Сырчина Анна Сергеевна
Тюменский государственный университет
Институт математики и компьютерных наук
Кафедра иностранных языков и межкультурной профессиональной коммуникации естественнонаучных направлений
Ассистент
bigglosha@mail.ru

ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ МЕТОДОВ КОРПУСНОЙ ЛИНГВИСТИКИ ПРИ СОСТАВЛЕНИИ АВТОРСКОГО ГЛОССАРИЯ
CORPUS LINGUISTICS AND CORPUS-BASED APPROACHES IN COMPILING THE AUTHOR GLOSSARY

АННОТАЦИЯ. Корпусная лингвистика – это относительно молодое и активно развивающееся направление в рамках компьютерной лингвистики. Целью данной статьи является определение и выявление методов корпусной лингвистики, а также практическое использование программы Word List Creator и AntConc для составления авторского глоссария. В статье рассмотрены такие понятия, как лингвистический корпус (совокупность текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой) и корпусная лингвистика (раздел языкознания, занимающийся разработкой, созданием и использованием текстовых корпусов).

ABSTRACT. Corpus linguistics is a relatively young and rapidly developing trend within computational linguistics. The purposes of this article are identification and detection of corpus-based approaches and practical use of the program Word List Creator and AntConc in compiling the author glossary. The article considers such notions as linguistic corpus (systematic collection of naturally occurring texts

providing specialized search system) and corpus linguistics (branch of linguistics dealing with eleboration, development and use of text corpora)

КЛЮЧЕВЫЕ СЛОВА: корпусная лингвистика, лингвистический корпус, глоссарий.

KEY WORDS: corpus linguistics, corpus-based approach, linguistic corpus, glossary.

Прежде, чем говорить о корпусной лингвистике, необходимо определить само понятие лингвистического корпуса. По-английски это будет **linguistic corpus** или **text corpus (linguistic corpora).** Существует довольно много определений, которые сходятся в одном: корпус есть «некоторый филологический объект» [2]. Пожалуй, наиболее полное определение дает Джон Синклер. Корпус - собрание отрывков текстов в электронной форме, отобранных в соответствии с внешними критериями, чтобы наиболее полно представлять язык или вариацию языка. Функционирует как источник данных для лингвистических исследований. [5]

В качестве примеров корпусов можно привести тексты конкретного писателя или писателей; тексты за конкретное десятилетие или столетие; современные тексты определённой тематики; современные тексты, адекватно представляющие язык или общество.

Понятно, что корпус — это набор текстов, с которыми можно что-то делать. Но что же может делать корпус? Ответ может показаться неожиданным: сам корпус не может делать ничего. Но мы можем использовать специальное программное обеспечение, чтобы искать в корпусе что-либо и производить некоторые вычисления. Что же мы можем искать? В первую очередь, это слова и фразы, которые имеют культурную или лингвистическую значимость.

Таким образом, с помощью корпусов можно изучать самые разные языковые явления. Примеры возможных запросов к текстовой базе данных приводит В.М. Андрющенко. Вот некоторые из них:

– Каковы все (или наиболее типичные) контексты употребления слова (конструкции, словосочетания, явления)?

– Выдать весь словарь определенного автора или определенной системы.

– Собрать из текстов все ситуации определенной структуры и т.д. [1]

Исследования в области словарного запаса – самые частые в корпусной лингвистике. Можно сказать, что корпуса совершили революцию в лексикографии. По крайней мере, все современные словари английского языка создаются на базе корпусов.

Корпусы позволяют получить данные по лексеме в целом (поиск по лемме) и по конкретной словоформе, выявить типичные/нетипичные употребления и характерные сочетания слов. Эти данные могут быть разными: контексты, частоты (абсолютные и относительные), частоты по коллокациям и т.д. При составлении словарей корпусы помогают: выявить новые значения; более точно упорядочить отдельные значения внутри словарных статей. [3]

Корпусная лингвистика – это раздел прикладной лингвистики, занимающийся разработкой общих принципов построения и использованием лингвистических корпусов (корпусов текстов). Это относительно молодое и активно развивающееся направление, тесно связанное с компьютерной лингвистикой. Корпусная лингвистика рассматривает текстовые массивы как поле изучения и как источник фактов для лингвистического описания и аргументации. Она сосредотачивается на «речи», а не на «языке». [6]

Лингвистическая информация из корпуса извлекается при помощи специальных компьютерных программ. Например, рассмотрим программу **Word List Creator.**

Word List Creator – это программа, позволяющая пользователям извлекать список слов из нужного текста. Слова могут быть извлечены по алфавиту, показывая каждое слово, которое было найдено в тексте, или по частоте употребления слов, начиная со слов которые встречаются реже всего, и заканчивая самыми, часто употребляемыми. Также предусмотрена опция, которая показывает, сколько раз было употреблено слово в тексте. [7]

Для поиска и извлечения информации из корпуса используется некоторое количество довольно стандартных процедур. Самый простой формат отображения информации о корпусе — **это простые списки**. Эти списки могут быть разных типов — **от простых глоссариев до конкордансов.** Давайте посмотрим на то, как всё это может быть представлено.

