале частотных лексем *год* и *говорить*, было выявлено несколько классов контекстов:

* типовые контексты, которые отражают «закономерное в языке» – структурную организацию и лексическое наполнение синтагм, содержащих опорные слова;
* типовые контексты, содержащие маргинальные единицы, которые, по мнению авторов, отражают «случайное в языке»;
* контексты, совмещающие случайное и закономерное – идиомы, связи между лексемами внутри которых являются одновременно случайными и закономерными.

При этом информация из специализированной лексической базы данных оказывается более разнообразной по сравнению со словарными статьями, но при этом более «концентрированной» и сбалансированной, чем случайные корпусные данные.

А. Килгарифф и коллеги в статье [Kilgarriff и др. 2008], посвящённой корпусному отбору иллюстраций (предложений) для словаря коллокаций Macmillan, также отмечают типичность как необходимую характеристику примера. Для ранжирования контекстов, автоматически найденных в корпусе, используются следующие признаки:

* длина предложения: контексты короче 10 и длиннее 25 слов штрафовались;
* частоты слов внутри предложения;
* наличие в предложении придаточного с *that*;
* является ли контекст законченным предложением;
* положение опорной фразы внутри предложения.

Веса для признаков подбирались на основе размеченной выборки 1000 положительных примеров, получившуюся метрику GDEX можно использовать для сортировки конкорданса в SketchEngine. Ранжирование не использовалось непосредственно для принятия решений о включении контекста в словарь, но, по словам авторов, значительно сократила усилия лексикографов.

## Практическая реализация иллюстративного блока в двуязычных словарях

Считается, что переводной словарь, составляемый в том числе и с учётом учебных целей, должен включать именно типовые контексты и иллюстрации к информации о сочетаемости слов. Поэтому от определения этих контекстов можно перейти к изучению реализации этой идеи в конкретных словарях.

### Русско-английский словарь под общим руководством проф. А.И. Смирницкого

Словарь содержит около 50 тысяч статей. Основной особенностью словаря, по словам составителей, является то, что «в нём в большей мере, чем это обычно делается в подобных словарях, уделено внимание фонетическому и грамматическому аспектам слова» [Ахманова 1948: 15]. В словаре сочетаемость слов указывается регулярно несколькими способами: например, для глаголов схематически указывается предложно-падежное управление, кроме того, для многих слов даётся информация о лексической сочетаемости. Важно отметить, что сведения о грамматической сочетаемости (управлении) даны как для русских слов, так и для английских эквивалентов. Таким образом реализуется ещё одна особенность словаря – ориентация одновременно на изучающих и русский язык, и английский. Об иллюстрациях авторы далее замечают, что они приводятся «в особо сложных случаях». На самом деле, словосочетания с переводами включены практически в каждую статью, например, при слове **вести**:

**вести**, повести **1**. … **2**. (*вн.; руководить*) conduct (d.), direct (d.); ~ кружок conduct a circle; ~ собрание preside over a meeting …

Кроме того, много внимания уделяется тому, что авторы обозначают как «фразеологические сочетания» – те, в которых данное слово переводится особо. Эти сочетания в словарных статьях приводятся наряду с иллюстрацией сочетаемости.

### Англо-русский словарь В.К. Мюллера

*Англо-русский словарь*, составленный В.К. Мюллером, – один из наиболее переиздаваемых переводных словарей в России – в первом издании [Мюллер 1935] включал 40 тысяч слов. Как и в словаре под руководством А.И. Смирницкого, основной акцент был сделан на произношении и грамматическом справочнике. Авторы уделяют значительное внимание идиомам, «специфическим и характерным выражениям» и включают их в словарь в максимально возможном количестве. При этом (учитывая время издания) теоретическая база для отбора этих сочетаний не разрабатывается, составители полагаются на лексикографическую интуицию.

### Большой англо-русский словарь под общим руководством И.Р. Гальперина

Словарь [Гальперин 1977] включает 150 тысяч статей и предназначен для активного использования. В предисловии отмечается, что «словарь должен дать как можно более полное описание функционирования каждого данного слова». Отбор иллюстраций основывается на принципе «предсказуемости лексического окружения», то есть на наличии устойчивой ассоциации заголовочного слова с определённым выражением, например, для слова **blame**:

**blame** I *n* **1**. порицание, упрёк; to deserve ~ заслуживать порицание; to incur ~ for smth. навлекать (на себя) упрёк за что-л. …

### The Oxford Russian dictionary : Russian-English, English-Russian

Объём словника словаря [Wheeler, Unbegaun 1997] составляет 180 тысяч слов и словосочетаний, причём один том включает одновременно англо-русский и русско-английский словари. К иллюстрациям относят словосочетания и предложения, которые показывают характерное и идиоматическое употребление слова (characteristic and idiomatic usage). Предложения в качестве иллюстраций приводятся только для описания модели управления глаголов:

**вырв|аться,усь,ешься** *pf*. (of вырываться) **1**. (*из+g.*) to tear oneself away (from); to break out (from) … ; в. из чьих-н. рук to break loose from someone’s grip; едва ли мне удастся до лета в. из Москвы I shall hardly manage to get away from Moscow before the summer. **2**. …

### Большой англо-русский словарь Abbyy Lingvo

Словарь [Бурнашева 2011] подготовлен лексикографической группой отдела прикладной лингвистики компании ABBYY. Впервые вышедший в 1990 году, *LingvoUniversal (En-Ru)* стал первым в России[[2]](#footnote-2) лексикографическим изданием совершенно нового типа, предназначенным для электронного поиска и показа и совмещающим в себе функции переводного и толкового словаря. Большинство лексических значений снабжено толкованиями, комментариями об использовании, примерами употребления; многие включены в синонимические ряды и антонимические пары. При подготовке издания широко использовались материалы обсуждений на специализированных лексикографических форумах, специально созданных для развития словарей ABBYY Lingvo: переводческих форумах, форумах Ассоциации лексикографов ABBYY Lingvo, многочисленные одно- и двуязычные словари и интернет-ресурсы.

Иллюстративный материал в электронном словаре Lingvo доступен не только в рамках конкретной словарной статьи, но и через полнотекстовый поиск. Словарь содержит 100 тысяч заголовочных слов и около 50 тысяч словосочетаний. В предисловии к бумажной версии словаря принципы отбора и источники иллюстраций и блока сочетаемости не указаны. Составители лишь отмечают, что «словосочетания позволяют показать употребление слова, например, несвободный выбор глагола или определения для передачи нужного смысла» [Бурнашева 2011]. Словосочетания авторы относят к блоку сочетаемости, тогда как полноценными иллюстрациями считаются предложения из английских текстов с переводом на русский язык. Фактически, иллюстрации-предложения приводятся при редких и «контекстно зависимых» переводах, например, в статье для слова **example**:

1)

а) пример, иллюстрация, типичный случай, аналогичный случай

*to cite / give / provide an example — приводить пример*

*classic example — классический пример*

…

2) урок, назидание; предостережение

*Let these unhappy examples be a warning to others. — Пусть эти прискорбные уроки послужат предостережением другим.*

### Новый большой англо-русский словарь под руководством Ю.Д. Апресяна

Словарь [Апресян 2000] построен на материале *Англо-русского словаря И.Р. Гальперина* по новым принципам новым коллективом авторов под руководством Ю.Д. Апресяна. В отличие от большинства рассмотренных словарей содержит подробное предисловие, в котором описываются принципы, положенные в основу словаря. Главным таким принципом является «установка на активность» - с этим связан и тщательный подход к отбору иллюстраций в блоке сочетаемости. По мнению составителей словаря, важнее всего отразить сочетаемость глаголов и существительных, поскольку она наименее симметрична в англо-русском направлении перевода. Словосочетания и предложения иллюстрируют два основных явления: модель управления заголовочного слова и специфические признаки, которые авторы называют синтаксическими, – фактически это лексико-грамматические разряды, особенности употребления которых проще проиллюстрировать, чем добавлять специальные обозначения. Так, например, авторы обращают внимание на явление лексикализации пассива некоторых глаголов в английском языке (*You are assigned a difficult task*). Эта особенность характерна для отдельных лексем, поэтому должна быть проиллюстрирована. Что касается модели управления, то отмечается, что иллюстрациями снабжаются обычно следующие случаи:

* число актантов и способы их выражения (практически для всех глаголов);
* типичные эпитеты существительного;
* глаголы, при которых заголовочное существительное является подлежащим;
* типичные ситуации, обозначаемые глаголом/существительным;
* виды и части объекта и т.п.

Внутри словарной статьи иллюстрации располагаются в порядке смысловой сложности: например, для существительных сначала приводятся сочетания с прилагательным, а затем модель управления с реализациями.

## Выводы

Во второй главе описываются некоторые подходы к составлению иллюстративного блока словарей, анализируется иллюстративный блок англо-русских переводных словарей. Мы выделили и охарактеризовали основные функции двуязычных словарей и принципы отбора иллюстраций в связи с функциональной направленностью этих словарей. Из анализа подходов к иллюстрированию двуязычных словарей, зафиксированных в предисловиях к словарям и реализованных в самих словарях, можно заключить, что составители в большинстве случаев не формулируют строгих критериев отбора, но полагаются на интуицию лексикографа. Кроме того, несмотря на важность типовых контекстов, которая часто акцентируется, многие словари уделяют большее внимание идиоматическим сочетаниям и специфическому употреблению лексем, когда сложно подобрать дословный переводной эквивалент. Тем самым, наше исследование существующего практического опыта формирования иллюстративного блока переводных словарей подводит к выводу о том, что на сегодняшний день отсутствует единая теория и методология отбора иллюстративных примеров употребления в контексте и типовой сочетаемости слов, включаемых в словники.

# Устойчивые сочетания различных типов и опыт их лексикографирования

## Основные концепции устойчивых сочетаний

Термин «коллокация» широко распространился как в отечественной, так и в зарубежной лингвистике не так давно. Обычно говорят об устойчивых словосочетаниях, фразеологизмах и т.п. Последние обычно включаются в специальный блок – фразеологическую («заромбовую») часть словарной статьи. Коллокации же, как мы видели в предыдущей главе, представляют собой основной материал для иллюстративного блока двуязычных словарей (особенно учебных).

Английский термин *collocation* ввёли в употребление Дж. Фёрс [Firth 1935, обычно цитируется по Firth 1957] и Г. Палмер [Palmer 1933]. Этот подход затем развивался в британской лингвистике как с теоретической, так и с практической стороны (в частности, при создании словарей с использованием корпусных методов – серии COBUILD и словарей издательства Macmillan). Широко известно высказывание Дж. Фёрса, которое можно отнести как к изучению коллокаций, так и к идеям дистрибутивной семантики: «you shall know a word by the company it keeps» [Firth 1957].

Г. Палмер и его коллега А.С. Хорнби [Hornby 1942] обращали внимание на коллокации в контексте методики преподавания языков и учебной лексикографии. Об этом свидетельствует, например, собственно определение коллокации из работы [Palmer 1933: 6]: «A collocation is a succession of two or more words that must be learnt as an integral whole and not pieced together from its component parts» (Коллокация – это последовательность двух или более слов, которые должны изучаться как единое целое, а не покомпонентно). На самом деле, по этому определению сложно провести границу между сочетаниями различной степени связанности (например, коллокациями и идиомами), однако в нем отражена главная, по мнению Г. Палмера, характеристика языковых единиц, которые он называет коллокациями – необходимость усвоения целиком, а не по частям при изучении языка. Кроме того, некоторая часть работы посвящена пояснению терминологии – почему автор не использует другие термины (idioms, heterosemes, phrases, locutions, formulas, compounds).

Поскольку нововведённый термин остаётся неясным, а явления, которые им описываются, весьма разнообразны, приводятся примеры различных классов коллокаций – «гетеросемы» (heterosemes, сочетания, в которых одно из слов используется в новом значении), характерные грамматические конструкции.

Работы Дж. Фёрса, как следует из приведённой выше цитаты, были в основном посвящены определению значения через сочетаемость лексемы. В статье [Firth 1935] он отмечает, что его исследования также вдохновлены практическими языковыми задачами, которые возникали при работе в Индии и Африке. Понятие коллокации и его значение обсуждаются в контексте идеи многоуровневого анализа языка (в отличие от структуралистского анализа по уровням). Контекст порождает значение; как пишет Дж. Фёрс, одно из значений слова *night* – его способность сочетаться с прилагательным *dark*, то есть *dark night*.

Первая классификация устойчивых сочетаний представлена в работах Ш. Балли [Bally 1951]. Он замечает, что некоторые слова более или менее ассоциируются между собой, причём их сочетание часто запоминается и прочно входит в употребление. Ассоциация или связь между элементами таких сочетаний располагается между пограничными случаями, когда

1. сочетание распадается сразу после создания, то есть его компоненты сочетаются свободно;
2. слова постоянно употребляются в конкретном сочетании так, что теряют независимость и самостоятельное значение.

Естественно, множество переходных случаев с трудом поддаётся классификации, поэтому Ш. Балли неформально выделяет три группы:

1. фразеологические обороты – сочетания, прочно вошедшие в язык (например, французское *tout de suite*);
2. фразеологические группы – сочетания, в которых отдельные слова сохраняют самостоятельность (*avoir de la chance*);
3. фразеологические единства – сочетания, в которых наблюдается полная связанность элементов (*avoir lieu*).

Среди фразеологических групп Ш. Балли отдельно отмечает группы с усиливающим определением (*chaleur suffocante*) и глагольные группы, в частности, перифразы самостоятельных глаголов (*remporter une victoire*), и так же, как и предыдущие авторы, подчёркивает важность этого языкового явления при изучении иностранного языка. Поскольку работа Ш. Балли посвящена стилистике, его в значительно большей степени интересуют оригинальные явления литературного языка, то есть фразеологические единства. Фразеологические группы называются «самой банальной частью словаря», поэтому им не уделяется большого внимания в этом исследовании фразеологии.

Подобным образом, исходя из степени трансформации значения, классифицирует фразеологизмы В.В. Виноградов [Виноградов 1977]. Он выделяет три класса фразеологизмов, и его схема стала классической [см., например, Шанский 1985]:

1. фразеологические сращения - устойчивые сочетания, значение которых не выводится из значения составляющих их компонентов, то есть является немотивированным в синхронической перспективе (*попасть впросак*).
2. фразеологические единства - устойчивые сочетания, значение которых отчасти связано с семантикой составляющих их компонентов (*зайти в тупик*).
3. фразеологические сочетания - устойчивые обороты с мотивированным значением, но значение одного из компонентов сочетания является связанным, а само слово имеет ограниченную сочетаемость (*щекотливый*).

Те сочетания, которые Ш. Балли называл фразеологическими группами, в отечественной фразеологии, таким образом, не получили специального названия. Классификация В.В. Виноградова расширяется, в частности, в работах Н.М. Шанского [Шанский 1985], который к трём перечисленным категориям добавляет фразеологические выражения – обороты, которые состоят из слов со свободным значением. Отличие таких выражений от обычных синтагм заключается в том, что они «извлекаются говорящим из памяти целиком (как отдельное слово или фразеологический оборот, эквивалентный слову)» [Шанский 1985: 69]. Среди них выделяются выражения коммуникативного и номинативного характера – предикативные словосочетания, соотносимые с предложением (*Без труда не вытащишь и рыбку из пруда*), и словосочетания с назывной функцией (*высшее учебное заведение*).

Вопрос о классификации и характеристике устойчивых сочетаний подробно рассматривается в [Мельчук, Иорданская 2007]. Авторы являются активными разработчиками теории «Смысл Текст», и их работа касается в том числе особенностей лексикографирования устойчивых сочетаний различных типов. Кроме того, именно в этой работе предпринята попытка систематизации терминологии и определений различных понятий фразеологии, поэтому следует остановиться на ней подробнее.

Первое, «неформальное» определение фраземы как предмета фразеологии вообще звучит так [Мельчук, Иорданская 2007: 226]:

«Фразему можно определить, как такое сочетание, означаемое и означающее которого НЕ МОГУТ быть построены без ограничений и регулярно.»

Строгое определение фраземы затем выводится через отрицание из определения свободного сочетания:

Свободное словосочетание удовлетворяет следующим двум условиям:

1. его означаемое представляет собой регулярную сумму означаемых его составляющих;
2. его означающее представляет собой регулярную сумму означающих его составляющих.

В рамках теории «Смысл Текст» [Мельчук 1999] авторы поясняют это определение таким образом: «Иначе говоря, свободное словосочетание – это такое словосочетание, которое может быть получено из данного Концептуального Представления с применением любых общих правил языка L без каких бы то ни было ограничений» [Мельчук, Иорданская 2007: 227].

Несвободные словосочетания, которые в общем случае называются *фраземами*, могут не удовлетворять одному из двух указанных выше условий. В соответствии с этим выделяются различные типы фразем.

*Прагматемы* – это словосочетания, в которых нарушается первое условие из определения свободного словосочетания, то есть для понятие, выражаемого данным словосочетанием, возможно только одно данное означаемое.

К *прагматемам* И.А. Мельчук и Л.Н. Иорданская также относят сочетания, означающие которых построены регулярно, но с ограничениями. Например, английское “*Best before*…” (*употребить до*) построено регулярно, но имеет прагматические ограничения в том смысле, что в ситуациях его использования (например, на упаковке с йогуртом) не возможен никакой синонимичный эквивалент. Кроме того, к *прагматемам* же причисляются те случаи, когда для одного означаемого возможны различные синонимичные сочетания (“*Please be quiet*”, “*No talking please*” и т.п.).

*Полуфраземы* или *коллокации* – это разновидность идиом, в которой означаемого одного из компонентов входит в сочетание без изменений, тогда как вместо означаемого другого компонента включается некоторое другое означаемое, обычно зависящее от первого: «*отпустить шутку1*», «*выдерживать сравнение1*». Следует сразу заметить, что в модели «Смысл Текст» полуфраземы описываются с помощью *лексических функций* [Мельчук 1999].

Означаемое *квазифразем* или *квазиидиом* включает в себя означаемые обеих лексем и «некую непредсказуемую добавку»: “*bacon and eggs*” – *яичница с беконом*, “*shopping center*” – *торговый центр*.

Поскольку нас прежде всего интересуют коллокации, рассмотрим также понятие лексических функций, с помощью которых они описываются в модели «Смысл Текст». И.А. Мельчук и коллеги предложили новый тип словаря – толково-комбинаторный, в котором главным способом описания и являются лексические функции – зависимости, связывающие заголовочное слово с его «лексическими коррелятами», синтагматически и парадигматически связанными с ним лексическими единицами. По этому признаку выделяются лексические функции-замены, описывающие парадигматические связи слова, и параметры, описывающие сочетаемость. Приведём несколько примеров таких функций-параметров:

* , где *i* – номер – типовое название актанта, которые нумеруются в порядке значимости для данного слова: .
* - типовое название сирконстанта:
* – глагол, связывающий данное слово с первым или вторым актантом (агенсом / пациенсом):

;

.

В Машинном фонде русского языка используется оригинальный подход к устойчивым сочетаниям, описанный в [Борисова 1990: 88-89]: «Основу пакета «Устойчивые словосочетания» составляют те единицы, которые традиционно включались в число фразеологических единиц, однако выделяли какими-либо особенностями, не позволяющими сводить их в один класс с идиомами, паремиями и другими «типичными» фразеологизмами». Такие сочетания характеризуются, как правило, сохранением лексического значения одного из компонентов и ослаблением значения образности. В Машинном фонде русского языка к устойчивым сочетаниям относят те фразеологизмы, один компонент которых выбирается в процессе речи на основании его свободного значения, а второй – на основании значения первого (*отпустить шутку*). Таким образом, они соответствуют полуфраземам или коллокациям в классификации И.А. Мельчука и Л.Н. Иорданской. Кроме того, Е.Г. Борисова упоминает фраземы в значении, введённом в работе [Амосова 1963] – словосочетания, один из компонентов которых употребляется только в сочетании с другим компонентом: «щурить глаза», «разбиться вдребезги». При этом значения обоих компонентов сохраняются. Автор также отмечает два важных свойства устойчивых сочетаний:

* невозможность предсказания выбора несвободного компонента на основании знаний о семантическом составе сочетающихся слов;
* необъяснимость сочетаемости на основании знаний о семантическом составе лексем.

Эти свойства связаны с тем, что для устойчивых словосочетаний нарушено условие свободной сочетаемости, когда смысловые компоненты каждого из слов совместимы.

В [Борисова 1990] устойчивые словосочетания классифицируются на основании зависимости выбора несвободного компонента: этот компонент может употребляться только в рамках данного сочетания, в ограниченном множестве сочетаний или свободно. Выделяются словосочетания, выражающие уникальные смыслы (*закадычный друг*), и словосочетания, выражающие регулярные смыслы (*вступать в бой, в соперничество*). Последние хорошо описываются в терминах уже упомянутых лексических функций.

Когда работы по созданию Машинного фонда русского языка были прекращены, Е.Г. Борисова продолжила заниматься коллокациями, что отражено в монографии [Борисова 1995а] и словаре сочетаемости русского языка с английскими эквивалентами [Борисова 1995б].

Задача анализа устойчивых сочетаний является частной по отношению к проблеме исследования сочетаемости лексических единиц, или лексической синтагматики. Именно исследования по семантической синтагматике могут объяснить механизмы взаимодействия значений, обуславливающие появление и свободных, и устойчивых сочетаний, и сочетаемостных аномалий. Упомянем здесь кратко концепцию В.Г. Гака [Гак 1971] и схожие идеи Ю.Д. Апресяна [Апресян 1974].

