Фрумкина Р.М. О законах распределения слов и классов слов // Структурно-типологические исследования. – М., 1962. С. 124-133.

Джубанов А.Х. Квантитативная структура казахского текста (опыт лингвистического анализа на ЭВМ). – Алма-Ата, 1987. С. 147.

Бектаев К.Б., Зубов А.В., Ковалевич Е.Ф. и др. К исследованию законов распределения лингвистических единиц // Статистика текста. – Минск, 1969. С. 131-162.

Лукьяненков К.Ф. Использование схем Пуассона и Гаусса в исследовании распределения лингвистических единиц текста // Вопросы лингвостатистики и автоматизации лингвистических работ. Вып. 2. – М., 1970. С. 3-14.

Феллер В. Введение в теорию вероятностей и ее приложения. Т.1. – М., 1967.– 498 с.

Резюме

В статье рассматриваются вопросы математического моделирования статистических распределений частей речи в текстах Национального корпуса казахского языка. Благодаря вероятностным моделям можно будет описать некоторые статистические свойства языковых единиц данного языка.

Ибрагимов Т.И., Сайхунов М.Р., г. Казань

ПИСЬМЕННЫЙ КОРПУС ТАТАРСКОГО ЯЗЫКА: СТРУКТУРНЫЕ И ФУНКЦИОНАЛЬНЫЕ ХАРАКТЕРИСТИКИ

Создание корпусов национальных языков становится актуальной задачей отечественной лингвистики, активно разрабатываются корпусы башкирского, бурятского, осетинского, калмыкского, лезгинского, языков, начаты работы в этом направлении хакасскими, шорскими, тувинскими языковедами [Бадмаева, Бадагаров, Цыдыпов 2008; Есипова 2013; Шеймович 2011, Салчак 2012; Куканова 2013; Муталов 2009; Сиразитдинов 2013]. Созданию корпуса татарского языка уделяется внимание и в Республики Татарстан [Сайхунов, Ибрагимов, 2014]. Объем Корпуса современного татарского литературного языка в настоящее время составляет более 116 млн. слов, число различных словоформ более 1 миллиона единиц, и по этому показателю наш корпус в отечественной корпусной лингвистике занимает второе место после корпуса русского языка. При создании Корпуса использованы материалы WEB-ресурсов. Более половины текстов (≈ 60% от общего объема Корпуса) представляют публицистический стиль татарского языка. Остальные стили письменного литературного татарского языка содержатся в Корпусе примерно в следующих объемах:

- стиль художественной прозы ≈ 35%,
- стиль научной литературы гуманитарного профиля ≈ 4%,
- стиль деловых бумаг ≈ 1%.

В качестве основных источников лингвистического материала были использованы тексты, представленные на наиболее распространенных Web-сайтах, таких, например, как “Татар – информ”, “Ислам – татарлар һәм мөселманнар (Ислам – татары и мусульмане)”, “Татар электрон китапханәсе (Татарская электронная библиотека)”, “Дуслык (Дружба)”, “Азатлык (Свобода)”, “Яңа гасыр (Новый век)”, “Татар әдипләре (Татарские писатели)” и т.д.

Стиль художественной прозы представлен отрывками из книг, включенных в “Таткнигафонд”, а также изданиями произведений известных татарских писателей – прозаиков, собранных из других источников.

В презентации научного стиля языка использованы статьи из журнала “Фән һәм тел (Наука и язык)”, а также монографии и учебники известных ученых, изданные на татарском языке.

Если учесть, что современный письменный татарский язык не отличается богатством стилевых разнообразий, а существующие стили распределены в языке примерно в указанных пропорциях, то можно полагать, что Письменный корпус

татарского языка сбалансирован. Поскольку общий объем Корпуса превышает 100 млн. слов, то можно считать, что Письменный корпус татарского языка удовлетворяет также и условию репрезентативности.

Основное предназначение Письменного корпуса татарского языка – исследование лексико-семантической системы татарского языка в рамках статистической лексикологии и когнитивной лингвистики.

Поисковая система Письменного корпуса позволяет проводить следующие операции:

- организовать поиск нужных слов, выявить частоту их употребления, находить примеры, подтверждающие употребление данного слова или данной формы в языке;
- определить, какие слова могут следовать впереди и за заданным словом (левый и правый контексты заданного слова), а также употребительность контекстно связанных слов в сочетании с заданным словом;
- находить примеры, подтверждающие употребление данного слова в сочетании с тем или другим словом из левого контекста;
- находить примеры, подтверждающие употребление данного слова в сочетании с каким-либо словом из правого контекста;
- находить полный текст, в котором было обнаружено данное слово в данном контексте.

Указанные функции поисковой системы демонстрируются на примере слова «фикер (мнение)». В результате ввода слова *фикер* в текстовое поле и нажатия кнопки “Найти” получаем, что

- слово *фикер* (мнение) в Корпусе встречается 39111 раз,
- перед словом *фикер* чаще (4429 раз) встречается послелог “*турында*”,
- за словом *фикер* чаще (3646 раз) следует глагол “*алышу*”,
- слово *фикер* может проявлять определенную избирательность не только в выборе контекстов, но и в выборе тематически связанных с ним слов. Так, в анализированных текстах, куда входит слово *фикер,* из существительных наиболее часто вне непосредственного контакта с ним (с *фикер*) встречаются слова: *Татарстан, татар, дәүләт (государство), мәсьәләләре (задачи)* и т.д.

