МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА НА РЕПУБЛИКА БЪЛГАРИЯ
КИРИЛО-МЕТОДИЕВСКИ НАУЧЕН ЦЕНТЪР ПРИ БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. М. Т. КАЛАШНИКОВА
НАУЧНОЕ СООБЩЕСТВО “ПИСЬМЕННОЕ НАСЛЕДИЕ”
DIGITAL MEDIEVALIST SCHOLARLY COMMUNITY
ФОНДАЦИЯ „УСТОЙЧИВО РАЗВИТИЕ НА БЪЛГАРИЯ “

Писменото наследство и информационните технологии

El’Manuscript–2014

Материали от V международна научна конференция
Варна, 15–20 септември 2014 г.

София · Ижевск
2014

Сборникът е издаден с финансовата подкрепа на Министерството на образованието и науката на Република България по процедура за подкрепа на международни научни форуми.

Отговорни редактори: проф. дфн В. А. Баранов
доц. д-р В. Желязкова
д-р А. М. Лаврентьев

Редактори: Нели Ганчева, Веселка Желязкова (български текст)
О. В. Зуга, В. А. Баранов (руски текст)
Кевин Хокинс (Kevin Hawkins) (английски текст)

Писменото наследство и информационните технологии [Текст] : материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. — София ; Ижевск, 2014. — 448 с.

Сборникът съдържа материали от конференция, посветена на разработването и създаването на съвременни средства за съхраняване, описване, обработка, анализ и публикуване на ръкописни и старопечатни книжовни паметници и исторически извори, а също и на въпросите за подготвянето на електронни ресурси в областта на хуманитаристиката и тяхното използване в научните изследвания и преподаването.

ISBN 978–954–9787–25–2

Формирование аннотированных баз данных семантического поля “агрессия” (на материале русскоязычных и англоязычных электронных интернет-источников)[1]

Р. К. Потапова, Л. Р. Комалова

Аннотированные базы данных, корпусная лингвистика, семантическое поле “агрессия”, интернет-коммуникация

The Creation of Annotated Databases of the Semantic Field of “Aggression” (Using Russian and English Internet Sources)

Rodmonga Potapova, Liliya Komalova

The article presents fulltext databases containing Russian- and English-language documents from the mass media containing verbal representatives of the concept “aggression”. Each database consists of 120 annotated text units. The annotation covers lexical, semantic and pragmatic levels. Special metrics and a local dictionary of the semantic field of “aggression” accompany each text. The database can be implemented in scientific research on speechology, for improving computer-aided Internet monitoring systems, in pedagogy, and for improving search systems based on the semantic field of “aggression”.

В среде интернет тексты средств массовой информации служат, как правило, для формирования общественного мнения и установок реципиентов на восприятие и оценку окружающей действительности. Отрицательные и положительные эмоции текста активизируют смежные эмоции и модальности реципиента.

В рамках исследования семантического поля “агрессия” (далее — СПА) в фокусе внимания находятся средства и механизмы реализации негативных эмоций и переживаний реципиента, побуждающие к переживанию и реализации агрессии, в том числе опосредованные текстами цифровых СМИ в Интернете.Анализ содержания современных СМИ подтверждает предположение о расширении набора слов СПА, представленных в СМИ, и о снижении порога критичности восприятия.

Доступность текстов в среде интернета и современные программные разработки позволяют проводить мониторинг интернет-контента и выявлять тексты, провоцирующие агрессивное поведение. Весьма актуальным и перспективным является разработка баз данных и баз знаний, позволяющих кластеризировать

[1] Исследование поддержано Российским фондом фундаментальных исследований (РФФИ), проект № 14−06−00363 (научный руководитель д-р филол. наук, профессор Р. К. Потапова).

современные языковые и речевые тенденции текстов цифровых СМИ в совокупности с возможностью оценки их воздействующей силы, в том числе потенциала побуждения реципиента к разным формам агрессии.

В настоящем докладе описываются принципы формирования языковых баз данных и представляются полнотекстовые базы русскоязычных и англоязычных письменных текстов цифровых СМИ в интернете.

