Жолобова Анна Олеговна

НАЦИОНАЛЬНЫЙ КОРПУС ИСПАНСКОГО ЯЗЫКА: CORDE И CREA

В данной статье дается характеристика двум корпусам испанского языка, созданным Испанской королевской академией: современному корпусу CREA и историческому корпусу CORDE. По каждому корпусу приводятся данные хронологического, географического, жанрово-тематического и статистического характера. Описывается принцип работы, анализируются возможности и недостатки системы поиска данных корпусов.

Адрес статьи: www.gramota.net/materials/2/2014/9-1/13.html

Источник

Филологические науки. Вопросы теории и практики

Тамбов: Грамота, 2014. № 9 (39): в 2-х ч. Ч. I. С. 56-58. ISSN 1997-2911.

Адрес журнала: www.gramota.net/editions/2.html

Содержание данного номера журнала: www.gramota.net/materials/2/2014/9-1/

УДК 811.134.2
Филологические науки

В данной статье дается характеристика двум корпусам испанского языка, созданным Испанской королевской академией: современному корпусу CREA и историческому корпусу CORDE. По каждому корпусу приводятся данные хронологического, географического, жанрово-тематического и статистического характера. Описывается принцип работы, анализируются возможности и недостатки системы поиска данных корпусов.

Ключевые слова и фразы: национальный корпус испанского языка; исторический корпус; современный корпус; текстовый массив; Испанская королевская академия.

Жолобова Анна Олеговна, к. филол. н.
Казанский (Приволжский) федеральный университет
anijola@yahoo.es

НАЦИОНАЛЬНЫЙ КОРПУС ИСПАНСКОГО ЯЗЫКА: *CORDE* И *CREA*©

Работа выполнена при поддержке Российского гуманитарного научного фонда: проект № 14-04-0061.

Корпус того или иного языка представляет собой собрание текстов с лингвистической разметкой в электронной форме. Доступность, быстрота обработки информации и выдачи результатов, представительность, разные типы лингвистической аннотации делают национальный корпус языка важнейшим источником эмпирических данных для проведения лингвистического исследования [1].

Бурное развитие новых компьютерных технологий в конце XX в. позволило Испанской королевской академии в 1995 году приступить к грандиозной работе по составлению двух национальных корпусов испанского языка: исторического корпуса CORDE (*Corpus diacr nico del español*) [3] и современного корпуса CREA (*Corpus de referencia del español actual*) [5]. Недостатки данных корпусов, о которых мы поговорим в нашей статье, и необходимость развития послужили толчком к созданию двух новых корпусов, работа над которыми ведется в последние годы: «Корпус Нового исторического словаря» *Corpus del Nuevo diccionario hist rico* (CDH) [2] и «Корпус испанского языка XXI века» *Corpus del español del siglo XXI* (CORPES) [4].

Начнем наш обзор с первого исторического корпуса испанского языка CORDE. Объем корпуса составляет 250 млн словоупотреблений. С хронологической точки зрения, в соответствии с историко-лингвистической ситуацией в CORDE представлена следующая периодизация с прогрессивным увеличением текстового наполнения [8]:

1. Средние века (16,5%): а) до 1250 года; б) 1251-1491.
2. Золотой век (30,5%): а) 1492-1598; б) 1599-1712.
3. Современная эпоха (53%): а) 1713-1812; б) 1813-1898; в) 1899-1939; г) 1940-1974.

Жанровое и тематическое распределение всего массива текстов выглядит так [Ibidem]:

1. художественные тексты (44%): поэзия (10%), проза (27%), драма (7%);
2. нехудожественные тексты (56%): дидактика (10%), наука и техника (14%), религия (6%), общество (8%), история (9%), юриспруденция (6%), пресса (3%).

Что касается географических характеристик, то 74% всех текстов были созданы в Испании и лишь 25% принадлежат латиноамериканским авторам, что связано с объективными историческими предпосылками (напомним, что испанская колонизация Америки началась в 1492 году). Оставшийся 1% текстов написан на сефардском языке.