Устойчивые словосочетания различных типов в контексте вопроса о синтаксической синтагматике упоминаются В.Г. Гаком в статье [Гак 1971]. В.Г. Гак рассматривает закономерности *синтагматически обусловленной номинации*, то есть номинации, зависящей от окружающих номинаций высказывания. В определении В.Г. Гак пользуется понятием *семы* – минимального компонента значения. Семантическое сочетание слов определяется *законом семантического согласования*: чтобы два слова составили правильное сочетание, они должны иметь одну общую сему, помимо специфических, присущих каждому слову. Эту сему французские лингвисты Б. Потье и А. Греймас назвали классемой. Классема должна повторяться в сочетающихся единицах; в примере В.Г. Гака «простуженная женщина» - правильное сочетание, поскольку

а) значение слова «простуженный» включает значение одушевлённости, поскольку признак присущ только одушевлённым лицам;

б) значение слова «женщина» включает значение одушевлённости.

Под классемой обычно подразумевается некий категориальный компонент – одушевленность, возраст, пол, действие или состояние и так далее. В.Г. Гак предлагает рассматривать также и те случаи, где в сочетании связующим является некоторый второстепенный, не категориальный компонент значения – «связующий семантический компонент» или *синтагмема*. Экспериментальные данные показывают, что возможны три варианта взаимодействия слов и в сочетании в зависимости от реализации синтагмемы .



*Семантическое согласование*, возникает в тех случаях, когда в значениях синтагматически связанных слов есть общий компонент, интегральная сема. Так, в сочетаниях *птица летит*, *змея ползёт* наблюдается семантическое согласование, поскольку «летать» предполагает «летающего» деятеля, а «ползать» - «ползающего».

*Семантическое несогласование*, возникающее вследствие семантической экономии: потенциально общий компонент опускается в одном из сочетающихся слов (*птица, змея приближается*). Это происходит лишь в тех случаях, где опущенный элемент предсказуем: например, предложение «*Возле стены стоял шкаф*» сохраняет то же значение, что и у «*Возле стены находился шкаф*», поскольку шкаф обычно занимает вертикальное положение.

*Семантическое рассогласование*, проявляется при наличии в синтагме несовместимых (противоположных или ненужных) компонентов значения (*птица ползёт, змея летит*). В таких сочетаниях некоторые семы могут быть утрачены, а сами сочетания переосмысляются или получают новое переносное значение: например, к значению *ползти* в сочетании с «неползающим» субъектом добавляется компонент «скорость». При исследовании коллокаций необходимо принимать во внимание именно этот тип выражений.

По мнению Ю.Д. Апресяна [Апресян 1974], в общем возможны два варианта сочетаемости. В случае семантической сочетаемости или семантических ограничений на сочетаемость, речь идёт о наборе семантических признаков, которыми должны обладать слова, замещающие данную валентность. Лексическая сочетаемость и лексические ограничения описываются набором лексем. В рамках предложенной Ю.Д. Апресяном модели управления эти ограничения легко записываются в условном виде. Кроме того, в модели «Смысл Текст» толкования должны формулироваться таким образом, чтобы с его помощью можно было проверить потенциальный актант. Например, «*грозить = A обещает причинить Y-у зло или неприятность X*». Далее, в случае лексической сочетаемости, может описываться набор слов, которые могут оказаться на месте А.

## Представление коллокаций в словарях сочетаний

Коллокации и устойчивые словосочетания различных типов включаются в соответствующие словари – словари коллокаций, фразеологические словари – которые по-своему формулируют критерии отбора материала.

В разделе, посвящённом синтагматическим и фразеологическим словарям, авторы «Истории английской лексикографии» [Cowie 2009] говорят о трёх критериях, которые указывают на значимость определённого словосочетания:

* частота словосочетания: статистически значимые сочетания включают коллокации [Herbst 1996], «кластеры» (*a bit*, *a little bit*… [Mittman 2004]), стеммы [Altenberg 1998]: *and then I*, *there is a*;
* степень вариации элементов словосочетания: фиксированные элементы, систематическое варьирование в пределах грамматических категорий, лексические варианты;
* степень мотивированности значения (semantic transparency), от полной аддитивности значений до идиом.

Среди существующих синтагматических и фразеологических словарей выделяют: словари пословиц (dictionaries of catchverbs and proverbs), собственно фразеологические словари (idiom dictionaries), словари фразовых глаголов (phrasal verb dictionaries), словари валентности (valency dictionaries), словари коллокаций (collocation dictionaries).

Для дальнейшей работы важно рассмотреть, чем характеризуются последние два типа словарей. Словари валентности обычно составляются для тех языков, где особенную сложность представляет глагольное управление (например, немецкий). Однако, понятие валентности относится не только к глаголам: так, к примеру, в *A Valency Dictionary of English* [Herbst 2004] включены рамки валентностей для глаголов, прилагательных и существительных. Статья состоит из трёх частей: сначала приводятся обобщённые синтаксические шаблоны, затем примеры их реализаций и указание на семантические характеристики лексем, которые заполняют актантные позиции:

**ability**

P1 Over 500 events are planned throughout the country, for people of all ages and *ability*. …

P2 **+ to-INF** (>30%) My memory is such that I used to win prizes in school because of my *ability* to remember names and dates. …

По сравнению со словарями идиом словари коллокаций встречаются редко. Это связано с недостаточной формальной определённостью самого термина «коллокация». *Комбинаторный словарь английского языка* [Бенсон 1990] впервые разграничивает в словаре лексические и грамматические коллокации, затем *Dictionary of English Collocations* и *Collins Cobuild English Collocations* включают просто статистически значимые словосочетания, а *Oxford Collocations Dictionary for Students* вместо грамматических коллокаций помещает в основном сочетания с предлогами. Последние упомянутые словари уже используют корпусные методы.

В этом разделе мы подробно рассмотрим принципы построения следующих словарей: *Толково-комбинаторного словаря русского языка*, *Словаря английских коллокаций*, *Оксфордского словаря коллокаций, Словаря коллокаций Collins Cobuild*, *Комбинаторного словаря английского языка*, *Англо-русского фразеологического словаря, Англо-русского словаря глагольных сочетаний, Словаря коллокаций Macmillan, словаря «Устойчивые словосочетания русского языка», справочника «Русские глаголы и предикативы», Пособия по лексической сочетаемости слов русского языка и Словаря сочетаемости слов русского языка*.

### Толково-комбинаторный словарь русского языка

*Толково-комбинаторный словарь русского языка* (далее – ТКС) разрабатывался как один из необходимых компонентов реализации модели «Смысл Текст», в печатном виде существует лишь версия с ограниченным словником (около 250 статей) – «Опыт семантико-синтаксического описания русской лексики» [Мельчук и др. 1984]. Поскольку издание было подготовлено в качестве первого образца словарных описаний подобного рода, «с целью ликвидировать пропасть между лексикографией и теоретической лингвистикой» [Там же: 74], заголовочные слова отобраны из разных лексико-семантических классов для демонстрации различных явлений. Словарь характеризуется как словарь активного типа, однако метод формального описания, разработанный в рамках модели «Смысл Текст» позволяет представить информацию в унифицированном виде (в частности, с использованием лексических функций). Внутри словарной статьи выделяется 10 зон: морфологические сведения, стилистическая помета, толкование, состоящее их постоянных единиц и переменных, модель управления, в которой используются переменные из толкования, ограничения к модели управления, примеры к модели управления, лексические функции (всего их вводится несколько десятков), примеры реализации лексических функций, энциклопедическая информация, идиомы. Надо отметить, что создатели опираются как на семантико-синтаксические исследования И.А. Мельчука, А.К. Жолковского, Ю.Д. Апресяна и коллег, так и на существующие толковые и прочие словари.

Внешне *ТКС* сильно отличается от современных толковых словарей и словарей сочетаемости и представляет собой уникальную разработку, реализацию одной из сложных лингвистических теорий в лексикографическом произведении. Попытки создания толково-комбинаторных описаний для отдельных лексем производились и на материале других языков [польский – Янус 1971, французский – Melcuk 1984]. Из последних прикладных разработок следует отметить проект *DiCo/LAF* [Polguere 2000]. По словам разработчиков, в определении коллокации они следуют [Hausmann 1979] и даже [Bally 1951], однако словарь представляет развитие идей толково-комбинаторного описания на французском материале – в терминах лексических функций. Для каждого заголовочного слова описываются значения лексических функций, которые к нему применимы.

### A Dictionary of English Collocations

Словарь коллокаций *A Dictionary of English Collocations* [Kjellmer 1994] представляет собой одно из первых лексикографических произведений, при составлении которого использовались методы корпусной лингвистики. В нем представлены все (по мнению автора) коллокации, встречающиеся в Брауновском корпусе, – корпусе текстов на американском английском объёмом около 1 миллиона словоупотреблений [Kučera, Francis 1964]. На момент создания словаря это один из наиболее представительных корпусов как с точки зрения объёма, так и с точки зрения жанра.

Коллокации определяются автором с практической точки зрения – как повторяющиеся последовательности грамматически корректных единиц («recurring sequences of grammatically well-formed items»). По мнению автора словаря, коллокации являются основными составляющими ментального лексикона носителя языка, тем самым формируя необходимый элемент владения языком. В качестве примеров во введении приводятся коллокации различных типов: *at the outset, could be expected to, not significantly different from, peaceful coexistence, powdered* *coffee,* and *with great difficulty*. В словаре представлено около 85 тысяч коллокаций. Каждая словарная статья помимо специальных помет также включает статистику о частоте встречаемости в корпусе, распределении употреблений коллокации в текстах различных жанров и степень значимости, то есть устойчивости связи внутри коллокации.

По словам составителей, словарь будет полезен как исследователям в области английского языка, лексикографии, стилистики, автоматического анализа языка, так и преподавателям: «It will be an invaluable reference source for researchers in linguistics, English-language teaching, lexicography, stylistics, and automatic language analysis» [Kjellmer 1994: 5].

### Oxford Collocations Dictionary

Словарь коллокаций *Oxford Collocations Dictionary* [McIntosh 2009] разработан для изучающих английский язык. Эта цель и обуславливает выбор материала и содержательное наполнение словаря.

Коллокации определяются как частотные словосочетания, которые являются необходимыми шаблонами в естественной речи, свойственной носителю языка. Словарь включает около 150 тысяч коллокаций для 9 тысяч заголовочных слов. Среди коллокаций встречаются слова, которые обычно употребляются в сочетании с заголовочным словом – существительные, прилагательные, глаголы, наречия, предлоги – а также общеупотребительные фразы. Источником для составления словаря послужил Британский национальный корпус; авторы также отмечают, что для поиска наиболее современных употреблений использовался поиск в интернете. Для многих коллокаций приводятся примеры употребления – целые предложения или фрагменты текста, а также грамматическая и стилистическая информация:

**HOUR**

1 VERB + HOUR **take** *It takes two hours to get to London*. | **spend** | **last** *The performance lasted three hours*. | **gain**, **lose** *You gain five hours when you fly from New York to London*.

### Collins Cobuild English Collocations

Словарь, как и большинство изданий серии Collins COBUILD, предназначен для изучающих язык. Представленный материал целиком взят из корпуса The Bank of English (170 миллионов слов) и отобран на основании статистики встречаемости. В описании говорится, что словарь предоставляет информацию об устойчивых сочетаниях, естественных выражениях, фразеологии, идиомах, клише и специальных формах (fixed phrases, natural phrases, natural phraseology, idioms, clichés, and specialized forms). Для 20 тысяч заголовочных слов приводится около 140 тысяч коллокаций, которые определяются как частотные последовательности слов внутри фиксированного окна. В этом заключается один из наиболее упоминаемых недостатков данного словаря, поскольку такой метод позволяет найти не только достаточное количество устойчивых сочетаний, но и свободные употребления (*such disaster*), а также грамматически не связанные слова (*nature because*). С другой стороны, корпусный анализ и электронный формат (словарь был выпущен на CD-ROM) позволяют, например, иллюстрировать использование коллокаций контекстами из корпуса:

***Small hours*** *('предрассветные часы'):* *They were arrested in the small hours of Saturday morning.*

### The BBI Combinatory Dictionary Of English

Целью данного словаря [Бенсон 1990] является представление необходимых повторяющихся грамматических и лексических сочетаний (коллокаций, «fixed, identifiable, non-idiomatic phrases and constructions», recurrent combinations, fixed combinations, collocations). Составители подчеркивают важность словаря подобного рода для изучающих английский язык, поскольку такие сочетания часто не имеют дословного перевода. В словаре выделяются грамматические и лексические коллокации. Первые представляют собой сочетание опорного слова (существительного, прилагательного, глагола) с предлогом или грамматической структурой (инфинитив, придаточное предложение). Лексические коллокации обычно не содержат предлогов, инфинитива или придаточного; напротив, они включают знаменательные слова.

Во введении подробно изложены принципы отбора материала и описаны типы коллокаций, которые включаются в словарь. Перечислим их с краткими пояснениями:

* Грамматические коллокации:
  + существительное + предлог (*in favor of*)

В эту группу авторы почти не включают сочетания с *of*, *by* и производными предлогами.

* + существительное + *to* + инфинитив *pleasure to*
  + существительное + придаточное с союзом that *an oath that he would do his duty*

Сюда не входят определительные придаточные предложения *(an agreement that would go into effect in a month)*.

* + предлог + существительное *by accident, in advance*
  + прилагательное + предлог *angry at*

В эту группу не включаются причастия и производные предлоги

* + прилагательное + *to* + инфинитив *necessary to (work)*
  + прилагательное + придаточное с союзом that *afraid that she would fail*
  + 19 глагольных групп: глагольные сочетания с двумя дополнениями без предлога (dative movement transformation ­– *send him the book*), глагол + *to* + инфинитив (*continue to write*) и т.д.
* Лексические коллокации:
  + глагол + существительное/местоимение/предложно-падежная форма *set an alarm*, *have a bath*;
  + глагол со значением уничтожения + существительное *repeal a law*, *denounce (abrogate) a treaty*;
  + прилагательное + существительное *fly a kite*;
  + существительное + глагол *bombs explode*;
  + количественная характеристика + существительное *a school of whales*;
  + наречие + прилагательное *sound asleep*;
  + глагол + наречие *affect deeply*.

Словарь содержит около 14 тысяч слов и более 75 тысяч сочетаний. Внутри словарной статьи коллокации расположены в том же порядке, в котором перечислены соответствующие типы, причём лексические коллокации предшествуют грамматическим. Как отмечается в русском издании словаря, «порядок следования словосочетаний отвечает современным требованиям методики преподавания английского языка» [Бенсон и др. 1990: 1].

### Англо-русский фразеологический словарь

В *Англо-русский фразеологический словарь* [Кунин 1984] включаются идиомы, идиофразематические и фразематические единицы (всего около 10 тысяч фразеологических единиц). К последним составители словаря относят случаи единичной сочетаемости одного из компонентов (*boon companion*), терминологические сочетания (*civil list*), вошедшие в литературное употребление, обороты со связанным значением (*pay attention*), необразные штампы (*all sorts and conditions of men*), предложно-именные сочетания с буквальным значением (*at (the) most*). Принципы отбора материала и название словаря указывают на то, что словарь ориентирован в большей степени на идиоматические сочетания, которые представляют трудность при переводе. Поэтому в словарь не включаются, например, типичные перифразы (*make an effort*). Большинство выражений иллюстрируется примерами из художественной литературы.

### Англо-русский словарь глагольных словосочетаний

Словник *Англо-русского словаря глагольных сочетаний* [Гинзбург и др. 1990] составляют 569 наиболее частотных глаголов английского языка, отобранных из 2 тысяч самых частых слов по данным нескольких частотных словарей. Идея словаря заключается в описании значения и сочетаемости глаголов по принципу моделирования, предложенному в учебной лексикографии Г. Палмером и А.С. Хорнби (Palmer 1933). Материал представляют 26 моделей глагольных словосочетаний: глагол + существительное-субъект (***begin*** *+ the lesson*), глагол + наречие (***run*** *+ fast*), глагол + прямое дополнение (***serve*** *smb*.) и так далее. Модель характеризуется следующими признаками:

* Представляет собой обобщённую структуру словосочетания (на морфосинтаксическом уровне);
* Все компоненты модели обязательны;
* Значение глагола в данном словосочетании обусловлено структурой модели;
* Данное значение регулярно реализуется именно в этой модели.

В словарь включаются именно типичные или слабо связанные словосочетания, а не фразеологизмы, так как последние не удовлетворяют признакам модели. Каждой модели приписан номер, который затем приводится в словарной статье наряду с реализациями модели:

**HIRE** <III, IV, VII, XI, XXI> **III** **1**. ~ *smb*. ~ a clerk (a servant, a day-labourer, etc.) нанимать клерка и т.д. …

### Macmillan Collocations Dictionary

Во введении к словарю *Macmillan Collocations Dictionary* [Rundell 2010], который, по-видимому, является одним из наиболее современных словарей сочетаемости английского языка, подробно описывается методика отбора словарного материала и концепция, лежащая в основе издания. Составители первого издания словаря были первыми пользователями системы WordSketch (https://www.sketchengine.co.uk), которая осуществляет поиск заданных лексико-грамматических шаблонов по корпусу и собирает информацию о частоте встречаемости слов в выражении. Благодаря использованию таких шаблонов, информация о сочетаемости внутри словарной статьи сгруппирована по грамматическим отношениям, в которые вступает опорное слово. Каждой коллокации приписывается вес, который показывает её устойчивость. Состав словника определяется также корпусными методами – с помощью специальной меры *collocationality*, которая характеризует степень устойчивости всех сочетаний с данным словом [Kilgariff 2006]. Таким образом, в словарь практически не включаются слова, встречающиеся по большей части в свободных сочетаниях, а общий объём словника составляет около 4,5 тысяч лексем. Как уже было сказано, коллокации группируются сначала по синтаксическому типу, затем по значению. Кроме того, даётся информация о том, к какой форме тяготеют данные выражение (например, если глагол употребляется в пассиве).

### Устойчивые словосочетания русского языка

Учебное пособие «*Устойчивые словосочетания русского языка*» [Регинина и др. 1983] содержит устойчивые словосочетания, характерные для газетно-публицистического, научного и разговорных стилей. Предназначено для студентов-иностранцев, изучающих русский язык. Пособие включает около 3000 словосочетаний, причём их список ограничен темами, рекомендованными для обучающихся.

Как устойчивые рассматривают фразеологические сочетания по терминологии В.В. Виноградова и фразеологические выражения по Н.М. Шанскому (см. выше). Включают те словосочетания, которые характеризуются:

* переосмыслением значения одного из компонентов (*заклятый враг*);
* ограниченной сочетаемостью одного из компонентов (*одерживать верх, победу*);
* не имеют дословных переводных эквивалентов (*дождь идёт*);
* устойчивостью, повторяемостью в речи.

Составители классифицируют материал по грамматической структуре:

* субъективно-предикативные словосочетания (*растёт безработица*);
* предикативные:
  + глагол + существительное в винительном падеже (*давать оценку*);
  + глагол + существительное в косвенном падеже (+ предлог) (*приводить в порядок*);
  + быть + именная группа (*быть во главе, не по силам, высокого мнения*);
* атрибутивные:
  + прилагательное + существительное (*глубокая мысль*);
  + существительное + существительное в родительном падеже (*круг вопросов*);
  + существительное + предлог + существительное в косвенном падеже (*воля к борьбе*);
* адвербиальные (*тепло поздравлять*).

Опорное слово выбирается в зависимости от типа словосочетания: так, в первых двух типах это глагол, в третьем – прилагательное, в последнем – наречие.

### Русские глаголы и предикативы

Словарь «*Русские глаголы и предикативы*» [Красных 1993] представляет собой пособие по сочетаемости, в котором содержатся описания около 600 глаголов и 200 предикативов. Словарь имеет учебную направленность, причём подчёркивается, что материал словаря – глаголы и предикативы – наиболее сложный в усвоении. В статьях отражается информация о сочетаемости изъяснительных глаголов и предикативов с союзными средствами:

**СОКРУШАТЬСЯ**, сокруша|юсь, -ешься, -ются, *нсв*. *Разг*.Сожалеть о чём-л., сильно огорчаться, печалиться.

Обычно с союзом *что*.

…

Мне показалось, что Лена и в самом деле *сокрушалась* о том, что не стала такой красивой, как прежде.

Включаются сочетания из газетного и журнального материала, а также художественной литературы.

### Пособие по лексической сочетаемости слов русского языка

Словарь-справочник [Анисимова и др. 1975] состоит из 40 тысяч словосочетаний, включённых в пособия для изучающих русский язык, для 1900 опорных слов. Авторы используют синтаксический подход при отборе материала: используются только следующие словосочетания:

* Для опорных существительных:
  + существительное + прилагательное – *глубокая мысль*,
  + существительное + сущ – *подвиг народа*,
  + существительное в роли дополнения + глагол – *совершить подвиг*,
  + существительное-субъект + глагол – *наступил вечер*;
* Для опорных прилагательных:
  + прилагательное + существительное (объект) – *холодный климат, ветер, вечер*,
  + прилагательное + существительное (дополнение) – *похожий на отца, мать*;
* Для опорных глаголов:
  + глагол + наречие – *бежать быстро*,
  + глагол + существительное в винительном падеже – *ждать весну*,
  + глагол + существительное в косвенном падеже – *смотреть на докладчика, на преподавателя*,
  + глагол + существительное-субъект действия – *трамвай, автобус, электричка ходит.*

В качестве опорных слов авторы выделяют только «общераспространённую стилистически нейтральную лексику».

