УДК 81-139, 81`32

Г.В. Садыкова Г.В., *канд. филол. наук, к.ф.н., доцент, Казанский (Приволжский) федеральный университет, г. Казань, gsadykova@yahoo.com*

МЕТОДЫ КОРПУСНОЙ ЛИНГВИСТИКИ В СРАВНИТЕЛЬНО-СОПОСТАВИТЕЛЬНЫХ ИССЛЕДОВАНИЯХ (НА МАТЕРИАЛЕ СОПОСТАВЛЕНИЯ ЧАСТОТНОЙ ЛЕКСИКИ РУССКОЯЗЫЧНЫХ И АНГЛОЯЗЫЧНЫХ ТЕКСТОВ ЭЛЕКТРОННЫХ СМИ)

В статье автор описывает этапы проведения исследования текстов больших объемов (корпусов) с помощью доступной компьютерной программы Simple Concordance Program. Описание сопровождается наглядной демонстрацией данного метода исследования на материале сопоставления частотной лексики русскоязычных и англоязычных текстов электронных СМИ. Описываемый метод используется лингвистами, работающими в русле корпусной лингвистики, признанной в современном научном сообществе.

Ключевые слова: копусная лингвистика, конкордансинг, методология лингвистических исследований, электронные тексты, сравнительно-сопоставительная лингвистика.

Sadykova G.V. **CORPUS LINGUISTICS METHODS IN COMPARATIVE STUDIES (BASED ON THE COMPARATIVE STUDY OF FREQUENTLY USED WORDS IN RUSSIAN AND ENGLISH TEXTS OF ELECTONIC NEWSPAPERS)**

The article describes steps in conducting research that involves large textual data (corpuses) with the help of user-friendly computer software Simple Concordance Program. To demonstrate this research method, the author uses her study of frequently used lexical units in Russian and English texts of electronic newspapers. The described research method is currently employed by linguists working in the area of corpus linguistics which is recognized in modern scientific circles.

Keywords: corpus linguistics, concordancing, methods of linguistic research, electronic texts, comparative linguistics.

Российское научное сообщество до настоящего момента слабо интегрировано в мировое научное пространство. Исследователи, в особенности гуманитарии, включая лингвистов, нередко замыкаются на традиционных для российской и советской науки методах исследования, игнорируя те методы, что получили распространение и признание у

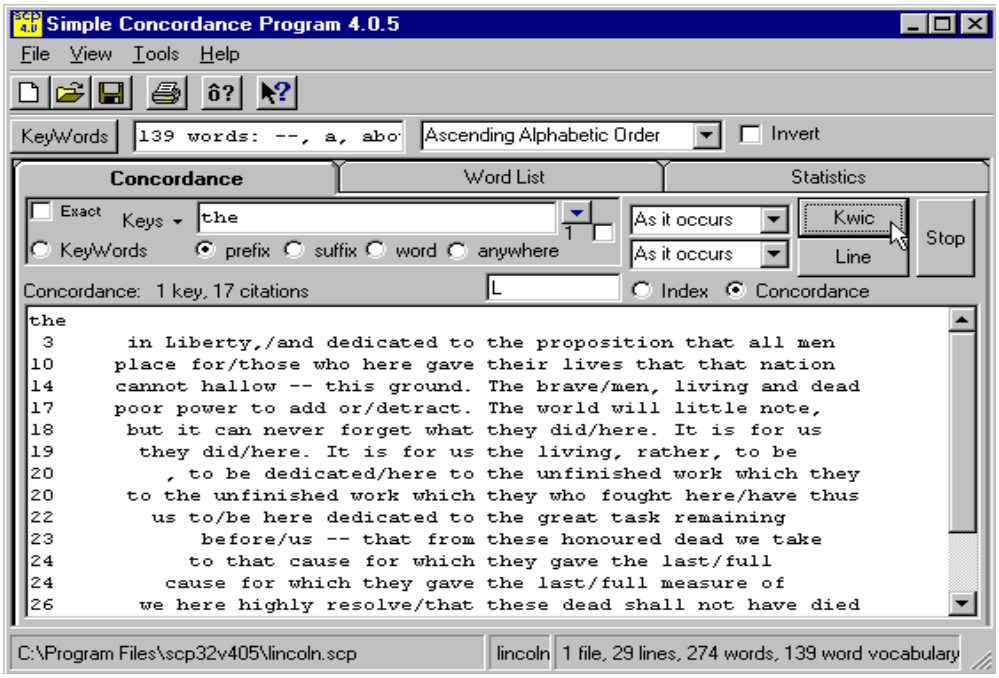
зарубежных коллег. В результате мы имеем значительные, иногда непреодолимые, трудности при публикации статей в качественных зарубежных журналах, так как «международный научный рынок более мобилен, новые направления быстрее овладевают массами и быстрее теснят традиционные» [Фрадков, 2003; 154]. Ситуация усугубляется также отставанием российской науки в сфере использования методов исследования, основанных на современных информационно-коммуникационных технологиях, применение которых, согласно 10-летнему социологическому мониторингу исследователей РАН, положительно коррелируется с профессиональной успешностью ученого [Мирская, 2005].