Часто нужно разобраться со словами, которые употребляются в тексте. Список слов в самой простой своей форме – это попросту список всех слов, содержащихся в исследуемом тексте. Часто этот список отсортирован по частоте встречаемости слов или по алфавиту.

Однако формат простого списка не даёт возможности снять полисемию и неоднозначность грамматического класса слова, поскольку это невозможно сделать без контекста. Чтобы разобраться с этим вопросом, нам нужно будет перейти к понятию **«конкорданс» (concordance).**

Конкорданс- [англ. concordance - согласие, соответствие < лат. conncordare - согласовываться, приводить к согласию] - филол. 1) расположенный в алфавитном порядке перечень встречающихся в книге слов (или сходных по содержанию мест) с минимальным контекстом(в несколько слов2) особый тип словаря, в котором каждое слово приводится с минимальным контекстом. Нем. Konkordanz. [4]

Конкорданс - это не просто список слов или словосочетаний. Его ценность в том, что он даёт контекст слова. То есть, мы можем запустить поиск и получить все появления данного конкретного слова в тексте. Результаты поиска показываются в формате, который называется **KWIC (key word in context).** Обычно при щелчке на строку программа-конкордансер выдаёт полный контекст.

AntConc является бесплатной, мультиплатформенной программой для проведения корпусных лингвистических исследований и управления данными. Она работает на любом компьютере под управлением Microsoft Windows. [8]

Программа содержит семь инструментов, к которым можно получить доступ, нажав на клавишу табуляции в меню инструментов, или используя функциональные клавиши F1-F7.

Основные функции:

Конкорданс. Данный инструмент показывает результаты исследования формата KWIC (ключевое слово в контексте). Он позволяет увидеть, как слова и фразы обычно используются в разных контекстах. **Расположение.** Инструмент «расположение» показывает расположение элемента поиска. Это позволяет исследовать непоследовательные модели в языке. **Список слов.** Данный инструмент подсчитывает все слова в корпусе и представляет их в упорядоченном списке. Это позволяет быстро найти, какие слова употребляются наиболее часто в корпусе. **Список ключевых слов.** Такой список даёт базу для терминологических исследований и позволяет составить глоссарий.

Материалом для нашего исследования являются пьесы У. Шекспира «Ромео и Джульетта», «Антоний и Клеопатра» и «Отелло», «Буря», «Венецианский купец», «Гамлет, принц датский», «Генрих VI», «Генрих VIII», «Двенадцатая ночь, или что угодно», «Зимняя сказка», «Король Лир», «Король Иоанн», «Король Ричард III», «Магбет», «Мадфогские записки», «Перикл», «Ричард II», «Сон в летнюю ночь».

«Ромео и Джульетту», «Антония и Клеопатру» и «Отелло» относят к любовным трагедиям. Эти трагедии отличаются от других тем, что рок преследует влюблённых не из-за какого-то их проступка (кроме решения Ромео и Джульетты совершить самоубийство), но из-за некоторых преград в мире вокруг в них. В этих трагедиях смерть предстаёт почти как высшее свершение их любви, поскольку в трагическом мире любовь победить не может. Особый интерес для шекспироведов, безусловно, представляют словари, регистрирующие и обрабатывающие любовную лексику из сочинений Шекспира.

Воспользуемся программой **AntCon.** На сайте http://ebookbrowsee.net/ представлен огромный корпус текстов У. Шекспира. Мы воспользуемся тремя вышеупомянутыми текстами: «Ромео и Джульетта», «Антоний и Клеопатра» и «Отелло».

Для начала выполним поиск по конкордансу. Т.к. нас интересует любовная терминология и лексика, необходимо вычленить лексему *love* и ее возможные вариации.

Лексема love была использована 329 раз. Этим подтверждается, что основная тематика трагедий – любовь.

Love (n) love (v),beloved, lover, loving.

Далее обратимся к крупнейшему online-словарю **http://www.macmillandictionary.com/.**

Ниже представлен основной список синонимов лексемы *love*, которая является наиболее частотной в текстах трёх трагедий.

1. **Love (n) -329 times**

a very strong emotional and sexual feeling for someone

2. **passion (n) 20 times**

a very strong feeling of sexual love

3. **respect (n) 17 times**

a feeling that something is important and deserves serious attention

4. **admiration (n) 12 times**

a feeling of respect and approval

5. **awe (n) 9 times**

a feeling of great respect and admiration, often combined with fear

6. **devotion (n) 6 times**

great love, admiration, or loyalty

7. **reverence (n) 3 times**

a strong feeling of respect and admiration for someone or something

8. **adoration (n) 2 times**

a feeling of great love and respect for someone

Лексема **love** использована в всех пьесах 1962 раза самостоятельно, а как часть сложного слова 2768 раз.