### Словарь сочетаемости слов русского языка

*Словарь сочетаемости слов русского языка* [Денисов, Морковкин 1983] представляет собой переиздание *Учебного словаря сочетаемости слов русского языка* [Денисов, Морковкин 1978], предварённое обширным введением, где излагается авторский подход к сочетаемости и её описанию. Во-первых, определяются понятия лексической и семантической сочетаемости (по Ю.Д. Апресяну), а также синтаксической сочетаемости – это набор имеющихся при слове семантико-синтаксических позиций, иначе говоря, его рамка валентности. Основной задачей словаря сочетаемости является определение этих семантико-синтаксических позиций для каждого слова и описание их заполнения. Основной единице словаря, таким образом, будет словосочетание – реализация валентностей опорного слова. Она может описываться тремя способами:

* полное перечисление ряда, заполняющего данную валентность (*бегать на коньках, на лыжах*);
* выборочное перечисление типичных представителей ряда (*начало чего: соревнований, спектакля …*);
* указание на характер ряда и перечисление наиболее типичных его представителей (*фотографировать кого-что: (о человеке) друга, сына, дочь, памятник …*).

## Выводы

В третьей главе рассматривались различные подходы к устойчивым словосочетаниям, в частности, к коллокациям, которые представляются важным материалом для иллюстративного блока переводного словаря. На основании приведённых теоретических определений коллокаций можно выделить два направления, которые отражаются и в лексикографической практике, в том числе и в описанных словарях сочетаемости. Первый подход (свойственный британской лингвистической школе) исходит из регулярного характера коллокаций и более практического критерия частотности сочетания. Второй подход, развивавшийся в отечественном языкознании, скорее следует назвать семантико-синтаксическим (такими, например, являются описания лексических функций). Таким образом, мы приходим к выводу, что в практической задаче поиска коллокаций в корпусе должны учитываться, как узуальные, так и семантические и синтаксические факторы.

# Поиск и ранжирование контекстов для англо-русского машинного словаря

Задачей практической части работы было построение алгоритма ранжирования переводных контекстов и его тестирование. Часть экспериментов описана в нашей статье [Protopopova и др. 2015]

Исследование лексикографических источников, представленное в Главах 1-3, показало, что существующие переводные словари, по-разному подходят к проблеме выбора иллюстраций, а словари коллокаций по-разному формулируют требования к включаемым словосочетаниям. Однако, можно выделить некоторые характеристики, на которые можно опираться при автоматическом подборе параллельных сочетаний. Очевидно, фразы-иллюстрации должны быть грамматически корректны и иметь правильный перевод. Двуязычные словари предъявляют следующие требования к иллюстративному материалу:

* Примеры представляют собой словосочетания (реже – предложения) на входном языке с переводом на выходной язык.
* Примеры представляют контрастивный материал, иллюстрирующий особенности употребления слов во входном языке в отличие от слов выходного языка (или наоборот).
* С другой стороны, примеры могут иллюстрировать типичные контексты заголовочного слова во входном / выходном языке.

Важную роль в иллюстрации сочетаемости играют коллокации, которые должны обладать следующими свойствами:

* Коллокация – словосочетание, часто и регулярно употребляемое в речи.
* Коллокации не создаются говорящим по частям, но воспроизводятся из памяти как целое.

## Материал и инструменты исследования

Иллюстрации подбирались к словарным статьям словаря сервиса «Яндекс.Переводчик» (<http://translate.yandex.ru>), создание которого описано в [Antonova, Misyurev 2014] (см. Главу 1).

В качестве основного материала в экспериментах использовался корпус параллельных текстов на английском и русском языках из архива коммерческой поисковой системы объёмом около 50 миллионов предложений. Разнообразие жанров, представленных в корпусе (публицистика, техническая документация, тексты рекламных объявлений, художественная литература, субтитры), позволяет находить иллюстрации, относящиеся к различным регистрам и содержащие разные переводы заголовочного слова. Кроме того, объём корпуса достаточно велик, чтобы извлекать иллюстрации для малоупотребительных лексем и переводов.

Параллельные тексты предобработаны и снабжены информацией о выравнивании по предложениям, пословном выравнивании и синтаксическом разборе каждого предложения. Выравнивание производится с помощью инструмента GIZA++ [Och, Ney 2003], а синтаксический разбор – с помощью парсера SyntAutom [Antonova, Misyurev 2012].

GIZA++ – один из инструментов выравнивания, который используется в большинстве систем машинного перевода, в частности, потому, что входит в состав системы Moses (http://www.statmt.org/moses/). Общий принцип работы GIZA++ описан в разделе о моделях IBM [см. Och, Ney 2003].

Из современных синтаксических анализаторов в рамках данной работы нам был доступен SyntAutom [Antonova, Misyurev 2012]. Он основан на правилах и позволяет получать синтаксический разбор предложения с высокой точностью, в частности, в соревновании, проведённом РОМИП в 2012 году, он занял третье место после анализаторов Compreno и ЭТАП-3. Сравнение разборов тестовой коллекции этого соревнования доступно на сайте соревнования[[3]](#footnote-3). SyntAutom позволяет получить разборы для предложений на русском и английском языках. Ниже приведём пример разбора предложения «*Мне нравилось смотреть на улицу через стекло.*» в табличном формате:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0 | \*Top\* | \*Top\* | 0 | \_ | / |
| 1 | мне | я | 3 | subj | /prn/sg/fem/msc/neu/dat/fst/ |
| 2 | нравилось | нравиться | 3 | auxd | /vrb/sg/neu/fin/fst/sec/trd/pst/ind/act/ |
| 3 | смотреть | смотреть | 0 | fin | /vrb/sg/neu/inf/fst/sec/trd/pst/act/ |
| 4 | на | на | 5 | prep | /prp/acc/ |
| 5 | улицу | улица | 3 | prepnp | /nn/sg/fem/acc/trd/ |
| 6 | через | через | 7 | prep | /prp/acc/ |
| 7 | стекло | стекло | 3 | prepnp | /nn/sg/neu/acc/trd/ |

В качестве одного из факторов ранжирования используется оценка фразы по модели языка (см. ниже). Для построения этих моделей использовались более надёжные корпуса текстов на английском и русском языках (одноязычные), собранные из интернета. Данные по объёму всех использованных корпусов приведены в таблице 1. Модели языка строятся с помощью утилит пакета SRILM [Stolcke 2002].

Таблица 1. Объём использованных корпусов в токенах (словоупотребления и знаки препинания).

|  |  |  |
| --- | --- | --- |
| **объём корпусов** | **английский** | **русский** |
| **параллельный Интернет-корпус** | 1,145 миллиарда | 1,132 миллиарда |
| **одноязычный Интернет-корпус** | 4,1 миллиарда | 3,16 миллиарда |

## Описание процедуры поиска контекстов

Для поиска потенциальных словарных иллюстраций из параллельного корпуса извлекаются параллельные контексты заданной длины. Сначала по параллельному корпусу с помощью инструмента GIZA++ строится модель перевода, полученные контексты затем обрабатываются парсером SyntAutom. Полученная таблица содержит следующие данные (см. Таблицу 2):

* входной контекст;
* перевод;
* синтаксический разбор входного контекста и перевода;
* пословное выравнивание;
* входной контекст и перевод в лемматизированном виде.

Таблица 2. Пример записей из модели перевода.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| a back panel | задняя панель | det\_/det/\_det\_3 nn\_/nn/sg/trd/\_gen\_3 nn\_/nn/sg/trd/\_homo\_0 | adj\_/adj/sg/fem/nom/trd/\_adj\_2 nn\_/nn/sg/fem/nom/trd/\_homo\_0 | 1-0 2-1 | a back panel | задний панель |
| a back part | арьергардная часть | det\_/det/\_det\_3 nn\_/nn/sg/trd/\_gen\_3 nn\_/nn/sg/trd/\_subj\_0 | adj\_/adj/sg/fem/nom/trd/\_adj\_2 nn\_/nn/sg/fem/nom/trd/\_subj\_0 | 1-0 2-1 | a back part | арьергардный часть |
| a back part | в заднюю часть | det\_/det/\_det\_3 nn\_/nn/sg/trd/\_gen\_3 nn\_/nn/sg/trd/\_prepnp\_0 | prp\_/prp/acc/\_prep\_3 adj\_/adj/sg/fem/acc/trd/\_adj\_3 prepnp\_/nn/sg/fem/acc/trd/\_prepnp\_0 | 0-0 1-1 2-2 | a back part | в задний часть |
| a back part | до задней части | det\_/det/\_det\_3 nn\_/nn/sg/trd/\_gen\_3 nn\_/nn/sg/trd/\_np\_0 | prp\_/prp/gen/\_prep\_3 adj\_/adj/sg/fem/gen/trd/\_adj\_3 prepnp\_/nn/sg/fem/gen/trd/\_prepnp\_0 | 0-0 1-1 2-2 | a back part | до задний часть |
| a back part | задний отдел | det\_/det/\_det\_3 nn\_/nn/sg/trd/\_gen\_3 nn\_/nn/sg/trd/\_sm\_0 | adj\_/adj/sg/msc/nom/trd/\_adj\_2 nn\_/nn/sg/msc/nom/trd/\_subj\_0 | 1-0 2-1 | a back part | задний отдел |
| a back part | задняя часть | det\_/det/\_det\_3 nn\_/nn/sg/trd/\_gen\_3 nn\_/nn/sg/trd/\_np\_0 | adj\_/adj/sg/fem/nom/trd/\_adj\_2 nn\_/nn/sg/fem/nom/trd/\_np\_0 | 1-0 2-1 | a back part | задний часть |

Как видно из примера, каждой последовательности слов приписывается синтаксический разбор. Из полученной модели перевода удаляются те фразы, которые не являются связным поддеревом синтаксического разбора всего предложения. Например, в предложении «*The result is an industry that works in a radically different way than it did a decade ago.*» последовательность “a radically different way” является связным поддеревом, а “*in a radically different*” – нет.

Кроме того, из модели удаляются пары переводных эквивалентов, включающих только одно знаменательное слово. Затем каждому контексту в модели перевода ставятся в соответствие возможные ключи (потенциальный заголовок статьи и его перевод). Для этого информация о пословном выравнивании контекстов объединяется со словником словаря. Таким образом, например, для пары биграмм на английском и русском языке “*digital image – цифровая фотография*”, где слова выровнены по порядку, мы получаем две записи – для перевода “digital - цифровой” и “image - фотография”.

Выделенные контексты хранятся в виде значений, разделённых табуляцией, в следующем порядке:

* ключ (заголовочное слово);
* переводной эквивалент;
* частота пары «контекст – перевод» в параллельном корпусе;
* контекст на исходном языке;
* синтаксический разбор и леммы контекста;
* перевод контекста;
* синтаксический разбор и леммы перевода;
* информация о позиции ключа в контексте;
* информация о выравнивании.

## Подготовка обучающей и тестовой выборки

Как уже было сказано, основной задачей эксперимента является упорядочивание иллюстрирующих контекстов в соответствии с допустимостью их использования в качестве материала для иллюстративного блока. Для решения этой задачи мы предлагаем ранжирующий классификатор – алгоритм, который по обучающей выборке приписывает элементам тестовой выборки оценки в заданном диапазоне таким образом, что элементы могут быть затем упорядочены по значению этой оценки. Цель ранжирующей модели — наилучшим образом (в некотором смысле) приблизить и обобщить способ ранжирования в обучающей выборке на новые данные.

Для обучения ранжирующего классификатора были размечены обучающий и тестовый наборы контекстов.

Первым шагом в подготовке выборок был отбор переводных эквивалентов, к которым затем подбирались контексты. Чтобы сделать этот «словарь» более репрезентативным, переводные эквиваленты выбирались случайно таким образом, чтобы распределение их частот соответствовало распределению частот слов в параллельном корпусе. Кроме того, необходимо было установить такое же соответствие с распределением частот запросов к словарю. Было замечено, что количество запросов на английском языке в значительной степени коррелирует с частотой соответствующих слов в корпусе [Antonova, Misyurev 2014], поэтому можно полагаться только на корпусную статистику. Также в выборку не включалось сто самых частых английских слов.

Для каждой пары переводных эквивалентов в словаре из параллельного корпуса извлекаются все возможные контексты (биграммы), как описано в предыдущем разделе. Случайная выборка из полученного множества контекстов могла бы быть ненадёжной, поскольку в ней сложно было бы обеспечить баланс между положительными и отрицательными примерами. Поэтому каждому контексту приписывается вес по эмпирической формуле, которая соответствует произведению прямой и обратной вероятности перевода

где

– частота выравнивания контекста на входном языке в контекст на выходном языке в параллельном корпусе;

– частота контекста на входном языке в том же параллельном корпусе;

– частота контекста на выходном языке в том же параллельном корпусе.

Затем для каждой пары переводных эквивалентов выбирается несколько (от одного до трёх в зависимости от общего числа кандидатов) с наибольшим весом. Таким образом была получена выборка объёмом 700 словосочетаний.

Разметка производилась вручную по пятибалльной шкале от 1 (неприемлемый контекст) до 5 (идеально подходящий контекст). В таблице приведены неформальные критерии, использованные при выставлении оценки. Стоит напомнить, что каждый параллельный контекст состоит из двух частей – входной и выходной. В качестве эксперимента примеры размечались в двух режимах – сначала оценка приписывалась обеим частям, затем каждой по отдельности. При составлении критериев использовался опыт группы аналитиков отдела машинного перевода компании «Яндекс».

Таблица 3. Принципы разметки контекстов-кандидатов.

|  |  |  |  |
| --- | --- | --- | --- |
| Оценка | разметка обеих частей | разметка одной части | пример |
| 1 | Обе части бессмысленны и грамматически неправильны; части не являются переводными эквивалентами | Фраза бессмысленна и грамматически некорректна | *\*pickled <loveliness> → \*маринованная <красота>* |
| 2 | Одна из частей соответствует оценке один по принципам разметки одной части примера; обе или одна из фраз грамматически некорректна | Фраза грамматически некорректна; фраза не является переводным эквивалентом | *caribbean <community> → \*караибское <содружество>* |
| 3 | Обе части грамматически корректны, но не отражают особенностей значения / употребления / перевода ключа | Фраза грамматически корректна, но не отражает особенностей значения / употребления / перевода ключа | *\*его <любовь> → \*his <fondness>* |
| 4 | Обе части грамматически корректны и частично иллюстрируют особенности значения / употребления / перевода ключа | Фраза грамматически корректна и частично иллюстрирует особенности значения / употребления / перевода ключа | *quit the company → покинуть компанию* |
| 5 | Идеально подходящий контекст | Идеально подходящий контекст | *ball lightning → шаровая молния* |

Как было сказано выше, контексты извлекались из корпуса для всех соответствий «английская лексема – русская лексема», полученных из машинного словаря. Машинный словарь в свою очередь содержит некоторое количество «шумных» (ошибочных) переводов: например, из недословных переводов можно извлечь перевод «*beautiful - красота*». Они были удалены из выборки после разметки. В результате размеченный набор составил более 600 примеров. Результаты разметки каждой из частей по отдельности приведены на рисунке 9.

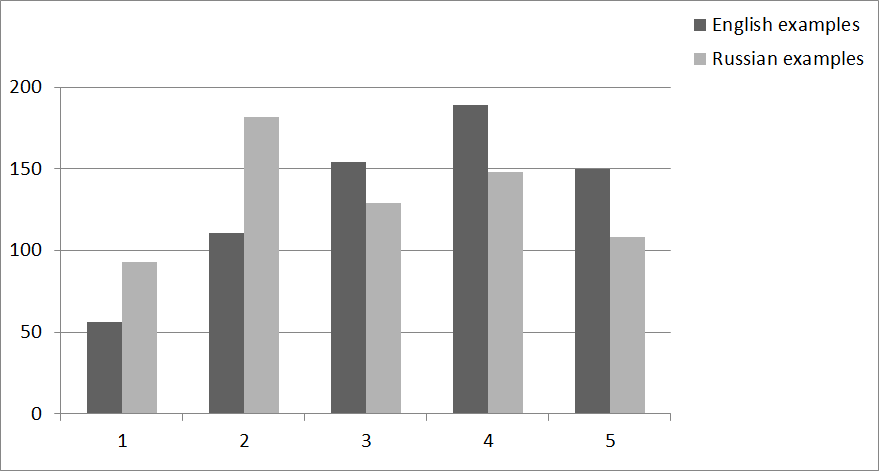


Рисунок 9. Результаты разметки контекстов по пятибалльной шкале: тёмным отмечены английские контексты, светлым - русские.

## Анализ ошибок в контекстах-кандидатах

Среди примеров с оценками 1 и 2 были выделены следующие группы ошибок (ключи выделены угловыми скобками, ошибки отмечены звёздочкой):

1. Ошибки в форме примера
   1. Грамматически некорректные фразы

*\*<preparation> enamel → <составление> эмали*

*<appreciate> acrobatics → \*<оценить> акробатика*

* 1. Незаконченные фразы

*county <detention> → деревенский <исправительный>*

* 1. Фразы в несловарной форме

*\*<created> tsunamis → \*<породило> цунами*

*monstrously <big> → \*чудовищно <огромная>*

*header files → \*заголовочных файлов*

* 1. Фразы, содержащие иностранные слова, транслитерацию

*<improve> resiliency → \*<улучшать> resiliency*

*unformatted <capacity> → \*unformatted <емкость>*

*\*<beginning> shvatyvanija → начало> схватывания*

* 1. Фразы, содержащие слово с опечаткой

*caribbean <community> → \*караибское <содружество>*

*burgundy <sole> → \*бардовая <подошва>*

1. Ошибки в значении
   1. Неинформативные фразы

*\*его <любовь> → \*his <fondness>*

*\*очень <глупый> → \*really <stupid>*

*\*nonpregnant <woman >→ \*небеременная <женщина>*

* 1. Фразы с несвязанными словами

*\*pickled <loveliness> → \*маринованная <красота>*

*\*<saving> neurotic → \*<спасение> невротиков*

*\*синхроничная <жизнь> → \*synchronistic <life>*

* 1. Труднопонятные фразы со специфическим значением

*\*sagittal <reconstruction> → \*сагиттальная <реконструкция>*

*\*threshold <panel> → \*пороговое <табло>*

* 1. Машинный / пословный перевод

*\*<soya> squirrels → <соевый> белок*

*\*<character> stitches → <символьные> строчки*

*\*harvest <control> → жмешь <контрол>*

*\*Berners-<whether> → Бернерс-<ли>*

*hi <camcorder> → \*привет <видеокамеры>*

* 1. Оскорбительные контексты для нейтральных слов

*naked <girl> → голая <девушка>*

*<Japanese> militarists → <японские> милитаристы*

*Hitlerite <Germany> → гитлеровская <Германия>*

*<become> a Shaheed → <стать> шахидом*

* 1. Фразы, не являющиеся переводными эквивалентами

*<saving> rolling → <спасение> утопающих*

Первую группу ошибок можно объяснить двумя взаимосвязанными причинами. Во-первых, контексты извлекаются из интернет-корпуса, который содержит «некачественные» тексты, полученные в результате машинного перевода, случайно и намеренно искажённые тексты. С другой стороны, на извлекаемые контексты не накладывается никаких ограничений, кроме связности поддерева. Таким образом, учитывая возможные ошибки парсера, находятся грамматически некорректные фразы. С качеством корпуса связаны и некорректные переводы контекстов (ошибки в значении), и пословные переводы: *soya squirrels*, *saving rolling*.

При обучении ранжирующих классификаторов из обучающего набора удалялись примеры с оценкой 3, поскольку они не представляют качественного материала для обучения, а скорее зашумляют его.

## Общий алгоритм ранжирования параллельных контекстов

Исходя из проведённого анализа ошибок, а также из изложенных выше принципов отбора иллюстраций и устойчивых сочетаний в словарях классического типа, мы предлагаем следующий алгоритм ранжирования переводных контекстов. Для каждого контекста вычисляются значения признаков, которые позволяют отделить релевантные фразы от некорректных (подробное описание этих признаков приведено ниже). В результате получаем матрицу размера , где *N* – количество переводных контекстов, а *M* – количество признаков. Эти данные подаются на вход ранжирующему классификатору, который приписывает каждому контексту некоторое значение (от 0 до 1 или от 1 до 5). Затем контексты ранжируются внутри группы, найденной для ключа (пары «английское слово – русский перевод»). В словаре для каждого ключа отображается только лучший пример из группы.

## Факторы (признаки) ранжирования

Перечислим признаки, которые подавались на вход ранжирующему классификатору, по группам. Для каждой группы в скобках приводится её сокращённое название, которое используется ниже.

### Оценка по языковой модели (группа LM)

Статистическая языковая модель (модель языка) – это вероятностное распределение последовательностей слов. Такая модель позволяет на основе данных о вероятности появления в тексте слов и их последовательностей определённой длины вычислить вероятность произвольной последовательности. Простейшая модель языка, основанная на n-граммах, работает следующим образом. Вероятность появления предложения оценивается следующим образом:

Здесь применяется предположение о том, что генерация предложения – случайный марковский процесс, то есть текущее состояние (выбор *i*-го слова) зависит лишь от нескольких предыдущих. Вероятность появления слова при условии предшествующего контекста длины вычисляется очевидным образом по значениям частот слов и их последовательностей в корпусе:

где *count(s)* – частота последовательности слов *s*.