В связи с вышеизложенным актуальным является описание и демонстрация несложной, но эффективной компьютерной программы, используемой лингвистами, работающими в русле признанной в современном научном сообществе корпусной лингвистики. Освоение данной программы и методов обработки полученных с его помощью данных позволяет современному лингвисту выйти за пределы описательного метода и метода систематизации материала посредством составления классификаций и словарей, традиционно используемых, к примеру, в российской фразеологии. Описываемый метод дает возможность производить количественный подсчет и делать статистический анализ, а использование количественных методов исследования высоко ценится во многих ведущих зарубежных журналах, ориентирумых на гуманитарные дисциплины, где качественные методы исследования обычно превалируют. Данный метод был использован автором статьи в ряде работ, две из которых [Meskill 2007; Sadykova 2009] опубликованы в рецензируемых журналах США, что говорит о его ценности как метода, признанного зарубежом.

Вниманию лингвистов предлагается программа Simple Concordance Program (SCP 4.0.9), размещенная в свободном доступе на сайте http://www.textworld.com/scp/. Данная программа позволяет обрабатывать тексты больших объемов (корпусы) и получать так называемый конкорданс – список слов в непосредственном контексте (см. Рис. 1). Это позволяет выявлять частотность слов и других лингвистических единиц, а также исследовать контекст использования отдельно взятой единицы. Методы корпусной лингвистики нередко используют в лексикографии, грамматике, дискурсивной лингвистике, исторической лингвистике и в прикладной лингвистике, в частности, в сфере изучения языка как иностранного [Biber, 2002]. Также данную программу можно использовать в смежных с лингвистикой науках – социолингвистике и лингвистической антрополигии, а также в педагогике, где лингвистический анализ текстов документов может дополнять традиционный для дисциплины метод наблюдения или опроса. Поскольку программа позволяет работать с

текстами на английском, русском, французском, немецком и других языках, она представляет интерес и для лингвистов-компаративистов.

Рис.1 Интерфейс Simple Concordance Program, демонстрирующий результаты конкорданса префиксального *the*.



Лингвистический анализ с помощью программы Simple Concordance Program можно условно поделить на 3 этапа: 1) подготовка корпусов, 2) конкордансинг, 3) обработка полученных данных. Для иллюстрации каждого из этапов воспользуемся проведенным ранее и опубликованным исследованием «Медиатопик «экономика» в российский и американских СМИ» [Садыкова, 2009].