Необходимым требованием к формированию баз данных для поисковых целей является предварительное проведение исследования речевого материала, которое позволяет выявить особенности функционирования концепта “агрессия” в текстах СМИ, специфику таких текстов в среде интернета, позволяющей представлять контент с большей “вольностью” подачи материала, а порой подавать позицию автора в весьма радикальной форме.

Следующим шагом является отбор речевого материала, который должен отражать определенный срез, быть взвешенным и представлять разные тематические составляющие концепта “агрессия”. Отбор источников (в данном случае цифровых СМИ, представленных в среде интернет в открытом доступе) определяет рамки так называемой позиции редакции и авторов, то есть круг установок и стереотипных представлений об описываемых событиях действительности.

Описание языкового материала полнотекстовой базы данных подразумевает описание тех социальных условий, в которых происходило формирование корпуса, что, в свою очередь, является важным социокультурным показателем, особенно для пользователей иноязычных культур. По мнению Р.К. Потаповой, разрабатываемые базы данных представляются открытой системой, доступной для пополнения и расширения в базы знаний и экспертные системы.

Немаловажным представляется и разработка форм унификации и аннотирования отобранного материала, которые позволяют представить каждую базу данных в доступной для дальнейших целей форме, а также содержат метаинформацию о каждой единице базы.

Представляемые русскоязычная и англоязычная базы данных письменных текстов цифровых СМИ содержат вербальные репрезентанты концепта “агрессия” и являются лингвистически паспортизированными полнотекстовыми базами данных. В БД представлены тексты из мировых, федеральных и региональных СМИ, имеющих цифровые аналоги в сети интернет. Репрезентативность корпуса обеспечивается за счет включения текстов широкого круга источников: печатные СМИ, новостные передачи телерадиовещательных компаний, сообщения новостных агентств, новостные ленты интернет-порталов, описывающие состояние дел в мире в целом и в РФ, США, Великобритании в частности.

Массив текстов для БД отбирался методом сплошной выборки текстов в течение двух лет (2011–2013 гг.). Принципы и методы отбора текстов, а также способы конструирования СПА описаны в работах [Потапова, Комалова 2012; 2013а; 2013б; Potapova, Komalova 2013].

Единицей базы данных является исходная текстовая единица, которая содержит *аннотацию*, включающую:

I. Выходные данные текста:

1. Название текста;
2. Название источника;
3. Дата публикации текста;
4. Адрес публикации в сети интернет;
5. Информацию об авторе текста;

II. Лингвистический паспорт текста:

6. Жанровая принадлежность текста (при описании выбрать из списка ниже):

- заметка,
- репортаж,
- статья,
- листовка,
- комментарий,
- фельетон,
- интервью,
- эссе,
- справка,
- коммюнике,
- рекламный текст;

7. Прагматическая составляющая текста (при описании выбрать из списка ниже):

- информирование,
- аналитика,
- полемика,
- разъяснение,
- заявление,
- обращение,
- агитация,
- пропагандирование,
- критика,
- воздействие;

8. Тема текста (при описании выбрать из списка ниже):

- политические конфликты,
- геополитические конфликты,
- военные действия,
- криминал,
- насилие над личностью,
- насилие в семье,
- агрессивная экономика,
- судебные разбирательства,
- проявления экстремизма,
- проявления расизма,
- проявления шовинизма,
- проявление межнациональной розни,
- проявление ксенофобии,
- межконфессиональные конфликты.
- призывы к насилию,
- призывы к дискриминации;

III. Каждый текст сопровождается метрикой, представляющей:

9. Общее количество знаменательных и незнаменательных слов в тексте;

10. Количество знаков (графем) без пробела в тексте;
11. Список слов текста, входящих в СПА;
12. Относительную величину — плотность СПА, т. е. отношение числа слов СПА к общему числу слов в тексте (в %).

В каждой единице БД приводится полный исходный текст публикации.

Каждая БД в настоящее время включает 120 аннотированных единиц (с учетом двенадцати параметров-помет), которые были отобраны из репрезентативной выборки в 2000 русскоязычных и 2000 англоязычных текстов цифровых СМИ.