Для оценки вероятности n-грамм, не встреченных в корпусе, применяется сглаживание различными методами от линейной интерполяции до методов Гуда-Тьюринга и back-off моделей (подробнее о способах сглаживания в статье [Chen, Goodman 1998]). Простейший метод сглаживания – линейная интерполяция – заключается представлении произведения вероятностей через сумму – например, когда по корпусу невозможно определить условную вероятность триграммы . В триграммной модели условная вероятность слова по двум предыдущим будет оцениваться по значениям вероятностей , , . Метод линейной интерполяции позволяет подобрать для каждого из множителя коэффициенты , , таким образом, чтобы вероятность была представима следующим образом:

Возможность построения различных языковых моделей предоставляет инструмент SRILM (SRI Language Modeling, [Stolcke 2002]). Мы использовали триграммную модель с настройками по умолчанию, собранную отдельно по английской и русской части параллельного корпуса, из которого извлекались контексты. Также были собраны модели по частеречным тегам на тех же корпусах.

Для оценки качества языковых моделей используется перплексия (perplexity) – мера, которая позволяет оценить, насколько хорошо модель описывает новые данные. Идея заключается в следующем: возьмём предложения, которые не входили в корпус для обучения языковой модели. Каждое «новое» предложение можно оценить с помощью имеющейся модели, и, кроме того, можно оценить вероятность всего набора предложений. Для всех *m* предложений общим объёмом *M* слов перплексия вычисляется так:

где

Чем меньше значение, тем лучше моделируются «новые» данные. С другой стороны, при фиксированной языковой модели можно оценить сложность текстов, не входящих в модель. Для одного предложения она представляет собой вероятность предложения, нормированную на количество слов в предложении и приведённую к диапазону

Таким образом, для каждого контекста вычислялись следующие значения (отдельно для английской и русской частей):

* перплексия по униграммной модели;
* перплексия по триграммной модели.

Кроме того, тексты были размечены синтаксическим анализатором, следовательно, каждому токену приписана морфологическая информация. Таким образом, можно оценить не только вероятность последовательности слов, но и вероятность последовательности граммем. Для этого была построена языковая модель по частеречным тегам и в группу признаков *LM* включалась также

* перплексия по триграммной частеречной модели.

### Относительная частота контекста (RelF)

Относительная частота примера вычисляется по статистической модели перевода. Используемая формула позволяет одновременно оценить как частоту примера в корпусе, так и его значимость относительно входящих в него слов:

где

– частота выравнивания английской части примера в русскую;

– максимальная частота по словам, не являющимся ключом, в английском контексте;

– максимальная частота по словам, не являющимся ключом, в русском контексте.

### Взаимная информация (MI)

Как неоднократно отмечалось выше, значительный интерес для иллюстративного блока двуязычного словаря представляют коллокации. Многочисленные меры оценки степени связи внутри коллокаций перечислены, например, в рамках интернет-ресурса collocations.de [Evert 2004] или в работах [Захаров, Хохлова 2010; Пивоварова, Ягунова 2010; Кощеева 2014]. Считается, что некоторые из мер отражают определённый тип сочетаемости; проводились различные попытки сравнения мер.

Таблица 4. Общий вид таблицы сопряжённости для наблюдаемых значений частот слов X и Y.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

Таблица 5. Общий вид таблицы сопряжённости для ожидаемых значений частот слов X и Y.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

В частности, предлагаются следующие способы оценки связи внутри словосочетания, основанные на таблице сопряжённости (*contingency table*, таблицы 4-5) и проверке гипотезы о значениях ожидаемых частот:

*Логарифм правдоподобия (Log-likelihood ratio test)*

*t-test*

*-test*

*z-score*

Каждая из этих мер имеет свои достоинства и недостатки, о которых подробно сказано, например, в [Manning, Schutze 1999: 141-172].

Для данной задачи была выбрана взаимная информация (mutual information, MI), которая является универсальной мерой связи и эффективность которой была показана, в частности, в нашем исследовании [Букия и др. 2015]. В общем случае взаимная информация для пары слов и оценивается так:

где *p(w)* – вероятность слова в корпусе.

Мы предлагаем расширение традиционного понятия взаимной информации для коллокации, поскольку частоты для оценки вероятности вычисляются по синтаксическим связям, а не по биграммам или совместной встречаемости в рамках предложения. Кроме того, поскольку контексты не ограничены биграммами, вычисляется среднее значение MI по всем парам связей слова-ключа. Для того, чтобы «выровнять» значение меры (не поощрять редкие слова и сочетания), используется линейное сглаживание: к частотам отдельных слов при вычислении вероятностей и добавляется константа, значение которой получено эмпирически.

### Векторные модели (WV)

В последнее время значительную популярность приобрели векторные семантические представления, которые позволяют описать различные языковые закономерности (например, [Baroni и др. 2014]). В [Mikolov и др. 2013a] описан способ получения таких векторов на основе рекуррентных нейронных сетей (о них ниже). Затем авторы [Mikolov 2013b] оценивают полученные вектора применительно к различным задачам, таким как нахождение семантически связанных слов, построение аналогий («*Какое слово так относится к Италии, как Париж к Франции?*»), выделение морфологических форм. Семантические вектора широко применялись при исследовании семантической близости на материале русского языка в рамках соревнования RUSSE [Панченко и др. 2015]. Кроме того, было показано, что косинусная мера близости между такими векторами может учитываться и для оценки синтагматической сочетаемости [Panicheva и др. 2016].

Одним из наиболее популярных инструментов для построения семантических векторов является **word2vec** (<https://code.google.com/archive/p/word2vec/>). Это реализация векторных моделей от авторов статьи [Mikolov и др. 2013a], в которой используются два описанных в статье алгоритма – continuous bag-of-words (CBOW) и skip-gram. Оба алгоритма предполагают использование рекуррентной нейронной сети, на вход которой подаются последовательности слов. На выходе сеть должна учиться максимизировать косинусную близость между векторами слов, которые часто встречаются рядом. Алгоритм CBOW обучается предсказывать слово по его контексту, skip-gram – контекст по слову. Авторы статьи [Mikolov и др. 2013а] уточняют, что первая модель лучше подходит для обучения на больших корпусах текстов и работы с частыми словами, а вторая – на меньших корпусах, но лучше работает с редкими словами.

Мы тренировали семантические вектора в 200-мерном пространстве с помощью инструмента word2vec на английской и русской частях параллельного корпуса, из которого извлекались контексты. При построении дистрибутивной модели использовалась архитектура skip-gram.

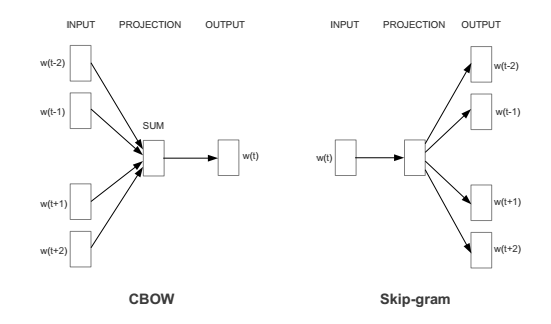


Рисунок 10. Схематическое изображение архитектур cbow и skip-gram.

В качестве признаков группы WV мы использовали конкатенацию всех векторов данного контекста (в порядке следования соответствующих слов) и бинарные признаки, которые указывали на положение ключевого слова в рамках контекста. Таким образом, для биграммы , где ключом является второе слово, вектор признаков WV выглядит следующим образом: .

### Семантическая близость (Sim)

Основным способом оценки семантической близости слов (ассоциации, association) по их семантическим представлениям, является косинус угла между векторами:

где и - вектора, соответствующие словам и .

Как и в случае с взаимной информацией, для контекста вычислялось среднее значение близости ключа к остальным словам примера.

## Методы классификации

Несмотря на детальную разметку обучающей выборки, были произведены эксперименты с бинарной классификаций методом случайного леса и с помощью нейронной сети прямого распространения с одним скрытым слоем. Для предсказания оценки от 1 до 5 использовался набор бинарных классификаторов (также основанных на методе случайного леса). Рассмотрим указанные методы классификации подробнее.

### Случайный лес (Random forest)

Метод случайного леса – алгоритм машинного обучения, предложенный в [Breiman 2001]. Классификатор представляет собой ансамбль решающих деревьев (отсюда и название). Алгоритм классификации таков:

Построим некоторое количество решающих деревьев следующим образом

Сгенерируем случайную подвыборку с повторами размера исходной выборки.

Построим решающее дерево, классифицирующее примеры данной подвыборки, причём при создании нового узла дерева будем выбирать признак, на основе которого производится разбиение, не из всех признаков, а лишь из случайно выбранных. Выбор наилучшего из этих признаков может осуществляться различными способами. В оригинальной статье используется критерий Джини, применяющийся также в алгоритме построения решающих деревьев CART – это мера, которая показывает, насколько часто случайный элемент из множества будет помечен неправильно, если следовать распределению классов в подмножестве элементов. Она вычисляется как сумма произведений вероятности выбора каждого элемента на вероятность ошибки при его классификации. Для элементов мера Джини вычисляется следующим образом при условии, что – доля элементов, размеченных как *i*-ый элемент множества:

Построение прекращается, когда заканчиваются элементы подвыборки.

Классификация каждого элемента производится затем с учётом решений всех деревьев: выбирается тот класс, к которому элемент относит большинство.

Количество деревьев определяется таким образом, чтобы минимизировать ошибку классификатора на тестовой выборке.

Одним из важных свойств метода случайного леса является возможность оценки значимости каждого признака. Для этого каждому элементу на этапе построения решающих деревьев приписывается так называемая ошибка *out-of-bag*: средняя вероятность ошибки на данном элементе, если не учитывать деревья, в которые входит данный элемент. Затем на всём тренировочном наборе перемешиваются (меняются по элементам) значения признака, который оценивается. Значимостью параметра будет среднее значение разности ошибки *out-of-bag* по всем элементам набора до и после перемешивания.

Возможность оценки вероятной ошибки с использованием только обучающего набора (*out-of-bag*) – одно из основных достоинств метода случайного леса. Можно также упомянуть нечувствительность к масштабированию значений признаков, возможность одновременной работы с дискретными и непрерывными значениями без потери в качестве.

В нашей работе использовались параметры из оригинальной статьи Бреймана (критерий Джини, оценка с помощью out-of-bag) в реализации библиотеки scikit-learn (<http://scikit-learn.org>, [Pedregosa и др. 2011]).

### Нейронная сеть прямого распространения

Искусственные нейронные сети изначально создавались как модель биологической нервной системы. Именно поэтому в этой области многие термины заимствованы из нейрофизиологии. Основной единицей этой модели является нейрон, состоящий из следующих частей:

* определённое количество входных сигналов;
* сумматор, складывающий входные сигналы с определенными весами в значение ;
* сеть синапсов (связей) соединяющая входные сигналы с сумматором, при этом у каждого ребра-синапса есть вес;
* функция активации , преобразующая результат суммирования в выходной сигнал – обычно в диапазоне ;

Обычно используются следующие функции активации:

* пороговая функция (модель Мак-Каллока – Питца)
* кусочно-линейная функция
* сигмоидальная функция

где – параметр наклона.

Множество таких нейронов составляет слой нейронной сети. Самая простая сеть будет состоять из трёх слоёв нейронов: входного, скрытого и выходного. Для входного и выходного слоя известны некоторые параметры: какой сигнал подаётся на вход и какой оказывается на выходе. В скрытом слое, как следует из названия, производятся неявные преобразования сигналов. Многослойные сети позволяют находить более сложные зависимости, особенно когда размер входного слоя достаточно велик.

Архитектура нейронной сети предполагает возможность обратной связи, когда выходной сигнал оказывает влияние на новый входной. В зависимости от наличия в сети обратной связи выделяют сети прямого распространения и рекуррентные сети. В первых информация из входного слоя передаётся на выходной слой нейронов; в случае многослойной сети – через один или несколько слоёв скрытых нейронов. Рекуррентные нейронные сети отличает наличие хотя бы одной обратной связи.

Математическая модель нейронной сети может применяться в обучении как с учителем, так и без него. Одним из наиболее распространённых алгоритмов обучения сети является метод обратного распространения ошибки. Он позволяет вычислить синаптические веса (веса связей) таким образом, что значение функции потерь (величина ошибки) стабилизируется и становится достаточно небольшим. При этом каждый новый обучающий пример вносит свой вклад в изменение весов, когда на каждом этапе выходное значение сравнивается с желаемым, и веса пересчитываются в обратном направлении – от выходного слоя к входному.

В экспериментах использовалась сеть прямого распространения с одним скрытым слоем. Такое решение было принято, с одной стороны, чтобы сократить вычислительные затраты, а с другой, поскольку множество признаков, которыми описываются входные данные, сравнительно невелико. Размер скрытого слоя составляет 10 элементов, а максимальное количество итераций – 150 (если сходимость не достигается раньше). Параметры были подобраны эмпирически для достижения наилучших показателей качества.

### Оценка мультиклассификации

Оценка примеров с использованием пятибалльной шкалы полезна, когда необходимо выделить наиболее релевантные примеры. Однако, классификация на более чем два класса – довольно сложная задача для одного классификатора. Поэтому были объединены бинарные классификаторы на основе метода случайного леса для каждого класса. Кроме того, при обучении каждого из них из обучающего набора удалялись примеры с ближайшей оценкой: например, при обучении классификатора, определяющего принадлежность примера к 5 классу, из выборки удалялись примеры с оценкой 4. При предсказании оценки на тестовом наборе выбиралась оценка, соответствующая классификатору с наибольшим значением. Кроме того, вычислялась надёжность такой оценки по формуле:

где – предсказание *i*-го классификатора.

## Оценка качества классификации

Основные эксперименты используют бинарную классификацию или мультиклассификацию. Для их оценки использовались стандартные метрики – точность (precision) и аккуратность (accuracy). В общем виде значения вычисляются следующим образом:

используемые обозначения проще всего представить в таблице (см. Таблица 6).

Точность вычислялась по отрицательным примерам, поскольку важнее было убедиться, что метод фильтрует некачественные контексты. Ниже такая метрика будет обозначаться .

Таблица 6. Варианты ошибок классификации.

|  |  |  |  |
| --- | --- | --- | --- |
| количество примеров | | исходная разметка | |
| 0 | 1 |
| разметка классификатора | 0 | *tn* | *fn* |
| 1 | *fp* | *tp* |

## Результаты классификации

Для оценки качества классификации размеченная выборка делилась в соотношении 2:1 на обучающую (416 примеров) и тестовую (206 примеров) соответственно. Проведена оценка пяти методов классификации по разным группам признаков:

* оценка каждой части примера по отдельности:
  + – случайный лес на группе признаков *WV*;
  + – мультиклассификатор на группе признаков *WV*;
  + – нейронная сеть на группе признаков *WV*;
* общая оценка параллельного контекста:
  + – случайный лес на группах *LM*, *MI, RelF* and *Sim*;
  + – случайный лес на всех признаках.

Результаты классификации приведены в таблицах 7а и 7b.

Из таблицы 7а можно заключить, что английские части параллельных контекстов классифицируются значительно лучше русских. Мы предполагаем, что это связано с качеством семантических векторов: и те, и другие были обучены на корпусах одного размера, однако языки с богатой морфологией (как русский), как правило, требуют большего объёма данных.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *P0(en)* | *A(en)* | *P0(ru)* | *A(ru)* |
| *RF1* | 0.71 | **0.74** | 0.62 | **0.64** |
| *RF2* | **0.83** | 0.65 | **0.63** | 0.62 |
| *NN* | 0.67 | 0.65 | 0.56 | 0.61 |

1. Предсказание качества одной из частей примера.

|  |  |  |
| --- | --- | --- |
|  | *P0* | *A* |
| *RF3* | **0.69** | 0.7 |
| *RF4* | 0.685 | **0.71** |

1. Предсказание качества параллельного контекста.

Таблица 7. Результаты классификации контекстов.

В случае с классификацией параллельных контекстов целиком, можно заметить, что добавление признаков группы WV практически не влияет на качество классификации. В связи с этим можно заключить, что информации о близости векторов слов достаточно для описания словосочетания.

## Сравнение ранжирования с простыми эвристическими методами

Для сравнения результатов ранжирования было выбрано 140 пар «английское слово – русский эквивалент», которые не входили в размеченную ранее выборку. Для каждой пары извлекались все возможные параллельные контексты, которые затем ранжировались следующим значениям:

* – абсолютная частота параллельного контекста в параллельном корпусе;
* – сумма значений взаимной информации ( в том виде, как описано выше) для английской и русской части примера;
* , , – значения, полученные от соответствующих классификаторов (фактически – вероятность отнесения соответствующего слова к положительному классу);
* – модифицированные значения классификатора . В случае, если обе части контекста оценивались как положительные, использовалось значение надёжности c. Если только одна из частей контекста была отнесена к классам (4, 5), то из значения надёжности для этой части вычиталось наименьшее значение надёжности для другой части.

Оценка производилась следующим образом. Из каждого упорядоченного списка для каждого ключа сохранялся лучший параллельный контекст. Затем примеры в полученных списках размечались как положительные (1) и отрицательные (0). Результаты разметки приведены в приложении А.

Мерой качества служило количество положительных примеров в списке. Это связано с тем, что, как было сказано выше, формат используемого машинного словаря позволяет показывать только один параллельный контекст для одного ключа, то есть важно оценить, как часто в верхней части списка оказываются отрицательные примеры. Результаты приведены в таблице 8.

Таблица 8. Оценка ранжирования иллюстративных примеров.

|  |  |  |
| --- | --- | --- |
|  | **положительные примеры** | **доля положительных примеров** |
| *MI* | 60 | 42.8 |
| *F* | 44 | 31.4 |
|  | 59 | 42.1 |
|  | 76 | 54.3 |
|  | **88** | **62.9** |
|  | 74 | 52.9 |

Результаты показывают, что использование машинного обучения значительно улучшает ранжирование по сравнению с простыми методами. Сравнение различных наборов признаков позволяет сказать, что наиболее удачной комбинацией является сочетание всех групп за исключением *WV*.

## Значимость признаков классификации

Некоторые выводы о важности той или иной группы признаков можно сделать по результатам оценки в предыдущих разделах. Так, заметно, что использование группы WV вместе с остальными избыточно, хотя возможно, что в отдельности классификация, основанная на этой группе, показала бы лучшие результаты на большем обучающем наборе.

Как было сказано, метод случайного леса позволяет вычислить значимость каждого признака. Данные о значимости групп признаков, использовавшихся при классификации , приведены в таблице 9.

Таблица 9. Значимость признаков для классификатора RF3.

|  |  |
| --- | --- |
| MI | 0.13 … 0.097 |
| LM | 0.1 … 0.04 |
| F | 0.079 … 0.05 |
| Sim | 0.063 … 0.04 |

Наиболее существенными оказываются признаки группы *MI*, то есть устойчивость сочетания. Таким образом, получается, что в обучающей выборке отражена тенденция к выбору в качестве лучших иллюстраций коллокаций, а классификатор стремится воспроизвести эту закономерность. Следующими по значимости выделяются признаки группы *LM*, причём больший вес имеют значения по триграммной языковой модели и модели по частеречным тегам. Следовательно, таким образом учитывается грамматическая корректность и «беглость» (*fluency*) словосочетания. Очевидно, что при использовании взаимной информации по синтаксическим связям, как меры устойчивости, косинусная близость векторов слов оказывается наименее значимым признаком.

## Выводы

В главе 4 был описан эксперимент по ранжированию контекстов для иллюстрации статей машинного переводного словаря. Были использованы два метода машинного обучения – метод случайного леса и нейронная сеть прямого распространения; в качестве ранжирующих признаков выбраны те, которые отражают принципы отбора иллюстраций, выявленные при анализе различных словарей в главах 2 и 3. Полученные результаты оценивались стандартными метриками (точность и аккуратность), лучшие значение составили 0.83 и 0.74 соответственно. Кроме того, было проведено сравнение с простыми эвристическими способами ранжирования и показано, что использование расширенного набора признаков и методов машинного обучения даёт значительное улучшение в результатах ранжирования.

# Заключение

Одной из наиболее существенных частей словарной статьи в двуязычном словаре, в частности, вследствие его учебных задач, является иллюстративный блок, включающий примеры употребления заголовочного слова. При этом принципы его создания не разработаны ни в классической лексикографии, ни в машинной. Анализ существующих двуязычных словарей (англо-русского направления) позволяет выделить несколько критериев отбора материала для иллюстративного блока: словосочетания предпочитаются предложениям, словосочетания содержат дополнительную информацию о сочетаемости, стилистических особенностях опорного слова и так далее.

Поскольку сочетаемость становится одним из важнейших критериев для двуязычных иллюстраций, в работе так же рассматриваются различные концепции устойчивых сочетаний – от классических до новейших. Из приведённого обзора словарей сочетаемости и словарей коллокаций можно также сделать выводы о некоторых принципах, на которых базируется понятие устойчивости.

Аналитический обзор, изложенный в главах 1-3, позволяет нам выделить признаки для автоматического ранжирования параллельных контекстов относительно их релевантности для иллюстративного блока машинного словаря. Проведены эксперименты с различными наборами признаков, методами машинного обучения, сделана оценка полученных результатов. Наилучшая комбинация методов и признаков показывает точность 84% на английских контекстах и 71% при ранжировании пар контекстов. Полученные результаты используются при составлении машинного словаря сервиса «Яндекс.Переводчик».