На первом этапе исследователь должен провести подчас кропотливую работу по отбору текстов и составлению корпуса или нескольких копусов (если предполагается сравнительно-сопоставительная работа). К примеру, лингвист, работающий в русле медиалингвистики, может отобрать значительное количество текстов электронных газет в целях сравнения особенности языка разных жанров или разных газет. Диахронисты могут поставить целью проследить языковые изменения в том же жанре или в той же газете, для чего им надо найти и отобрать (а иногда отсканировать и перевести в электронных текстовый формат) тексты разных лет. Анализ и сопоставление может также касаться текстов, написанных на двух или более языках, для чего исследователь должен отобрать

сопоставляемые (и сопоставимые) тексты, составив из них несколько корпусов. Вместе с тем, поскольку корпусная лингвистика имеет немало приверженцев и является популярной и уважаемой среди лингвистов на протяжении нескольких десятилетий, в открытом доступе сети Интернет можно найти корпусы, составленный ранее. Некоторые из данных корпусов имеют так называемые тэги, то есть обозначения лингвистических характеристик слов, таких как часть речи, что расширяет спектр возможных лингвистических исследований. Среди доступных лингвистам корпусов можно отметить следующие:

Collins WordBank Online http://www.collinslanguage.com/content-solutions/wordbanks

The Corpus of Contemporary American English http://corpus.byu.edu/coca/

British National Corpus (corpus demo) http://info.ox.ac.uk/bnc/

Hong Kong Virtual Language Centre http://vlc.polyu.edu.hk/concordance/

WordNet http://wordnet.princeton.edu/

Russian WordNet http://wordnet.ru/

Visual Interactive Syntax Learning (VISL) http://beta.visl.sdu.dk/visl2/corpus_linguistics.html

В демонстрационном исследовании [Садыкова, 2009] было составлено 2 корпуса, один из которых состоял из текстов семи российских электронных СМИ, а второй их текстов шести электронных СМИ США. Тексты отбирались методом сплошной выборки в течении 7 дней. Все тексты относились к медиатопику «экономика». Оба корпуса в результате имели приблизительно одинаковый объем — около 20 тысяч слов в каждом. Таким образом при составлении корпусов соблюдался принцип сопоставимости текстов: они отбирались из одного типа СМИ, состояли из текстов того же жанра и медиатопика, написанных в тот же временной отрезок и имеющих одинаковый общий объем. Соблюдение этого принципа позволило минимизировать возможность погрешности в результате исследования, что особенно важно в количественных исследованиях.

Вторым этапом исследования является обработка каждого из корпусов непосредственно с помощью программы и получение конкорданса, то есть списка слов (или других лингвистических единиц) в соответствии с их частотностью или в окружении непосредственного контекста. Здесь особенно важно, чтобы исследователь понимал, какие результаты могут дать ответ на поставленный им исследовательский вопрос. Целью демонстрационного исследования было «выявить и сравнить картины мира в области экономики, создаваемые сетевыми СМИ двух стран» [Садыкова 2009; 318]. Составление списка и сопоставление наиболее частотных слов решало данную задачу. Кроме того, было

принято решение отдельно рассмотреть частотность слов, обозначающих ключевые медиафигуры – президенты, главы правительства и т.п., что также позволило дополнить представление о картине мира в области экономики, создаваемой журналистами. Таким образом была составлена таблица двадцати наиболее употребимых слов в обоих корпусах с указанием их частотности, а также выделены слова, вошедшие в списки топ-20 как в русских, так и в американских электронных СМИ (см. Таблица 1).

Кроме частотности слов и изучения контекстуального окружения отдельных слов, Simple Concordance Program позволяет рассчитать среднее количество слов в предложениях, что может быть важно для исследований, например, по стилистике. Для этого достаточно дать программе подсчитать количество точек в тексте и разделить полученный результата на количество слов. Также можно изучать контекстуальное употребление отдельных лексем, фразеологических единиц или пунктуационных знаков. Можно также рассмотреть частотность и контекст отдельных грамматических единиц, например пассива, но для сложного синтаксического или морфологического анализа потребуется наличие тэгированных корпусов, так как программа сама по себе не может отличать части речи или части слов и, например, отличить существительное *wave* (волна) от глагола *wave* (помахать) или приставку *при* в слове *пришел* от части корня в слове *приз*. Хотя программа позволит найти эти слова в огромном корпусе, дифференциацию нужного исследователю материала в подобных случаях придется делать вручную.