Как мы видим, с помощью программы **AntConc** можно легко извлечь информацию из корпуса и сделать необходимые статистические данные, составить мини-глоссарий по теме **“love”.**

Вновь мы воспользуемся программой-конкордансером AntConc. Теперь нас интересует редкие архаичные лексемы. На этот раз мы добавили в программу наиболее популярные пьесы У. Шекспира.

В конкордансе указано, что в произведениях Шекспира насчитывается 29 066 лексем и 884 647 слов в общей сложности. Но нас интересует редкие архаичные лексемы.

С помощью функции Список слов (Word list) программа сделала выборку слов в алфавитном порядке, а с помощью конкорданса мы легко вычленили контекст. При помощи англо-английского и русско-английского онлайн словарей AbbyLingvo мы осуществили толкование и перевод лексем.

Word	Definition	Translation	KWIC
abodement (noun):	Omen; portent	предзнаменование	"**Abodements** must not now affright us" (*Henry VI Part III*).
abroach (adverb and adjective):	Opened or tapped to release contents; active.	Активный	Example: "Who set this ancient quarrel new **abroach**?" (*Romeo and Juliet*).
absey-book (noun):	any elementary textbook	учебник	"Then comes [the] answer like an **absey-book**" (*King John*).
accite (verb)	excite; cause	вызывать	And what **accites** your most worshipful thought to think so?" (*Henry IV Part II*).

accomplement (noun)	Armor; gear; equipment	Обмундирование, броня	"[The soldiers were] arrayed . . . in all **accomplements**" (*Edward III*).
advocation (noun)	Act of advocating, promoting, or making a plea.	Жалоба, подавать жалобу	"My **advocation** is not now in tune" (*Othello*).
acquittance (noun)	Release from obligation or debt.	Освобождение от обязательства или долга	"Now must your conscience my **acquittance** seal" (*Hamlet*).
ague (noun)	Fever with chills, sweating, and shivering	Лихорадочный озноб	"Here let them lie / Till famine and the **ague** eat them up" (*Macbeth*).
Almsdrink (noun)	Dregs of a beverage, remains of drink set aside for the poor.	Остатки напитка	"They have made him [Lepidus] drink **almsdrink**" (*Antony and Cleopatra*).
attainture (noun)	Dishonor; disgrace	Бесчестие, позор	Her **attainture** will be Humphrey's fall" (*Henry VI Part II*,).

Как мы видим, с помощью **Concordance** можно одновременно просматривать полный список слов, найденные соответствия и исходный текст, а также просматривать оригинальный текст, просто нажав на любое из слов, после чего будет показан контекст.

На основании полученного авторского глоссария с помощью программы Word List Creator и AntConc можно сделать вывод, что программы имеют обучающую цель, а также является идеальным вариантом при составлении списка слов и авторского глоссария. Программа позволяет выделить ключевые слова исследуемого текста, таким образом, появляется возможность проанализировать исходный текст, выявить основную мысль. Программа удобна для писателей и исследователей-лингвистов. Она поможет пользователям в создание глоссариев или индексов, или позволит писателям увидеть, не перегружен ли их рассказ отдельными словами.

Активное использование программ позволяет сделать вывод, что помимо конкордансов программы анализа корпусов отображают и базовую статистическую информацию о корпусе: соотношение числа словоформ и словоупотреблений, среднюю длину предложения, количество предложений и их распределение по длине, индекс исключительности (каков процент слов, употреблявшихся лишь один раз), индекс постоянства (каков процент частых слов) и так далее.

Таким образом, подводя итоги, следует подчеркнуть, что корпус служит надежным источником фактического материала для составления словарей, грамматик, учебников, справочных пособий, выполнять функции справочного пособия для выяснения вопросов о современном русском литературном словоупотреблении, т.е. служить эффективным помощником для всех, работающих со словом (лингвисты, литературоведы, журналисты, писатели, переводчики, преподаватели русского языка или иностранного языка и др.).

СПИСОК ЛИТЕРАТУРЫ

1. Андрющенко В.М. Концепция и архитектура машинного фонда русского языка / Отв. ред. А.П. Ершов. – М., 1989. – 235 с.
2. Баранов А.Н. Корпусная лингвистика // Баранов А.Н. Введение в прикладную лингвистику. – М., 2001. – С.112-137.

3. Колпакова Г.В. Корпусная лингвистика и лексикография. // Электронный научно-образовательный журнал ВГПУ «Грани познания». – №2 (12). – Июнь 2011[Электронный ресурс]. URL: www.grani.vspu.ru
4. Комлев Н.Г. Словарь иностранных слов. – Москва, ЭКСМО-Пресс, 2000. - 1308 с.
5. Sinclair, J. Corpus, Concordance, Collocation, Oxford University Press. 1991. -171 p.
6. Фонд знаний «Ломоносов» [Электронный ресурс]. URL: http://www.lomonosov-fund.ru/enc/ru/
7. Word List Creator. URL: http://www.wordlistcreator.com/
8. AntConc. URL: http://www.antlab.sci.waseda.ac.jp/software.html