Современный корпус CREA (последняя версия 3.2., июнь 2008 года) включает 140 тыс. документов и более 160 млн словоупотреблений. В хронологическом плане корпус охватывает период с 1975 года до 2004 года. В корпусе выделены два блока: письменный корпус (90%) и устный корпус (10%). 49% письменного корпуса составляют книги, 49% представлены периодическими изданиями, остальные 2%, включенные в категорию *miscelánea* «разное», отобраны из брошюр, блогов, электронной почты и т.п. Тексты блока «книги и периодические издания» распределены по двум группам:

1. Художественные тексты: роман, рассказы и театр.
2. Нехудожественные тексты представлены 6 гиперполями, каждое из которых содержит до 20 подтем: наука и техника; социальные науки, убеждения и мышление; политика, экономика, торговля и финансы; искусство; досуг и быт; здоровье.

Устный корпус включает 9 млн словоупотреблений и более 1600 документов. Устные тексты делятся на две группы: 1) транскрибированные и кодифицированные записи радио- и телевизионных передач; 2) транскрипции выступлений политиков, телефонных разговоров, сообщений на автоответчике, бытовые диалоги и т.п.

С географической точки зрения, корпус CREA охватывает США и 6 зон Латинской Америки:

1. Андская (Перу, Эквадор, Боливия).
2. Антильская (Пуэрто Рико, Доминиканская республика, Куба).
3. Континентальные страны Карибского бассейна (Колумбия, Венесуэла).
4. Чилийская (Чили).

5. Мексика и Центральная Америка (Мексика, Сальвадор, Гватемала, Коста Рика, Панама, Никарагуа, Гондурас).

6. Рио де ла Плата (Аргентина, Уругвай, Парагвай).

Самой значительной по объему представленных текстов является пятая группа: ей принадлежит более 40% от общего количества латиноамериканских текстов [7].

Перейдем к описанию системы поиска в CREA и CORDE. На странице поиска кроме окна *Consulta* «запрос» представлены следующие критерии поиска: *Autor* «автор», *Obra* «произведение», *Cronol gico* «хронология» (позволяет задать определенный период времени), *Medio* «средство коммуникации» (позволяет выбрать из списка одну из 6 категорий: *todos* «все», *libros* «книги», *peri dicos* «газеты», *revistas* «журналы», *miscelánea* «разное», *orales* «устное»), *Geográfico* «география» (позволяет выбрать из списка одну из 22 стран), *Tema* «тема» (позволяет при желании выбрать из списка определенный жанр или тему).

Система запроса рассматриваемого корпуса хорошо подходит для поиска точных форм слов и фраз, однако трудности возникают при более сложном запросе. Главная проблема корпусов CREA и CORDE заключается в отсутствии лемматизации, что не позволяет получить все словоформы запрашиваемого слова, значительно ограничивая таким образом поисковые и, следовательно, исследовательские возможности. Так, если нас интересует вся парадигма определенного глагола, нам придется в окне поиска вводить последовательно все его формы, что существенно затруднит нашу задачу. Рассмотрим подробнее параметры формулировки запроса:

1. Система запроса разграничивает прописные и строчные буквы (например, слово в начале предложения и в середине), а также знаки с графическим ударением и без него (например, sí «да» и si «если»).

2. С помощью вопросительного знака «?» можно заменить один другой знак в запросе, что позволит найти в корпусе все возможные варианты с подстановкой данного знака. Так, запрос в форме *p?so* даст следующие варианты ответов: *peso* «вес», *paso* «шаг», *piso* «квартира», *puso* «положил», *poso* «осадок».

