НАЦИОНАЛЬНЫЙ МНОГОЯЗЫЧНЫЙ КОРПУС ИМЕНИ АБУСУПЬЯНА АКАЕВА

Д.А. Темирова
МГУ имени М.В. Ломоносова
Москва

Несмотря на то что корпусная лингвистика — относительно молодая наука, в настоящее время существует большое количество определений понятия «корпус». Так, у Т. МакЭнери и Э. Уилсона можно найти определение корпуса как массива языковых фрагментов, собранных на основе определенных языковых критериев, для использования их как модели языка [3, с. 2]. В работе С. Кублер и Х. Зинсмейстер дается следующее определение: современный лингвистический корпус — доступная в электронном виде коллекция текстов и расшифровки аудио записей, которые отбираются для представления определенного языка, языкового многообразия, или других лингвистических категорий [2, с. 4].

К предпосылкам создания и использования корпусов следует отнести большой объем корпуса, (позволяющий гарантировать типичность данных и обеспечить представление всех языковых явлений в полной мере), естественную контекстную форму различных типов данных, находящихся в корпусе, которая дает возможность для их всеобъемлющего изучения, а также многократность использования одного массива данных различными исследователями в различных целях [1, с. 8].

Национальный Многоязычный Корпус (НМК) назван именем Абусупьяна Акаева в связи с той ролью, которую он сыграл в развитии просветительского и просвещенческого движения в Дагестане I четверти XX в.

Абусупьян Акаев — ученый-тюрколог, поэт, педагог, публицист и общественный деятель, фигура которого известна не только народам Дагестана, но и всему Северному Кавказу благодаря впервые созданной им по европейскому образцу ново-методной («усул-и джадид») школе. «Абусупьян Акаев издал первый печатный букварь и учебно-методическое пособие на кумыкском языке, грамматические пособия по арабскому языку, серию многоязычных словарей. А.А. Акаев был инициатором создания первой национальной типолитографии в Дагестане, которая превратилась, по сути, в главного поставщика книг на языках народов Северного Кавказа, издававшихся на арабской графической основе. Он автор многочисленных публицистических и научных статей, кроме них издал в 1902–1928 гг. более 40 отдельных книг по самым различным отраслям» [4].

Целью данного проекта является создание лингвистической разметки текстов, которая поможет при обучении математическим моделям языка, а также при создании программ для автоматической обработки текстов. В соответствии с поставленной целью будут решаться следующие задачи:

- проанализировать тексты на основных уровнях языка (морфологический, синтаксический, стилистический, семантический, прагматический);
- пронаблюдать процессы языковых изменений с точки зрения лексики, грамматики и др. в определенный период;
- определить возможность влияния социального статуса и социальной роли того или иного автора при написании текста.

Лингвистическая разметка текстов будет представлять собой 2 вида: метатекстовую (сведения об авторе и общие характеристики текстов: жанр, тематика, стиль) и языковую / лингвистическую (соответствующую упомянутым уровням языка).

Необходимость создания такого корпуса обусловлена проведением лингвистических исследований для сравнительного анализа данных, который позволит проследить непрерывные процессы языковых изменений, происходящих в языках. С этой точки зрения особый интерес к настоящему проекту будет представлять для исследователей-лингвистов разного профиля. Однако на этом их круг не ограничивается, т.к. надежные статистические данные о языке в определенный период его развития или его пользования определенным автором может стать объектом изучения представителей многих областей гуманитарных наук (литературоведов, историков и др.). Одной из ключевых функций национальных корпусов является их возможность дать ответы на вопросы об устройстве и функционировании языка как носителям, так и всем, изучающим его в качестве иностранного [5].

НМК планируется разработать таким образом, чтобы он включал в себя работы, относящиеся к тюркской группе языков, а именно:

- кумыкский (Къумукъ Тил / Qumuq til) – половецко-кыпчакский;
- карачаево-балкарский (Къарачай-малкъар тил / Qarachay-malqar til) – половецко-кыпчакский;
- крымско-татарский (Къырымтатар тили / Qirimtatar tili) – половецко-кыпчакский;
- ногайский (Ногай тили / Nogay tili) – кыпчакский.

НМК будет представлять собой массив, который вберет в себя работы ученых-тюркологов, начиная с периода творчества Абусупьяна Акаева вплоть до текстов современных авторов. Предполагается, что корпус будет включать в себя следующие разделы:

- общий;
- исторический;
- религиозный;
- поэтический;
- диалектный;
- газетный/журнальный;
- сетевой;
- учебный.

Вычисление частоты использования словоформ (types) и словоупотреблений (tokens), встречающихся в текстах, будет производиться с помощью программы-конкордансера AntConc [6], разработанной японскими учеными, которая позволит изучить не только контекст слова, но и выявить типичные случаи употребления слов в одной коллокации.

Создание такого корпуса в дальнейшем должно послужить отправной точкой к разработке различных типов словарей (словарей-конкордансов, частотных словарей, терминологических, исторических и др.).

Так как язык — явление динамичное, которое претерпевает определенные изменения на пути своего развития, ставится вопрос о сохранении языковых стандартов, а также о развитии такого подхода к текстам, который помог бы улучшить их понимание и восприятие не только носителям, но и представителям близкородственных языков, благодаря разметкам и метаданным, представленным непосредственно в корпусе.

ЛИТЕРАТУРА

1. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник для студентов гуманитарных вузов. Иркутск: ИГЛУ, 2011. 161с.
2. Kübler S., Zinsmeister, H. Corpus linguistics and linguistically annotated corpora. Bloomsbury Publishing Pic, 2015.
3. McEnery T., Wilson, *A.* Corpus Linguistics. Edinburgh: Edinburgh University Press, 2001.
4. http://kumukia.ru/person?pid=3004.
5. http://www.ruscorpora.ru/corpora-intro.html.
6. http://www.laurenceanthony.net/software.html.