# Лексикографические источники

1. Амосова Н.Н. Большой англо-русский словарь / Сост. Амосова, Н.Н., Апресян, Ю.Д., Гальперин И.Р.; под общ. руководством Гальперина И.Р. - 2-е изд., стер. – М. : Русский язык, 1977.
2. Англо-русский русско-английский словарь = English-russian Russian-English dictionary : 35573 слова, значения и словосочетания / [ред. и рук. проекта: В.Ю. Бурнашева]. – М. : ABBYY Press, 2011. – 603 с.
3. Анисимова, Т.И. Пособие по лексической сочетаемости слов русского языка: Словарь-справочник / Сост. Т.И. Анисимова, З.Э. Иванова, Р.В. Ульянко; Под ред. Т.П. Плещенко и Л.Ф. Саковец. - Минск : Вышэйш. школа, 1975. – 303 с.
4. Апресян, Ю.Д. Новый большой англо-русский словарь = New English-Russian Dictionary : Ок. 250000 лексич. единиц : В 3 т / Сост. Ю.Д. Апресян, Э.М. Медникова, А.В. Петрова и др.Под общ. руководством акад. Ю.Д. Апресяна и д.филол.н., проф. Э.М. Медниковой. - 5-е изд., стер. – М. : Рус. яз., 2000.
5. Ахманова, О.С. Русско-английский словарь : Около 50000 слов / Сост. О.С. Ахманова, Т.П. Горбунова, Н.Ф. Ротштейн [и др.]; Под общ. рук. проф. А.И. Смирницкого. – М. : Гос. изд-во иностр. и нац. словарей, 1948 (16-я тип. треста "Полиграфкнига"). – 988 с.
6. Бенсон, М. Комбинаторный словарь английского языка. / Бенсон, М., Бенсон, Э., Илсон, Р. – М.: Русский язык, 1990 – 286 с.
7. Большой англо-русский словарь = New English-Russian dictionary : [В 2 т. : Ок. 160000 слов] / Под общ. руководством И.Р. Гальперина, Э.М. Медниковой. - 4-е изд., испр., с доп. – М. : Рус. яз., 1987.
8. Борисова, Е.Г. Слово в тексте. Словарь коллокаций (устойчивых словосочетаний) русского языка с англо-русским словарём ключевых слов. – М.: 1995. (Борисова 1995б).
9. Братусь, Б.В. Russian word-collocations = Словосочетания русского языка : Учеб. словарь для говорящих на англ. яз / Б.В. Братусь, И.Б. Братусь, Е.А. Данциг и др. – М. : Рус. яз., 1979. – 367 с.
10. Гинзбург Р.С. Англо-русский словарь глагольных словосочетаний = English-Russian dictionary of verbal collocations / Р.С. Гинзбург и др.; Под общ. руководством Э.М. Медниковой. - 2-е изд., испр. – М. : Рус. яз., 1990. – 667 с.
11. Козырева, М.Н., Англо-русский словарь = English-Russian dictionary : Около 14000 слов и 63000 словосочетаний / М.Н. Козырева, И.В. Федорова. – М. : Рус. яз., 1998. – 1118 с.
12. Красных, В.И., Русские глаголы и предикативы : Слов. сочетаемости. – М. : Арсис лингва, 1993. – 226 с.
13. Кунин, А.В. Англо-русский фразеологический словарь – М. Русский язык, 1984. – 944 с.
14. Мельчук, И.А., Толково-комбинаторный словарь современного русского языка : Опыты семантико-синтаксического описания русской лексики. – Вена, 1984. – 992 с
15. Мюллер, В.К. Англо-русский словарь, 1928.
16. Ожегов, С. И. Словарь русского языка. Изд. 21-6. – М., 1989.
17. Регинина, К.В. Устойчивые словосочетания русского языка / К.В. Регинина, Г.П. Тюрина, Л.И. Широкова; Под ред. Л.И. Широковой. – 3-е изд. – М : Рус. яз., 1983. – 300 с.
18. Kjellmer, G. A dictionary of English collocations : based on the Brown corpus : in three volumes. – Oxford; New York: Clarendon Press: Oxford University Press, 1994.
19. Longman dictionary of contemporary English. / Сост. Procter, P. – Harlow ; London : Longman, 1978. - XXXIX, 1303 с.
20. McIntosh, C. Oxford Collocations Dictionary. – Oxford University Press, 2009. – 963 с.
21. Rundell, M. Macmillan Collocations Dictionary. – Macmillan, 2010.
22. Sinclair, J. Collins COBUILD English collocations on CD-ROM / Sinclair, J., Fox, G., Seymour, D., Clear, J. – London: HarperCollins Publishers, 1995.
23. The BBI combinatory dictionary of English : A guide to word combinations / Сост. Benson, M., Benson, E., Ilson R. – Amsterdam ; Philadelphia : Benjamins, 1986. – XXXVI, 286 с.
24. Wheeler, M., Unbegaun, B. The Oxford Russian dictionary : Russian-English, English-Russian. – Oxford ; New York : Oxford univ. press, 1997. – XVII, 1340 c.

# Литература

1. Азарова, И.В. Принципы построения wordnet-тезауруса RussNet / Азарова, И.В., Синопальникова А.А., Яворская М.В. // Компьютерная лингвистика и интеллектуальные технологии. – М., 2004.
2. [Амосова, Н. Н.](https://ru.wikipedia.org/wiki/%D0%90%D0%BC%D0%BE%D1%81%D0%BE%D0%B2%D0%B0,_%D0%9D%D0%B0%D1%82%D0%B0%D0%BB%D0%B8%D1%8F_%D0%9D%D0%B8%D0%BA%D0%BE%D0%BB%D0%B0%D0%B5%D0%B2%D0%BD%D0%B0) Основы английской фразеологии. – Л., 1963.
3. Апресян, Ю.Д. Лексическая семантика (синонимические средства языка). – М.: Наука, 1974.
4. Баранов, А.Н. Введение в прикладную лингвистику. Учебное пособие. – М.: Эдиториал УРСС, 2001. – 360 с.
5. Берков, В. П. Двуязычная лексикография. – АСТ, 2006
6. Берков, В.П. Вопросы двуязычной лексикографии. – Л., 1973.
7. Борисова, Е.Г. К типологии составляющих пакета Устойчивые сочетания // Фразеография в Машинном фонде русского языка. М., 1990. – с. 88-103.
8. Борисова, Е.Г. Коллокации. Что это такое и как их изучать. – М.: 1995. (Борисова 1995a).
9. Букия, Г.Т. Корпусная оценка степени близости единиц в лексических конструкциях / Букия, Г.Т., Протопопова, Е.В., Митрофанова, О.А. // Структурная и прикладная лингвистика. Межвузовский сборник. №11. Под. ред. А.С. Герда. – СПб, 2015.
10. Буторова, У.В. Структурная типология словарных статей в словарях русского языка и способы их формального представления / Буторова У.В., Герд А.С., Захаров В.П., Панков Д.И., Пурицкая Е.В., Хохлова М.В. // НТИ. Сер. 2. ИНФОРМ. ПРОЦЕССЫ И СИСТЕМЫ 2016. № 2. – М., 2016.
11. Бюллетень машинного фонда русского языка /отв. редактор В.М. Андрющенко / – М., 1996, Вып.3
12. Виноградов, В.В. Об основных типах фразеологических единиц в русском языке // Виноградов В. В. Избранные труды. Лексикология и лексикография. – М., 1977. - С. 140-161.
13. Гак, В.Г. К проблеме семантической синтагматики // Проблемы структурной лингвистики. – М.: Наука, 1972. – С. 367-395.
14. Захаров, В.П. Электронный обменный формат проекта TEI (Text Encoding Initiative) для словарей. Учебное пособие. – СПб.: СПбГУ. РИО. Филологический факультет, 2013. – 80 с.
15. Захаров, В.П., Богданова, С.Ю. Корпусная лингвистика: Учебник для студентов направления Лингвистика. 2-е изд., перераб. и дополн., – СПб.: СПбГУ. РИО. Филологический факультет, 2013. – 148 с.
16. Захаров, В.П., Хохлова, М.В. Анализ эфффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции Диалог’2010. – М., 2010
17. Караулов, Ю.Н. Анализ метаязыка словаря с помощью ЭВМ / Караулов Ю.Н., Молчанов В.И., Афанасьев В.А., Михалев Н.В. – М., 1982.
18. Кобозева, И.М. Лингвистическая семантика: Учебник для вузов. – М.: УРСС, 2000. 350 с.
19. Кощеева, С.С. Сравнение методов автоматического выделения глагольно-именных словосочетаний // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Труды XVII Всероссийской объединённой конференции «Интернет и современное общество» (IMS-2014), Санкт-Петербург, 19-20 ноября 2014 г. – СПб: Университет ИТМО, 2014. – С. 298-303.
20. Крижановский, А.А. Оценка использования корпусов и электронных библиотек в Русском Викисловаре // Труды Международной научной конференции Корпусная лингвистика-2011. – СПбГУ, Филологический факультет Санкт-Петербург, 2011. С. 217-222.
21. Крижановский, А.А. Построение машинно-читаемого словаря на основе русского викисловаря. Тр. СПИИРАН, 2009, в. 11. С. 228-233
22. Крылов, С.А., Митрофанова, О.А. «Типовой контекст»: случайность или закономерность? // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог’2006 (Бекасово, 31 мая – 4 июня 2006 г.). – М.: Издательство РГГУ, 2006. – C. 382-388.
23. Крылов, С.А., Старостин, С.А. Металингвистическая разметка текстовых баз данных в системе STARLING и современные задачи корпусной лингвистики // Прикладная лингвистика в поиске новых путей. Международная конференция Megaling’2005. 27 июня - 2 июля 2005. – Симферополь, Таврический национальный университет им. В.И. Вернадского, 2005. С. 33.
24. Марчук Ю.Н. Вычислительная лексикография. – М.: ВЦП, 1976, 175 с.
25. Мельчук, И.А. Опыт теории лингвистических моделей "Смысл-Текст" : Семантика, синтаксис / И.А. Мельчук. – [Переизд.]. – М : Шк. "Языки рус. культуры", 1999. – 345 с.
26. Мельчук, И.А., Иорданская, Л.Н. Смысл и сочетаемость в словаре. – М., 2007.
27. Панченко, А. RUSSE: семинар по оценке семантической близости для русского языка. / Панченко, А., Лукашевич Н.В., Усталов Д., Паперно Д., Мейер К.М., Константинова Н. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции Диалог (Москва, 27 - 30 мая 2015 г.). Вып. 14 (21). – М.: Изд-во РГГУ, 2015. – C. 89
28. Телия, В. Н. Русская фразеология: Семантический, прагматический и лингвокультурологический аспекты. – М., 1996.
29. Телия, В.Н., Типы языковых значений : Связан. значение слова в яз. – М : Наука, 1981. – 269 с.
30. [Шанский, Н.М.](https://ru.wikipedia.org/wiki/%D0%A8%D0%B0%D0%BD%D1%81%D0%BA%D0%B8%D0%B9,_%D0%9D%D0%B8%D0%BA%D0%BE%D0%BB%D0%B0%D0%B9_%D0%9C%D0%B0%D0%BA%D1%81%D0%B8%D0%BC%D0%BE%D0%B2%D0%B8%D1%87) Фразеология современного русского языка / Н. М. Шанский. – 3-е изд., испр. и доп. – М., 1985. – 160 с.
31. Янус, Э. Пять польских словарных статей // Научно-техническая информация, Серия 2, N 11. С. 21-24.
32. Altenberg, B. On the phraseology of spoken English: the evidence of recurrent word-combinations. // A. P. Cowie (ред.), Phraseology: theory, analysis and applications (pp. 101-122). – Oxford: Oxford University Press, 1998.
33. Amsler, R.A. Computational lexicology: a research program // Proceedings of the June 7-10, 1982, national computer conference, 657-63. – ACM, 1982.
34. Antonova, A., Misyurev, A. Automatic Creation of Human-Oriented Translation Dictionaries // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2014” [Komp’yuternaya Lingvistika I Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2014”]. – М., 2014.
35. Antonova, A., Misyurev, A. Russian dependency parser SyntAutom at the DIALOGUE-2012 parser evaluation task // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2012” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2012”]. – М., 2012.
36. Antonova, A. The influence of different data sources on finding and ranking synonyms for a large-scale vocabulary / Antonova, A., Kobernik, T., Misyurev, A. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2016” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2016”]. – М., 2016. [в печати]
37. Atkins, S. Theoretical Lexicography and its relation to Dictionary-making”. // Dictionaries: the Journal of the Dictionary Society of North Americaю – DSNA, Cleveland Ohio, 1993. С. 4-43.
38. Atkins, S., Rundell M. The Oxford Guide to Practical Lexicography. – Oxford University Press, 2008.
39. Bai M. DOMCAT: a bilingual concordancer for domain-specific computer assisted translation / Bai, M., Hsieh Y., Chen K., Chang J.S. // Proceedings of the ACL 2012 System Demonstrations, 55–60. – Association for Computational Linguistics, 2012.
40. Bally Ch. Traité de stylistique française. – Paris, 1951.
41. Bansal M. Unsupervised translation sense clustering / Bansal, M., DeNero J., Lin D. // Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 773–82. – Association for Computational Linguistics, 2012.
42. Barlow, M. Parallel Concordancing and Translation // Translating and the Computer, 2004.
43. Baroni M. Frege in space: A program of compositional distributional semantics. / Baroni, M., Bernardi, R., Zamparelli, R. // Linguistic Issues in Language Technology 9, 2014.
44. Bouamor Dh. Building Specialized Bilingual Lexicons Using Word Sense Disambiguation / Bouamor, D., N. Semmar, P. Zweigenbaum. // – IJCNLP, 2013. С. 952–956.
45. Breiman L. Random Forests. // Machine Learning 45 (1): 5–32. 2001.
46. Brown P.F. The mathematics of statistical machine translation: Parameter estimation / Brown P.F., Della Pietra V., Della Pietra S., Mercer R. // Computational linguistics 19/ 2 (1993). – C. 263–311.
47. Butina-Koller E. Kollokationen im zweisprachigen Wörterbuch. Zur Behandlung lexikalischer Kollokationen in allgemeinsprachlichen Wörterbüchern des Sprachenpaares Französisch/Russisch. Lexicographica. Series Maior 124. – De Gruyter, 2005
48. Chatterjee D. Co-occurrence graph based iterative bilingual lexicon extraction from comparable corpora / Chatterjee D., Sarkar S., Mishra A. // Proceedings of the 4th Workshop on Cross Lingual Information Access, COLING 2010 workshop, 2010. С. 25-42.
49. Chen S.F., Goodman J. An Empirical Study of Smoothing Techniques for Language Modeling // Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.
50. Church, K. W., Hanks P. Word association norms, mutual information, and lexicography. // Computational linguistics. V.16, No. 1 (1990). С. 22–29.
51. Collocations, corpus, dictionnaires / P. Blumenthal, F.J. Hausmann. - Paris : Larousse, 2006. – 128 с.
52. Cowie A.P. The Oxford History of English Lexicography. – Oxford: Clarendon Press, 2009.
53. Dunning, T. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1). – MIT Press, 1999. С. 61–74
54. Evert, S. The Statistics of Word Cooccurences Word Pairs and Collocations. PhD thesis. – Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart: 2004.
55. Firth, J. R. The Technique of Semantics. – Transactions of the Philological Society, 1935. С. 36-72.
56. Firth, J.R. Modes of Meanings. Reprinted in Papers in Linguistics 1934-1951. – London: Oxford University Press, 1957. С. 190-215.
57. Fišer, D., Ljubešic, N. Bilingual lexicon extraction from comparable corpora for closely related languages // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP’11), 2011. С. 125–31.
58. Fung, P. A statistical view on bilingual lexicon extraction // Parallel Text Processing. – Springer, 2000. C. 219-236.
59. Grundfragen der elektronischen Lexikographie : elexico - das Online-Informationssystem zum deutschen Wortschatz / Сост. Ulrike Haβ. – Berlin ; New York : de Gruyter, cop. 2005. – 334 c. – (Schriften des Instituts für deutsche Sprache ; Bd. 12).
60. Halliday, M.A.K. Lexis as a linguistic level // In Memory of J.R. Firth. – London: Longman, 1966. C. 150-61.
61. Hanks, P. Definitions and Explanations // Looking Up – An account of the Cobuild project in lexical computing. – London: William Collins Sons & Co. Ltd., 1987.
62. Hanks, P. How people use words to make meanings: Semantic types meet valencies // Input, Process and Product: Developments in Teaching and Language Corpora, 2012. C. 54–69.
63. Hausmann, F.J. Un dictionnaire des collocations est-il possible? // Travaux de linguistique et de litterature 17/1. 1979. C. 187-195.
64. Hazem, A., Morin, E. Adaptive Dictionary for Bilingual Lexicon Extraction from Comparable Corpora // LREC, 2012. С. 288–92.
65. Herbst, Th. A Valency Dictionary of English: A Corpus-based Analysis of the Complementation Patterns of English Verbs, Nouns, and Adjectives‬ // [Topics in English linguistics (V. 40)‬‬‬‬](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=bibliogroup:%22Topics+in+English+linguistics%22&source=gbs_metadata_r&cad=9), 2004.‬‬‬‬‬‬‬‬
66. Herbst, Th. What are collocations: Sandy Beaches or False Teeth? // English Studies V.4, 1996. C. 379-93.
67. Herbst, Th. Why Construction Grammar Catches the Worm and Corpus Data can Drive you Crazy: Accounting for Idiomatic and Non-Idiomatic Idiomaticity // Journal of Social Sciences 11, V. 3, 2015. C. 91–110.
68. Hjelm, H. Using a Third Language to Improve Extraction of Bilingual Term Correspondences. Term Paper, 2006.
69. Hornby, A.S. Idiomatic and syntactic English dictionary. – Tokyo, Institute for Research in Language Teaching, 1942.
70. Huck, M. Lexicon models for hierarchical phrase-based machine translation / Huck, M., Mansour S., Wiesler S., Ney H. // IWSLT, 2011. С. 191–98.
71. Ismail, A., Manandhar S. Bilingual lexicon extraction from comparable corpora using in-domain terms // Proceedings of the 23rd International Conference on Computational Linguistics: Posters. –Association for Computational Linguistics, 2010. С. 481–89.
72. Jian, J. TANGO: Bilingual collocational concordancer / Jian, J., Chang Y., Chang J.S. // Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, 19. – Association for Computational Linguistics, 2004.
73. Kilgariff, A. GDEX: Automatically finding good dictionary examples in a corpus / Kilgariff A., Rychlý, P, Husák M., Rundell, M., Mcadam K. // Proceedings of the XIII EURALEX International Congress. 1. – Barcelona: Institut Universitari de Lingüística Aplicada, 2008. С. 425-432.
74. Kilgarriff A. Collocationality (and how to measure it) // Proceedings of the Euralex International Congress. – Torino, 2006.
75. Kilgarriff, A. Using corpora as data source for dictionaries. The Bloomsbury Companion to Lexicography. – London: Bloomsbury, 2013. С. 77–96.
76. Kitamura, M., Matsumoto Y. Automatic extraction of word sequence correspondences in parallel corpora // Proceedings of the 4th Workshop on Very Large Corpora, 1996. С. 79–87.
77. Kjaersgaard, P. S. RefTex – a context-based translation aid // Third conference of the European Chapter of the Association for Computational Linguistics: Proceedings of the conference. – Copenhagen, 1987.
78. Koehn, P. Statistical Machine Translation. – Cambrige University Press, 2010.
79. Kučera H. , Francis, W. N. Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. – Providence, Rhode Island: Department of Linguistics, Brown University, 1964.
80. Langlois, L. Bilingual concordancers: a new tool for bilingual lexicographers // Expanding MT horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas. – Montreal, Quebec, Canada, 1996.
81. Laws, F. A Linguistically Grounded Graph Model for Bilingual Lexicon Extraction. / Laws F., Heid U., Michelbacher L., Christian Scheible, Beate Dorow, и Hinrich Sch¨utze. // Coling Poster Session. – COLING, 2010.
82. Li, L. Cyberdictionaries // English Today. V. 14/2. – Cambrige University Press, 1998.
83. Liu, Z. Two-word collocation extraction using monolingual word alignment method / Liu Z., Wang H., Wu H., Li Sh. // ACM Transactions on Intelligent Systems and Technology (TIST) 3/1. – ACM, 2011. C. 16.
84. Luo J., Lepage Y. Extraction of Potentially Useful Phrase Pairs for Statistical Machine Translation // Journal of Information Processing. V. 23, N. 3, 2015.
85. Melamed, I. D. Automatic construction of clean broad-coverage translation lexicons. 1996.
86. Mickievicz, A. Computational Lexicography of Multi-Word Units: How Efficient Can It Be? / Mickiewicz, A., Savary, A., Czerepowicka, M. , Makowiecki F. // 23rd International Conference on Computational Linguistics, 2. – ACL, 2010.
87. Mikolov, T. Efficient Estimation of Word Representations in Vector Space / Mikolov T., Chen K., Corrado G., Dean J. // Proceedings of Workshop at ICLR, 2013. (Mikolov 2013a)
88. Mikolov, T. Linguistic Regularities in Continuous Space Word Representations / Mikolov T., Yih W., Zweig G. // Proceedings of NAACL HLT, 2013. (Mikolov 2013b).
89. Mittman, B. Pragmatik und Wörterbücher: Präfabrizierte Einheiten der gesprochenen Sprache. // Th. Herbst, G. Lorenz, B. Mittmann, M. Schnell (ред.): Lexikografie, ihre Basis- und Nachbarwissenschaften. (Englische) Wörterbücher zwischen 'common sense' und angewandter Theorie (= Lexicographica. Series Maior 118). – Tübingen: Niemeyer, 2004. C. 137-148.
90. Morin, E., Prochasson E. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora // Proceedings of the 4th workshop on building and using comparable corpora: comparable corpora and the web. – Association for Computational Linguistics, 2011. С. 27–34.
91. Och F.J., Ney H. Improved Statistical Alignment Models // Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. – ACL, 2000. C. 440-447.
92. Och, F. J., Ney, H. A Systematic Comparison of Various Statistical Alignment Models // Computational linguistics 29 (1), 2003. C. 19-51.
93. P5: Guidelines for Electronic Text Encoding and Interchange. Chapter 9. Dictionaries. URL: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html> (дата обращения 18.05.2016)
94. Palmer, H. E. Second Interim Report on English Collocations. – Tokyo: Institute for Research in English Teaching, 1933.
95. Panicheva, P. Evaluating Distributional Semantic Models with Russian Noun-Adjective Compositions / Panicheva P., Bukiya G., Protopopova E., Mitrofanova O. // Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016. – Yekaterinburg, Russia, 2016, Revised Selected Papers. [в печати]
96. Polguere A., Kahane S. Formal foundation of lexical functions. В Proceedings of ACL/EACL 2001 Workshop on Collocation, 8–15, 2001.
97. Polguère, A. Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French // Proceedings of the Ninth EURALEX International Congress, EURALEX 2000. – Stuttgart, Germany, August 8th-12th, 2000. C. 517–27.
98. Protopopova, E. Acquiring relevant context examples for a translation dictionary / Protopopova, E., Antonova, A., Misyurev, A. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2015” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2015”]. – М., 2015.
99. Rees, N. W., J. D. Riding. Automatic Concordance Creation for Texts in Any Language // Proceedings of Translating and the Computer. 2009.
100. Sadat, F. A combination of models for bilingual lexicon extraction from comparable corpora / Sadat, F., Dejean H., Gaussier E. // Proceedings of the Séminaire Papillon 2002, 2002.
101. Sinclair, J. Beginning the study of lexis // In memory of J.R. Firth. – London: Longman, 1966. С. 410-430.
102. Soderland, St. Compiling a massive, multilingual dictionary via probabilistic inference / Soderland, St., Etzioni O., Weld D., Skinner M., Bilmes J. // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1. – Association for Computational Linguistics, 2009. C. 262–70.
103. Stolcke, A. SRILM – An Extensible Language Modeling Toolkit. Proc. Intl. Conf. on Spoken Language Processing, vol. 2. – Denver, 2002. C. 901-904
104. Storjohann, P., Mohrs, Ch. Das elexiko-Wörterbuch im Wandel - Typische Verwendungen neu beleuchtet // Sprachreport 23.– Mannheim, 2007. C. 12-16
105. Tamura, A. Bilingual lexicon extraction from comparable corpora using label propagation / Tamura A, Watanabe T., Sumita E. // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. – Association for Computational Linguistics, 2012. C. 24–36.
106. Tomaszczyk, J. The bilingual dictionary under review // Zurilex’86 Proceedings: Papers Read at the Euralex International Congress. – University of Zurich, 1986. C. 289–97.
107. Tsunakawa, T. Building a Bilingual Lexicon Using Phrase-based Statistical Machine Translation via a Pivot Language / Tsunakawa T., Okazaki N., и Tsujii T. // COLING (Posters), 2008. C. 127–30.
108. Turcato, D. Automatically creating bilingual lexicons for machine translation from bilingual text // Proceedings of the 17th international conference on Computational linguistics-Volume 2. – Association for Computational Linguistics, 1998. C. 1299–1306.
109. Tyers, F. M., Pienaar J.A. Extracting bilingual word pairs from Wikipedia // Collaboration: interoperability between people in the creation of language resources for less-resourced languages 19, 2008. C. 19–22.
110. Uszkoreit, J. Large scale parallel document mining for machine translation / Uszkoreit J., Ponte J.M., Popat A.C., Dubiner M. // Proceedings of the 23rd International Conference on Computational Linguistics. – Association for Computational Linguistics, 2010. C. 1101–9.
111. Wang, M. A critical evaluation of bilingual Chinese/English dictionaries for elementary and intermediate Mandarin learners at Stellenbosch University. – University of Stellenbosch, 2012.
112. Wu, J.-C. TOTALrecall: A bilingual concordance for computer assisted translation and language learning / Wu J.-C., Yeh K.C., Chuang T.C., Shei W.-Ch., Chang J.S. // Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2. – Association for Computational Linguistics, 2003. C. 201–4.
113. Yamamoto, K., Matsumoto, Y. Acquisition of phrase-level bilingual correspondence using dependency structure // Proceedings of the 18th conference on Computational linguistics-Volume 2. – Association for Computational Linguistics, 2000. C. 933–39.
114. Yu, K., Tsujii J. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity // Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. – Association for Computational Linguistics, 2009. C. 121–24.