Таблица 1. Наиболее частотная лексика СМИ России и США

Российские СМИ		**СМИ США**	
Слово (лемма)	**кол-во**	**Слово (лемма)**	**количество**
Банк	272	Say	221
Год	203	Percent	118
Россия	179	$ (dollar)	108
Кредит	166	Year	105
Миллиард, млд.	131	Sale	68
Рубль, руб.	145	Company	64
Доллар	114	Loss	61
Финансы	111	Market/marketplace	57
Рынок	109	Work/workforce	57
Капитал	89	Economy	55
Кризис	78	Last	53
Экономика	75	Employ/unemployment	50
Автомобиль/авто	75	Fall/fell (verb)	48
Инвестор, инвестировать	70	Intel	48

Акция, акционеры	67	Job/jobless	45
Цена, обесценивать	70	Bank	44
Процент	58	Spend	40
Эксперт	55	State (=country)	40
Правительство	52	Million	40
Государство	52	Billion	39

На третьем этапе работы исследователь делает выводы по полученным даннам и уточняет результаты. К примеру в демонстрационном исследовании были сделаны выводы по лингвистическим и эктралингвистическим факторам влияющим на частотность употребления слов в текстах российских и американских газет по экономике. Наблюдение за частотностью упоминания медиафигур, в частности, позволило сделать выводы о том, что место наиболее значимого лица в российской экономике в 2009 году делили В.Путин и Д.Медведев, причем первый, будучи в должности премьер-министра, упоминался в экономических текстах чаще, чем второй; в то же время в экономических текстах СМИ США место наиболее значимой медиа фигуры единолично занимал президент Б.Обама.

Естественно, за третим этапом исследования может вновь повториться второй, так как обработка полученных первых данных может привести к необходимости более тщательного анализа отдельных языковых единиц. Опыт показывает, что после получения первых списков частотности слов, возникает вопрос, как именно используются те или иные слова. Иногда интерес лингвиста может вызывать не столько наиболее частотные слова, сколько малочастотные и казалось бы неожиданные для данного текста.

Таким образом, представленная программа позволяет существенно облегчить труд лингвиста по обработке объемных текстов (корпусов) за счет составления списков частотности лексических и других лингвистических единиц и выявлению контекста их использования. Программа позволяет производить количественный анализ, что особенно продуктивно при его сочетании с качественными методами исследования. Данный метод признан зарубежом, и исследования, проведенные с использованием программ, подобным Simple Concordance Program, публикуются в качественных реферируемых изданиях, в том числе индексируемых в Scopus, то есть признаваемых ВАК. Данный факт позволяет рекомендовать изучение и использование метода при ведении лингвистических и других исследований, объектом которых являются тексты больших объемов.

Литература

1. Мирская, Е. З. Наука в информационном обществе: новые возможности и проблемы / Е.З. Мирская // Информационное общество. – 2005. – Вып. 5. – С. 4-7.

2. Садыкова, Г. В. Медиатопик «экономика» в российских и американских электронных СМИ / Г. В. Садыкова // Слово и текст: коммуникативный, лингвокультурный и исторический аспекты. Материалы международной научной конференции. – Ростов н/Д: НМЦ «Логос», 2009. – С. 318-320 .

3. Фрадков, А. Л. Как опубликовать хорошую статью и отклонить плохую. Заметки рецензента / А. Л. Фрадков // Автоматика и телемеханика. – 2003. – №10. – С. 149-157.

4. Biber, D. Corpus Linguistics: Investigating Language Structure and Use / D. Biber, D., S. Conrad and R. Reppen. – New York: Cambridge University Press, 2002. – 312 p.

5. Meskill, C. The presentation of self in everyday ether: A corpus analysis of student self-tellings in online graduate courses / C. Meskill, G. Sadykova // Journal of Asynchronous Learning Networks. – 2007. - №11(3). – P. 123-138.

6. Sadykova, G. The language of digital learning objects: A cross-disciplinary study / G. Sadykova, C. Meskill // Journal of Online Learning and Teaching. – 2009. - №5(2). – P. 239-252.