3. Звездочка «*» позволяет заменить сразу несколько знаков. Так, запрос в форме *abuel** даст следующие результаты: *abuelo* «дедушка», *abuela* «бабушка», *abuelos* «бабушка с дедушкой», *abuelito* «дедуля» и т.д. Однако если мы зададим короткую основу слова из распространенной комбинации букв (например, *sal**), то появится сообщение: «La consulta introducida es demasiado compleja, por favor simplifíquela» («Запрос слишком сложный, пожалуйста, упростите его»). Такая же ситуация обстоит с суффиксами, окончаниями: невозможно получить результаты по таким запросам как: *azo, *ieran, *ísim* и т.д. [6].

4. При помощи команды *dist/* можно задать максимальное расстояние, на котором должны находиться друг от друга два запрашиваемых слова: команда *árbol* «дерево» *dist/4 cereza* «черешня» будет означать, что в тексте между данными словами должно быть максимум четыре слова.

5. Команда Y «и» позволит найти контекст с двумя заданными словами: *árbol Y cereza* «дерево И черешня». Команда *O* «или» позволит найти контекст с одним из заданных слов: *árbol O cereza* «дерево ИЛИ черешня». Команда *NO* «не» исключает из поиска второе слово: *árbol Y NO cereza* «дерево И НЕ черешня».

После введения критериев запроса и команды «поиск» открывается окно *Resultado de la consulta al banco de datos* «результат запроса в банке данных». В нем высвечивается количество случаев, а также количество документов, и дается ссылка *Ver estadística* «посмотреть статистику». Статистика приводится по годам, странам и темам в процентном и количественном соотношении.

Для просмотра примеров имеется окошко *Obtenci n de ejemplos* «получение примеров», в котором можно задать параметры представления полученных результатов. Из выпадающего списка можно выбрать форму представления: *Documentos* «документы» (результаты приводятся в виде таблицы, в которой указывается количество случаев, год, автор, произведение, страна, тема, издательство), *Concordancias* «конкордансы» (результаты выдаются с минимальным контекстом со всеми выходными данными), *Párrafos* «абзацы» (в качестве результата выводится расширенный контекст со всеми выходными данными), *Agrupaciones* «группирование» (результаты по умолчанию сведены в таблицы в зависимости от сочетания в два, три или пять слов с указанием процентного и количественного соотношения, однако есть возможность в дополнительной графе указать другое интересующее число сочетаний данного слова). Кроме того, все полученные результаты можно упорядочить, выбрав из выпадающего списка один из критериев: *casos* «случаи» (по количеству случаев в документе), *autor* «автор» (по алфавиту), *año* «год» (по возрастанию), *país* «страна» (по алфавиту), *tema* «тема» (по алфавиту), *título* «название» (по алфавиту).

Итак, несмотря на достоинства корпусов Испанской королевской академии CREA и CORDE, которые выражаются в значительном объеме текстового массива, обширной географической и хронологической представленности, разнообразии тематики, возможности классификации результатов, имеется весомый недостаток, а именно отсутствие лемматизации, что значительно затрудняет осуществление запроса и получение необходимой лингвистической информации.

Список литературы

1. **Жолобова А. О.** Фразеологические единицы библейского происхождения в английском, испанском и русском языках: дисс. ... к. филол. н. Казань, 2005. 267 с.
2. **CDH** [Электронный ресурс]. URL: http://www.rae.es/recursos/banco-de-datos/cdh (дата обращения: 20.05.2014).
3. **CORDE** [Электронный ресурс]. URL: http://www.rae.es/recursos/banco-de-datos/corde (дата обращения: 20.05.2014).
4. **CORPES** [Электронный ресурс]. URL: http://www.rae.es/recursos/banco-de-datos/corpes-xxi (дата обращения: 20.05.2014).