# Приложение А. Сравнение и оценка ранжирования иллюстрирующих контекстов с помощью разных методов

## Ранжирование контекстов на основе меры MI

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ключ en** | **ключ ru** | **пример en** | **пример ru** | **оценка** |
| area | зона | dining <area> | обеденная <зона> | 1 |
| area | местность | mountainous <area> | гористая <местность> | 1 |
| balance | балансирование |  |  |  |
| balance | балансировать | <balance>d market | балансирующий рынок | 1 |
| balance | балансировка |  |  |  |
| balance | балансовый | <balance> sheet value | балансовая стоимость | 1 |
| balance | вес |  |  |  |
| balance | весы | false <balance> | неверные <весы> | 1 |
| balance | остаток | unspent <balance> | неизрасходованный <остаток> | 1 |
| balance | противовес |  |  |  |
| balance | равновесие | emotional <balance> | душевное <равновесие> | 1 |
| balance | сбалансировать | well <balance>d diet | хорошо сбалансированная диета | 1 |
| capacity | качество | in various capacities | в различном качестве | 1 |
| capacity | мощность | nominal <capacity> | номинальная <мощность> | 1 |
| capacity | потенциал | statistical <capacity> building | наращивание статистического <потенциал>а | 0 |
| care | беречь | <care> of Russia | <беречь> Россию | 0 |
| care | внимательность | prenatal <care> | пренатальная <внимательность> | 0 |
| care | волновать |  |  |  |
| care | забота | fatherly <care> | отеческая <забота> | 0 |
| care | заботить |  |  |  |
| care | заботиться | really <care> | действительно <заботиться> | 1 |
| care | осторожность | utmost <care> | предельная <осторожность> | 1 |
| care | плевать |  |  |  |
| care | позаботиться | take <care> | <позаботиться> о том |  |
| carrier | переносчик |  |  |  |
| carrier | теплоноситель |  |  |  |
| carrier | транспортер | caterpillar <carrier> | гусеничный <транспортер> | 1 |
| character | признак | morphological <character>s | морфологические <признак>и | 0 |
| character | символ | unprintable <character>s | непечатные <символ>ы | 1 |
| community | общество | civil <community> | гражданское <общество> | 1 |
| concert | выступление | solo <concert> | сольное <выступление> | 1 |
| concert | концертировать | <concert> pianist | концертирующий пианист | 1 |
| concert | концертный | chamber <concert> hall | камерный <концертный> зал | 1 |
| containment | удержание | <containment> systems | система удержания | 0 |
| control | борьба | pest <control> | <борьба> с вредителями | 1 |
| control | диспетчерский | <control> tower | диспетчерская вышка | 1 |
| control | контролирование |  |  |  |
| control | контролировать | remotely <control> | дистанционно <контролировать> | 1 |
| control | надзор | fire <control> | пожарный <надзор> | 1 |
| control | проверка | audit <control> | аудиторская <проверка> | 1 |
| control | проконтролировать |  |  |  |
| control | регулирование | infinitely variable <control> | бесступенчатое <регулирование> | 1 |
| control | регулировать |  |  |  |
| control | регулировка | automatic volume <control> | автоматическая <регулировка> громкости | 1 |
| detention | лишение | arbitrary <detention> | произвольное <лишение> | 1 |
| detention | содержание | <detention> on remand | <содержание> под стражей | 1 |
| eventually | впоследствии | <eventually> become | <впоследствии> стать | 1 |
| eventually | наконец |  |  |  |
| eventually | окончательн |  |  |  |
| eventually | окончательно | <eventually> come | <окончательно> прийти | 0 |
| eventually | с |  |  |  |
| feast | банкет |  |  |  |
| feast | застолье | celebratory <feast> | праздничное <застолье> | 1 |
| feast | пир | wedding <feast> | брачный <пир> | 1 |
| feast | пировать |  |  |  |
| feast | пиршество | wedding <feast> | свадебное <пиршество> | 1 |
| feast | празднество |  |  |  |
| feast | праздник | <feast> of tabernacles | <праздник> кущей | 0 |
| guy | мужчина |  |  |  |
| guy | паренек |  |  |  |
| guy | парень | skinny <guy> | тощий <парень> | 1 |
| guy | парнишка |  |  |  |
| guy | пацан |  |  |  |
| guy | тип | type of <guy> | <тип> ванты | 0 |
| guy | человек | young <guy> | молодой <человек> | 1 |
| guy | чувак |  |  |  |
| history | краеведческий | regional <history> museum | <краеведческий> музей | 0 |
| history | хронология | file <history> | <хронология> файлов | 1 |
| however | вместе |  |  |  |
| however | же |  |  |  |
| however | ни | <however> strange | <ни> странно | 0 |
| however | но |  |  |  |
| however | однако |  |  |  |
| improve | повысить | dramatically <improve> | существенно <повысить> | 1 |
| improve | повыситься |  |  |  |
| improve | повышать | greatly <improve> | значительно <повышать> | 1 |
| improve | улучшаться | continually <improve> | постоянно <улучшаться> | 1 |
| improve | улучшиться | significantly <improve> | существенно <улучшиться> | 1 |
| industry | сфера | banking <industry> | банковская <сфера> | 1 |
| industry | хозяйство | agricultural <industry> | сельское <хозяйство> | 1 |
| insurance | страхование | compulsory pension <insurance> | обязательное пенсионное <страхование> | 1 |
| insurance | страхсбор | pet <insurance> | <страхсбор> любимчика | 0 |
| integration | включение | possible <integration> | возможное <включение> | 0 |
| integration | внедрение | easy <integration> | легкое <внедрение> | 0 |
| integration | встраивание |  |  |  |
| integration | интеграционный | European <integration> process | европейский <интеграционный> процесс | 1 |
| integration | интеграция | Eurasian economic <integration> | Евразийская экономическая <интеграция> | 1 |
| less | низко |  |  |  |
| less | поменьше |  |  |  |
| less | редко |  |  |  |
| manufacturer | промышленник | association of <manufacturer>s | ассоциация <промышленник>ов | 1 |
| manufacturer | товаропроизводитель | domestic <manufacturer> | отечественный <товаропроизводитель> | 1 |
| manufacturer | фабрикант | textile <manufacturer> | текстильный <фабрикант> | 0 |
| mode | вид | different transport <mode>s | различные <вид>ы транспорта | 1 |
| mode | время | real time <mode> | режим реального времени | 1 |
| mode | метод | safe <mode> | невредимый <метод> | 0 |
| mode | мода | waveguide <mode>s | волноводные моды | 0 |
| mode | способ | capitalist <mode> | капиталистический <способ> | 0 |
| report | донесение | intelligence <report> | разведывательное <донесение> | 1 |
| report | заключение | auditor's <report> | аудиторское <заключение> | 1 |
| report | отчет | stenographic <report> | стенографический <отчет> | 1 |
| report | рапортовать |  |  |  |
| report | репортаж | televised <report>s | телевизионные <репортаж>и | 1 |
| report | сводка | meteorological <report> | метеорологическая <сводка> | 1 |
| report | сообщать | promptly <report> | незамедлительно <сообщать> | 1 |
| report | сообщение | unconfirmed <report>s | неподтвержденные сообщения | 1 |
| save | экономия | toner <save> mode | режим экономии тонера | 0 |
| sitting | сидячий | <sitting> meditation | сидячая медитация | 1 |
| sitting | усаживание |  |  |  |
| size | величина | maximal <size> | максимальная <величина> | 1 |
| size | габарит | small <size> | малые <габарит>ы | 1 |
| size | крупность |  |  |  |
| size | масштаб |  |  |  |
| size | объем | total <size> | суммарный <объем> | 1 |
| size | площадь | actual <size> | реальная <площадь> | 1 |
| size | размерный | <size> range | <размерный> ряд | 1 |
| size | типоразмер | frame <size> | <типоразмер> корпуса | 1 |
| size | формат | custom <size> paper | бумага нестандартного <формат>а | 1 |
| small | малогабаритный | <small> apartment | малогабаритная квартира | 0 |
| small | маломерный | <small> vessels | маломерных судов | 0 |
| small | малочисленный | numerically <small> peoples | малочисленные народы | 1 |
| small | малый | microscopically <small> | микроскопически <малый> | 1 |
| sole | единственный | <sole> proprietorship | единственное право собственности | 1 |
| sure | точно | know for <sure> | <точно> знать | 1 |
| sure | убедиться | always make <sure> | всегда <убедиться> | 0 |
| sure | уверенн |  |  |  |
| sure | уверенный | not quite <sure> | не совсем <уверенный> | 1 |
| sure | удостовериться |  |  |  |
| threaten | грозить |  |  |  |
| together | воедино |  |  |  |
| together | сочетание |  |  |  |
| too | излишне |  |  |  |
| too | очень | <too> bad | <очень> жаль | 0 |
| too | чересчур | <too> risky | <чересчур> рискованный | 1 |
| too | чрезмерно | <too> high | <чрезмерно> высокий | 1 |
| too | чрезмерный | <too> much alcohol | чрезмерное употребление алкоголя | 0 |
| wheat | пшеница | <wheat> bran | отруби пшеницы | 1 |
| wheat | пшеничный | <wheat> bran | пшеничные отруби | 1 |
| wireless | беспроводный | <wireless> LAN | беспроводная локальная сеть | 1 |

## Ранжирование контекстов на основе частоты перевода

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ключ en** | **ключ ru** | **пример en** | **пример ru** | **оценка** |
| area | зона | dining <area> | обеденная <зона> | 1 |
| area | местность | rural <area> | сельская <местность> | 1 |
| balance | балансирование | careful <balance> | тщательное <балансирование> | 0 |
| balance | балансировать | <balanced> market | <балансирующий> рынок | 1 |
| balance | балансировка | load <balance> | <балансировка> нагрузки | 1 |
| balance | балансовый | <balance> sheet | <балансовая> ведомость | 1 |
| balance | вес | <balance> law | закон <весов> | 1 |
| balance | весы | analytical <balance> | аналитические <весы> | 1 |
| balance | остаток | unspent <balance> | неизрасходованный <остаток> | 1 |
| balance | противовес | additional <balances> | дополнительные <противовесы> | 0 |
| balance | равновесие | ecological <balance> | экологическое <равновесие> | 1 |
| balance | сбалансировать | <balanced> approach | <сбалансированный> подход | 0 |
| capacity | качество | business <capacity> | деловые <качества> | 1 |
| capacity | мощность | production <capacity> | производственная <мощность> | 1 |
| capacity | потенциал | national <capacity> | национальный <потенциал> | 0 |
| care | беречь | <care> cartridges | <беречь> патроны | 0 |
| care | внимательность | dental <care> | зубоврачебная <внимательность> | 0 |
| care | волновать | really <care> | особо <волновать> | 1 |
| care | забота | special <care> | особая <забота> | 1 |
| care | заботить | <care> more | <заботить> больше | 0 |
| care | заботиться | really <care> | действительно <заботиться> | 1 |
| care | осторожность | great <care> | большая <осторожность> | 1 |
| care | плевать | <care> deeply | глубоко <плевать> | 0 |
| care | позаботиться | best <care> | лучше <позаботиться> | 0 |
| carrier | переносчик | oxygen <carrier> | <переносчик> кислорода | 1 |
| carrier | теплоноситель | <carrier> temperature | температура <теплоносителя> | 0 |
| carrier | транспортер | horizontal <carrier> | горизонтальный <транспортер> | 0 |
| character | признак | sexual <character> | половой <признак> | 0 |
| character | символ | special <character> | специальный <символ> | 1 |
| community | общество | local <community> | местное <общество> | 1 |
| concert | выступление | live <concert> | живое <выступление> | 1 |
| concert | концертировать | <concert> organist | <концертирующий> органист | 1 |
| concert | концертный | <concert> hall | <концертный> зал | 1 |
| containment | удержание | <containment> systems | система <удержания> | 0 |
| control | борьба | biological <control> | биологическая <борьба> | 0 |
| control | диспетчерский | <control> center | <диспетчерский> центр | 1 |
| control | контролирование | <control> costs | <контролирование> расходов | 0 |
| control | контролировать | <controlled> substance | <контролируемое> вещество | 0 |
| control | надзор | state <control> | государственный <надзор> | 1 |
| control | проверка | internal <control> | внутренние <проверки> | 1 |
| control | проконтролировать | <control> flood | <проконтролировать> поток | 0 |
| control | регулирование | <control> measure | мера <регулирования> | 1 |
| control | регулировать | <controlled> substance | <регулируемое> вещество | 0 |
| control | регулировка | automatic <control> | автоматическая <регулировка> | 1 |
| detention | лишение | unlawful <detention> | незаконное <лишение> | 0 |
| detention | содержание | <detention> conditions | условия <содержания> | 1 |
| eventually | впоследствии | <eventually> become | <впоследствии> стать | 0 |
| eventually | наконец | lessen <eventually> | уменьшать <наконец> | 0 |
| eventually | окончательн | <eventually> lead | <окончательн> вести | 0 |
| eventually | окончательно | <eventually> become | <окончательно> стать | 0 |
| eventually | с | <eventually> defeat | <со> счетом | 0 |
| feast | банкет | wedding <feast> | свадебный <банкет> | 1 |
| feast | застолье | georgian <feast> | грузинское <застолье> | 0 |
| feast | пир | wedding <feast> | свадебный <пир> | 1 |
| feast | пировать | <feast> together | <пировать> вместе | 0 |
| feast | пиршество | great <feast> | большое <пиршество> | 1 |
| feast | празднество | winter <feast> | зимние <празднества> | 1 |
| feast | праздник | great <feast> | великий <праздник> | 1 |
| guy | мужчина | white <guy> | белый <мужчина> | 0 |
| guy | паренек | poor <guy> | бедный <паренек> | 1 |
| guy | парень | bad <guy> | плохой <парень> | 1 |
| guy | парнишка | little <guy> | маленький <парнишка> | 0 |
| guy | пацан | normal <guy> | нормальный <пацан> | 0 |
| guy | тип | bald <guy> | лысый <тип> | 1 |
| guy | человек | young <guy> | молодой <человек> | 1 |
| guy | чувак | cool <guy> | крутой <чувак> | 1 |
| history | краеведческий | <history> museum | <краеведческий> музей | 0 |
| history | хронология | job <history> | <хронология> задания | 1 |
| however | вместе | <however> great | <вместе> великий | 0 |
| however | же | <however> cold | <же> холод | 0 |
| however | ни | <however> strange | <ни> странно | 0 |
| however | но | <however> sooner | <но> рано | 0 |
| however | однако | <however> much | <однако> много | 0 |
| improve | повысить | significantly <improve> | значительно <повысить> | 1 |
| improve | повыситься | <improved> capacity | <повысившийся> потенциал | 0 |
| improve | повышать | significantly <improve> | значительно <повышать> | 1 |
| improve | улучшаться | gradually <improve> | постепенно <улучшаться> | 1 |
| improve | улучшиться | <improve> significantly | значительно <улучшиться> | 1 |
| industry | сфера | banking <industry> | банковская <сфера> | 1 |
| industry | хозяйство | fish <industry> | рыбное <хозяйство> | 0 |
| insurance | страхование | health <insurance> | медицинское <страхование> | 1 |
| insurance | страхсбор | auto <insurance> | автоматический <страхсбор> | 0 |
| integration | включение | possible <integration> | возможное <включение> | 0 |
| integration | внедрение | system <integration> | <внедрение> системы | 0 |
| integration | встраивание | perfect <integration> | идеальное <встраивание> | 0 |
| integration | интеграционный | <integration> process | <интеграционный> процесс | 1 |
| integration | интеграция | european <integration> | европейская <интеграция> | 1 |
| less | низко | be <less> | быть <ниже> | 1 |
| less | поменьше | <less> nervous | <поменьше> нервничать | 0 |
| less | редко | meet <less> | встречаться <реже> | 0 |
| manufacturer | промышленник | russian <manufacturer> | российский <промышленник> | 0 |
| manufacturer | товаропроизводитель | domestic <manufacturer> | отечественный <товаропроизводитель> | 1 |
| manufacturer | фабрикант | textile <manufacturer> | текстильный <фабрикант> | 0 |
| mode | вид | different <modes> | различные <виды> | 1 |
| mode | время | animation <mode> | шкала <времени> | 0 |
| mode | метод | safe <mode> | невредимый <метод> | 0 |
| mode | мода | wave <mode> | волновая <мода> | 0 |
| mode | способ | capitalist <mode> | капиталистический <способ> | 0 |
| report | донесение | intelligence <report> | разведывательное <донесение> | 1 |
| report | заключение | medical <report> | медицинское <заключение> | 1 |
| report | отчет | annual <report> | годовой <отчет> | 1 |
| report | рапортовать | regular <reports> | периодически <рапортуя> | 0 |
| report | репортаж | special <report> | специальный <репортаж> | 1 |
| report | сводка | weather <report> | <сводка> погоды | 1 |
| report | сообщать | also <report> | также <сообщать> | 0 |
| report | сообщение | official <report> | официальное <сообщение> | 1 |
| save | экономия | <save> time | <экономия> времени | 0 |
| sitting | сидячий | <sitting> position | <сидячее> положение | 1 |
| sitting | усаживание | <sitting> area | зона <усаживания> | 0 |
| size | величина | average <size> | средняя <величина> | 0 |
| size | габарит | small <size> | малый <габарит> | 1 |
| size | крупность | boundary <size> | граничная <крупность> | 0 |
| size | масштаб | small <size> | малый <масштаб> | 1 |
| size | объем | total <size> | общий <объем> | 1 |
| size | площадь | total <size> | общая <площадь> | 1 |
| size | размерный | <size> range | <размерный> ряд | 1 |
| size | типоразмер | different <sizes> | различных <типоразмеров> | 0 |
| size | формат | different <sizes> | различных <форматов> | 0 |
| small | малогабаритный | <small> apartment | <малогабаритная> квартира | 0 |
| small | маломерный | <small> vessels | <маломерных> судов | 0 |
| small | малочисленный | <small> people | <малочисленный> народ | 0 |
| small | малый | <small> business | <малый> бизнес | 1 |
| sole | единственный | <sole> purpose | <единственная> цель | 1 |
| sure | точно | entirely <sure> | сказать <точно> | 0 |
| sure | убедиться | make <sure> | сначала <убедиться> | 0 |
| sure | уверенн | <sure> pakistan | <уверенн> пакистан | 0 |
| sure | уверенный | absolutely <sure> | абсолютно <уверенный> | 1 |
| sure | удостовериться | make <sure> | окончательно <удостовериться> | 0 |
| threaten | грозить | <threaten> latvia | <грозить> латвии | 0 |
| together | воедино | bring <together> | свести <воедино> | 1 |
| together | сочетание | bad <together> | неплохое <сочетание> | 0 |
| too | излишне | <too> long | <излишне> длинный | 1 |
| too | очень | <too> often | <очень> часто | 1 |
| too | чересчур | be <too> | быть <чересчур> | 0 |
| too | чрезмерно | <too> high | <чрезмерно> высокий | 1 |
| too | чрезмерный | <too> busy | <чрезмерная> занятость | 0 |
| wheat | пшеница | winter <wheat> | озимая <пшеница> | 1 |
| wheat | пшеничный | <wheat> flour | <пшеничная> мука | 1 |
| wireless | беспроводный | <wireless> network | <беспроводная> сеть | 1 |