Анализ текстов и выявление структурных компонентов, репрезентирующих концепт "агрессия" выполнялся по комплексной методике контент-анализа текста/дискурса Р. К. Потаповой и В. В. Потапова [Потапова, Потапов 2004; 2006].

Плотность закодированных вложений в анализируемом тексте рассчитывалась при помощи компьютерной программы с открытой лицензией "Textus Pro 1.0" [Каплунов].

В отдельной ячейке исходной текстовой единицы представлены все присутствующие в данном тексте лексические единицы, входящие в состав СПА (локальный словарь). Данные единицы выделены в исходном тексте подчеркиванием и/или цветом.

Базы данных применимы на практике в качестве:

- баз данных словоформ семантического поля "агрессия" для обучения автоматизированных систем мониторинга текстов цифровых СМИ на предмет выявления потенциально провокативных / конфликтогенных сообщений, а также для определения "горячих" точек в мире (например, для обучения системы Europe Media Monitor http://emm.newsbrief.eu);
- баз данных словоформ для составления тематического частотного словаря семантического поля "агрессия" на русском и английском языках;
- баз данных словоформ для создания информационно-поискового тезауруса семантического поля "агрессия";
- баз данных текстов для проведения тематического морфолого-синтаксического анализа с целью выявления структуры текстов, порождающих состояние агрессии у реципиента;
- баз данных текстов для использования в составе учебно-методических комплексов дисциплин "Лингвоконфликтология", "Лингвокриминалистическая экспертиза текста", "Фундаментальное и прикладное речеведение", "Социолингвистика";
- баз данных текстов в составе комплексного исследования языковых и речевых признаков передачи и порождения состояния агрессии посредством сообщений СМИ в сети интернет.

Дальнейшее развитие базы данных предполагается связать с разработкой внутренней поисковой системы на основе гипертекстовых технологий с использованием XML-разметки.

Литература

Каплунов — *Каплунов Д. А.* Программа “TEXTUS PRO 1.0”. Режим доступа: http://www.blog-kaplunoff.ru/programmy-dlya-kopirajterov.html.

Потапова, Комалова 2012 — *Потапова Р. К., Комалова Л. Р. и др.* Промежуточный отчет (1-й этап 2012г.) по проекту: 6.4411.2011 “Исследование лингвокогнитивного механизма становления и развития состояния агрессии в межъязыковой и межкультурной коммуникации (применительно к многоязыковому дискурсу). М.: МГЛУ, 2012. 310 с.

Потапова, Комалова 2013а — *Потапова Р. К., Комалова Л. Р. и др.* Промежуточный отчет (2-й этап 2013г.) по проекту: 6.4411.2011 “Исследование лингвокогнитивного механизма становления и развития состояния агрессии в межъязыковой и межкультурной коммуникации (применительно к многоязыковому дискурсу). М.: МГЛУ, 2013. 170 с.

Потапова, Комалова 2013б — *Потапова Р. К., Комалова Л. Р.* Лингвокогнитивное исследование состояния “агрессия” в межъязыковой и межкультурной коммуникации: письменный текст // Семиотическая гетерогенность языковой коммуникации: теория и практика. Часть II. М.: ИПК МГЛУ “Рема”, 2013. С. 164–175. (Вестн. Моск. гос. лингвист. ун-та; вып. 15 (675). Сер. Языкознание).

Потапова, Потапов 2004 — *Потапова Р. К., Потапов В. В.* Семантическое поле “наркотики”: Дискурс как объект прикладной лингвистики. М.: УРСС, 2004. 190 с.

Потапова, Потапов 2006 — *Потапова Р. К., Потапов В. В.* Язык, речь, личность. М.: Языки славянской культуры, 2006. 496 с.

Potapova, Komalova 2013 — *Potapova R. K., Komalova L. R.* Lingua-Cognitive Survey of the Semantic Field “Aggression” in Multicultural Communication: Typed Text // SPECOM 2013, LNAI 8113 / M. Železný et al. (Eds.). Springer International Publishing Switzerland, 2013. Pp. 227-232.