5. **CREA** [Электронный ресурс]. URL: http://www.rae.es/recursos/banco-de-datos/crea (дата обращения: 20.05.2014).
6. **Davies M.** Un corpus anotado de 100.000.000 palabras del español histórico y moderno // Procesamiento del Lenguaje Natural. 2002. № 29. P. 21-27.
7. **Manual de consulta** [Электронный ресурс]. URL: http://corpus.rae.es/ayuda_c.htm (дата обращения: 20.05.2014).
8. **Sánchez Sánchez M., Domínguez Cintas C.** El banco de datos de la Real Academia Española: CREA y CORDE // Per Abbat. 2007. № 2. P. 137-146.

THE NATIONAL CORPUS OF THE SPANISH LANGUAGE*: CORDE* AND C*REA*

Zholobova Anna Olegovna, Ph. D. in Philology
Kazan (Volga Region) Federal University
anijola@yahoo.es

The article characterizes two corpuses of the Spanish language created by the Royal Spanish Academy: the modern corpus CREA and the historical corpus CORDE. On each corpus the author presents the data of chronological, geographical, genre and subject and statistical nature. The researcher describes the principle of work, and analyzes the possibilities and shortcomings of the data retrieval system of the corpuses.

Key words and phrases: National Corpus of the Spanish Language; historical corpus; modern corpus; text corpus; the Royal Spanish Academy.

УДК 821
Филологические науки

В статье автор обращается к символике солнца, выявленной в романе французского мифотворца М. Турнье «Пятница, или Тихоокеанский Лимб». В тексте прослеживается значение солярного культа в различных мифологических системах, христианской традиции и обнаруживается трансформирующее влияние архетипа солнца на мышление Робинзона, главного героя повествования. Основные идеи в статье излагаются через интертекстуальную призму преподнесения материала, ономастические игры с именами собственными в произведении, а также изучение проблематики в русле пространственно-временных реалий романа. Размышления автора преподнесены на фоне широкого культурологического контекста.

Ключевые слова и фразы: аллюзия; символика; культ солнца; христианская традиция; мифология; темпоральные характеристики.

Завадская Анна Иосифовна
Белорусский государственный педагогический университет имени Максима Танка
Annie_Za@mail.ru

АРХЕТИП СОЛНЦА В РОМАНЕ М. ТУРНЬЕ «ПЯТНИЦА, ИЛИ ТИХООКЕАНСКИЙ ЛИМБ»©

Солнце – древнейший символ в мифологии множества народностей, со всеми своими смыслами перекочевавший в мировую литературу. У японцев, кельтов, индейцев Америки, представителей Африки и Океании, а также в фольклорной традиции древних германцев оно воплощает в себе женское начало, во всех других языческих верованиях – мужское, поскольку символизирует верховную власть, свет и энергию жизни, а также источник мудрости, например, у даосов солнце – это ян, мужской элемент гендерной философской диады, у перуанских индейцев оно персонифицируется в образе человека [3].

В Египте культ солнца с древних времен являлся государственным, верховным божеством считался бог солнца Ра, хотя боги Озирис и Гор имели непосредственное отношение к символу, в них воплощался образ солнца на разных стадиях его движения. В скандинавских сказаниях оно олицетворяет собой всеведение, представляя из себя глаз бога Одина (в древнегреческой мифологии солнце – глаз Зевса, а также одна из ипостасей Аполлона) [4]. Наиболее часто в фольклорной традиции солнце представляется в облике прекрасного юноши, а его диск, наподобие феникса, призван символизировать огненное колесо жизни с мистерией умирания и неизменного возрождения из пепла (культ бога Митры в поздней Римской империи).

В христианской символике солнце указывает на Бога Отца и Его божественную благодать, так, Данте полагал, что данный образ служит для восприятия нами Творца. В иконографии некоторые святые и отцы церкви (Фома Аквинский) изображались с солнцем, расположенным на груди. В алхимической традиции последний день недели – воскресенье – считалось Днем Солнца [3].

Наряду с фениксом, солнце изображали в виде орла, змеи, петуха и быка, свастики и блистающей колесницы, управляемой богом солнца в той или иной мифологической традиции, например, скандинавской, древнегреческой, индуистской.