## Ранжирование

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ключ en** | **ключ ru** | **пример en** | **пример ru** | **оценка** |
| area | зона | priority <area> | приоритетная <зона> | 0 |
| area | местность | vast <area> | обширная <местность> | 1 |
| balance | балансирование | constant <balance> | постоянное <балансирование> | 0 |
| balance | балансировать | <balanced> document | <балансирующий> документ | 0 |
| balance | балансировка | exceptional <balance> | исключительная <балансировка> | 0 |
| balance | балансовый | <balance> part | <балансовая> часть | 1 |
| balance | вес | <balance> interface | интерфейс <весов> | 0 |
| balance | весы | family <balance> | семейные <весы> | 0 |
| balance | остаток | life <balance> | <остаток> жизни | 0 |
| balance | противовес | application <balances> | применение <противовесов> | 0 |
| balance | равновесие | optimal <balance> | оптимальное <равновесие> | 1 |
| balance | сбалансировать | <balanced> approach | <сбалансированный> подход | 0 |
| capacity | качество | structural <capacity> | структурные <качества> | 0 |
| capacity | мощность | total <capacity> | общая <мощность> | 1 |
| capacity | потенциал | total <capacity> | общий <потенциал> | 1 |
| care | беречь | <care> sister | <беречь> сестру | 0 |
| care | внимательность | palliative <care> | паллиативная <внимательность> | 0 |
| care | волновать | really <care> | особо <волновать> | 1 |
| care | забота | loving <care> | любящая <забота> | 0 |
| care | заботить | <care> more | <заботить> больше | 0 |
| care | заботиться | daily <care> | ежедневно <заботиться> | 0 |
| care | осторожность | maximum <care> | максимальная <осторожность> | 0 |
| care | плевать | even <care> | вообще <плевать> | 0 |
| care | позаботиться | daily <care> | <позаботиться> ежедневно | 0 |
| carrier | переносчик | information <carrier> | информационный <переносчик> | 0 |
| carrier | теплоноситель | transportation <carrier> | транспортировка <теплоносителя> | 0 |
| carrier | транспортер | tracked <carrier> | гусеничный <транспортер> | 1 |
| character | признак | main <character> | главный <признак> | 0 |
| character | символ | control <character> | управляющий <символ> | 1 |
| community | общество | international <community> | международное <общество> | 1 |
| concert | выступление | live <concert> | живое <выступление> | 1 |
| concert | концертировать | <concerts> russia | <концертировать> россией | 0 |
| concert | концертный | <concert> organization | <концертная> организация | 1 |
| containment | удержание | <containment> systems | система <удержания> | 0 |
| control | борьба | <control> programme | программа <борьбы> | 1 |
| control | диспетчерский | <control> panel | <диспетчерский> пульт | 1 |
| control | контролирование | <control> costs | <контролирование> расходов | 0 |
| control | контролировать | <control> system | <контролирующая> система | 1 |
| control | надзор | police <control> | полицейский <надзор> | 1 |
| control | проверка | <control> system | система <проверки> | 1 |
| control | проконтролировать | <control> situation | <проконтролировать> ситуацию | 1 |
| control | регулирование | <control> system | система <регулирования> | 1 |
| control | регулировать | <control> system | <регулирующая> система | 1 |
| control | регулировка | <control> system | система <регулировки> | 1 |
| detention | лишение | illegal <detention> | незаконное <лишение> | 1 |
| detention | содержание | illegal <detention> | незаконное <содержание> | 1 |
| eventually | впоследствии | <eventually> reach | <впоследствии> достичь | 0 |
| eventually | наконец | <eventually> see | увидеть <наконец> | 0 |
| eventually | окончательн | <eventually> lead | <окончательн> вести | 0 |
| eventually | окончательно | work <eventually> | работать <окончательно> | 0 |
| eventually | с | <eventually> defeat | <со> счетом | 0 |
| feast | банкет | greatest <feast> | большой <банкет> | 0 |
| feast | застолье | family <feast> | семейное <застолье> | 1 |
| feast | пир | mental <feast> | <пир> разума | 0 |
| feast | пировать | <feast> forever | <пировать> вечно | 0 |
| feast | пиршество | family <feast> | <пиршество> семьи | 1 |
| feast | празднество | big <feast> | большое <празднество> | 1 |
| feast | праздник | favorite <feast> | любимый <праздник> | 1 |
| guy | мужчина | big <guy> | крупный <мужчина> | 1 |
| guy | паренек | young <guy> | молодой <паренек> | 1 |
| guy | парень | big <guy> | здоровенный <парень> | 1 |
| guy | парнишка | small <guy> | маленький <парнишка> | 1 |
| guy | пацан | solid <guys> | солидные <пацаны> | 0 |
| guy | тип | big <guy> | важный <тип> | 1 |
| guy | человек | big <guy> | большой <человек> | 0 |
| guy | чувак | arab <guy> | арабский <чувак> | 0 |
| history | краеведческий | <history> museum | <краеведческий> музей | 0 |
| history | хронология | <history> data | данные <хронологии> | 0 |
| however | вместе | <however> great | <вместе> великий | 0 |
| however | же | <however> possible | <же> способы | 0 |
| however | ни | <however> far | <ни> далеко | 0 |
| however | но | <however> sooner | <но> рано | 0 |
| however | однако | <however> own | <однако> иметь | 0 |
| improve | повысить | significantly <improve> | значительно <повысить> | 1 |
| improve | повыситься | <improved> demand | <повысившийся> спрос | 1 |
| improve | повышать | significantly <improve> | значительно <повышать> | 1 |
| improve | улучшаться | continuously <improve> | постоянно <улучшаться> | 1 |
| improve | улучшиться | significantly <improve> | значительно <улучшиться> | 1 |
| industry | сфера | construction <industry> | строительная <сфера> | 1 |
| industry | хозяйство | agricultural <industry> | сельское <хозяйство> | 1 |
| insurance | страхование | <insurance> policy | политика <страхования> | 1 |
| insurance | страхсбор | <insurance> policy | <страхсбор> политика | 0 |
| integration | включение | psychological <integration> | психологическое <включение> | 0 |
| integration | внедрение | numerical <integration> | численное <внедрение> | 0 |
| integration | встраивание | russian <integration> | <встраивание> россии | 0 |
| integration | интеграционный | <integration> organization | <интеграционная> организация | 0 |
| integration | интеграция | economic <integration> | экономическая <интеграция> | 1 |
| less | низко | buy <less> | покупать <низко> | 0 |
| less | поменьше | pay <less> | платить <поменьше> | 0 |
| less | редко | pay <less> | платить <реже> | 0 |
| manufacturer | промышленник | russian <manufacturer> | российский <промышленник> | 0 |
| manufacturer | товаропроизводитель | domestic <manufacturer> | отечественный <товаропроизводитель> | 1 |
| manufacturer | фабрикант | russian <manufacturer> | русский <фабрикант> | 0 |
| mode | вид | control <mode> | <вид> управления | 1 |
| mode | время | day <mode> | дневное <время> | 0 |
| mode | метод | improved <mode> | усовершенствованный <метод> | 0 |
| mode | мода | standby <mode> | бай <мод> | 0 |
| mode | способ | proper <mode> | верный <способ> | 1 |
| report | донесение | police <report> | полицейское <донесение> | 1 |
| report | заключение | police <report> | <заключения> полиции | 1 |
| report | отчет | entitled <report> | озаглавленный <отчет> | 0 |
| report | рапортовать | regular <reports> | периодически <рапортуя> | 0 |
| report | репортаж | big <report> | большой <репортаж> | 0 |
| report | сводка | police <report> | полицейская <сводка> | 1 |
| report | сообщать | immediately <report> | немедленно <сообщать> | 1 |
| report | сообщение | immediately <report> | незамедлительное <сообщение> | 1 |
| save | экономия | <save> money | <экономия> средств | 0 |
| sitting | сидячий | <sitting> accommodation | <сидячие> места | 1 |
| sitting | усаживание | pretty <sitting> | милое <усаживание> | 0 |
| size | величина | great <size> | крупная <величина> | 0 |
| size | габарит | total <size> | суммарные <габариты> | 1 |
| size | крупность | maximum <size> | максимальная <крупность> | 0 |
| size | масштаб | physical <size> | физический <масштаб> | 1 |
| size | объем | engine <size> | рабочий <объем> | 1 |
| size | площадь | internal <size> | внутренняя <площадь> | 1 |
| size | размерный | <size> relation | <размерное> соотношение | 1 |
| size | типоразмер | large <sizes> | большие <типоразмеры> | 0 |
| size | формат | large <size> | большой <формат> | 1 |
| small | малогабаритный | <small> module | <малогабаритный> модуль | 1 |
| small | маломерный | <small> boats | <маломерные> суда | 1 |
| small | малочисленный | <small> organization | <малочисленная> организация | 1 |
| small | малый | <small> part | <малая> часть | 1 |
| sole | единственный | <sole> owner | <единственный> владелец | 1 |
| sure | точно | <sure> need | <точно> надо | 0 |
| sure | убедиться | <sure> arguments | <убедиться> рассуждениями | 0 |
| sure | уверенн | <sure> pakistan | <уверенн> пакистан | 0 |
| sure | уверенный | <sure> way | <уверенный> шаг | 0 |
| sure | удостовериться | <sure> first | сначала <удостовериться> | 0 |
| threaten | грозить | <threaten> america | <грозить> сша | 0 |
| together | воедино | put <together> | складывать <воедино> | 1 |
| together | сочетание | bring <together> | достичь <сочетания> | 0 |
| too | излишне | <too> fast | <излишне> голодать | 0 |
| too | очень | <too> far | <очень> далеко | 1 |
| too | чересчур | <too> fast | <чересчур> быстрый | 1 |
| too | чрезмерно | care <too> | <чрезмерно> опекать | 0 |
| too | чрезмерный | <too> fast | <чрезмерная> торопливость | 0 |
| wheat | пшеница | buy <wheat> | покупать <пшеницу> | 0 |
| wheat | пшеничный | <wheat> money | <пшеничные> деньги | 0 |
| wireless | беспроводный | <wireless> device | <беспроводное> устройство | 1 |

## Ранжирование

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ключ en** | **ключ ru** | **пример en** | **пример ru** | **оценка** |
| area | зона | beach <area> | пляжная <зона> | 1 |
| area | местность | rural <area> | сельская <местность> | 1 |
| balance | балансирование | careful <balance> | тщательное <балансирование> | 0 |
| balance | балансировать | <balanced> market | <балансирующий> рынок | 1 |
| balance | балансировка | <balance> quality | качество <балансировки> | 0 |
| balance | балансовый | <balance> sheet | <балансовая> ведомость | 1 |
| balance | вес | <balance> law | закон <весов> | 1 |
| balance | весы | analytical <balance> | аналитические <весы> | 1 |
| balance | остаток | account <balance> | <остаток> счета | 1 |
| balance | противовес | additional <balances> | дополнительные <противовесы> | 0 |
| balance | равновесие | ecological <balance> | экологическое <равновесие> | 1 |
| balance | сбалансировать | <balanced> food | <сбалансированное> питание | 1 |
| capacity | качество | moral <capacities> | моральные <качества> | 1 |
| capacity | мощность | full <capacity> | полная <мощность> | 1 |
| capacity | потенциал | trade <capacity> | торговый <потенциал> | 1 |
| care | беречь | <care> cartridges | <беречь> патроны | 0 |
| care | внимательность | hospital <care> | <внимательность> стационара | 0 |
| care | волновать | really <care> | особо <волновать> | 1 |
| care | забота | skin <care> | <забота> кожи | 0 |
| care | заботить | <care> more | <заботить> больше | 0 |
| care | заботиться | watchful <care> | чутко <заботиться> | 0 |
| care | осторожность | due <care> | должная <осторожность> | 1 |
| care | плевать | even <care> | вообще <плевать> | 0 |
| care | позаботиться | watchful <care> | хорошенько <позаботиться> | 0 |
| carrier | переносчик | healthy <carrier> | здоровый <переносчик> | 0 |
| carrier | теплоноситель | <carrier> temperature | температура <теплоносителя> | 0 |
| carrier | транспортер | caterpillar <carrier> | гусеничный <транспортер> | 1 |
| character | признак | morphological <characters> | морфологические <признаки> | 0 |
| character | символ | <character> class | класс <символов> | 1 |
| community | общество | kazakhstan <community> | казахстанское <общество> | 0 |
| concert | выступление | next <concert> | ближайшее <выступление> | 1 |
| concert | концертировать | <concert> organist | <концертирующий> органист | 1 |
| concert | концертный | <concert> program | <концертная> программа | 1 |
| containment | удержание | <containment> systems | система <удержания> | 0 |
| control | борьба | effective <control> | эффективная <борьба> | 1 |
| control | диспетчерский | <control> area | <диспетчерский> район | 0 |
| control | контролирование | erosion <control> | <контролирование> эрозии | 1 |
| control | контролировать | <controlled> substance | <контролируемое> вещество | 0 |
| control | надзор | technical <control> | технический <надзор> | 1 |
| control | проверка | customs <control> | таможенная <проверка> | 1 |
| control | проконтролировать | personally <control> | лично <проконтролировать> | 1 |
| control | регулирование | thermal <control> | термостатическое <регулирование> | 1 |
| control | регулировать | <controlled> substance | <регулируемое> вещество | 0 |
| control | регулировка | level <control> | <регулировка> уровня | 0 |
| detention | лишение | illegal <detention> | незаконное <лишение> | 1 |
| detention | содержание | illegal <detention> | незаконное <содержание> | 1 |
| eventually | впоследствии | <eventually> use | <впоследствии> использовать | 0 |
| eventually | наконец | arrive <eventually> | <наконец> дойти | 0 |
| eventually | окончательн | <eventually> lead | <окончательн> вести | 0 |
| eventually | окончательно | <eventually> reach | <окончательно> достигнуть | 1 |
| eventually | с | <eventually> defeat | <со> счетом | 0 |
| feast | банкет | wedding <feast> | свадебный <банкет> | 1 |
| feast | застолье | wedding <feast> | свадебное <застолье> | 1 |
| feast | пир | wedding <feast> | свадебный <пир> | 1 |
| feast | пировать | <feasting> priests | <пирующие> жрецы | 0 |
| feast | пиршество | real <feast> | настоящее <пиршество> | 1 |
| feast | празднество | winter <feast> | зимние <празднества> | 0 |
| feast | праздник | church <feast> | церковный <праздник> | 1 |
| guy | мужчина | white <guy> | белый <мужчина> | 0 |
| guy | паренек | italian <guy> | итальянский <паренек> | 0 |
| guy | парень | big <guy> | большой <парень> | 1 |
| guy | парнишка | nice <guy> | милый <парнишка> | 1 |
| guy | пацан | normal <guy> | нормальный <пацан> | 0 |
| guy | тип | big <guy> | важный <тип> | 1 |
| guy | человек | bad <guy> | плохой <человек> | 1 |
| guy | чувак | normal <guy> | нормальный <чувак> | 0 |
| history | краеведческий | <history> department | <краеведческий> отдел | 0 |
| history | хронология | <history> file | файл <хронологии> | 0 |
| however | вместе | <however> great | <вместе> великий | 0 |
| however | же | <however> cold | <же> холод | 0 |
| however | ни | <however> strange | <ни> странно | 0 |
| however | но | <however> sooner | <но> рано | 0 |
| however | однако | <however> much | <однако> много | 0 |
| improve | повысить | significantly <improve> | существенно <повысить> | 1 |
| improve | повыситься | <improve> moderately | умеренно <повыситься> | 1 |
| improve | повышать | <improve> efficiency | <повышать> эффективность | 1 |
| improve | улучшаться | also <improve> | также <улучшаться> | 0 |
| improve | улучшиться | significantly <improve> | значительно <улучшиться> | 1 |
| industry | сфера | banking <industry> | банковская <сфера> | 1 |
| industry | хозяйство | water <industry> | водное <хозяйство> | 0 |
| insurance | страхование | <insurance> money | <страхование> деньги | 0 |
| insurance | страхсбор | pet <insurance> | <страхсбор> любимчика | 0 |
| integration | включение | progressive <integration> | постепенное <включение> | 0 |
| integration | внедрение | easy <integration> | легкое <внедрение> | 0 |
| integration | встраивание | easy <integration> | простое <встраивание> | 0 |
| integration | интеграционный | <integration> initiative | <интеграционная> инициатива | 0 |
| integration | интеграция | regional <integration> | региональная <интеграция> | 1 |
| less | низко | <less> anticipations | <ниже> ожиданий | 0 |
| less | поменьше | drink <less> | пить <поменьше> | 1 |
| less | редко | appear <less> | появляться <реже> | 0 |
| manufacturer | промышленник | domestic <manufacturers> | отечественные <промышленники> | 1 |
| manufacturer | товаропроизводитель | russian <manufacturers> | российский <товаропроизводитель> | 0 |
| manufacturer | фабрикант | textile <manufacturer> | текстильный <фабрикант> | 0 |
| mode | вид | alternative <mode> | альтернативный <вид> | 1 |
| mode | время | animation <mode> | шкала <времени> | 0 |
| mode | метод | treatment <mode> | <метод> лечения | 1 |
| mode | мода | wave <mode> | волновая <мода> | 0 |
| mode | способ | alternative <mode> | альтернативный <способ> | 1 |
| report | донесение | combat <reports> | боевые <донесения> | 1 |
| report | заключение | audit <report> | аудиторское <заключение> | 1 |
| report | отчет | final <report> | итоговый <отчет> | 1 |
| report | рапортовать | regular <reports> | периодически <рапортуя> | 0 |
| report | репортаж | exclusive <report> | эксклюзивный <репортаж> | 1 |
| report | сводка | weather <report> | <сводка> погоды | 1 |
| report | сообщать | also <report> | также <сообщать> | 0 |
| report | сообщение | media <report> | <сообщение> сми | 1 |
| save | экономия | <save> money | <экономия> средств | 0 |
| sitting | сидячий | <sitting> position | <сидячее> положение | 1 |
| sitting | усаживание | <sitting> position | положение <усаживания> | 0 |
| size | величина | average <size> | средняя <величина> | 0 |
| size | габарит | small <size> | небольшой <габарит> | 1 |
| size | крупность | feed <size> | начальная <крупность> | 0 |
| size | масштаб | project <size> | <масштаб> проекта | 1 |
| size | объем | total <size> | общий <объем> | 1 |
| size | площадь | total <size> | общая <площадь> | 1 |
| size | размерный | <size> group | <размерная> группа | 1 |
| size | типоразмер | valve <size> | <типоразмер> клапана | 1 |
| size | формат | page <size> | <формат> страницы | 1 |
| small | малогабаритный | <small> apartment | <малогабаритная> квартира | 0 |
| small | маломерный | <small> boats | <маломерные> суда | 1 |
| small | малочисленный | <small> nation | <малочисленный> народ | 0 |
| small | малый | <small> business | <малый> бизнес | 1 |
| sole | единственный | <sole> source | <единственный> источник | 1 |
| sure | точно | entirely <sure> | сказать <точно> | 0 |
| sure | убедиться | make <sure> | сначала <убедиться> | 0 |
| sure | уверенн | <sure> pakistan | <уверенн> пакистан | 0 |
| sure | уверенный | absolutely <sure> | абсолютно <уверенный> | 1 |
| sure | удостовериться | make <sure> | окончательно <удостовериться> | 0 |
| threaten | грозить | <threaten> latvia | <грозить> латвии | 0 |
| together | воедино | come <together> | сойтись <воедино> | 1 |
| together | сочетание | bring <together> | достичь <сочетания> | 0 |
| too | излишне | <too> restrictive | <излишне> ограничительный | 0 |
| too | очень | ask <too> | требовать <очень> | 0 |
| too | чересчур | <too> active | <чересчур> активный | 1 |
| too | чрезмерно | <too> high | <чрезмерно> высокий | 1 |
| too | чрезмерный | <too> busy | <чрезмерная> занятость | 0 |
| wheat | пшеница | winter <wheat> | озимая <пшеница> | 1 |
| wheat | пшеничный | <wheat> field | <пшеничное> поле | 1 |
| wireless | беспроводный | <wireless> keyboard | <беспроводная> клавиатура | 1 |