В случае необходимости расширения левого контекста, т.е. знания того, какое слово предшествует сочетанию *турында фикер,* достаточно активизировать (нажать курсором мышки) слово *турында*. Выполнение данной операции позволит определить грамматическую категорию предшествующего слова, а также значения слова *фикер*. Расширение правого контекста, т.е. выявление слова, следующего за сочетанием *фикер алышу*, внесет дополнительные уточнения в значение просмотренного отрезка предложения и поможет снять все оставшиеся неопределенности относительно рассматриваемого слова. Отмеченные особенности Корпуса будут полезными в решении проблемы полисемии, при разработке систем машинного перевода и речевого общения с компьютером.

Объем каждого – правого, левого и тематического – контекстов включает до 100 наиболее часто наблюдаемых в окружении рассматриваемого слова словоформ. Предполагается, что данное количество часто встречающихся словоформ до и после рассматриваемого слова (в нашем случае перед и после слова *фикер)* достаточно для оценки сочетаемостных особенностей рассматриваемой словоформы.

В Корпусе предусмотрена также возможность ознакомления пользователя с предложениями, в которых используется заданное слово. При этом пользователь может ограничиться первыми 50 примерами – предложениями или ознакомиться со всеми предложениями, в которые входит слово *фикер* путем перехода на следующие страницы.

Кроме того, по указанному в конце примера – предложения адресу сайта пользователь может также прочитать и сам текст – оригинал, содержащий данное предложение.

В последней версии Корпуса путем вычисления показателя "Log-likelihood" *(LL)* – логарифмического правдоподобия [Dunning 1993] предусмотрена возможность получения информации об устойчивости или неустойчивости связей слов в данном словосочетании. Так, в Корпусе словосочетания *фикер дөрес (мнение правильное)* и *фикер уята (мысль пробуждает*) встречаются 42 и 41 раза, соответственно (правый контекст). Вычисление логарифмического правдоподобия показывает, что LL в первом случае составляет всего 36,6, а во втором случае, т.е относительно словосочетания *фикер уята – 106,* Следовательно, в текстах, содержащих слово фикер, распределение относительных частот словосочетаний *фикер дөрес* является более равномерным, чем распределение относительных частот словосочетания *фикер уята.* Это означает, что связь между словами *фикер* и *дөрес* более стабильна и устойчива, чем между словами *фикер* и *уята.*

Письменный корпус татарского языка – это собрание огромного количества текстов, вышедших из-под пера людей, достаточно хорошо владеющих языком и знающих культуру народа. Существенным, на наш взгляд, является и то, что в отличие от устных, создание письменных текстов не требует непосредственного контакта с ситуацией. Они пишутся на злобу дня в удобное для авторов время и в удобных условиях. Вместе с тем текст – это описание конкретного события конкретным человеком или «речь, погруженная в жизнь» [Лингвистический энциклопедический словарь 1990].

Совокупность большого количества текстов, имеющихся в Internet и представляющих в Корпусе публицистический стиль татарского языка, а также поисковые ресурсы Корпуса позволяют ученым (языковедам, культурологам, этнографам) изучать особенности видения мира как отдельного носителя татарского языка, так и этнической общности в целом.

Письменный корпус представляет современный татарский язык и может оказаться несомненно полезным в определении этнокультурных ценностей, в изучении их современного состояния, а также культурной ориентации этноса. Корпус может внести существенный вклад в определение статистических характеристик словаря, грамматики, фразеологии татарского языка. Данная версия Корпуса, как и предыдущая, найдет применение в лексикографии, в деле обучения языку и в качестве справочника по татарскому языку. Работы по расширению функциональных возможностей Корпуса будут продолжены.

Литература

Бадмаева Л.Д., Бадагаров Ж.Б., Цыдыпов Б.З. Общие проблемы формирования корпуса бурятского языка /Труды международной конференции «Корпусная лингвистика – 2008» 6–10 октября 2008 г., Санкт-Петербург. – Санкт-Петербург, 2008. С. 24-30

Куканова В.В. О национальном корпусе калмыцкого языка // Актуальные проблемы диалектологии языков народов России: материалы XIII международной конференции. – Уфа, 2013. С. 209-212.

Лингвистический энциклопедический словарь / под редакцией В.Н.Ярцева. – М.: «Советская энциклопедия»,1990.

Муталов Р.О. Опыт создания корпусов Дагестанских языков // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.). Вып. 8 (15). – М.: РГГУ, 2009. С. 329 – 332.

Сайхунов М.Р., Ибрагимов Т.И. Письменный корпус татарского языка [Электронный ресурс]. – Казань, 2014. Режим доступа: http://corpus.tatfolk.ru

Салчак А.Я. Электронный корпус текстов тувинского языка // Новые исследования Тувы. 2012, № 3. (Электронный журнал). URL:http://www.new-tuva.info/journal/issue_15/5231-salchak.html (дата обращения: 17.06.2013).

Сиразитдинов З.А. Корпусные проекты лаборатории лингвистики и информационных технологий // Известия Уфимского научного центра Российской академии наук, №4. 2013, С. 104-112.

Шеймович А.В. (2011) Морфологическая разметка корпуса хакасского языка // Российская тюркология. № 2 (5). С. 48–61.

Dunning T. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 1993, 19/1, pp. 61-74.