## Ранжирование

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ключ en** | **ключ ru** | **пример en** | **пример ru** | **оценка** |
| area | зона | conflict <area> | <зона> конфликта | 1 |
| area | местность | mountainous <area> | гористая <местность> | 1 |
| balance | балансирование | constant <balance> | постоянное <балансирование> | 0 |
| balance | балансировать | deftly <balance> | ловко <балансировать> | 1 |
| balance | балансировка | load <balance> | <балансировка> нагрузки | 1 |
| balance | балансовый | <balance> account | <балансовый> счет | 1 |
| balance | вес | <balance> sensitivity | чувствительность <весов> | 1 |
| balance | весы | electronic <balance> | электронные <весы> | 1 |
| balance | остаток | account <balance> | <остаток> счета | 1 |
| balance | противовес | additional <balances> | дополнительные <противовесы> | 0 |
| balance | равновесие | delicate <balance> | хрупкое <равновесие> | 1 |
| balance | сбалансировать | <balanced> budget | <сбалансированный> бюджет | 1 |
| capacity | качество | sea <capacity> | мореходные <качества> | 0 |
| capacity | мощность | maximum <capacity> | максимальная <мощность> | 1 |
| capacity | потенциал | production <capacity> | производственный <потенциал> | 1 |
| care | беречь | special <care> | особенно <беречь> | 0 |
| care | внимательность | <care> tips | концы <внимательности> | 0 |
| care | волновать | really <care> | особо <волновать> | 1 |
| care | забота | special <care> | особая <забота> | 1 |
| care | заботить | <care> more | <заботить> больше | 0 |
| care | заботиться | particularly <care> | особенно <заботиться> | 1 |
| care | осторожность | special <care> | особая <осторожность> | 1 |
| care | плевать | <care> deeply | глубоко <плевать> | 0 |
| care | позаботиться | best <care> | лучше <позаботиться> | 0 |
| carrier | переносчик | energy <carrier> | <переносчик> энергии | 1 |
| carrier | теплоноситель | <carrier> temperature | температура <теплоносителя> | 0 |
| carrier | транспортер | tracked <carrier> | гусеничный <транспортер> | 1 |
| character | признак | morphological <characters> | морфологические <признаки> | 0 |
| character | символ | special <character> | специальный <символ> | 1 |
| community | общество | human <community> | человеческое <общество> | 1 |
| concert | выступление | live <concert> | живое <выступление> | 1 |
| concert | концертировать | <concert> organist | <концертирующий> органист | 1 |
| concert | концертный | <concert> tour | <концертный> тур | 1 |
| containment | удержание | <containment> systems | система <удержания> | 0 |
| control | борьба | rigid <control> | жесткая <борьба> | 1 |
| control | диспетчерский | <control> point | <диспетчерский> пункт | 1 |
| control | контролирование | obstacle <control> | <контролирование> препятствий | 1 |
| control | контролировать | <controlled> study | <контролируемое> исследование | 0 |
| control | надзор | fire <control> | пожарный <надзор> | 1 |
| control | проверка | fiscal <controls> | налоговые <проверки> | 1 |
| control | проконтролировать | <control> situation | <проконтролировать> ситуацию | 1 |
| control | регулирование | temperature <control> | <регулирование> температуры | 1 |
| control | регулировать | automatically <control> | автоматически <регулировать> | 1 |
| control | регулировка | level <control> | <регулировка> уровня | 0 |
| detention | лишение | unlawful <detention> | незаконное <лишение> | 0 |
| detention | содержание | <detention> conditions | условия <содержания> | 1 |
| eventually | впоследствии | <eventually> become | <впоследствии> стать | 0 |
| eventually | наконец | <eventually> turn | <наконец> обратить | 0 |
| eventually | окончательн | <eventually> lead | <окончательн> вести | 0 |
| eventually | окончательно | <eventually> come | <окончательно> прийти | 0 |
| eventually | с | <eventually> defeat | <со> счетом | 0 |
| feast | банкет | wedding <feast> | свадебный <банкет> | 1 |
| feast | застолье | abundant <feast> | обильное <застолье> | 1 |
| feast | пир | marriage <feast> | брачный <пир> | 1 |
| feast | пировать | <feasting> priests | <пирующие> жрецы | 0 |
| feast | пиршество | grandiose <feast> | грандиозное <пиршество> | 1 |
| feast | празднество | sumptuous <feast> | пышное <празднество> | 1 |
| feast | праздник | great <feast> | великий <праздник> | 1 |
| guy | мужчина | burly <guy> | дородный <мужчина> | 1 |
| guy | паренек | puny <guy> | маленький <паренек> | 1 |
| guy | парень | bad <guy> | плохой <парень> | 1 |
| guy | парнишка | funny <guy> | веселый <парнишка> | 1 |
| guy | пацан | normal <guy> | нормальный <пацан> | 0 |
| guy | тип | nasty <guy> | мерзкий <тип> | 1 |
| guy | человек | nice <guy> | порядочный <человек> | 0 |
| guy | чувак | coolest <guy> | прикольный <чувак> | 1 |
| history | краеведческий | <history> museum | <краеведческий> музей | 0 |
| history | хронология | brief <history> | краткая <хронология> | 1 |
| however | вместе | <however> great | <вместе> великий | 0 |
| however | же | <however> also | <же> тоже | 0 |
| however | ни | <however> regrettable | <ни> прискорбно | 0 |
| however | но | <however> sooner | <но> рано | 0 |
| however | однако | <however> there | <однако> есть | 0 |
| improve | повысить | <improve> performance | <повысить> производительность | 1 |
| improve | повыситься | <improved> demand | <повысившийся> спрос | 1 |
| improve | повышать | <improve> performance | <повышать> производительность | 1 |
| improve | улучшаться | gradually <improve> | постепенно <улучшаться> | 1 |
| improve | улучшиться | <improve> considerably | значительно <улучшиться> | 1 |
| industry | сфера | energy <industry> | энергетическая <сфера> | 1 |
| industry | хозяйство | fishing <industry> | рыбное <хозяйство> | 0 |
| insurance | страхование | personal <insurance> | личное <страхование> | 1 |
| insurance | страхсбор | art <insurance> | <страхсбор> искусства | 0 |
| integration | включение | full <integration> | полноценное <включение> | 0 |
| integration | внедрение | <integration> strategy | стратегия <внедрения> | 0 |
| integration | встраивание | easy <integration> | простое <встраивание> | 0 |
| integration | интеграционный | <integration> process | <интеграционный> процесс | 1 |
| integration | интеграция | successful <integration> | успешная <интеграция> | 1 |
| less | низко | <less> zero | <ниже> нуля | 0 |
| less | поменьше | talk <less> | <поменьше> говорить | 0 |
| less | редко | occur <less> | происходить <реже> | 0 |
| manufacturer | промышленник | major <manufacturer> | крупный <промышленник> | 1 |
| manufacturer | товаропроизводитель | domestic <manufacturer> | отечественный <товаропроизводитель> | 1 |
| manufacturer | фабрикант | french <manufacturer> | французский <фабрикант> | 0 |
| mode | вид | alternative <mode> | альтернативный <вид> | 1 |
| mode | время | zoom <mode> | шкала <времени> | 0 |
| mode | метод | safe <mode> | невредимый <метод> | 0 |
| mode | мода | demo <mode> | демонстрация <мод> | 0 |
| mode | способ | manual <mode> | ручной <способ> | 1 |
| report | донесение | battle <reports> | боевые <донесения> | 1 |
| report | заключение | medical <report> | медицинское <заключение> | 1 |
| report | отчет | view <reports> | просмотр <отчетов> | 0 |
| report | рапортовать | regular <reports> | периодически <рапортуя> | 0 |
| report | репортаж | special <report> | специальный <репортаж> | 1 |
| report | сводка | weather <report> | <сводка> погоды | 1 |
| report | сообщать | immediately <report> | незамедлительно <сообщать> | 1 |
| report | сообщение | official <report> | официальное <сообщение> | 1 |
| save | экономия | <save> energy | <экономия> энергии | 0 |
| sitting | сидячий | <sitting> position | <сидячее> положение | 1 |
| sitting | усаживание | done <sitting> | сделанное <усаживание> | 0 |
| size | величина | physical <size> | физическая <величина> | 1 |
| size | габарит | maximum <size> | максимальные <габариты> | 1 |
| size | крупность | maximum <size> | максимальная <крупность> | 0 |
| size | масштаб | immense <size> | огромные <масштабы> | 1 |
| size | объем | maximum <size> | максимальный <объем> | 1 |
| size | площадь | actual <size> | жилая <площадь> | 1 |
| size | размерный | <size> range | <размерный> ряд | 1 |
| size | типоразмер | different <sizes> | различных <типоразмеров> | 0 |
| size | формат | standard <size> | стандартный <формат> | 1 |
| small | малогабаритный | <small> apartment | <малогабаритная> квартира | 0 |
| small | маломерный | <small> vessels | <маломерных> судов | 0 |
| small | малочисленный | <small> nation | <малочисленный> народ | 0 |
| small | малый | <small> weight | <малый> вес | 1 |
| sole | единственный | <sole> reason | <единственная> причина | 1 |
| sure | точно | <sure> need | <точно> надо | 0 |
| sure | убедиться | quite <sure> | окончательно <убедиться> | 0 |
| sure | уверенн | <sure> pakistan | <уверенн> пакистан | 0 |
| sure | уверенный | absolutely <sure> | абсолютно <уверенный> | 1 |
| sure | удостовериться | quite <sure> | окончательно <удостовериться> | 0 |
| threaten | грозить | <threaten> israel | <грозить> израилю | 0 |
| together | воедино | gather <together> | собрать <воедино> | 1 |
| together | сочетание | bad <together> | неплохое <сочетание> | 0 |
| too | излишне | <too> impressionable | <излишне> впечатлительный | 1 |
| too | очень | <too> weak | <очень> слабый | 1 |
| too | чересчур | prove <too> | оказаться <чересчур> | 0 |
| too | чрезмерно | <too> high | <чрезмерно> высокий | 1 |
| too | чрезмерный | <too> emotional | <чрезмерная> эмоциональность | 0 |
| wheat | пшеница | <wheat> variety | сорт <пшеницы> | 1 |
| wheat | пшеничный | <wheat> beer | <пшеничное> пиво | 1 |
| wireless | беспроводный | <wireless> connection | <беспроводное> соединение | 1 |

## Ранжирование

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ключ en** | **ключ ru** | **пример en** | **пример кг** | **оценка** |
| area | зона | geographic <area> | географическая <зона> | 1 |
| area | местность | scenic <area> | живописная <местность> | 1 |
| balance | балансирование | constant <balance> | постоянное <балансирование> | 0 |
| balance | балансировать | <balanced> market | <балансирующий> рынок | 1 |
| balance | балансировка | dynamic <balance> | динамическая <балансировка> | 0 |
| balance | балансовый | <balance> account | <балансовый> счет | 1 |
| balance | вес | <balance> calibration | поверка <весов> | 1 |
| balance | весы | analytical <balance> | аналитические <весы> | 1 |
| balance | остаток | outstanding <balance> | непогашенный <остаток> | 1 |
| balance | противовес | additional <balances> | дополнительные <противовесы> | 0 |
| balance | равновесие | ecological <balance> | экологическое <равновесие> | 1 |
| balance | сбалансировать | <balanced> approach | <сбалансированный> подход | 0 |
| capacity | качество | moral <capacity> | нравственные <качества> | 1 |
| capacity | мощность | maximum <capacity> | максимальная <мощность> | 0 |
| capacity | потенциал | strengthening <capacity> | укрепление <потенциала> | 0 |
| care | беречь | <care> cartridges | <беречь> патроны | 0 |
| care | внимательность | <cares> foundation | учредительство <внимательностей> | 0 |
| care | волновать | really <care> | особо <волновать> | 1 |
| care | забота | maternal <care> | материнская <забота> | 1 |
| care | заботить | <care> more | <заботить> больше | 0 |
| care | заботиться | special <care> | особо <заботиться> | 1 |
| care | осторожность | special <care> | особая <осторожность> | 1 |
| care | плевать | <care> deeply | глубоко <плевать> | 0 |
| care | позаботиться | best <care> | лучше <позаботиться> | 0 |
| carrier | переносчик | glucose <carrier> | <переносчик> глюкозы | 1 |
| carrier | теплоноситель | <carrier> temperature | температура <теплоносителя> | 0 |
| carrier | транспортер | tracked <carrier> | гусеничный <транспортер> | 1 |
| character | признак | distinctive <character> | отличительный <признак> | 1 |
| character | символ | control <character> | управляющий <символ> | 1 |
| community | общество | civil <community> | гражданское <общество> | 1 |
| concert | выступление | live <concert> | живое <выступление> | 1 |
| concert | концертировать | <concert> organist | <концертирующий> органист | 1 |
| concert | концертный | <concert> venue | <концертная> площадка | 1 |
| containment | удержание | reliable <containment> | надежное <удержание> | 1 |
| control | борьба | <control> tactics | тактика <борьбы> | 1 |
| control | диспетчерский | <control> center | <диспетчерский> центр | 1 |
| control | контролирование | regular <control> | регулярное <контролирование> | 1 |
| control | контролировать | <controlled> trial | <контролируемое> исследование | 0 |
| control | надзор | veterinary <control> | ветеринарный <надзор> | 1 |
| control | проверка | thorough <control> | тщательная <проверка> | 1 |
| control | проконтролировать | <control> situation | <проконтролировать> ситуацию | 1 |
| control | регулирование | <control> system | система <регулирования> | 1 |
| control | регулировать | <control> valve | <регулирующая> арматура | 0 |
| control | регулировка | <control> pressure | <регулировка> давления | 0 |
| detention | лишение | illegal <detention> | незаконное <лишение> | 1 |
| detention | содержание | illegal <detention> | незаконное <содержание> | 1 |
| eventually | впоследствии | <eventually> become | <впоследствии> стать | 0 |
| eventually | наконец | <eventually> turn | <наконец> обратить | 0 |
| eventually | окончательн | <eventually> lead | <окончательн> вести | 0 |
| eventually | окончательно | <eventually> destroy | <окончательно> погубить | 1 |
| eventually | с | <eventually> defeat | <со> счетом | 0 |
| feast | банкет | wedding <feast> | свадебный <банкет> | 1 |
| feast | застолье | wedding <feast> | свадебное <застолье> | 1 |
| feast | пир | marriage <feast> | брачный <пир> | 1 |
| feast | пировать | <feast> forever | <пировать> вечно | 0 |
| feast | пиршество | wedding <feast> | свадебное <пиршество> | 1 |
| feast | празднество | sumptuous <feast> | пышное <празднество> | 1 |
| feast | праздник | religious <feast> | религиозный <праздник> | 1 |
| guy | мужчина | old <guy> | пожилой <мужчина> | 1 |
| guy | паренек | young <guy> | молодой <паренек> | 1 |
| guy | парень | big <guy> | здоровенный <парень> | 1 |
| guy | парнишка | young <guy> | молодой <парнишка> | 1 |
| guy | пацан | tough <guy> | крутой <пацан> | 1 |
| guy | тип | nasty <guy> | мерзкий <тип> | 1 |
| guy | человек | smart <guy> | умный <человек> | 1 |
| guy | чувак | cool <guy> | крутой <чувак> | 1 |
| history | краеведческий | <history> museum | <краеведческий> музей | 0 |
| history | хронология | brief <history> | краткая <хронология> | 1 |
| however | вместе | <however> great | <вместе> великий | 0 |
| however | же | provide <however> | <же> предоставить | 0 |
| however | ни | <however> regrettable | <ни> прискорбно | 0 |
| however | но | <however> sooner | <но> рано | 0 |
| however | однако | <however> own | <однако> иметь | 0 |
| improve | повысить | significantly <improve> | значительно <повысить> | 1 |
| improve | повыситься | <improved> demand | <повысившийся> спрос | 1 |
| improve | повышать | significantly <improve> | значительно <повышать> | 1 |
| improve | улучшаться | gradually <improve> | постепенно <улучшаться> | 1 |
| improve | улучшиться | significantly <improve> | значительно <улучшиться> | 1 |
| industry | сфера | banking <industry> | банковская <сфера> | 1 |
| industry | хозяйство | agricultural <industry> | сельское <хозяйство> | 1 |
| insurance | страхование | health <insurance> | <страхование> здоровья | 1 |
| insurance | страхсбор | personal <insurance> | приватный <страхсбор> | 0 |
| integration | включение | smooth <integration> | плавное <включение> | 0 |
| integration | внедрение | gradual <integration> | постепенное <внедрение> | 1 |
| integration | встраивание | easy <integration> | простое <встраивание> | 0 |
| integration | интеграционный | <integration> process | <интеграционный> процесс | 1 |
| integration | интеграция | <integration> process | процесс <интеграции> | 1 |
| less | низко | <less> zero | <ниже> нуля | 0 |
| less | поменьше | <less> smoke | <поменьше> надо | 0 |
| less | редко | <less> common | <редко> встречающийся | 0 |
| manufacturer | промышленник | russian <manufacturer> | российский <промышленник> | 0 |
| manufacturer | товаропроизводитель | domestic <manufacturer> | отечественный <товаропроизводитель> | 1 |
| manufacturer | фабрикант | wealthy <manufacturer> | богатый <фабрикант> | 1 |
| mode | вид | widespread <mode> | распространенный <вид> | 0 |
| mode | время | day <mode> | дневное <время> | 0 |
| mode | метод | safe <mode> | невредимый <метод> | 0 |
| mode | мода | high <mode> | высокая <мода> | 0 |
| mode | способ | capitalist <mode> | капиталистический <способ> | 0 |
| report | донесение | investigation <reports> | <донесения> разведки | 1 |
| report | заключение | audit <report> | аудиторское <заключение> | 1 |
| report | отчет | <report> preparation | составление <отчетов> | 1 |
| report | рапортовать | regular <reports> | периодически <рапортуя> | 0 |
| report | репортаж | special <report> | специальный <репортаж> | 1 |
| report | сводка | police <reports> | милицейские <сводки> | 1 |
| report | сообщать | immediately <report> | немедленно <сообщать> | 1 |
| report | сообщение | official <report> | официальное <сообщение> | 1 |
| save | экономия | <save> money | <экономия> денег | 0 |
| sitting | сидячий | <sitting> position | <сидячее> положение | 1 |
| sitting | усаживание | baby <sitting> | <усаживание> младенца | 0 |
| size | величина | physical <size> | физическая <величина> | 1 |
| size | габарит | compact <size> | компактные <габариты> | 1 |
| size | крупность | maximum <size> | максимальная <крупность> | 0 |
| size | масштаб | enormous <size> | огромные <масштабы> | 1 |
| size | объем | total <size> | суммарный <объем> | 1 |
| size | площадь | actual <size> | жилая <площадь> | 1 |
| size | размерный | <size> range | <размерный> ряд | 1 |
| size | типоразмер | <size> range | ассортимент <типоразмеров> | 0 |
| size | формат | custom <size> | нестандартный <формат> | 1 |
| small | малогабаритный | <small> apartment | <малогабаритная> квартира | 0 |
| small | маломерный | <small> vessels | <маломерное> судно | 0 |
| small | малочисленный | <small> nation | <малочисленный> народ | 0 |
| small | малый | <small> part | <малая> толика | 1 |
| sole | единственный | <sole> shareholder | <единственный> акционер | 1 |
| sure | точно | <sure> indication | <точно> определить | 0 |
| sure | убедиться | <sure> beforehand | <убедившись> предварительно | 0 |
| sure | уверенн | <sure> pakistan | <уверенн> пакистан | 0 |
| sure | уверенный | <sure> step | <уверенный> шаг | 0 |
| sure | удостовериться | <sure> first | сначала <удостовериться> | 0 |
| threaten | грозить | <threaten> israel | <грозить> израилю | 0 |
| together | воедино | merge <together> | слиться <воедино> | 1 |
| together | сочетание | bad <together> | неплохое <сочетание> | 0 |
| too | излишне | speak <too> | говорить <излишне> | 0 |
| too | очень | <too> loud | <очень> шумный | 1 |
| too | чересчур | <too> addicted | <чересчур> увлекаться | 0 |
| too | чрезмерно | worry <too> | <чрезмерно> беспокоиться | 0 |
| too | чрезмерный | <too> optimistic | <чрезмерный> оптимизм | 0 |
| wheat | пшеница | food <wheat> | продовольственная <пшеница> | 1 |
| wheat | пшеничный | <wheat> bran | <пшеничные> отруби | 1 |
| wireless | беспроводный | <wireless> device | <беспроводное> устройство | 1 |

1. <http://cfrl.ru/outline.shtml> [↑](#footnote-ref-1)
2. см. описание словаря:

   <http://www.lingvo-online.ru/ru/LingvoDictionaries/Details?dictionary=LingvoUniversal%20(En-Ru>) [↑](#footnote-ref-2)
3. <http://otipl.philol.msu.ru/~soiza/testsynt/res00/duo.php> [↑](#footnote-ref-3)